J. R. Statist. Soc. B (1978),
40, *No.* 2, *pp.* 113–146

Density Estimation, Stochastic Processes and Prior Information

By TOM LEONARD

Department of Statistics, University of Warwick

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 7th, 1977, Professor J. F. C. KINGMAN in the Chair]

SUMMARY

A method is proposed for the non-parametric estimation of a probability density, based upon a finite number of observations and prior information about the smoothness of the density. A logistic density transform and a reproducing inner product from the first-order autoregressive stochastic process are employed to represent prior information that the derivative of the transform is unlikely to change radically within small intervals. The posterior estimate of the density possesses a continuous second derivative; it typically satisfies the frequentist property of asymptotic consistency. A direct analogy is demonstrated with a smoothing method for the time-dependent Poisson process; this is similar in spirit to the normal theory Kalman filter. A procedure for grouped observations in a histogram provides an alternative to the histospline method of Boneva, Kendall and Stefanov. Five practical examples are presented, including two investigations of normality, an analysis of pedestrian arrivals at a Pelican crossing and a histogram smoothing method for mine explosions data.

Keywords: NON-PARAMETRIC DENSITY ESTIMATION; PRIOR LIKELIHOOD; STOCHASTIC PROCESS; LOGISTIC DENSITY TRANSFORM; LIPSCHITZ CONDITION; GAUSSIAN PROCESS; REPRODUCING HILBERT SPACE; INNER PRODUCT; CALCULUS OF VARIATIONS; CONTROL THEORY; ROUGHNESS PENALTY; GREEN'S FUNCTION; HISTOGRAM SMOOTHING; HISTOSPLINE; ORDER STATISTICS; WINDOW FUNCTION; POISSON PROCESS; KALMAN FILTER; CHANGE-POINT INFERENCE; TESTING FOR NORMALITY

1. DISCUSSION AND BACKGROUND

CONSIDER the situation where observations $x_1, \ldots, x_m$ constitute a random sample from a distribution with density f which is concentrated on a bounded interval $[a, b]$ of the real line, and suppose that f is not restricted to belong to any particular parameterized family. The assumption of a bounded interval is slightly restrictive, but should usually lead to reasonable approximations in the unbounded case.

One possible method for estimating f might be to maximize the log-likelihood functional

$$L(f) = \sum_{i=1}^{m} \log f(x_i) \tag{1.1}$$

over the space of integrable functions satisfying the constraints

$$\int_a^b f(t)\,dt = 1 \tag{1.2}$$

and

$$f(t) \geqslant 0 \quad \text{for } t \in [a, b]. \tag{1.3}$$

This yields the limiting solution $f = \phi$ where

$$\phi(t) = m^{-1} \sum_{i=1}^{m} \delta_{x_i}(t) \tag{1.4}$$

denotes the average of the Dirac-delta functions at the m data points. The estimate in (1.4) therefore possesses a spike at each data point, and would not be viewed as particularly sensible in the presence of any prior information that f was likely to be a well-behaved function.

The idea of relating the density estimation problem to the theory of stochastic processes has been discussed by several authors. For example, Ferguson (1973) has developed the novel concept of *Dirichlet* processes; these introduce a prior estimate of the shape of the density; the corresponding posterior estimate takes an interesting weighted average form of the function ϕ and the prior estimate. Note that Dirichlet processes do not introduce any prior correlations between $f(s)$ and $f(t)$ for $s \neq t$ beyond those induced by the constraint in (1.2). Therefore the posterior estimate will still possess spikes at the data points, though these will be somewhat "less infinite".

The method presented here will be somewhat similar in spirit to the important pioneer work of Whittle (1958), which seems to have been completed many years in advance of its time. Whittle assumes that different values of the density are related by a prior covariance structure, and then obtains integral equations for his posterior estimates. His method just involves the specification of the first two moments of the prior distribution, and attention is restricted to a class of "window functions" which provides linear estimates for f. As indicated by Dickey (1969), Whittle's estimates for f can sometimes be negative, although they will always integrate to unity. Our approach will remove this drawback, and will be based upon a complete prior specification of a distribution over function space, leading to posterior estimates more general than the class of window functions. Some conceptually related work is by Kimeldorf and Wahba (1970) who obtain a beautiful analogy between Bayesian estimation for stochastic processes, and density smoothing by splines; this is pursued by Wahba (1975, 1976). Also, Parzen (1977) employs the transformations $xF(x_i)$ of the observations where F denotes the distribution function, and obtains an analogy with testing for white noise in a normal time series, under an information theoretic approach. Some recent Bayesian work in the field of non-parametrics is by Goldstein (1975), who suggests an ingenious method for obtaining posterior moments in terms of prior moments.

Our approach is also related to the method of Good and Gaskins (1971) who, in a paper of considerable conceptual importance, subtract a *roughness penalty* $\Phi(f)$ from the functional in (1.1), and then maximize $L(f)-\Phi(f)$, using a method based upon Hermite polynomials. Their posterior estimate for f is typically a well-behaved smooth function for all $t \in [a, b]$, even though it is only based upon a finite number of observations. It however seems difficult to interpret their choice of roughness penalty as formally representing a particular set of prior beliefs. They do not, for example incorporate a prior estimate for f into their formulation. We will be able to obtain an integral equation for our posterior estimate, leading to some precise results about its behaviour and properties.

Our analysis will be based upon the formalism of reproducing Hilbert spaces; these have been previously used by Boneva *et al.* (1971), who provide an elegant histospline method for estimating f in situations where the observations are grouped in a histogram. Some possible advantages of our approach are discussed in Section 6. For example, our derivatives of f will appear to be subjected to somewhat greater smoothing, when compared with histosplines. Hilbert space methods have also been discussed by Montricher *et al.* (1975) who suggest that they could be used to rigorize the approach of Good and Gaskins.

A few of the conceptual ideas employed here were discussed by Parzen after the paper by Boneva *et al.*; we make some rather different assumptions, e.g. concerning a logistic transformation introduced in Section 2, and our formulation of the estimation problem in Section 4, which will enable us to solve a difficult analysis, and to get some posterior estimates satisfying well-defined regularity conditions, as well as incorporating subjective prior information about the shape of the density.

We follow Leonard (1973) who obtains smoothed estimates for the multinomial probabilities in a histogram, which provides a restricted type of estimate for f. He employs a

multivariate normal prior distribution for the logits, with a covariance structure from the first-order autoregressive process. For example, the data for Fig. 1 were discussed by Snedecor and Cochran (1967, p. 71) and concern the frequency distribution of 511 means of samples of 10 pig gains. The smoothed frequencies are based on similar prior assumptions as employed for the numerical example in the 1973 paper. They adjust the corresponding observed frequencies by allowing for values in neighbouring intervals, and the smoothed histogram is better behaved than the observed histogram.

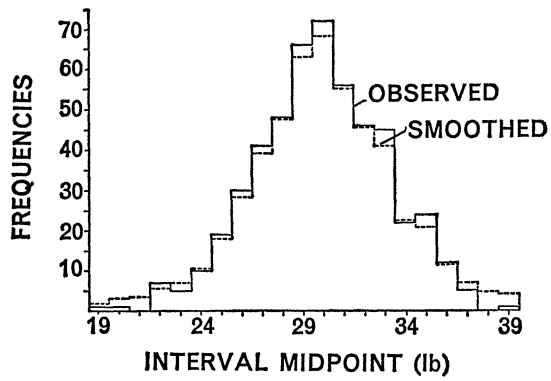

FIG. 1 Observed and smoothed frequency distributions of pig-gains.

The method in the present paper is similar in spirit, but provides an estimate for f which is twice-differentiable rather than resembling the histogram-type estimate described in the 1973 paper. It will be appropriate for both grouped and ungrouped observations, and is based upon a logistic transformation of the probability density. Note that whilst the posterior estimates will be consistent as $m \to \infty$ they will depend substantially upon the prior information which is available; we view it as an advantage that we are able to formally incorporate this information into the analysis since the problem would not otherwise appear to be well enough defined to provide the possibility of a sensibly formulated solution.

2. THE PRIOR MODEL

The construction of an appropriate prior model for f is inconvenienced by the constraints in (1.2) and (1.3). This problem may however be minimized by introducing a *logistic density transform* of f to a function g satisfying

$$f(t) = e^{g(t)} \Big/ \int_a^b e^{g(s)}\, ds, \tag{2.1}$$

so that the inverse transformation is given by

$$g(t) = \log f(t) + D(g), \tag{2.2}$$

where

$$D(g) = \log \int_a^b e^{g(s)}\, ds. \tag{2.3}$$

Note from (2.1) that the constraints in (1.2) and (1.3) will be satisfied for any real valued integrable function g. It will therefore be possible to estimate g in unrestricted fashion without needing to concern ourselves with these constraints; our device helps us to avoid the difficulty of negative estimates for f, experienced, for example, by Whittle and by Boneva *et al.*

It should also be noted from (2.1) that g is not uniquely defined, since f is unaffected by the addition of the same arbitrary constant to each value of g; our analysis will also be unaffected by this value.

Silverman (1978) employs a logarithmic transformation when estimating the ratio of two densities using roughness penalties, and indicates that the log-density has considerable intrinsic appeal. Our logistic density transformation is slightly different technically, and appears to provide a powerful tool for the non-parametric estimation of a single density.

Some distributional assumptions are now introduced to represent possible prior information that the derivative $g^{(1)} = f^{(1)}/f$ is unlikely to change radically within small intervals. This could be viewed as a stochastic type of "Lipschitz" condition. A prior estimate μ is also introduced for the function g, suggesting the estimate $\xi(t) = e^{\mu(t)}/\int e^{\mu(s)}\,ds$ for f. One possibility might be to take ξ to denote the density from a particular parameterized family (note that we estimate f by the obvious transformation of the estimate of g; this could be justified by assuming a zero-one loss function for g throughout this paper).

It is possible to proceed under a formal Bayesian approach, in which case we would suggest the prior assumptions that the derivative $g^{(1)}$ possesses the probability structure over differentiable function space of a Gaussian process with mean value function $\mu^{(1)}$ and covariance kernel K satisfying

$$K(s,t) = \operatorname{cov}\{g^{(1)}(s), g^{(1)}(t)\} = \sigma^2 e^{-\beta|s-t|} \quad (0 < \sigma^2, \beta < \infty;\ t \in [a,b]). \tag{2.4}$$

The possible assumption in (2.4) is based upon the covariance kernel of the first-order autoregressive process, and represents prior information that the closeness of $g^{(1)}(s)$ and $g^{(1)}(t)$ is likely to increase as $|s-t|$ decreases. Other choices could be made for the form of the kernel, with possibly different results [see Section (12.4)]. Our results will depend upon the particular choice in (2.4), but we feel that this assumption will often be plausible.

Ferguson's Dirichlet process corresponds in spirit to a covariance kernel for the original function g from the normal white noise process, whilst under the above assumptions the quantity

$$(2\beta\sigma^2)^{-\frac{1}{2}}[g^{(2)}(t) - \mu^{(2)}(t) + \beta\{g^{(1)}(t) - \mu^{(1)}(t)\}]$$

possesses the probability structure from the standard white noise process. If the covariance kernel in (2.4) were assumed for g instead of $g^{(1)}$ then this would lead to a similar reproducing inner product to that employed by Boneva *et al.*, and our consequent posterior estimate for g would in fact possess a discontinuous first derivative. Note that the parameter σ^2 in (2.4) measures the degree of belief in the closeness of $g^{(1)}$ to its prior estimate $\mu^{(1)}$, whilst β provides a measure of the closeness of neighbouring values of $g^{(1)}$.

Under the prior assumptions described above it is possible to proceed by a Bayesian approach based upon some pure mathematical limit theorems. This leads to a posterior estimate for g which is optimal under a zero-one loss function. Since this is the limit of an estimate based upon a bounded and continuous loss function, our approach could be viewed as approximately "coherent" within a Bayesian framework. It is also possible to show that our estimate approximates the Bayes estimates under other loss functions, e.g. quadratic loss.

Since our Bayesian arguments are rather complicated, the details are omitted from this paper. We instead follow Edwards (1972) by proceeding under a likelihood approach. The prior information about g will be represented by a prior likelihood; the corresponding log-likelihood would be referred to by Edwards as the prior "support". This will be added to the sample log-likelihood to give the posterior support for g, which may then be maximized to provide an estimate for g.

Note that the estimate obtained in this way will not correspond to a posterior mode for g under the Bayesian method mentioned above. Owing to technical difficulties indicated at the beginning of the next section, the posterior mode cannot be uniquely defined over function

space, and there appears to be no specific way of defining it to yield the estimate obtained in this paper. However, the Bayes estimate under zero-one loss may still be uniquely defined and this in fact turns out to be exactly the same as the estimate described in Section 5.

3. A Prior Likelihood Approach†

A technical difficulty associated with the Bayesian approach lies in the strong dependence of the prior density, i.e. Radon–Nikodym derivative, upon the particular measure with respect to which its distribution is taken to be absolutely continuous, e.g. any Gaussian measure with the same covariance kernel but different mean value function. This difficulty is, however, easily circumvented by proceeding under a prior and posterior likelihood approach.

Suppose instead that $\mu^{(1)}$ represents the realization of a continuum of hypothetical prior observations which possess the probability structure of a Gaussian process with mean value function $g^{(1)}$ and the covariance kernel K previously described for $g^{(1)}$ in (2.4). This provides a technically simpler way of representing similar prior information. We may now more readily refer to the theory of reproducing Hilbert spaces.

Parzen (1961, p. 477) indicates that the reproducing Hilbert space for the first-order autoregressive scheme consists of all differentiable functions. The reproducing inner product is related to the covariance kernel K in (2.4) by a general result involving eigenfunctions of K. It is given in this instance by

$$(\mu^{(1)}, g^{(1)}) = \frac{1}{2\beta\sigma^2}\int_a^b \{\mu^{(2)}(t)g^{(2)}(t) + \beta^2\mu^{(1)}(t)g^{(1)}(t)\}\,dt$$

$$+\frac{1}{2\sigma^2}\{\mu^{(1)}(a)g^{(1)}(a) + \mu^{(1)}(b)g^{(1)}(b)\}, \tag{3.1}$$

where $\mu^{(1)}$ and $g^{(1)}$ should be interpreted as non-random realisations.

Let P denote the probability measure for $\mu^{(1)}$ associated with the Gaussian distribution with mean value function $g^{(1)}$ and covariance kernel K in (2.4), and let P^* denote a similar probability measure, but with zero mean value function. Then, on p. 479, Parzen (1961) shows that the Radon–Nikodym derivative of P, with respect to P^*, is given by

$$dP/dP^* = \exp\{L_0(\mu^{(1)}, g^{(1)})\}, \tag{3.2}$$

where

$$L_0(\mu^{(1)}, g^{(1)}) = (\mu^{(1)}, g^{(1)}) - \tfrac{1}{2}(g^{(1)}, g^{(1)}) \tag{3.3}$$

with the inner product contributions on the right-hand side of (3.3) defined in (3.1).

The quantity in (3.3) represents the prior log-likelihood functional of $g^{(1)}$, given $\mu^{(1)}$. As g in (2.1) is not completely identified in terms of f, there can be no useful information about g beyond that already specified for $g^{(1)}$. Therefore (3.3) provides the prior log-likelihood functional of g, which we will now denote by $L_0(g)$.

If we changed P^* to any other measure with respect to which P is absolutely continuous, then this would only affect the expression in (3.3) by the addition of a value not depending upon g. Note that under the Bayesian assumptions of Section 2, the prior density of g would depend substantially upon the choice of measure corresponding to P^* and hence cause severe technical difficulties with this type of analysis, e.g. in uniquely defining a posterior mode for g.

The functional in (3.3) includes quadratic terms in $g^{(1)} - \mu^{(1)}$ and $g^{(2)} - \mu^{(2)}$. In the next section we see how it is affected by the information from the sample.

† Readers unacquainted with measure theory may wish to omit Sections 3 and 4. The main result of the paper is described at the beginning of Section 5.

4. The Posterior Analysis

The posterior log-likelihood functional or support for g satisfies

$$L_1(g) = L(g) + L_0(g), \tag{4.1}$$

where $L_0(g)$ is equal to the functional in (3.3), and $L(g)$ is the sample log-likelihood functional of g,

$$L(g) = m\int_a^b g(t)\,\phi(t)\,dt - mD(g), \tag{4.2}$$

with ϕ and D defined in (1.4) and (2.3) respectively. Note that the quantity $-L_0(g)$ could be interpreted as a "roughness penalty".

The function g may be estimated by maximizing the posterior functional in (4.1), using the Calculus of Variations. We refer the reader to Bellman (1967) for an account in the context of control theory of the general techniques which should be employed.

In the Appendix we show that any optimal function $\tilde{g}$ for g satisfies the fourth-order non-linear differential equation

$$\tilde{f}(t) = \phi(t) - (2m\beta\sigma^2)^{-1}\,\omega(t) \quad \text{for } t \in [a,b], \tag{4.3}$$

where ϕ satisfies (1.4) and with D defined in (2.3),

$$\tilde{f}(t) = \exp\{\tilde{g}(t) - D(\tilde{g})\}, \tag{4.4}$$

and

$$\omega(t) = \tilde{g}^{(4)}(t) - \mu^{(4)}(t) - \beta^2\{\tilde{g}^{(2)}(t) - \mu^{(2)}(t)\}. \tag{4.5}$$

The solution is restricted by the boundary conditions

$$\lambda^{(3)}(b) - \beta^2\lambda^{(1)}(b) = 0, \quad \lambda^{(3)}(a) - \beta^2\lambda^{(1)}(a) = 0, \quad \lambda^{(2)}(b) + \beta\lambda^{(1)}(b) = 0$$

and

$$\lambda^{(2)}(a) - \beta\lambda^{(1)}(a) = 0, \tag{4.6}$$

where

$$\lambda = \tilde{g} - \mu.$$

The quantity in (4.4) provides a posterior estimate for f. Note from (4.3) that $\tilde{f}$ adjusts the sample maximum likelihood estimate ϕ by introducing a term ω in (4.5) based upon the second and fourth derivatives of g. Whilst ϕ in (1.4) is the average of the Dirac-delta functions at the data points our analysis will show that this is essentially absorbed in the contribution $(2m\beta\sigma^2)^{-1}\tilde{g}^{(4)}$ to ω. Our solution will possess a continuous second derivative, a discontinuous third derivative, and a fourth derivative with spikes at the data points.

In the Appendix we use a Green's function method to convert our fourth-order differential equation into an integral equation involving only one-dimensional integrations. The result is stated at the beginning of the next section.

5. An Integral Equation

Using the method outlined in the previous section, we have the result that an optimal function $\tilde{g}$ for g satisfies the equation

$$\tilde{g}(t) = \mu(t) + m\tau\{Q_t(\tilde{f}) - Q_t(\phi)\} \quad \text{for } t \in [a,b], \tag{5.1}$$

where ϕ and $\tilde{f}$ are given in (1.4) and (4.4) respectively, and

$$Q_t(\tilde{f}) = \int_a^b \{\exp(-\beta|t-u|)\}\tilde{f}(u)\,du + 2\beta\int_a^t (t-u)\tilde{f}(u)\,du \tag{5.2}$$

with

$$\tau = \sigma^2/\beta^2 \tag{5.3}$$

Equations (5.1) and (5.2) provide the main result of this paper. Note that any fixed value may be added to the right-hand side of (5.1) without affecting the values of $\tilde{f}$ in (4.4). The properties of the equations are examined in the next five sub-sections.

5.1. *Conditions on the Derivatives of* $\tilde{f}$

Let $x_{(1)}, \ldots, x_{(m)}$ denote the increasing order statistics of the observations, and consider the intervals $I_k = [x_{(k)}, x_{(k+1)}]$ for $k = 0, 1, \ldots, m$, with $x_{(0)} = a$ and $x_{(m+1)} = b$. Then we find from (1.4) and (5.2) that whenever $t \in I_k$ the function $Q_t(\phi)$ satisfies

$$mQ_t(\phi) = \sum_{i=1}^{m} \exp(-\beta|t - x_{(i)}|) + 2\beta\left(kt - \sum_{i=1}^{k} x_{(i)}\right). \tag{5.4}$$

The function in (5.3) is continuous for all $t \in [a, b]$, *in particular at the data-points* $x_{(1)}, \ldots, x_{(m)}$, and possesses continuous first and second derivatives, but has a third derivative which is discontinuous at the data-points.

It follows obviously that the contribution $Q_t(\tilde{f})$ to (5.2) will possess these same regularity properties, even if $\tilde{f}$ is taken to possess spikes at the data-points. The total expression for $\tilde{g}$ in (5.1) and any solution for $\tilde{f}$ to (4.4) and (5.1) will therefore possess a continuous second derivative, as long as this is assumed for the prior mean value function μ. This provides an obvious advantage over the methods of Ferguson and Boneva *et al.*

5.2. *Comparison with Window Functions*

The first term on the right-hand side of (5.4) is a *window function*, i.e. it assumes the same form as the estimates considered by many authors for the density f. For example, Whittle (1958) and Parzen (1962) take their estimates for f to belong to the class

$$f_m(t) = \frac{1}{mh} \sum_{i=1}^{m} \Omega\left(\frac{t - x_{(i)}}{h}\right), \tag{5.5}$$

where h and the function Ω are suitably chosen.

Our final estimates from (4.4), (5.1) and (5.2) will not be restricted to this class, and appear to take a less restrictive form than those recommended by many previous authors.

5.3. *Consistency of Posterior Estimate*

We now present an intuitive "physicist's" proof which shows that, under certain regularity conditions on the true function f, the estimate $\tilde{f}$ is consistent for f as $m \to \infty$. We specifically suppose that there exists $\varepsilon > 0$ and $M < \infty$ such that $\varepsilon < f(t) < M$ for all $t \in [a, b]$, so that any function g satisfying (2.2) is confined to an interval of finite width $\log M - \log \varepsilon$.

It is necessary to rescale $\tilde{g}$ in (5.1) to a function whose origin is fixed as $m \to \infty$. The equation

$$m^{-1}\{\tilde{g}(t) - \mu(t)\} = \tau\{Q_t(\tilde{f}) - Q_t(\phi) - (b-a)^{-1}\int_a^b Q_t(\tilde{f})\,dt + (b-a)^{-1}\int_a^b Q(\phi)\,dt\} \tag{5.6}$$

provides exactly the same solution for $\tilde{f}$, but with the condition

$$\int_a^b \tilde{g}(t)\,dt = \int_a^b \mu(t)\,dt. \tag{5.7}$$

It is easily checked from (5.4) that, as $m \to \infty$, the statistic $Q_t(\phi)$ will converge with (sampling) probability one to its expectation $Q_t(f)$. We therefore see that as $m \to \infty$ there is a solution for $\tilde{f}$ to (4.4) and (5.6) satisfying $Q_t(\tilde{f}) = Q_t(f)$ and $\tilde{f} = f$. This follows because the contribution $\tilde{g}$ corresponding to $\tilde{f} = f$ will remain bounded in view of (5.7) and the regularity conditions described above, so that the left-hand side of (5.6) will tend to zero. There must therefore exist a solution $\tilde{f} = f$ as $m \to \infty$ since the right-hand side of (5.6) would otherwise assume non-zero limiting values. This shows intuitively speaking that $\tilde{f}$ is typically consistent for f.

5.4. *Computation of Solution*

Equations (4.4) and (5.1) may be solved by substituting a trial function for $\tilde{f}$ into the right-hand side of (5.1), then obtaining a new function from (4.4) and cycling until convergence. The latter typically takes under 15 iterations for accuracy to 3 decimal places, and about 30 seconds on a Burroughs 6700 computer. It is often adequate to take a normal curve with mean and variance estimated from the data, as first trial estimate for f and, by trying different initial estimates, we have always obtained convergence to the degree of accuracy permitted by the numerical integration. There are some computational problems associated with exponential overflow and numerical integration which require special subroutines. A prototype computer package is available on request. Another has been prepared for the histogram method of Section 6.

5.5. *General Remarks*

In summary we feel that the main strengths of our method are as follows:

(a) The posterior estimate for f in (4.4) satisfies the defining properties of a density, is typically consistent for f as $m \to \infty$ and possesses a continuous second derivative.
(b) The theory is exact and provides a formal way of introducing prior information about both the shape of the density and its regularity conditions.
(c) The posterior estimate for f is not restricted a priori to belong to any particular class.
(d) The maximization procedure can be completed using the Calculus of Variations and by solving a simple integral equation in straightforward fashion.

We shall see in Section 7 that the method is analogous to smoothing the mean value function of the time-dependent Poisson process, using an approach which is similar in spirit to the normal theory Kalman filter.

6. Grouped Observations

Consider now the situation where the observations $x_1, \ldots, x_m$ are grouped into a histogram, with p intervals $J_1, \ldots, J_p$ comprising a partition of the bounded interval $[a, b]$; let y_j denote the number of observations falling in interval J_j, with $\sum y_j = m$, and suppose that the raw observations x_i are not available to the statistician. We provide here an alternative to the histospline method of Boneva *et al.*

Since $y_1, \ldots, y_p$ possess a multinomial distribution with cell probabilities equal to the corresponding interval probabilities of f, the sample log-likelihood functional of the density f is now given by

$$L^*(f) = \times \log\{C(\mathbf{y})\} + \sum_{j=1}^{p} y_j \log \int_{J_j} f(t)\, dt, \tag{6.1}$$

where $C(\mathbf{y}) = m! / \prod_j y_j!$.

It therefore follows from (2.3) that the corresponding functional for g is given by

$$L^*(g) = \times \log\{C(\mathbf{y})\} + \sum_{j=1}^{p} y_j \log \int_{J_j} \exp\{g(t)\}\, dt - m \log \int_a^b \exp\{g(t)\}\, dt. \tag{6.2}$$

If g is taken to possess the prior log-likelihood functional in (2.7), it is possible to show, using the Calculus of Variations, that any function g maximizing the posterior log-likelihood functional must satisfy the equation

$$\tilde{f}(t) = \tilde{\phi}(t) - (2m\beta\sigma^2)^{-1} \omega(t) \quad \text{for } t \in [a, b], \tag{6.3}$$

where $\tilde{f}$ and ω are defined in (4.4) and (4.5), and

$$\tilde{\phi}(t) = m^{-1} y_j \exp\{\tilde{g}(t)\} \Big/ \int_{J_j} \exp\{\tilde{g}(s)\}\, ds, \quad (t \in J_j; j = 1, \ldots, p). \tag{6.4}$$

The discontinuous function $\tilde{\phi}$ replaces ϕ from (1.4) in equation (4.3). Any maximum $\tilde{g}$ should also satisfy the boundary conditions in (4.6). By an analogous method to that indicated in Section 4 it follows that the optimal function for $\tilde{f}$ is given by the expression in (4.4) where

$$\tilde{g}(t) = \mu(t) + m\tau Q_t(u) \quad \text{for } t \in [a, b], \tag{6.5}$$

with Q_t defined in (5.2), and

$$u = \tilde{f} - \tilde{\phi}. \tag{6.6}$$

Equations (4.4), (6.5) and (6.6) may be solved by a similar method to that described in Section 5.4 for equations (4.4) and (5.1). Any solution will possess a continuous third derivative, and a fourth derivative which is discontinuous at the boundary points of the interval J_j.

The method described above appears to possess the following possible advantages when compared with the important and elegant approach of Boneva *et al.*:

(a) The estimate $\tilde{f}$ can never be negative, whilst Boneva *et al.* note that this is a moderately frequent property of histosplines.
(b) Our prior likelihood functional introduces a prior estimate for f; whilst Boneva *et al.* do not incorporate a prior estimate into their analysis.
(c) Our method is readily generalizable to the situation where the observations are ungrouped; Boneva *et al.* mention a generalization of their method to this situation but this restricts their estimates to the subclass in (5.2).
(d) The method of Boneva *et al.* constrains the interval probabilities of their histogram to be equal to the observed proportions y_j/m, whereas our approach will also smooth these proportions.
(e) Our method provides smooth estimates for f, whilst histosplines tend to possess a number of modes and fluctuate a fair amount under small changes in the data.
(f) Our intervals can be as narrow or unequal as the user likes, whilst the estimates of Boneva *et al.* tend to be rougher for narrower intervals.

A numerical example comparing the alternative approaches is described in Section 11.

7. The Time-dependent Poisson Process

Consider next the situation described by Cox and Miller (1970, p. 153) where observations occur in a Poisson process with time-dependent rate $\rho(t)$. Suppose that the process is observed on a fixed interval $[a, b]$; the number of observations falling in any time interval $[t_1, t_2]$ is Poisson distributed with mean

$$\int_{t_1}^{t_2} \rho(t)\, dt. \tag{7.1}$$

Kalman (1960) pioneered a method for filtering a process with normal observations. We describe here an approach which is similar in spirit, but which provides a smoothed estimate for ρ over the whole interval $[a, b]$.

Snyder (1975, p. 63) shows that the joint density associated with m observations in $[a, b]$ at the time-points $x_1, \ldots, x_m$ is given by

$$p(m; x_1, \ldots, x_m | \rho) = \begin{cases} \exp\left\{-\int_a^b \rho(t)\, dt\right\} & \text{for } m = 0 \\ \exp\left\{-\int_a^b \rho(t)\, dt\right\} \prod_{i=1}^{m} \rho(x_i) & \end{cases} \tag{7.2}$$

for $m = 1, 2, \ldots$.

Therefore, when $m \geqslant 1$, the sample log-likelihood functional of $q = \log \rho$ is given by

$$L(q) = m\int_a^b q(t)\,\phi(t)\,dt - \int_a^b e^{q(t)}\,dt, \tag{7.3}$$

where ϕ is defined in (1.4).

Suppose that the derivative $q^{(1)} = \rho^{(1)}/\rho$ possesses the prior log-likelihood functional described for $g^{(1)}$ in (3.3) but that there is no further prior information about the function q (note that q is only defined by $q^{(1)}$ up to the addition of the same arbitrary constant η to each value of q). One way of expressing this mathematically is by supposing that

$$q(t) = \eta + g(t) \quad \text{for } t \in [a, b], \tag{7.4}$$

where there is no prior information about η, and d possesses the prior log-likelihood functional $L_0(g)$ in (3.3).

The posterior log-likelihood for η and g is now given by

$$L_1(\eta, g) = L(\eta + g) + L_0(g) = m\eta + m\int_a^b g(t)\,\phi(t)\,dt - e^{\eta}\int_a^b \exp\{g(t)\}\,dt + L_0(g). \tag{7.5}$$

Using the Calculus of Variations to maximize the functional in (7.5) shows that the optimal values $\tilde{\eta}$ and $\tilde{g}$ satisfy

$$\exp(\tilde{\eta}) = m \Big/ \int_a^b \exp\{\tilde{g}(t)\}\,dt \tag{7.6}$$

and

$$m^{-1}\exp\{\tilde{\eta} + \tilde{g}(t)\} = \phi(t) - (2m\beta\sigma^2)^{-1}\,\omega(t), \tag{7.7}$$

where ω is defined in (4.5). Elimination of $\tilde{\eta}$ from (7.6) and (7.7) gives

$$\tilde{f}(t) = \phi(t) - (2m\beta\sigma^2)^{-1}\,\omega(t), \quad \text{for } t \in [a, b], \tag{7.8}$$

where $\tilde{f}$ is defined in (2.1). Note the pleasing result that (7.8) is identical to (4.3) in the density estimation situation. The boundary conditions in (4.6) still apply, and ρ may be estimated by

$$\tilde{\rho}(t) = \exp\{\tilde{\eta} + \tilde{g}(t)\} = m\exp\{\tilde{g}(t)\} \Big/ \int_a^b \exp\{\tilde{g}(s)\}\,ds = m\tilde{f}(t). \tag{7.9}$$

We therefore come to the conclusion that ρ may be estimated by evaluating $\tilde{f}$ in the density estimation situation as the solution to (4.4) and (5.1), and then simply multiplying by the number of observations m. This is not as surprising as it might at first sight appear, in view of a result described by Snyder (p. 65). The latter tells us that the conditional distribution of $x_1, \ldots, x_m$, given that m observations occur in $[a, b]$ is the same as the distribution of the order statistics of a random sample of size m from the distribution with density

$$f(t) = \rho(t) \Big/ \int \rho(s)\,ds.$$

Our conclusion is, however, difficult to prove without the representation in (7.4).

Note that almost identical results to those indicated above may be obtained by supposing instead that the occurrences cease after the mth observation (rather than assuming that the m occurrences are observed over a fixed interval). The stopping rule for the process is not really important; this is relevant to the numerical example of Section 10.

It might be slightly more reasonable to introduce also prior information about η into the analysis. This could be represented by a normal prior likelihood; the corresponding log-likelihood may be added to the right-hand side of (7.5) before carrying out the maximization

procedure. The present assumption leads us to the condition

$$\int_a^b \tilde{\rho}(t)\,dt = m$$

for our posterior estimate. We feel, however, that the prior information about η would have to be rather strong if it were to have a serious effect on our results. Another possibility would be to assume a second-order autoregressive process in constructing a prior likelihood for g, rather than our present first-order assumptions for $g^{(1)}$.

Our approach in the Poisson process situation may be viewed as a possibly simpler alternative to the existing methods of Snyder (1975), Vere-Jones (1975) and others. These are based upon conceptually similar assumptions but employ rather different technicalities, e.g. involving characteristic functions and Taylor Series expansions.

Clevenson and Zidek (1977) extend Whittle's density estimates to this Poisson situation. This leads to a similar restriction to a class of window functions, whereas our own posterior estimates are not restricted like this.

One interesting result from the theory of stochastic processes is that the stationary first-order autoregressive (or Orstein–Uhlenbeck) process is well known to be the most general Gaussian process which is both stationary and Markov; this provides some justification for our choice of prior model.

8. Change-point Inference

The method described in Section 7 shows promise of capability of generalization to many other problems involving point processes, e.g. forecasting future arrivals, or simultaneous estimation for several processes.

One possible extension of our method is concerned with change-point inference. This has, for example, been considered by Smith (1975) who adopted a Bayesian approach for discrete time models. Suppose here that attention is now restricted to estimates ρ^* for the function ρ which take the form

$$\rho^*(t) = \begin{cases} \rho_1^* & \text{for } t \leqslant t^*, \\ \rho_2^* & \text{for } t > t^*, \end{cases} \qquad \text{for } t \in [a, b], \tag{8.1}$$

where ρ_1^* and ρ_2^* are scalars representing the level of the process before and after the change-point t^*. We recommend choosing ρ_1^*, ρ_2^* and t^* by minimizing the functional

$$l(\tilde{\rho}, \rho^*) = \int_a^b \{\tilde{\rho}(t) - \rho^*(t)\}^2, \tag{8.2}$$

where $\tilde{\rho}$ is obtained by the method described in Section 7, e.g. with a constant prior estimate for ρ, and ρ^* is defined in (8.1). The squared error criterion in (8.2) enables us to find the estimate of the simplified form in (8.1) which is "closest" to the optimal function $\tilde{\rho}$. It yields the equations

$$\rho_1^* = \int_a^{t^*} \tilde{\rho}(t)\,dt/(t^* - a) \tag{8.3}$$

and

$$\rho_2^* = \int_{t^*}^b \tilde{\rho}(t)\,dt/(b - t^*), \tag{8.4}$$

where

$$\tilde{\rho}(t^*) = \tfrac{1}{2}(\rho_1^* + \rho_2^*) \tag{8.5}$$

for the optimal values of ρ_1^*, ρ_2^* and t^*. Equations (8.3), (8.4) and (8.5) may be solved iteratively, using the obvious substitutional procedure for t^*. This provides a conceptually simple alternative to Smith's approach, and is easily generalized to related models, e.g. in discrete time, or to situations where there is more than one change point.

9. Investigating Normality

The numerical examples in the following sections are intended to illustrate our density estimation and Poisson process approaches and to give the flavour of some of the many ways in which they can be applied. Many sensible conclusions can be reached using subjective priors, though these will not of course usually have an objective or frequentist interpretation.

The data in Table 1 were reported by Hald (1967, p. 329) and denote the values of the ranges in terms of percentage concentration of calcium carbonate, for 52 sets of 5 samples each, taken from a mixing plant for raw meal.

Table 1

Ranges for calcium carbonate data

0·32	0·30	0·18	0·38	0·41	0·21	0·23	0·37
0·42	0·56	0·29	0·19	0·41	0·41	0·39	0·37
0·27	0·12	0·12	0·28	0·17	0·39	0·34	0·55
0·21	0·32	0·38	0·26	0·55	0·16	0·31	0·31
0·37	0·30	0·10	0·24	0·17	0·40	0·24	0·13
0·18	0·54	0·68	0·61	0·14	0·49	0·28	0·36
0·32	0·28	0·22	0·21				

On p. 322, Hald introduces a limit thorem to justify theoretically a normal distribution of the ranges under appropriate assumptions for the distribution of the original observations. Investigation is confined here to the distribution of the ranges, without referring to the original observations.

Firstly, suppose that the prior mean value function μ is identically zero, with f concentrated on the interval [0, 1], implying a uniform density over this interval as prior estimate for f. This preliminary choice is made to concentrate attention on the data and on the effect of choices of the prior parameters σ^2 and β on the shape of the smoothed density.

The curves in Fig. 2 depict the posterior estimates for f for various choices of β, keeping $\tau = \sigma^2/\beta^2$ constant. When $\beta = 10$ the estimated density is fairly flat, since substantial account is taken of the covariance kernel in (2.4). As β increases the smoothing becomes less predominant, but the smoothed estimate is still a remarkably well-behaved function. The thickness of the right tail of the density is a consequence of the particular choice of prior estimate for f.

In Fig. 3 the parameter β is kept fixed and equal to 15. As τ increases the pulling in effect towards the prior estimate decreases; when for example $\tau = 0{\cdot}03$, the smoothed estimate again appears to be a fairly sensible function.

In Fig. 4 we investigate the consequences of taking the prior estimate

$$\xi(t) = e^{\mu(t)} \Big/ \int e^{\mu(s)}\, ds$$

for f to represent a normal curve, but truncated to the interval [0, 1]. The mean and variance of the prior estimate were estimated empirically from the data. (It might be possible to introduce prior information about the mean and variance but this would cause extra complications.) One posterior estimate takes $\beta = 50$ and $\tau = 0{\cdot}01$, i.e. $\sigma^2 = 25$; note from Fig. 2

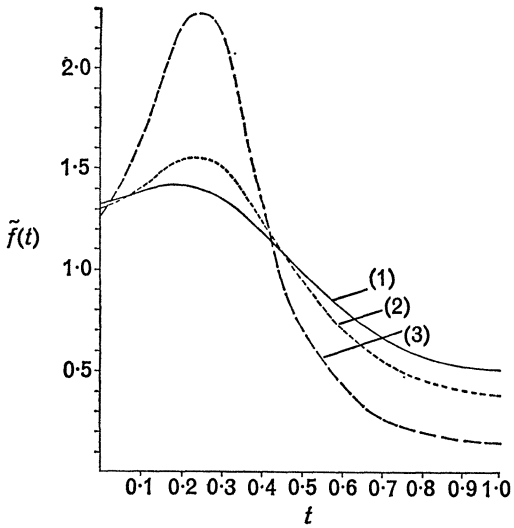

FIG. 2. Smoothed estimates for calcium carbonate data with values (1) 10, 0·01, (2) 15, 0·01 and (3) 50, 0·01 for β and τ.

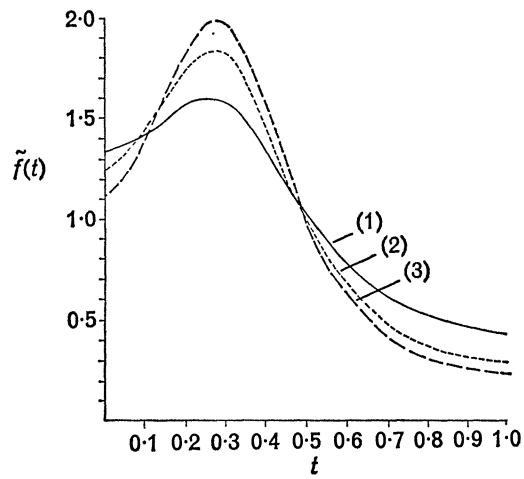

FIG. 3. Smoothed estimates for calcium carbonate data with values (1) 15, 0·01, (2) 15, 0·02 and (3) 15, 0·03 for β and τ.

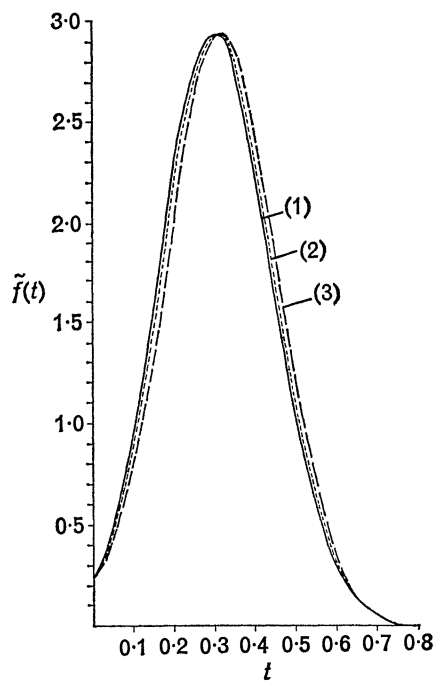

FIG. 4. Investigating normality for calcium carbonate data with values (1) 50, 0·01 and (2) 15, 0·02 for β and τ compared with normal curve (3).

that this fairly weak choice leads to substantial differences from the prior estimates assumed there.

Under our normal choice of prior estimate for f the posterior estimates are very close to the prior estimate, in particular the tails of the smoothed estimates are much thinner than in Fig. 1. Similar sorts of results were obtained for several other choices of β and σ^2, as long as β was not chosen to be too small, in which case over-flattening took place.

Our results suggest to us that there is virtually nothing in this particular data set to dispute a prior assumption of normality. However, similar conclusions were reached under prior assumptions of, for example, a Gamma or Beta curve for f. This suggests that the data set under consideration is simply unable to distinguish between sensibly chosen alternatives, although it is able to discard choices like the uniform distribution over [0, 1]. It is therefore necessary to introduce substantive prior information in order to obtain particular posterior conclusions from this data set.

A second chemical example concerns 22 silica assays of chondrite meteors, reported in graphical form by Ahrens (1965), and previously analysed by Burch and Parsons (1976), using squeeze statistics. The observations were all confined to the interval [20, 36]; for ease of analysis the original observations were re-scaled to the interval [0, 1], providing the 22 entries to Table 2.

TABLE 2

Silica assays of chondrite meteors

0·04	0·15	0·16	0·18	0·40	0·44	0·45	0·46	0·47	0·49	0·54
0·59	0·64	0·75	0·81	0·83	0·84	0·84	0·85	0·87	0·87	0·92

Burch and Parsons investigate the possibility that these data are normally distributed, and their squeeze significance test rejects this null hypothesis. We again proceed under the assumption that the prior estimate ξ for f is a normal curve, with mean and variance estimated empirically from the data.

In Fig. 5 the posterior estimate for $\beta = 50$ and $\tau = 0{\cdot}01$ is rather interesting. It is fairly skew, slightly bimodal and possesses a thick left tail. The alternative choice $\beta = 15$ and

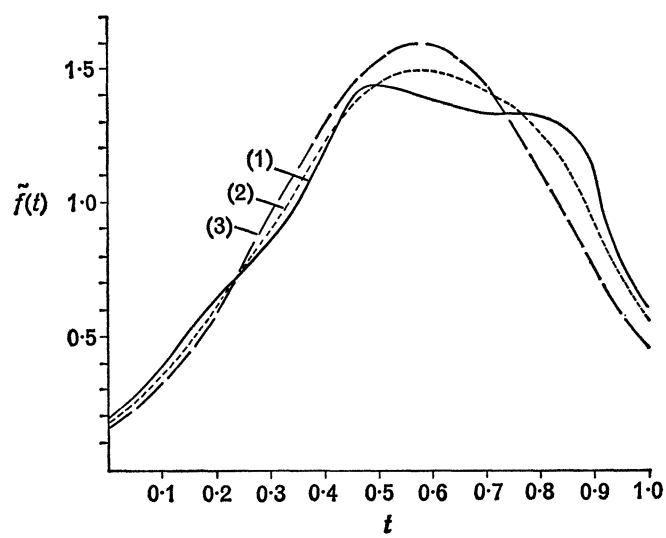

FIG. 5. Investigating normality for chondrite meteor data with values (1) 50, 0·01 and (2) 15, 0·02 for β and τ compared with normal curve (3).

$\tau = 0{\cdot}02$ leads to stronger smoothing, in particular removing the bimodality, but has a prominent bulge instead of the second mode. Our analysis with these and other prior parameter values suggests to us that unless there is strong prior information to the contrary, we would support Burch and Parsons in the sense that the data provide noticeable evidence that they may not be normally distributed.

Good and Gaskins recently analysed the above data using their method based upon roughness penalties. With their own prior parameter β equal to 0·0375 they obtained a trimodal density resembling a mixture of three normal densities. With $\beta = 0{\cdot}4$ they obtained a bumpy unimodal density which slightly resembles the bimodal curve in Fig. 5. Good and Gaskins hope to report their work in detail in a separate paper; one point worth noting is that they did not incorporate a normal prior estimate for f into their analysis.

9. Pedestrian Arrivals at a Pelican Crossing

The data in Table 3 were collected by Griffiths and Cresswell as background to their recent development (Griffiths and Cresswell, 1976) of a mathematical model for Pelican crossings. One assumption underlying their model is that pedestrians arrive in a Poisson process with constant rate. We employ here the more general assumptions of Section 7 which permit the rate $\rho(t)$ to vary with time. The arrivals reported here took place during a period of 348 seconds; the entries in Table 3 are normalized to lie between zero and one. (This is in fact the interesting part of a slightly larger data set from which we have selected the first 45 observations; as indicated in Section 7, the stopping rule is not really important.)

Table 3

Forty-five pedestrian arrival times
(as proportions of 348 seconds)

0·03	0·05	0·07	0·11	0·13	0·20
0·28	0·28	0·30	0·30	0·31	0·33
0·35	0·35	0·36	0·36	0·36	0·39
0·39	0·41	0·42	0·42	0·46	0·47
0·49	0·63	0·64	0·64	0·66	0·70
0·70	0·70	0·73	0·75	0·76	0·76
0·78	0·78	0·79	0·82	0·87	0·87
0·93	0·93	1·00			

The arrivals are concentrated in two groups, with a long uninterrupted period between 0·49 and 0·63. The smoothed estimates of the rate $\rho(t)$ take account of this in sensible fashion; the main feature of the curves in Fig. 6 (based on a constant prior estimate for ρ) is their

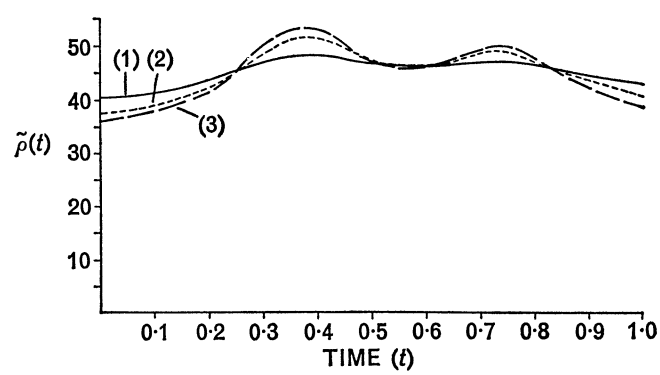

Fig. 6. Smoothed rates for Pelican crossing data with values (1) 15, 0·01, (2) 15, 0·02 and (3) 15, 0·03 for β and τ.

appodality. The fluctuations of ρ will hopefully not be pronounced enough to affect the local bimlicability of Griffith's and Cresswell's model.

11. Histogram Smoothing for Mine Explosions Data

On p. 17, Boneva *et al.* consider data previously analysed by Maguire *et al.* (1952) and concerning inter-arrival times between severe explosions in mines, in Great Britain between December 6th, 1875 and May 29th, 1951. Maguire *et al.* fit an exponential density to a 55-cell histogram, with interval width 30 days, since the largest inter-arrival time recorded was 1,650 days. Boneva *et al.* reduce this to a 20-cell histogram, presumably in order to increase the smoothness of their histospline, but still end up with a fluctuating curve which assumes negative values in three different cells, and possesses about six different modes. We consider the 55-cell histogram of Maguire *et al.*; again for comparative purposes the interval [0, 1,650] was rescaled to [0, 1] before commencing the analysis.

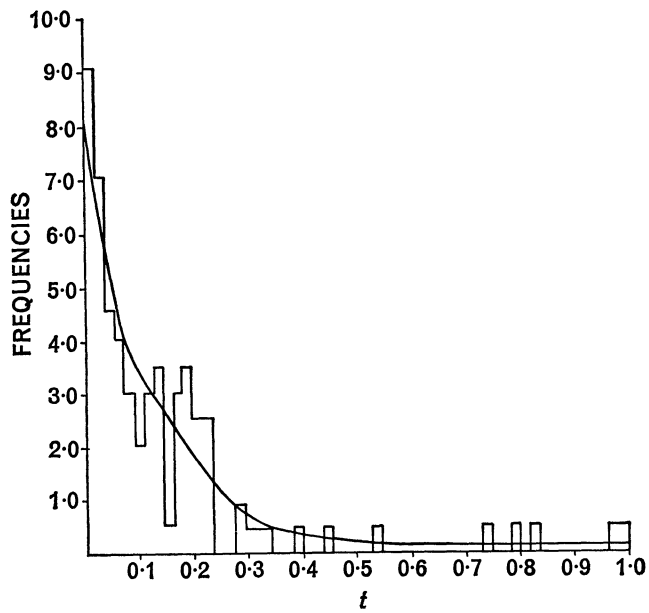

FIG. 7. Distribution of inter-arrival times for mine explosions data.

As prior estimate for f, we took the exponential density with mean estimated empirically from the data, and truncated to the interval [0, 1]. Various values were tried for β and σ^2, with similar qualitative results. In Fig. 7 we report the posterior estimate for f under the choice $\beta = 50$ and $\tau = 0{\cdot}02$, and compare this estimate with the histogram of observations. The estimate is positive even for cells with zero cell frequencies, and it is a smooth, well-behaved function. It has a higher peak and thicker tail than an exponential density, together with a noticeable bulge around $t = 0{\cdot}15$. Note that the smoothness does not depend upon our particular choice of prior estimate for f; a similarly well-behaved, though appreciably flatter, function was obtained under a uniform prior estimate.

12. Concluding Remarks

12.1. *Choice of Prior Parameters*

Our method is subjective and the posterior estimates depend substantially upon the particular choice of prior parameters μ, β and σ^2.

We feel that the statistician should often be able to use the background of his problem to make an appropriate choice of the mean value function μ, or the corresponding prior estimate ξ for f, e.g. this could be based upon a density from a parameterized family, with its parameters estimated empirically from the data, providing an alternative to standard hypothesis testing techniques. The estimate ξ may correspond to a null hypothesis for f; our method helps the statistician to reconsider this hypothesis in the light of the data. The choices of β and σ^2 may need more experience (remember that β measures the closeness of neighbouring values of $g^{(1)}$ and governs the smoothness of the posterior estimate, whilst σ^2 measures the closeness of $g^{(1)}$ to the prior estimate $\mu^{(1)}$.) These parameters are best chosen after experience with a number of related data sets; we find that the choices $\beta = 15$ and $\sigma^2/\beta^2 = 0{\cdot}02$ often provide a sensible starting point when f is concentrated on $[0, 1]$.

We find it helpful to rescale $[a, b]$ to $[0, 1]$ since this enables us to compare prior parameters from different data sets. It is, however, sometimes useful to also think in terms of the prior parameters corresponding to the original interval $[a, b]$, since these may yield an interpretation in the real world.

12.2. *Estimation of Prior Parameters*

It might be possible to assign distributions to β and σ^2 in the prior model, leading to a hierarchical prior approach. Alternatively, β and σ^2 could be estimated empirically from the data. It is rather difficult to obtain estimates which are consistent for β and σ^2 as $m \to \infty$. The author is currently researching this problem by investigating the marginal likelihood of β and σ^2, and hopes to report the results at a future date.

12.3. *Discussion of an Approach by Wahba*

Wahba (1976) develops approximate estimates for f based upon Fourier expansions and a Bayesian argument involving a covariance kernel (and a uniform prior estimate for f). She is able to estimate one of the prior parameters in her covariance kernel using a non-Bayesian cross-validatory argument. Her method parallels ours; readers may wish to compare the two approaches for themselves.

12.4. *Generalization to Autoregressive Processes of Higher Order*

The reproducing inner product for an autoregressive process of general order n is for example described by Parzen (1961, p. 476). If this is employed instead of (3.1), a more complicated analysis ensues, and the corresponding posterior estimate for f possesses a continuous $2n$th derivative for ungrouped observations, and a continuous $(2n+1)$th derivative for grouped observations. We however feel that the first-order process will be adequate for many needs, since a requirement of continuity beyond the second derivative is not of obvious importance.

12.5. *Miscellaneous Remarks*

There are possibilities of extending our approach to deal with circular data; this would avoid the necessity for the choice of an interval $[a, b]$ and would provide an alternative to the circular histospline method of Boneva *et al.*

Our main method depends of course upon the choice of the interval $[a, b]$. This may in particular affect the tail area behaviour of the estimate $\tilde{f}$; this does not unduly concern us, because tail area behaviour will anyway mainly depend upon the particular form of the prior estimate assumed for f. Whilst letting $a \to -\infty$ or $b \to \infty$ in our analysis would affect the rigour of some of our arguments, this would appear to be fairly reasonable intuitively, but fresh problems would be caused in computing a numerical solution. The author is currently investigating this problem further in the hope of using the method for coping with outlying observations.

ACKNOWLEDGEMENTS

The author wishes to thank Professor D. V. Lindley for originally suggesting the histogram problem, and for passing on the idea that autoregressive covariance kernels for derivatives might be appropriate. Professor P. J. Harrison introduced the author to the general concepts underlying prior likelihood, and the Kalman filter, and Dr J. K. Ord and Dr A. O'Hagan provided helpful comments. Thanks are also due to Professor J. M. Dickey for his inspiration and encouragement during the early days of the research, and to Professor I. J. Good for a fruitful correspondence. The author is grateful to Dr J. D. Griffiths; and Mr I. T. Parsons and Dr C. R. Burch for providing the original data sets for the Pelican crossings and chondrite meteor examples. Two referees were particularly helpful in suggesting improvements to the exposition and theoretical content of the paper. Many comments have been received from members of the Statistics Departments at Warwick and University College London, and during seminar visits to Aberystwyth and UMIST; my thanks to everybody who has shown an interest.

REFERENCES

AHRENS, L. A. (1965). Observations on the Fe–Si–Mg relationship in chondrites. *Cosmochem. Geochim. Acta*, **29**, 801–806.

BELLMAN, R. (1976). *Introduction to the Mathematical Theory of Control Processes*, Vol. 1. New York and London: Academic Press.

BONEVA, L. I., KENDALL, D. G. and STEFANOV, I. (1971). Spline transformations: three new diagnostic aids for the data analyst (with Discussion). *J. R. Statist. Soc.* B, **33**, 1–72.

BURCH, C. R. and PARSONS, I. T. (1976). "Squeeze" significance tests. *Appl. Statist.*, **25**, 287–290.

CLEVENSON, M. L. and ZIDEK, J. W. (1977). Bayes linear estimators of the intensity function of the nonstationary Poisson process. *J. Amer. Statist Ass.*, **72**, 112–120.

COX, D. R. and MILLER, H. D. (1970). *The Theory of Stochastic Processes.* London: Methuen.

DICKEY, J. M. (1969). Smoothing by cheating. *Ann. Math. Statist.*, **40**, 1477–1482.

EDWARDS, A. W. F. (1972). *Likelihood.* London: Cambridge University Press.

FERGUSON, T. S. (1973). A Bayesian analysis for some non-parametric problems. *Ann. Statist.*, **1**, 209–230.

GOLDSTEIN, M. (1975). A note on some Bayesian non-parametric estimates. *Ann. Statist.*, **3**, 736–740.

GOOD, I. J. and GASKINS, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

GRIFFITHS, J. D. and CRESSWELL, C. (1976). A mathematical model of a Pelican crossing. *J. Inst. Maths. Applics*, **18**, 381–394.

HALD, A. (1967). *Statistical Theory with Engineering Applications.* New York and London. Wiley.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D*, **82**, 35–45.

KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation in stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41**, 495–562.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika*, **60**, 297–308.

MAGUIRE, B. A., PEARSON, E. S. and WYNN, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika*, **39**, 168–180.

MONTRICHER, E. F. DE, TAPIA, R. A. and THOMPSON, J. R. (1975). Non-parametric maximum likelihood estimation of probability densities by roughness penalty methods. *Ann. Statist.*, **3**, 1329–1348.

PARZEN, E. (1961). Regression analysis of continuous parameter time series *Proc. 4th Berkeley Symp.*, Vol. 1. pp. 469–489. Berkeley, Calif.: University of California Press.

—— (1962). On estimation of a probability density and mode. *Ann. Math. Statist.*, **33**, 1065–1076.

—— (1977). Non-parametric data science: a unified approach based on density estimation. *SUNY at Buffalo*, unpublished report.

SILVERMAN, B. W. (1978). Density ratios, empirical likelihood and co death. *Appl. Statist.*, **27** (in press).

SMITH, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, **62**, 407–416.

SNEDECOR, G. W. and COCHRAN, W. G. (1967). *Statistical Methods.* Ames, Iowa: Iowa State University Press.

SNYDER, D. L. (1975). *Random Point Processes.* New York: Wiley.

VERE-JONES, D. (1975). On updating algorithms and inference for stochastic point processes. In *Perspectives in Probability and Statistics* (J. Gani, ed.), pp. 239–259. London: Academic Press.

WAHBA, G. (1975). Interpolating spline methods for density estimation. *Ann. Statist.*, **3**, 30–48.

—— (1976). Optimal smoothing of density estimates. *Department of Statistics, University of Wisconsin–Madison.* Technical Report No. 469.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. R. Statist. Soc.*, B, **20**, 334–343.

APPENDIX

Derivation of Integral Equation—Mathematical Details

Suppose that $\tilde{g}$ provides a maximum of $L_1(g)$ in (4.1), and that $\tilde{g}$ is perturbed to $\tilde{g}+\varepsilon u$ where ε is a scalar and u is a function on $[a,b]$. Then it follows from (3.1), (3.3) and (4.2) that

$$\left.\frac{\partial L(\tilde{g}+\varepsilon u)}{\partial \varepsilon}\right|_{\varepsilon=0} = \int_a^b u(t)\,\phi(t)\,dt - m\int_a^b u(t)\exp\{\tilde{g}(t)-D(\tilde{g})\}\,dt - (2\beta\sigma^2)^{-1}\int_a^b u^{(2)}(t)\,\lambda^{(2)}(t)\,dt$$

$$-(\beta/2\sigma^2)\int_a^b u^{(1)}(t)\,\lambda^{(1)}(t)\,dt - (2\sigma^2)^{-1}u^{(1)}(a)\,\lambda^{(1)}(a) - (2\sigma^2)^{-1}u^{(1)}(b)\,\lambda^{(1)}(b), \qquad (1)$$

where $\lambda = \tilde{g}-\mu$.

Repeated integration by parts for the third and fourth contributions to the right-hand side of (1) gives, after some rearrangement,

$$\left.\frac{\partial L(g+\varepsilon u)}{\partial \varepsilon}\right|_{\varepsilon=0} = \int_a^b u(t)\,\Lambda(t)\,dt - G_{ab}, \qquad (2)$$

where

$$\Lambda(t) = m\phi(t) - m\exp\{\tilde{g}(t)-D(\tilde{g})\} - (2\beta\sigma^2)^{-1}\lambda^{(4)}(t) + (\beta/2\sigma^2)\,\lambda^{(2)}(t) \qquad (3)$$

and

$$2\beta\sigma^2 G_{ab} = -u(b)\,r_a(b) + u(a)\,r_a(a) + u^{(1)}(b)\,r_1(b) - u^{(1)}(a)\,r_1(a) \qquad (4)$$

with $r_0(b)$, $r_0(a)$, $r_1(b)$ and $r_1(a)$ denoting the quantities on the respective left-hand sides of (4.6).

For $\tilde{g}$ to provide a maximum of $L_1(g)$ the quantity in (2) must be zero for all choices of u. It therefore follows from (4) that the boundary conditions in (4.6) must be satisfied, and more importantly from (3) that $\tilde{g}$ must satisfy (4.3), with f and ω defined in (4.4) and (4.5) respectively.

We now show how to convert the differential equation in (4.3) into the integral equation in (5.1). In terms of ω in (4.5), the second derivative $g^{(2)}$ satisfies

$$\tilde{g}^{(2)}(t) - \mu^{(2)}(t) = A_1 e^{\beta t} + A_2 e^{-\beta t} - \int_a^b M(t,u)\,\omega(u)\,du, \qquad (5)$$

where A_1 and A_2 are constants to be determined from the boundary conditions, and

$$M(t,u) = (2\beta)^{-1}\exp\{-\beta|t-u|\} \qquad (6)$$

is the Green's function associated with the operator on the left-hand side of the second-order differential equation

$$Z^{(2)}(t) - \beta^2 Z(t) = \omega(t) \quad \text{for } t\in[a,b]. \qquad (7)$$

The expression on the right-hand side of (5) provides the general solution for Z to (7), and hence the solution for $g^{(2)}-\mu^{(2)}$ for the relationship in (4.5).

Integrating (5) twice with respect to t gives

$$\tilde{g}(t) - \mu(t) = \beta^{-2}A_1 e^{\beta t} + \beta^{-2}A_2 e^{-\beta t} + B_1 t + B_2 - \int_a^b M^*(t,u)\,\omega(u)\,du, \qquad (8)$$

where

$$2\beta^3 M^*(t,u) = \exp(-\beta|t-u|) - \exp\{-\beta(u-a)\} - \beta(t-a)\exp\{-\beta(u-a)\} + 2\beta\{\max(t,u)-u\}. \qquad (9)$$

The integration constants A_1, A_2 and B_1 in (8) may be determined from the boundary conditions in (4.6) giving $A_1 = A_2 = 0$, and

$$2\beta^2 B_1 = -\int_a^b \exp\{-\beta(u-a)\}\,\omega(u)\,du.$$

The value of B_2 will not affect $\tilde{f}$ in (4.4), and we make the particular choice $B_2 = (a-\beta^{-1})B_1$. From (8) and (9) we therefore have after some manipulation that

$$\tilde{g}(t)-\mu(t) = -(2\beta^3)^{-1} Q_t(\omega), \tag{10}$$

where Q_t is defined in (5.2). Note from (4.3) that $\omega = 2m\beta^2(\phi-\tilde{f})$. Substituting for ω in (10) yields the integral equation in (5.1).

Discussion of Dr Leonard's Paper

Professor J. M. Dickey (University College of Wales, Aberystwyth): This paper by Dr Leonard reports his work over many years on Bayesian inference about a probability density. Other workers have considered the problem. Dr Leonard's efforts appear to have been successful.

I recall discussions with him early in this decade on the need to model uncertainty about an unknown probability density in terms of a random process $f(t)$ subject to the following requirements:

1. $f(t) \geqslant 0$ for all t, with probability one.
2. $\int f(t)\,dt = 1$, with probability one.
3. High local dependence, for smooth realizations.
4. Tractability of Bayes's theorem acting on the probability structure of the process.
5. Availability of descriptive quantities on the process, such as the mean function for estimation of $f(t)$ and/or for prediction of t.

Following earlier work of Professor Lindley and of Dr Leonard on inference about a probability mass function, the device of working with a process g having a many-to-one mapping to f has now met requirements 1 and 2. The process g (or rather $g^{(1)}$ is then given a Gaussian structure, to satisfy 3. Item 4 follows then essentially by combining two quadratic forms. And 5 now has a function-space analogue of the posterior mode.

In considering this paper it may be useful to separate the essentials of the inference problem from the technicalities of infinite dimensions. A finite-dimensional analogue would involve a histogram with unknown cell probabilities, the prior structure for which is induced by Dr Leonard's prior structure on an underlying unknown density. The resulting posterior structure on the cell probabilities should be the same as that induced by Dr Leonard's posterior structure on the density, since the likelihood from histogram data depends on the unknown density only through the cell probabilities. One would like to know how this relates to Dr Leonard's previous work on Bayesian inference for histograms. Also, what happens as the cell size is made finer?

An estimation utility structure (negative of loss) can be defined as the indicator of a fixed-width hyperinterval whose centre is the estimate. Then by an application of the Mean Value Theorem for integrals and a smoothness condition on the posterior density, the posterior expected utility can be approximated uniformly by the posterior density multiplied by the fixed hypervolume (when volume exists). Hence, the posterior mode gives an approximate maximum of the posterior expected utility, with the approximation improved by decreasing the size of the hyperinterval. Note the influence of the choice-of-variable on the utility structure through the fixed-width requirement.

Traditionally, the posterior density is defined with respect to a volume or Haar measure as natural dominating measure. The Haar-group invariance implies that this too is equivalent to a choice of variable. However, the posterior distribution and expected utility exist independently of the dominating measure, and hence the posterior density maximization, or posterior mode, is not strictly necessary to the Bayesian analysis. In infinite dimensions the volume measure does not exist, and so interest centres on the expected utility analysis. What happens in the finite-dimensional case as the dimensionality is increased, that is, as the cell size is made finer?

The Bayesian interpretation of maximum likelihood is as an approximate posterior mode from an approximately constant prior density. (The dominating measure might seem arbitrary here, but it should relate to the utility structure as aforesaid, in order that the posterior mode and the maximal expected utility relate.) Use of "prior likelihood" merely calls for a constant "prior prior density".

In high dimensions constant prior densities are known to cause trouble. In infinite dimensions the Bayesian interpretation of maximum likelihood appears to be lost by the non-existence of volume. If so, why use a prior-likelihood approach here?

The author seems to use simultaneously two different prior distributions when treating the change-point problem. And again, when treating the hypothesis-testing problem. There are also suggestions to "cheat" by using a single data set in more ways than in a single application of Bayes's theorem. Much research is needed in general on the practical import of such usages.

Perhaps Dr Leonard could provide help on the choice of β and σ values, say in terms of interpretation of the prior process, since these prior parameters are crucial to the amount of smoothing.

In the light of the intrinsic interest of the topic of this paper and the impressive results therein, I have great pleasure in proposing the vote of thanks.

Mr B. W. SILVERMAN (University of Oxford): I should very much like to add my congratulations to Dr Leonard for his excellent paper which has put density estimation and prior information so firmly on the same map. None of the available density estimation methods allows the explicit incorporation of prior information, and so Dr Leonard has made a most valuable contribution.

One point I should like to take up is Dr Leonard's remark that his method will work well for smaller samples than other methods can sensibly cope with. There is, of course, a large amount of information (speaking loosely) in a density curve, and so one cannot hope to be able to reconstruct the whole function from as small a sample as 22 observations, as in the chondrite meteor example. However, there is no difficulty if we have prior information *and* 22 observations; to be honest, therefore, Dr Leonard is solving a different problem. If all we have are the 22 observations, with no prior information, I think that what we should do for diagnostic purposes is not to attempt to estimate the density as a curve but to present the data as a stem-and-leaf plot as in Table D1; constructing this table took only a few seconds and required no computing. With such a small sample it would perhaps be unwarranted to proceed any further with diagnostics.

TABLE D1

Stem and leaf plot of chondrite meteor data

0	4
1	5 6 8
2	
3	
4	0 4 5 6 7 9
5	4 9
6	4
7	5
8	1 3 4 4 5 7 7
9	2

If the sample size is larger, then density estimation comes into its own as a diagnostic technique. Dr Leonard's largest sample is the data of calcium carbonate ranges. For comparison I have computed a kernel estimate for these data, choosing the window width using my "test graph" method. This rule of thumb involves examination of test graphs of the second derivative of the density estimate for various window widths; the best window width for estimating the density produces a test graph where there is a marked amount of random fluctuation but the systematic variation is still clearly visible. Using terms suggested by Professor Kendall and Dr O'Hagan, the ideal test graph contains "rabbits" but not "hedgehogs"! For further details see Silverman (1978b); I have used the kernel defined in that paper.

The estimate for the calcium carbonate data is given in Fig. D1. It can be seen that the thickness of the right-hand tail in Fig. 2 is, in part at least, due to the data rather than the prior estimate. The kernel estimate shows that there is some skewness or even bimodality evident in the data, though these do not seem marked enough to suggest a significant departure from normality. This last remark is moving away from diagnosis towards analysis, for which procedures other than density estimation should be considered.

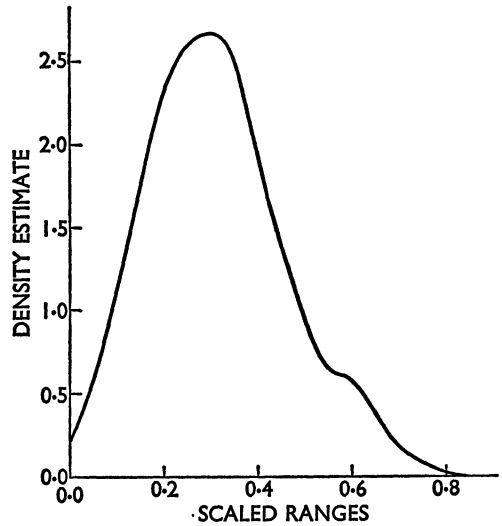

FIG. D1. Kernel density estimate for calcium carbonate ranges data, window width 0·093.

One of the advantages of the kernel method over Dr Leonard's method is, of course, that much less computer time is required to calculate the estimates. Against this must be weighed the fact that Dr Leonard's method stands on a much firmer logical basis than other methods. There are two choices to be made when estimating probability densities, or when performing any other kind of smoothing; those of choosing *how* to smooth and *how much* to smooth. In practice the first choice may not matter all that much. The work of Epachenikov (1969) and Good and Gaskins (1971), among others, shows that any reasonable kernel or roughness penalty can be used, and indeed that even the choice of method is relatively unimportant. The problem of how much to smooth remains, at least implicitly, whatever method we choose. In the absence of any stronger reasons, it may be best to use the method which is quickest and easiest, even if this means sacrificing logical foundation. In the multivariate case our choice of methods is severely restricted and almost the only feasible method at present is the kernel method. Dr Leonard's integral equation (5.2) looks as though it could be generalized to deal with multivariate data; could his method be thus extended?

A very important use of density estimation is in presentation. Probability densities are widely understood by non-statisticians, and so density estimates are a valuable way of presenting data, together with results which may possibly have been validated by other methods. If we wish to incorporate prior information into our presentation, we now have an excellent method of doing so. The use of density estimation in this context should not be under-estimated; I am sure that most of us would explain (for example) the Normal distribution to a layman by drawing a bell-shaped curve.

I am sure that we will all want to thank Dr Leonard for a most interesting and stimulating paper, and I have very great pleasure in seconding the vote of thanks to him tonight.

The vote of thanks was carried by acclamation.

Dr A. O'HAGAN (University of Warwick): This is a very important paper, showing as it does how we can employ prior information in the estimation of densities—or does it? Dr Leonard begins by defining a Gaussian prior distribution $P_{\mu,K}$ to give $g^{(1)}$ a prior mean $\mu^{(1)}$ and covariance kernel K. But he then adopts instead a "prior likelihood" based on the measure $M_{g,K}$ with mean $\mu^{(1)}$ and kernel K over the space inhabited by $\mu^{(1)}$, which he then regards as "representing similar prior information". I wish to examine this claim: Dr Leonard certainly interprets $\mu^{(1)}$ and K as representing prior means and covariances, but with what justification?

His "prior likelihood" is not $M_{g,K}$, which he calls P, but its Radon–Nikodym derivative with respect to the corresponding measure $M_{0,K}$ with zero mean function,

$$dM_{g,K}/dM_{0,K} = \exp\{(\mu^{(1)}, g^{(1)}) - \tfrac{1}{2}(g^{(1)}, g^{(1)})\}.$$

He uses this likelihood because it is analytically simpler than a Bayesian analysis based on the prior density

$$dP_{\mu,K}/dP_{0,K} = \exp\{(\mu^{(1)}, g^{(1)}) - \tfrac{1}{2}(\mu^{(1)}, \mu^{(1)})\}.$$

If the two approaches are equivalent, i.e. if Leonard's posterior likelihood is proportional to the posterior density with respect to $P_{0,K}$ for a Bayesian whose prior is $P_{\mu,K}$, then Leonard's estimator is the corresponding posterior mode.

But the dependence of $P_{0,K}$ on K casts doubt on the value of this mode. Use of a dominating measure which depends on the prior reduces the impact of the prior on the posterior density, as can be seen if we take $P_{\mu,K}$ itself as the dominating measure, since then the prior and posterior densities are independent of $\mu^{(1)}$ and K.

In fact it can be shown that Dr Leonard's posterior likelihood is proportional to a genuine posterior density obtained from the prior $P_{\mu,K}$ but it is the density with respect to a measure Q_K defined by

$$Q_K(G) = \int_G \exp\{\tfrac{1}{2}(g^{(1)}, g^{(1)})\}\, dP_{0,K}$$

for appropriate "Borel" sets G in the $g^{(1)}$-space. So a Bayesian *can* use the estimates of this paper, interpreting them as posterior modes, but what of the fact that Q_K apparently still depends on K? Are we still discarding prior information? I do not have a complete answer, but insight is obtained by pursuing the analogous arguments in R^n. The Gaussian measures are multivariate normal distributions and Q_K turns out to be Lebesgue measure (independent of K), whereas use of $P_{0,K}$ as dominating measure leads to strange posterior modes. Although I am told that no Lebesgue measure exists for the problem at hand, I imagine that the dependence of Q_K on K is at most minimal.

Dr J. D. GRIFFITHS (UWIST, Cardiff): Since the time I supplied the data given in Table 3 to Dr Leonard, some new information has come to light. We have recently found from a much larger set of similar data that definite periodicities in pedestrian inter-arrival times are observed, with high flows of pedestrians corresponding to particular parts of the Pelican crossing cycle. As an example, at one site where observations were taken, the "Green Man" signal, which allows pedestrians to start to cross, occupied only 12 per cent of the total cycle time, but the number of pedestrians actually arriving during this period averaged 29 per cent of the total. The reason (we think!) is that pedestrians in the vicinity of the crossing are attracted by the audible bleep, which coincides with the "Green Man" signal, and hurry to cross before the signal finishes, thus inflating the value of ρ. In contrast the "Flashing Green Man" signal, during which pedestrians are not allowed to start to cross, occupied 20 per cent of the cycle time but received only 4 per cent of pedestrian arrivals on average.

This explanation may account at least partially for the bi-modality found by Dr Leonard, although it should be pointed out that the data he quotes covers about seven cycles of the Pelican crossing signals and one might thus have expected more than two modes to be present.

It would thus appear that, even if the mean arrival rate per cycle remained constant, there would still be evidence of variation in ρ with time because of the changes in flow rates within a cycle. Presumably, it would not be difficult to take account of these findings in the prior information used to provide the smoothest estimate for ρ.

Mr R. BIRCH (University College of Wales): In Section 9, numerical examples concerned with the investigation of normality are given. If we are to investigate specifically for normality, it would seem reasonable to assume *a priori* that we have some information that f is likely to be a symmetric function. However, if f is to be a symmetric function then the following holds: Let m be the symmetry point and k a constant such that $[m-k, m+k] \subseteq [a, b]$ then

$$g^{(1)}(m-k) = \frac{f^{(1)}(m-k)}{f(m-k)} = \frac{-f^{(1)}(m+k)}{f(m+k)} = -\,g^{(1)}(m+k). \tag{1}$$

Equation (1) shows that the covariance kernel has the value $K(m-k, m+k) = -\sigma^2 \neq \sigma^2 \exp(-2\beta k)$. The above holds equally for the prior estimates ξ and μ.

The obvious extension to Dr Leonard's covariance kernel takes the form

$$K(s, t) = \sigma^2 \operatorname{sgn}\left(\frac{s-m}{t-m}\right) \exp\{-\beta \,\|\, s-m \,|-|\, t-m \,\|\}. \tag{2}$$

This reduces to Dr Leonard's covariance kernel if both s and t lie on the same side of m, otherwise it is negative and related to the exponential of their relative distances from the symmetry point.

This and any other sensible covariance kernel must, however, have a discontinuity at m, because of the sign change, unless it is zero at the point m. Parzen (1961b, p. 979) shows that equation (3.2) will only hold if K is continuous, since $C[a, b]$ is a separable metric space. Thus we can no longer find the covariance kernel in the manner of the paper.

I should be interested to hear any comments Dr Leonard may have on the above.

Dr J. K. ORD (University of Warwick): I should like to join earlier speakers in complimenting Dr Leonard upon his stimulating paper. Having seen the development of this work through workshop, seminar and discussions over coffee I can safely say that after a long pregnancy a sound and robust infant has emerged, further helped by the manner of its delivery.

However, there is one thing which concerns me about the procedure. If we consider any monotone transformation of the random variable then the logarithm of the density function, $g(t)$, changes in an obvious way, but the covariance kernel loses its simple form. Thus, for the transform $t = h(x)$, $s = h(y)$ we have a new kernel of the form

$$K^*(x, y) = \sigma(x, y) \exp\{-\beta\,|\,h(x)-h(y)\,|\,\}$$

so that we may lose both the constant variance and the simple exponential decay structure. Turning to Section 3, we see that the prior information is incorporated by means of a particular density function. With a normal prior, the kernel in equation (2.4) seems appropriate, but is this necessarily so for other prior functions? Do we need to think in terms of an initial transformation of the random variable to justify the type of kernel used or can we take refuge in Dr Silverman's comment to the effect that this is unimportant?

When analysing a set of data, if I start with initial views about appropriate distributions, I would want to pursue these ideas rather than finish up with a non-parametric estimate. Perhaps Dr Leonard could indicate situations where he feels that non-parametric estimates are more appropriate?

Professor P. WHITTLE (Cambridge University): I find Dr Leonard's suggestion of formulating one's prior hypotheses in terms of the logarithm of frequency (or of intensity) a most attractive and natural one. However, his derivation of his integral equation (5.1) is quite unnecessarily cumbersome —one can obtain the analogue of this equation for a general prior covariance structure without any operator inversion whatsoever.

Let us consider the case of the non-homogeneous Poisson process, to avoid normalization troubles, and discretize the problem, by breaking the x-axis up into cells, with $\rho_j = e^{\theta_j}$ being the the expected number of observations in the jth cell. Let $\mathbf{g}$, the vector of g_j's, have a prior distribution which is normal, with mean $\boldsymbol{\mu} = (\mu_j)$ and covariance matrix $\mathbf{V} = (v_{jk})$. The negative logarithm of the joint density of the g_j and of the numbers n_j sampled in the different cells is then

$$\text{const} + \tfrac{1}{2}(\mathbf{g}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathbf{V}^{-1}(\mathbf{g}-\boldsymbol{\mu}) + \sum e^{\theta_j} - \sum n_j g_j.$$

The value $\tilde{\mathbf{g}}$ maximizing this is readily found to satisfy

$$\tilde{g}_j = \mu_j + \sum_k v_{jk}\{n_k - \exp(\tilde{g}_k)\}, \qquad (*)$$

which is exactly the analogue of Dr Leonard's equation (5.1), for the discrete case with general $\mathbf{V}$. (In forming his "reproducing inner product" (3.1) and converting his differential equation to an integral equation in the Appendix, Dr Leonard essentially first calculates $\mathbf{V}^{-1}$ for a particular $\mathbf{V}$, and then re-inverts it; neither of these operations is necessary. Dr Leonard's 1973 paper also soldiers through this unnecessary double inversion, for a particular case.)

Equation (*) represents $\tilde{g}_j - \mu_j$ as a linear smoothing of $n_j - \tilde{\rho}_j$, with the smoothing kernel being just the covariance function v_{jk}. It is plain that, in the continuous version, differentiability properties of $\mathbf{V}$ will reflect themselves in differentiability properties of $\tilde{\mathbf{g}}$. Smoothing by $\mathbf{V}$ will not necessarily preserve positivity, but it is the point of working in terms of $\mathbf{g}$ (rather than ρ) that positivity no longer matters; this is Dr Leonard's thesis.

Since there is no implication that the labelling j corresponds to a linear ordering of the cells, it is plain that these methods work just as well in the multivariate case, or the case where x is a

co-ordinate on a circle or sphere: equation (*), or its continuous analogue, is just as valid for these cases.

There is an element of automatic pre-whitening (followed by post-colouring) in (*), in that one subtracts $\tilde{\rho}_j$ from n_j, and then adds μ_j to the smoothed result. However, it is a consequence of the non-linear character of the procedure that one operation is not quite the functional inverse of the other.

In my 1958 paper I made considerable play of the fact that, as a consequence of the Bayesian approach, the smoothing kernel automatically contracted to a delta-function as sample size increased—this had not been a feature of earlier approaches. I was somewhat surprised, then, to see that the smoothing by the covariance kernel in (*) does not change at all with sample size. However, the presence of the non-linear terms has a profound influence, and one can show, much on the lines of Dr Leonard's discussion in Section 5.3, that the dependence of $\tilde{g}_j$ upon the n_k will become increasingly "local" as sample size increases, in a subtle, non-linear kind of way. I find these ideas quite fascinating and compliment Dr Leonard upon a most interesting paper.

J. B. Copas (University of Salford): One of the difficulties with the discussion of the problem of density estimation is the lack of any clear statement of what one is about, i.e. what is one really trying to estimate, and how is the estimate going to be used?

I would like to mention a simple situation in which smooth density estimates occur naturally as the solution to a clearly stated problem. Let $y_1, \ldots, y_n$ be n fixed numbers, and suppose one wishes to estimate what proportion of them fall in some arbitrary interval (perhaps they are net taxable incomes of n particular people, and we are interested to know how many pay tax at any given marginal rate). We suppose the y's are unknown but are measured with a known error distribution f by observations $x_1, x_2, \ldots, x_n$ (perhaps we know the gross incomes but can only estimate what the tax allowances are). Thus x_i has density $f(x_i - y_i)$ and, assuming only vague knowledge about y_i, this also gives the posterior density of y_i. Thus the posterior probability

$$P(a < y_i < b) = \int_a^b f(x_i - y)\, dy$$

and averaging over the n cases one obtains

$$n^{-1} \sum P(a < y_i < b) = \int_a^b n^{-1} \sum f(x_i - y)\, dy.$$

The left-hand side is the posterior expectation of the proportion of y's in (a, b), the right-hand side is the integral of a kernel-type smooth density estimate like equation (5.5) of the paper with $h = 1$ and $\Omega = f$. Thus what is the smoothed density estimate? It is simply the minimum mean squared error estimate of the histogram of the underlying values y, with window function taken to equal the error distribution. Dr Leonard points out that his estimate is not of the kernel or window type, and therefore cannot be the solution to the problem given above. But how does it differ from a practical point of view when applied to data?

Suppose now the y's were to be considered, *a priori*, to be a random sample with density $g(y)$. Then the solution above is modified in only a trivial way, by simply multiplying the kernel-smoothed estimate by the prior function $g(y$. But there is a close correspondence between my $g(y$)and Dr Leonard's mean function $\mu(t)$. His prior function enters into the method in a much more complicated way—but how different is it from a practical point of view? Both procedures presumably imply a differential weighting on different parts of the real line.

Dr J. A. Anderson (University of Newcastle upon Tyne): The objective of all these methods is to obtain some understanding of the data. In the Pelican crossing example it seems to me that these methods obscure something rather important. In the paper there are smooth histograms which start at the "Green Man" and end at the "Green Man". Since the sample space contains this periodic effect, the data would be better represented on a circle. It may be that having spent some time at Leeds, I have been over-influenced by Professor Mardia's ideas. However, since it is quite likely, *a priori*, that there will be a single mode of the frequency distribution at the "Green Man" the representation here risks splitting this into a U-shaped distribution. Clearly there is a need for this method of empirical density estimation to be developed for distributions on circles and spheres.

Dr A. C. ATKINSON (Imperial College): In Section 11 Dr Leonard examines the distribution of the times between severe explosions in mines. In studying the data in the form shown in Fig. 7 he is ignoring the information on the time ordering of the observations. This is, of course, a reasonable procedure as part of an exploratory analysis. But it does, I feel, reinforce the point made by Mr Silverman, that density estimates are often required only as a preliminary to more formal analysis. In such cases it may not much matter how the smoothing is done.

A plausible model for this particular set of data is a Poisson process with time-dependent rate. In the course of his presentation, which I enjoyed, Dr Leonard said that his method could provide smoothed estimates of such rates. I hope in his reply he will let us see the estimates for these data. One question of interest is whether the rate of the process decreases smoothly or whether there is a jump such as might result from an alteration in the legislation on mines or from a change in the technology employed.

Professor A. M. WALKER (University of Sheffield): I have one comment and one short query.

My query has to do with the treatment of the Poisson process with non-uniform rate described in Section 7. Is it not obvious that the density estimates should take the form that they do? For if L_s denotes the log-likelihood of the sample, it can be expressed as the sum of L_1, the log likelihood of the total number of observations, m (equal to $-\lambda+m\log\lambda-\log m!$, where $\lambda=\int_a^b \rho(t)\,dt$), and L_2, the log-likelihood of the observations conditional on m (equal to that of an ordered random sample size m from a distribution with density $\rho(t)/\lambda$): compare Dr Leonard's remarks immediately following equation (7.9). But if I interpret Dr Leonard's prior log-likelihood functional L_0 correctly, this gives no information at all about the distribution of m. Moreover, the condition $m=\int_a^b \tilde{\rho}(t)\,dt$ given in Section 7 seems to correspond exactly to the fact that maximizing L_1 with respect to λ gives $\hat{\lambda}=m$.

My comment is as follows. The stochastic process used by Dr Leonard as a prior for $g^{(1)}(t)$ in Equation (2.4)—the Gaussian first-order autoregressive continuous time process—certainly has local smoothness in terms of covariance, but if we consider its sample functions, they are not very smooth. They are admittedly continuous with probability 1, but are not differentiable, so that they have a certain degree of irregularity (of a rather strange form). By proceeding to higher order autoregressive schemes it is possible to obtain differentiability up to a prescribed order, but the variation of derivatives with time will still be rather irregular. I wonder whether there might be any future in trying to look for some alternative process which had, in some sense, smoother sample functions. Of course, this might be a difficult matter and would perhaps not produce significantly better results, judging by what Dr Leonard has achieved by using his Gaussian process.

The following contributions were received in writing, after the meeting.

Professor I. J. GOOD (Virginia Polytechnic Institute and State University): It may be useful to list similarities and contrasts between Leonard's approach, with his remarkable integral equation, and that taken by G^2 (Good and Gaskins, 1970–77). L and G^2 both object to pure ML in that it leads to a Dirac catastrophe, so we both subtract a roughness penalty Φ from the loglikelihood before maximizing. We both achieve a non-negative f by using an auxiliary function; L uses $\log f$, and G^2 use $\gamma=f^{\frac{1}{2}}$. He chooses Φ as $-L_0$ and we choose $\Phi=\beta R$, where the *roughness* R is the integral of the square of a linear differential operator applied to γ, especially the second derivative, and β is a positive *hyperparameter*. It is called a hyperparameter because of the Bayesian interpretation which has been used for the evaluation of bumps and dips. The prior density $\exp(-\beta R_1)$ of f in function space is improper in the curious sense that it integrates to 0 (not infinity). This is because all realistic f's are of finite roughness whereas almost all mathematical f's are infinitely rough. This has prevented G^2 from using a hyperprior for β analogous to that for the flattening constant or Dirichlet hyperparameter in my work (much joint with J. F. Crook) on multinomial distributions and contingency tables (1965, 1967, 1974, 1976, 1977).

To maximize the penalized likelihood, G^2 now use a mixture of Fourier and Hermite series, and we determine β in terms of tests of goodness of fit whose tail-area probabilities are combined by means of Good's harmonic-mean rule of thumb (1958). We have applied our methods both to raw data (the chondrite data) and to histogram data that occur in this form in high-energy physics, where significant bumps reveal new particles. For the chondrite data the trimodal density is about twice as good as the unimodal density with one bulgy cheek. This suggests that the chondrites should

be put into three categories and it will be interesting to see if the chemists find independent evidence for this. In any case, density estimation is an interesting method for the botryologist or clusterer.

We hope to apply our method also to the estimation of parameters of a distribution, for the prior $\exp(-\beta R)$ *induces* a prior for the parameters as one may say.

Professor B. M. HILL (Michigan University, U.S.A.): Dr Leonard has presented an interesting and somewhat Bayesian approach to density estimation in a fairly general context. It may prove useful. However, I should like to call attention to some difficulties. First, it would seem to be important, particularly from a subjective Bayesian point of view, to be clear as to the purposes of such estimation. I presume (or at any rate, dearly hope) that Dr Leonard views probability densities as providing certain types of approximations, rather than as inherently meaningful. It would be well to think carefully as to the uses to which they might be put, both inferential and decision-theoretic. For example, they might be useful in specifying a posterior distribution for future observations. According to the theory of de Finetti, parameters can ordinarily be defined as functions of potential future observations, and much of statistics can in a sense be incorporated into the framework of specifying a posterior distribution for future observations. How would Dr Leonard do this, even for a few future observations? Can he integrate out, with respect to his posterior distribution for the density function, the conditional distribution of future observations, given the density? I rather doubt that this is possible in a meaningful sense. I also doubt that his procedures incorporate realistic forms of prior knowledge. Smoothness of a mathematical fiction does not seem terribly convincing to me.

Some years ago (Hill, 1968) formulated a general Bayesian approach to inductive inference, flexible enough to include both parametric and non-parametric models. It incorporated subjective prior knowledge, such as, for example, the opinion that given n observations from a sequence regarded as exchangeable, a next observation should be equally likely to fall in any of the open intervals into which the data partition the real line, in my opinion, a subjectively plausible assumption for certain types of data. The model was also in accord with Zipf's Law, a law known to provide a good approximation to many populations. Finally, recent work by Lane and Sudderth (1977) has shown that the model is coherent in the sense of de Finetti. It seems to me that these are the type of things one desires of a Bayesian procedure.

Dr D. M. TITTERINGTON (University of Glasgow): The unsatisfactory nature of the maximum likelihood estimate (1.4) also arises when kernel functions are used. Suppose the estimate (5.5) is denoted by $f_m(t \mid h, D)$, where h is to be chosen suitably and D denotes the data set. At first sight it might seem reasonable to choose h to maximize the "likelihood"

$$\prod_{i=1}^{m} f_m(x_i \mid h, D), \tag{1}$$

but this leads back to (1.4). Habbema *et al.* (1974) seem to have been the first to suggest, as an alternative to penalty functions, maximizing instead

$$\prod_{i=1}^{m} f_m(x_i \mid h, D_i),$$

where D_i denotes D with x_i removed. This can be derived as a formal cross-validation technique rather on the lines of Stone (1977) and it does seem to smooth the data appropriately according to the size of the data set, thereby meeting a need expressed by Professor Priestley in the discussion of Boneva *et al* (1971) for a completely data-based procedure. The computations can be quite heavy if there are many or multivariate observations, but continuous and discrete data can be coped with, although complications arise in the former case if there are duplicate observations. Alternatively some prior idea of smoothness can be introduced in the form of a prior density for h and "Bayesian" estimates for h or $f_m(\cdot \mid D)$ obtained, using (1) as a likelihood. If the prior is chosen suitably these estimates can be written down explicitly and, although actual evaluation is difficult, the feasibility of approximations is being investigated. I am in agreement with Mr Silverman that the technique used for density estimation is usually not critical and I feel strongly that we should develop simple but reliable *ad hoc* recipes for obtaining suitable values for the

smoothing parameter(s), thereby avoiding unfruitful heavy computation. Mr Peter Rundell of Imperial College has also been looking at the Bayesian approach to estimating h, using the posterior mode.

Professor G. WAHBA (University of Wisconsin, U.S.A.): I would like to thank Dr Leonard for giving us a number of interesting things to think about. The posterior log likelihood functional (4.1) as a form of penalized likelihood is interesting to note. It is also interesting to note that the consistency argument of Section 5.3 is apparently independent of the behaviour of either of the parameters σ^2 and β. Usually there is at least one smoothing parameter (e.g. h in (5.5)) which must behave in a certain way as $m \to \infty$ in order to have pointwise convergence, and then mean square error at a point goes to 0 at a certain rate if the parameter behaves suitably as $m \to \infty$. See Parzen (1962) and Wahba (1975).

I want to thank Dr Leonard for his remarks in Section 12.3. That density estimate has now appeared in *Classification and Clustering* (J. Van Ryzin, ed., Academic Press, 1977). The value of the smoothing parameter (analogous to h) which minimizes the integrated mean square error is estimated from the data there. This parameter, intuitively speaking, controls the half power point of a low-pass filter of which the smoothed density estimate is the output. There is another parameter there (called m) which appears roughly to correspond in Dr Leonard's method to choosing the order of the autoregressive scheme. It corresponds in a kernel estimate to the shape of the window. It was taken as 2 there, but can be estimated from the data in the same manner as the first parameter by minimizing the cross-validation function. It corresponds to the "steepness" of the cut-off of the low pass filter. I feel that further parameterization of the prior covariance (beyond the equivalent of half power point and steepness) is of negligible value unless of course real (i.e. physical) information is available. A good deal of subjectivity is thus eliminated from that method. In order to compare with the density estimates proposed here, one will need an automatic procedure for choosing μ, β and σ^2. I look forward to seeing further work on Dr Leonard's estimate, along the lines of choosing the parameters in the prior, as well as more detailed results on convergence properties.

Professor G. A. WHITMORE (McGill University, Montreal): I congratulate Dr Leonard on his interesting and significant paper. As an applied statistician who has often faced the problem of obtaining a smooth estimate of a probability density from limited sample data, I welcome his contribution to the available methodology. His paper introduces a new direction for research which is likely to produce additional important contributions.

I shall restrict my commentary on Dr Leonard's paper to a single suggestion. Suppose the density function $g(x)$ of random variable X is to be estimated. Let $G(x)$ denote the distribution function of X and assume density g is concentrated on an interval (a, b) of the real line (where possibly $a = -\infty$ and $b = +\infty$). Suppose G_0 is one's best guess about G prior to sampling. Setting $T = G_0(X)$, Dr Leonard's estimation procedure may be applied to $f(t)$, the density function of T. Given that $G_0(X)$ is an estimate of the probability integral transform of X, a reasonable prior estimate of $f(t)$ is the uniform density function. In the notation of the paper, a uniform density implies that $\mu(t)$ is constant and $\mu^{(1)}(t) \equiv 0$. Using the transformed sample observations $t_i = G_0(x_i)$, $i = 1, 2, \ldots, m$, the posterior density function $\tilde{f}(t)$ may be found by means of the maximum-likelihood procedure described in the paper and from it the posterior distribution function $\tilde{F}(t)$ may be derived. The composite distribution function $\tilde{F}G_0(x)$ is, therefore, a posterior estimate of distribution function G. The desired estimate of g may be obtained from $\tilde{F}G_0$.

My suggested approach leaves Dr Leonard's estimation procedure intact but increases its applicability in several ways. Firstly, the approach removes the restriction in the paper that the density function be concentrated on a bounded interval. Secondly, the parameters σ^2 and β of the covariance kernel can be interpreted more readily with this approach because the transformed prior density is always uniform. Thirdly, the approach permits some simplification and standardization of computational procedures and allows the sensitivity of the final estimate to choices of σ^2 and β to be examined systematically.

The author replied later, in writing, as follows.

I greatly appreciate the many helpful and constructive contributions to the discussion of my paper. I think that I can safely induce an enthusiastic interest in prior informative density estimation;

the comments have stimulated many ideas for further research, and have given me some fresh understanding of the results in my paper. All contributions were valuable, but, in particular, I think that Professor Whittle's is probably one of the most magnificent to have been made to an R.S.S. discussion; it includes two exciting new scientific contributions to the subject.

A generalization of the integral equation

Professor Whittle describes an equation for the discretized time-dependent Poisson process; this is well known to me in the discrete case and holds for a general covariance structure. The subtle point of interest is his implication that it is possible to go to the limit as the interval widths get finer and obtain a general integral equation in the continuous case. In fact the equation becomes

$$\tilde{g}(t) = \mu(t) + \sum_{i=1}^{m} K(x_i, t) - \int_a^b \exp\{\tilde{g}(s)\} K(s, t)\, ds \quad \text{for } t \in [a, b],$$

where g now denotes the log-intensity of the Poisson process, and possesses mean value function μ and completely general covariance kernel K, under a Bayesian formulation. The estimate for f in density estimation may be obtained by dividing the solution for $e^{\tilde{g}}$ to the above equation by a normalizing factor ensuring that the new function integrates to unity. This provides a generalization of my equation (5.1), which corresponds to the autoregressive covariance kernel in (2.4) for the derivative of g.

One problem with the above approach is that the solution for $\tilde{g}$ is not a Bayesian posterior mode in the limiting continuous case, and does not immediately possess an interpretation as a meaningful and sensible function with a mathematically rigorous justification. My approach based upon Radon–Nikodym derivatives and prior and posterior likelihoods probably provides the easiest way of giving a justification in particular cases, though it would be possible to use another more complicated method to show that $\tilde{g}$ is a Bayesian estimate for g under zero–one loss.

Also, the prior and posterior likelihood approach enables us to provide a much fuller statistical analysis by breaking away from the restriction of simple point estimates for g. It is, for example, possible to consider posterior likelihood bands, in the spirit suggested by Edwards, and consisting of functions whose likelihood is greater than a fixed scalar value. Alternatively, likelihood ratios may be used to compare different hypothesized values for g, or the posterior likelihood could be maximized amongst a sub-class of suitably simple functions.

I think that my approach based upon Radon–Nikodym derivatives is important, interpretatively speaking, from a roughness penalty viewpoint. In connection with this, it may be worth mentioning that equations (3.1) and (3.3) combine to give the prior log-likelihood functional

$$L_0(g) = \frac{1}{2\beta\sigma^2}\int_a^b \{g^{(2)}(t) - \mu^{(2)}(t)\}^2\, dt + \frac{\beta}{2\sigma^2}\int_a^b \{g^{(1)}(t) - \mu^{(1)}(t)\}^2\, dt + \frac{1}{2\sigma^2}\{g^{(1)}(a) - \mu^{(1)}(a)\}^2 + \frac{1}{2\sigma^2}\{g^{(1)}(b) - \mu^{(1)}(b)\}^2$$

This highlights the derivatives of g, which play such an important role in the posterior smoothing, and also enables us to interpret roughness penalty methods within a prior-informative framework.

Whilst the likelihood approach leads to a fuller statistical analysis and provides the user with a fair amount of conceptual understanding of the elements of the method, Professor Whittle's suggestion leads to many exciting generalizations of the results obtained in my paper. It does indeed readily cope with multivariate and circular situations, and goes some way towards answering points raised by Mr Silverman, Mr Birch and Dr Anderson. It has also stimulated me to obtain some new equations for approximations to the marginal likelihood estimates of my hyperparameters σ^2 and β; I hope to report these results at a future date.

Asymptotic behaviour of estimates

Professors Whittle and Wahba discuss the point that it is unnecessary to change the prior parameters σ^2 and β with sample size in order to prove consistency of my estimates. Professor Whittle's suggestion that this is caused by the non-linear terms in equation (5.1) is particularly valuable, and provides one emphatic justification for considering non-linear estimates rather than the slightly restrictive linear estimates employed by various previous authors. We are essentially saying that the linearity restriction prevents the right sort of asymptotic properties as $m \to \infty$. This result may turn out to be viewed, by people who actually believe in asymptotics, as one of the

most far-reaching conclusions to be drawn from my paper. However, perhaps the philosophy "the greater the amount of information the greater the chance of contradiction" may provide some insight into the relative importance of asymptotics.

In view of the above points, the reader will probably realize that the proof of consistency in Section 5.3 is mathematically rigorous rather than intuitive. It had appeared to be unrigorous since it differs from previous results for linear estimates, but I am now completely happy with it.

A variety of fundamental points were raised in Mr Silverman's stimulating contribution and I would therefore like to respond to them in some detail.

Prior ignorance and the kernel method

I find Mr Silverman's recent contributions to the literature of density estimation (roughness penalties and estimating bandwidths for kernel methods) to be particularly refreshing and imaginative; they will provide helpful methods for practitioners for a good many years.

I am mildly surprised by his suggestion that we might be in a state of prior ignorance about a probability density. I think that even ardent non-Bayesians will possess some idea that the density is likely to be, say, continuous, or to possess a continuous first derivative; this is the main type of prior information which I try to incorporate into the analysis.

It is interesting that kernel methods, by restricting attention to a linear combination of smooth functions, implicitly assume much more prior information about smoothness than is inherent in my own approach. Therefore, I do think that kernel methods are tackling the same problem under similar conditions; for small sample sizes they are well known to tend either to oversmooth or to possess bumps in the tails. I wonder if we would ever in fact be in a practical situation where the sample size was large enough to justify a kernel method, and where all the observations still constituted a random sample from the same distribution? If we were, then it seems that the linearity restriction would anyway lead to unusual sorts of asymptotic properties, in the light of the results described above.

I think that any non-parametric method should cover the important situation where the modeller thinks that the density might belong to a particular parametrized family, but where he is not quite sure about this. My method readily copes with this situation, thus providing another possible advantage when compared with frequentist kernel methods.

I do not think that the computer time required for my method is particularly disadvantageous when compared with kernel methods. If time is anyway being used to read in the data and then print out the results, surely a few more seconds do not really matter if it leads to estimates with a firmer logical basis? Why may it be best to "use the method which is quickest and easiest, even if this means sacrificing logical foundation"? This is a daunting suggestion which could result in the demise of mathematical statistics!

Whilst the suggestion by various authors that "any reasonable kernel or roughness penalty can be used, and indeed that even the choice of method is relatively unimportant" has considerable emotive appeal, I am afraid that I cannot really agree with it. For example, the roughness penalties of Silverman, and Good and Gaskins do not incorporate prior estimates for the density. Their methods are in fact roughly equivalent to a Bayesian approach where the prior estimate is a uniform density. Consequently, the tails of their posterior estimates will tend to be relatively thicker than under more reasonable prior estimates. Another example is that a roughness penalty based upon only the first derivative of f will lead to quite different results, with "jags" at the data points. I agree that the problem of how much to smooth is also important; I intend to solve this by computing estimates for my prior parameters σ^2 and β.

Perhaps I could take this opportunity to thank Mr Silverman for his comments, and for this helpful encouragement during the weeks before the meeting.

A reply to fellow Bayesians

Professor Dickey asks how the present approach relates to my 1973 method for histograms. Well, in my 1973 paper, I assumed a first-order autoregressive covariance structure for the multivariate logits; but also mentioned that a similar structure for the log contrasts $\log \theta_j - \log \theta_{j+1}$ of the multinomial probabilities $\theta_1, \ldots, \theta_s$ might be plausible. If I had made this latter assumption, then letting the cell size get finer, with the covariance matrix depending on the (equal) interval width in an obvious way, would provide the same results as given in this paper. This brings us back to my remarks (in connection with Professor Whittle's comments) concerning the necessity for a meaningful interpretation of the estimates, since they are not posterior modes.

Professor Dickey asks why I use a prior likelihood approach when it corresponds to a constant "prior prior" density.The main reason is that I know this will not cause trouble because I have further results to show that my posterior likelihood/distribution is approximately Gaussian. Also, it is not actually necessary to justify maximum likelihood by Bayesian arguments. Another, more general, point here is that likelihoods provide reasonable measures of information, whilst it would be difficult, say, to combine two probability distributions; i.e. the latter cannot really be viewed as representing information in a completely sensible way.

Professor Dickey suggests that there may be an element of "cheating" by using a single data set in more ways than in a single application of Bayes's theorem. In fact I got this idea from his 1969 paper! The empirical estimation of, say, the parameters of a density represented by the prior estimate will not really matter as long as only one or two parameters are involved. This might be justified theoretically by a two-stage hierarchical prior. I have enough previous experience to realize that, certainly in the examples considered, the results obtained would be very similar to empirical estimation.

If β and σ^2 are chosen subjectively, rather than empirically, then perhaps it would be best for the statistician to try to obtain his own practical experience from a number of data sets.

My discussions with Professor Dickey in 1971 and 1972 were invaluable in providing me with perception about the field of density estimation. Without these discussions, or the powerful conceptual ideas suggested to me by Professor Lindley, my research in this field would probably not have progressed.

Dr O'Hagan is essentially saying that if we divide my prior likelihood by my prior density then we get a quantity which, if interpreted as the derivative of a dominating measure, leads to another prior density which is the same as my prior likelihood; therefore I am a Bayesian. This is nice to know, but I would need convincing that his measure actually dominates the prior distribution, or that his resultant prior density is a density in my mathematical sense. I am much more interested in the pragmatic properties of my method, and in the formulation of the prior covariance structure; one key point here is that by smoothing the derivatives we avoid the kinks and squiggles experienced by many previous smoothing methods. I in fact used a prior likelihood approach simply because it led me to a meaningful conclusion.

Mr Birch suggests that in my investigations of normality we should incorporate some idea of symmetry into the specification of the covariance kernel. I am not completely convinced about this, since a symmetric prior estimate may well be adequate, but I accept that statisticians may sometimes possess beliefs like his. In such circumstances, I think that my choice of covariance kernel would anyway provide reasonable practical results as it concentrates on local smoothing. Under Mr Birch's choice of prior, the posterior estimates may however be easily obtained using the general integral equation described at the beginning of this reply. The reference to another paper by Parzen is useful because it gives conditions under which the Radon–Nikodym derivative exists.

The comments from Professor Good are welcome; the G_{ij}^2 approach is very powerful perceptively, and bases its choice of roughness penalty on an information measure rather than a prior distribution or likelihood; G_{ij} suggests an ingenious way of estimating his hyperparameter, using Good's harmonic-mean rule of thumb. He clearly has green fingers; I prefer to proceed more formally in the light of my distributional assumptions in the prior model. I think that $G_{ij} \gg L$ because of his ability to work over function spaces without distributional assumptions. I would find this a bit difficult myself, but G_{ij} has certainly been able to reach many interesting conclusions by following his own philosophy.

My thanks to Professor Wahba for her comments; her development is slightly in advance of my own, in the sense that she is able to estimate empirically one of her hyperparameters; when I have computerized the estimates of my two hyperparameters then it will be interesting to compare our approaches.

Mathematical fiction and the axioms of coherence

Professor Hill, as represented by the fictitious mathematical symbol H_{bm}, suggests that I am smoothing a mathematical fiction. I have been an ardent fan of his work on the tails of distributions, and would be sorry if this all turned out to be fiction too!

In a sense, the whole of mathematical statistics is a fiction; we simply employ mathematica theory as a device to try to obtain some sort of logical explanation and inductive understanding of the real-life process; I think that a sampling density may well be as useful as a predictive density

for doing this. I would be the first to agree that neither of these distributions exists in any real sense. We can, however, still incorporate quite meaningful prior information via these subjective devices.

H_{bm}, being coherent, thinks that there is little evidence of coherence in my approach. This is probably because I was not coherent enough to view the philosophy surrounding the axioms of coherence as sufficiently coherent to warrant the provision of too much coherent evidence in connection with it. However, if we take the axioms in, say, De Groot (1970, Chapter 6) then my method certainly satisfies the first three since these lead to a relative likelihood approach. The fourth axiom is simply a regularity condition leading to a condition of countable additivity. The fifth axiom, involving an auxiliary experiment, is rather strong and may or may not be inductively appropriate. However, it only leads to the requirement that I should act like a Bayesian, and I think that I have done this. If H_{bm} coheres to the extent that he thinks that my approach is incoherent then perhaps he should provide a counter-example to demonstrate this in a coherent way.

Perhaps I should add that I do not think that the axioms of coherence should be used by the true blue Bayesian establishment in an attempt to force people into the faith. It is much more important to stress the practical advantages of prior informative approaches, and to model the prior structure in a meaningful way.

There are various further assumptions implicit in the fifth axiom, for example that the first four axioms should also relate to a comparison of the parameter space and the auxiliary experiment; these are virtually tantamount to the final result that the statistician should act like a Bayesian. Therefore these axiomatics should be viewed more as a description of the Bayesian approach than a justification of it. Bayesianism would probably have progressed much further over the years since the great practical breakthrough of Lindley and Smith (1972) if Bayesians had continued to concentrate on practical advantages rather than esoteric axiomatic justifications.

The Primary Objectives of Non-parametric Density Estimation

(i) *Inducing real-life conclusions*

Several contributors (e.g. Silverman, Ord, Copas, Atkinson, and Hill) discuss possible objectives of prior-informative non-parametric density estimation. I think that the primary purpose of such procedures is to induce real-life conclusions from the data, and prior information, in situations where a parametric approach might conceal a point of real-life interest. For example, in the mine explosions example we might induce (since the bulge in the smoothed curve in Fig. 7 makes it look like a mixture of exponentials) that the explosion rate is not constant over time. It would be rather too much to expect the modeller to formulate a parametrized sampling distribution illustrating all possible points of interest, when he might not know about these points in advance. Also, before a non-parametric procedure has been carried out it may not be obvious whether features in the data are due to random fluctuations or to some real-life aspect of the problem.

In the Pelican crossing example, Dr Griffiths had originally assumed a parametrized model which appeared quite reasonable. However, by referring to a non-parametric approach he induced an important real-life conclusion, i.e. the possibility that pedestrians are attracted by an audible bleep. I think that Dr Griffith's explanation provides an admirable example of the inductive advantages of non-parametric procedures. These advantages are highlighted when prior information is also injected into the modelling procedure, since induction from data sets can be greatly improved if the data are related to other relevant information.

(ii) *Finding parametrized families and predicting future observations*

My approach is multi-purpose and could be employed with a variety of other objectives in mind. For example, we might wish to use a non-parametric analysis to find a suitable parametrized form for the sampling distribution. This may be carried out by finding approximate parametrized densities which approximate the posterior estimate, i.e. possess a high though sub-optimal posterior likelihood. Alternatively, we might wish to follow Professor Hill's suggestion of predicting future observations. A predictive density could, for example, be based upon a parametric sampling density approximating our posterior density estimate. Alternatively, it is possible to obtain an approximation to the predictive density under full non-parametric assumptions. The latter could be based upon the Gaussian approximation to my posterior likelihood/distribution discussed earlier, this will be reported elsewhere.

(iii) *Avoiding tests of fit and their bureaucratic idiosyncrasies*

Probably the second most important objective of my approach is to cope with situations where many frequentist statisticians would test the null hypothesis that the distribution belongs to a particular parametrized family, by, say, using a fixed-size test. I can see no practical, theoretical or inductive reason for using such tests in such circumstances; they are, for example, extremely dependent upon significance levels and sample sizes. As shown by Leonard (1977) they can lead to rejection of the null hypothesis (at any sensible significance level) in situations where sensibly formulated alternatives would lead to acceptance, or to overwhelming acceptance of the null hypothesis when rejection might be more sensible. Fixed-size significance tests could be viewed as a bureaucratic intrusion on the inductive nature of our discipline! By introducing the hypothesized density as a prior estimate, and then investigating differences from the posterior estimate, I think that we provide the practical statistician with the sort of method that he really wants.

Other important points

My thanks to my colleague Dr Ord for his contribution and for his expertise in the role of midwife. He suggests that if we transform the data then we should also transform the prior. The best course of action would therefore be to take the transformation of the sampling density for which we could most reasonably induce our Gaussian/logistic prior with a meaningful covariance kernel, and then refer to the general integral equation described at the beginning of this reply.

Professor Copas describes an ingenious example where an intuitive procedure provides reasonable estimates for the interval probabilities. I think that my approach would provide more natural smoothing, and also different asymptotic properties for large sample sizes. It would also be a bit more flexible to extend to different situations.

I am sorry that the deadline for submission of my reply came before I was able to follow Dr Atkinson's suggestion of computing the smoothed intensity function for the Poisson process underlying the mine explosions data. My existing results for the inter-arrival times certainly suggest that the intensity has changed in time; I think that it may be best to wait for my estimates of the prior parameters until I can judge whether the change was a jump or a smooth transition; or perhaps it would be better to refer back to the practical context of the data?

Professor Walker has a helpful query and an excellent comment. I agree that it is intuitively obvious that my estimates for the Poisson process should take the form they do; the exposition in Section 7 describes the only way I have been able to find of proving this precisely in terms of prior distributional assumptions relating to the intensity function.

Professor Walker's comment is interesting; aspects like this have been niggling my mathematical conscience. It seems that the numerical realizations (of the Gaussian process) which appear in my Radon–Nikodym derivative can only occur with small probability since they are twice differentiable; therefore some idea of, say, denseness over function space might be appropriate. I agree that it would be helpful to look for processes with smoother sample functions, though the process employed seems to give me precisely the level of smoothing that I need.

Dr Anderson makes a good point about the pedestrians; a cyclic interpretation appears quite plausible. Returning to a suggestion by Dr Griffiths, cycles like this could easily be represented when choosing the prior estimate. I am looking forward to examining the obvious extensions of my method to the analysis of circular data; this will provide an alternative to the existing histospline methods for circular situations.

Dr Titterington's first comment highlights another disadvantage of the kernel method. The cross-validation technique he describes is interesting, but I would be worried about a joint maximization with respect to both f and h. My intuition tells me that the estimate for h would be inferior —there are analogies with some well-known counter-examples to maximum likelihood estimation.

I would like to thank Professor Whitmore for his helpful contribution. I find it encouraging to think that my methods may benefit applied statisticians working on a variety of practical problems. I have already received a number of enquiries in this direction and would be happy to be as helpful as possible in communicating my approach to applied areas. Professor Whitmore's suggested transformation should be very useful for practitioners utilizing my method.

Finally, perhaps I could mention that there are possible extensions of my approach to cope with similar problems in spatial processes, survival data, non-Poisson point processes and non-linear stochastic processes, forecasting, non-linear regression models, and so on. Essentially we have a

general method for smoothing non-linear functions. I would welcome hearing from people who develop superior computational methods for solving my integral equation; I am also working on this.

References in the Discussion

DE GROOT, M. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

EPACHENIKOV, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theor. Prob. Appl.*, **14**, 153–158.

GOOD, I. J. (1958). Significance tests in parallel and in series. *J. Amer. Statist. Ass.*, **53**, 799–813.

—— (1965). *The Estimation of Probabilities.* M.I.T. Press.

—— (1967). A Bayesian significance test for multinomial distributions (with Discussion). *J. R. Statist. Soc.* B, **29**, 399–431. Corrigendum: **36** (1974), 109.

—— (1971a). Contribution to the discussion of a paper by Orear and Cassell in *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 284–286. Toronto and Montreal: Holt, Rinehart and Winston of Canada.

—— (1971b). Nonparametric roughness penalty for probability densities. *Nature, Phys. Sci.* **229**, 29–30. (Contains 21 misprints owing to dock strike.)

—— (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.*, **4**, 1159–1189.

GOOD, I. J. and CROOK, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Ass.*, **69**, 711–720.

—— (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Math.*, **19**, 23–45.

GOOD, I. J. and GASKINS, R. A. (1972). Global nonparametric estimation of probability densities. *Virginia J. Sci.*, **23**, 171–193.

HABBEMA, J. D. F., HERMANS, J. and VAN DEN BROEK, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat 1974* (G. Bruckmann, ed.), pp. 101–110. Vienna: Physica Verlag.

HILL, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *J. Amer. Statist. Ass.*, **63**, 677–691.

LANE, D. A. and SUDDERTH, W. D. (1977). Diffuse models for sampling. Technical Report No. 287, School of Statistics, University of Minnesota.

LEONARD, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Ass.*, **72**, 869–874.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

PARZEN, E. (1961b). An approach to time series analysis. *Ann. Math. Statist.*, **32**, 951–989.

SILVERMAN, B. W. (1978b). Choosing the window width when estimating a density. *Biometrika*, **65** (in press).

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B, **39**, 44–47.

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.*, **3**, 15–29.