J. R. Statist. Soc., B (1984)
46, *No.* 2, *pp.* 193–227

Monte Carlo Methods of Inference for Implicit Statistical Models

By PETER J. DIGGLE and RICHARD J. GRATTON†

University of Newcastle-upon-Tyne, UK

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, January 11th, 1984*, Professor J. B. Copas *in the Chair*]

SUMMARY

A prescribed statistical model is a parametric specification of the distribution of a random vector, whilst an implicit statistical model is one defined at a more fundamental level in terms of a generating stochastic mechanism. This paper develops methods of inference which can be used for implicit statistical models whose distribution theory is intractable. The kernel method of probability density estimation is advocated for estimating a log-likelihood from simulations of such a model. The development and testing of an algorithm for maximizing this estimated log-likelihood function is described. An illustrative example involving a stochastic model for quantal response assays is given. Possible applications of the maximization algorithm to *ad hoc* methods of parameter estimation are noted briefly, and illustrated by an example involving a model for the spatial pattern of displaced amacrine cells in the retina of a rabbit.

Keywords: DENSITY ESTIMATION; MAXIMUM LIKELIHOOD ESTIMATION; QUANTAL RESPONSE ASSAYS; SIMULATION; SPATIAL PATTERN; STOCHASTIC MODELLING

1. PRESCRIBED AND IMPLICIT STATISTICAL MODELS

We use the term *prescribed statistical model* to mean a parametric specification of the distribution of a random vector $\mathbf{Y}$, leading to a log-likelihood function

$$L(\boldsymbol{\theta}) = \ln f(\mathbf{y}; \boldsymbol{\theta}). \tag{1.1}$$

In (1.1), $f(\cdot)$ is a class of distributions, $\mathbf{y}$ a set of data and $\boldsymbol{\theta}$ a vector of parameters. Problems of inference associated with (1.1) will include estimation of $\boldsymbol{\theta}$ and goodness-of-fit testing for $f(\cdot)$. Sometimes a model is defined at a more fundamental level in terms of a stochastic mechanism whereby the data are generated. In principle, a unique log-likelihood function underlies any such *implicit statistical model* (but not *vice versa*) and if this can be written down explicitly, inference can proceed along conventional lines. The purpose of the present paper is to develop methods of inference which can be used when a conceptually attractive implicit model leads to intractable distribution theory.

Our preferred approach is to estimate the log-likelihood function from simulated realizations of an implicit model. Maximization of this function over the parameter space then gives estimators $\boldsymbol{\theta}^*$ whose properties are similar to those of the maximum likelihood estimators $\hat{\boldsymbol{\theta}}$, albeit with some loss of efficiency arising from the additional random variation introduced by the simulations.

To provide a specific illustration, we generate data y_i: $i = 1, \ldots, 25$ as an independent random sample from an exponential distribution with probability density function (pdf) $f(y) = \theta e^{-\theta y}$ and $\theta = 1$, and treat the exponential distribution as an implicit model. For each of $\theta = 0.60\ (0.02)\ 1.80$ we simulate an independent random sample x_k: $k = 1, \ldots, 200$ from the appropriate exponential

† *Present address*: Fisons Pharmaceutical Division, Science and Technology Labs., Loughborough, England.

 0035–9246–84–46193 $2.00

distribution, use a kernel method as described in Section 3 to estimate each $f(y;\theta)$ from the simulated x_k and construct an estimated log-likelihood.

$$L^*(\theta) = \sum_{i=1}^{25} \ln \hat{f}(y_i;\ \theta).$$

Fig. 1 compares the true log-likelihood function $L(\theta)$, the estimate $L^*(\theta)$ and a seven point simple moving average of $L^*(\theta)$. The smoothed estimate successfully recovers both the location of the maximum and the general shape of $L(\theta)$, with the consequence that confidence intervals for θ based on the familiar asymptotic result $2\{L(\hat{\theta}) - L(\theta_{\text{true}})\} \sim \chi^2_1$ would be in close agreement, whether based on the true $L(\theta)$ or on its smoothed estimate.

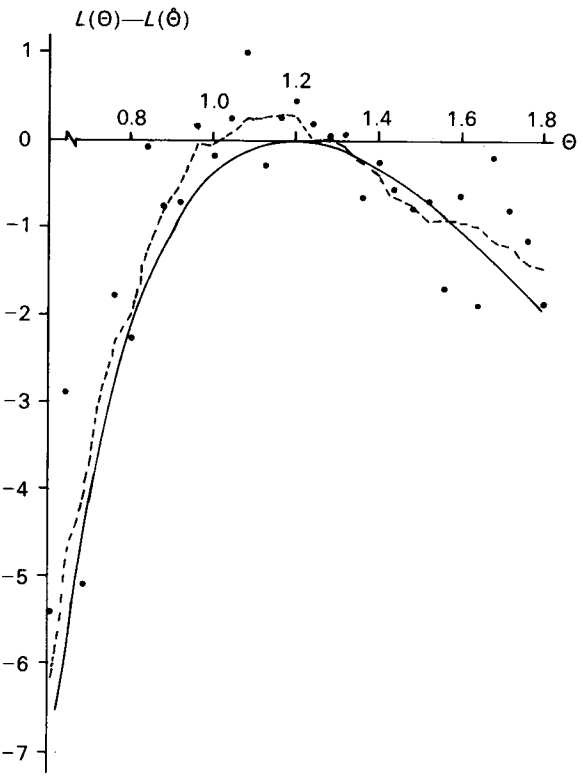

Fig. 1. Log-likelihood functions for exponential example. ——, $L(\theta)$; • • •, $L^*(\theta)$; – – –, smoothed $L^*(\theta)$.

We shall develop the likelihood approach explicitly for the univariate case in which $\mathbf{Y}$ consists of mutually independent, identically distributed components Y_i: $i = 1, \ldots, n$. The extension to multivariate structures is straightforward in principle, but unworkable in practice unless the dimensionality is small. From a theoretical point of view, the log-likelihood function (1.1) can always be written as

$$L(\boldsymbol{\theta}) = \Sigma \ln f_i(y_i \mid y_1, \ldots, y_{i-1};\ \boldsymbol{\theta}) \tag{1.2}$$

and if the univariate conditional distributions in (1.2) can be simulated our methods are directly applicable. When this is not possible, we shall be forced to abandon maximum likelihood in favour

of a more *ad hoc* approach based on minimizing an intuitively reasonable measure of the discrepancy between a model and a set of data.

In the remainder of the paper, we first discuss estimation of the log-likelihood function, distinguishing between discrete and continuous data. For discrete data, we use the obvious method of estimating probabilities by relative frequencies, but with a first-order correction for the bias introduced by the log-transformation. In the continuous case, we use a kernel method (Rosenblatt, 1956) for non-parametric estimation of a probability density function. We then describe an algorithm for function maximization in the presence of controllable random noise, which we use to maximize $L^*(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ has dimension $p > 1$. This algorithm combines the simplex method of Nelder and Mead (1965) with a quadratic fitting procedure in the vicinity of the maximum, and incorporates criteria for automatically increasing the number of simulations as the maximum is approached. The reference to quadratic fitting serves to emphasize a general point which applies equally to classical maximum likelihood estimation, namely that a good choice of parameterization can save a great deal of computing in practical cases. We investigate the performance characteristics of the proposed algorithm by experiments using artificial Normal and Weibull data. We then present two applications. In the first, a stochastic model for quantal response assays proposed by Puri and Senturia (1972) is fitted to data from Hewlett (1974) on the response of flour-beetles to poisons. In the second, an *ad hoc* method of parameter estimation is used to fit a model to the spatial pattern of displaced amacrine cells in the retinal ganglion cell layer of a rabbit.

Although the idea of using simulation to analyse intractable models is obviously not new, we are aware of no other systematic investigations along the lines of the present paper. Ross (1972) presents a number of examples, but without methodological details. Hoel and Mitchell (1971) discuss a particular application in detail, using standard response surface methodology to minimize an *ad hoc* optimization criterion. An intermediate case between classical inference and the methods of the present paper arises when point estimates can be obtained without recourse to simulation, but their associated sampling distributions are intractable. See, for example, Chambers (1977, pp. 146-147), Buckland (1980) and Diggle (1983, Ch. 5). The use of simulation to test a simple hypothesis is of wide applicability (Barnard, 1963; Hope, 1968; Marriott, 1979), and is particularly useful in the analysis of spatial pattern (Besag and Diggle, 1977). The research reported here was stimulated by an unpublished manuscript of Plackett (1977).

2. ESTIMATION AND GOODNESS-OF-FIT FOR DISCRETE DATA

In this section, we suppose that the data $y_i: i = 1, \ldots, n$ are an independent random sample from a discrete distribution $p_j(\boldsymbol{\theta}) = P\{Y = j\}$.

2.1. *Estimation of the Log-likelihood*

The log-likelihood is

$$L(\boldsymbol{\theta}) = \sum_{j=0}^{\infty} f_j \ln p_j(\boldsymbol{\theta}),$$

where f_j denotes the number of $y_i = j$. Simulations of the model for a particular value of $\boldsymbol{\theta}$ generate observations $x_k: k = 1, \ldots, s$ and corresponding frequencies $g_j(\boldsymbol{\theta})$. The obvious estimator for $p_j(\boldsymbol{\theta})$ is $\hat{p}_j(\boldsymbol{\theta}) = s^{-1} g_j(\boldsymbol{\theta})$, which is unbiassed with variance $s^{-1} p_j(\boldsymbol{\theta})\{1 - p_j(\boldsymbol{\theta})\}$. Suppressing the dependence on $\boldsymbol{\theta}$, this implies an estimated log-likelihood

$$L^*(\boldsymbol{\theta}) = \sum_{j=0}^{\infty} f_j \ln \hat{p}_j. \tag{2.1}$$

A Taylor series approximation then gives

$$E[L^*(\boldsymbol{\theta})] = \sum_{j=0}^{\infty} f_j E[\ln \hat{p}_j]$$

$$= L(\boldsymbol{\theta}) - \sum_{j=0}^{\infty} f_j (1-p_j)/(2sp_j) + O(s^{-2}), \tag{2.2}$$

where the expectation is with respect to the simulation-induced random variation. Similarly,

$$\mathrm{Var}\{L^*\{\boldsymbol{\theta}\}\} = \sum_j f_j^2 \mathrm{Var}\{\ln \hat{p}_j\} + \sum_{i \neq j} \sum f_i f_j \mathrm{cov}\{\ln \hat{p}_i, \ln \hat{p}_j\}$$

$$= \sum_j f_j^2 (1-p_j)/(sp_j) - \sum_{i \neq j} \sum f_i f_j / s + O(s^{-2}). \tag{2.3}$$

Strictly, the expectations in (2.2) and (2.3) do not exist, but the Taylor series approximations nevertheless give some insight into the behaviour of $L^*(\boldsymbol{\theta})$ when s is large. The form of (2.3) emphasizes the difficulty of estimating $L(\boldsymbol{\theta})$ when any of the $p_j(\boldsymbol{\theta})$ are small whilst (2.2) suggests a modification of (2.1) in which $\ln \hat{p}_j$ is replaced by

$$l_j = \ln \hat{p}_j + (1-\hat{p}_j)/(2s\hat{p}_j). \tag{2.4}$$

Fig. 2 shows the effect of the correction (2.4) for an example involving data $f_j = (2, 5, 8, 8, 2)$ generated as an independent random sample from a binomial distribution $B(4; \theta)$ with $\theta = 0.5$.

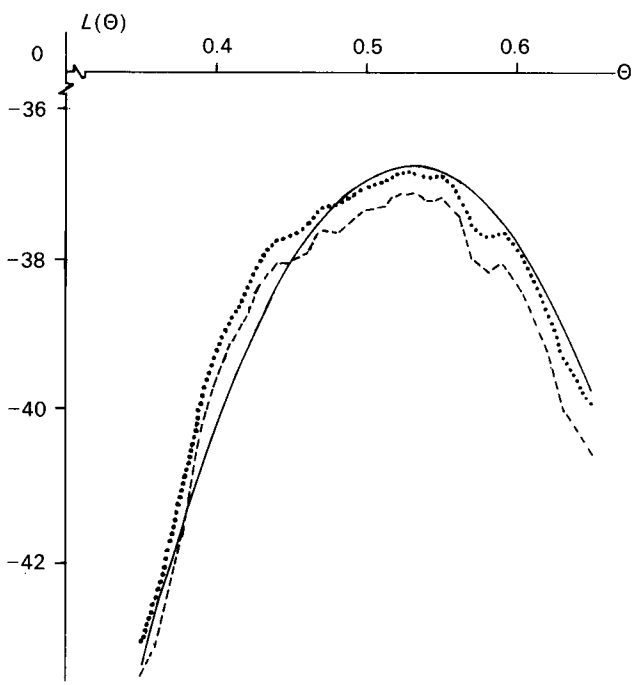

Fig. 2. Log-likelihood functions for binomial example. ——, $L(\theta)$; – – –, $L^*(\theta)$ without correction for log-bias; . . . . . , smoothed $L^*(\theta)$ with correction for log-bias.

For each of $\theta = 0.32\,(0.01)\,0.68$ we use $s = 200$ simulations to estimate the binomial probabilities and hence the log-likelihood function, which we then smooth by a seven-point simple moving average as in Fig. 1; the correction (2.4) has an almost negligible effect on the shape of the estimated log-likelihood function within the plotted range of θ.

Titterington (1980) discusses more sophisticated estimators for $p_j(\boldsymbol{\theta})$, for example using kernel-smoothing of the empirical proportions $\hat{p}_j$. The need for this seems less pressing than in the continuous case but we have not explored the method in any detail.

2.2. *Goodness-of-fit*

Goodness-of-fit testing for discrete or grouped continuous data can proceed along conventional lines since, for a single value $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, there is usually no difficulty in reducing the simulation-induced random variation to negligible proportions. For example, suppose that frequencies $f_j : j = 1, \ldots, k$ in each of k cells are to be compared with frequencies $g_j : j = 1, \ldots, k$ from s simulations of an implicit model. The contribution from the jth cell to the X^2-statistic for testing equality of the two distributions is

$$(f_j s - g_j n)^2 / \{ns(f_j + g_j)\} = \{(f_j - ns^{-1} g_j)^2 / (ns^{-1} g_j)\}\{1 + O(ns^{-1})\}$$

and the $O(ns^{-1})$ term represents the difference between this and the corresponding contribution to X^2 for testing observed frequencies f_j against expected frequencies $ns^{-1} g_j$.

3. ESTIMATION OF THE LOG-LIKELIHOOD FOR CONTINUOUS DATA

We now suppose that the data $y_i : i = 1, \ldots, n$ are an independent random sample from a continuous distribution with pdf $f(y;\ \boldsymbol{\theta})$. One approach to such data would be to fit a discrete distribution

$$p_j(\boldsymbol{\theta}) = \int_{a_{j-1}}^{a_j} f(y;\ \boldsymbol{\theta})\, dy,$$

for suitable class-intervals defined by the a_j. The arbitrary choice of the a_j is an unattractive feature, and any grouping into discrete classes results in a loss of information. We therefore prefer to estimate the log-likelihood directly as

$$L^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \hat{f}(y_i;\ \boldsymbol{\theta}), \tag{3.1}$$

where $\hat{f}(\cdot)$ denotes a non-parametric estimator of $f(\cdot)$ obtained from simulations $x_k : k = 1, \ldots, s$ for any specified value of $\boldsymbol{\theta}$.

3.1. *Non-parametric Probability Density Estimation*

Classes of non-parametric estimators of a pdf $f(\cdot)$ include kernel estimators (Rosenblatt, 1956), k-nearest neighbour estimators (Loftsgaarden and Quesenberry, 1965), orthogonal series estimators (Kronmal and Tarter, 1968) and penalized maximum likelihood estimators (Good and Gaskins, 1971). Details of these various approaches can be found in the above papers, in the review articles of Wegman (1972a, b), Fryer (1977) and Bean and Tsokos (1980), or in the monograph by Tapia and Thompson (1978). In general terms, each class of estimators involves a constant which determines the degree of smoothing and whose value critically affects the behaviour of the estimator, but with this proviso the different approaches can all give good results.

In our context, we require an estimator which is easy to compute, which is itself a pdf and which can cope with underlying distributions of restricted or unrestricted support. A further requirement is that the smoothing constant should be capable of automatic up-dating when the simulation size s changes in the course of the numerical maximization of $L^*(\boldsymbol{\theta})$. These

considerations lead us to the kernel method, which we now describe in some detail.

3.2. *Kernel Estimation of the Log-likelihood*

Rosenblatt's (1956) kernel estimator of a pdf is defined as

$$\hat{f}(y) = (sh)^{-1} \sum_{k=1}^{s} \delta\{h^{-1}(y - x_k)\}, \tag{3.2}$$

where $\delta\{\cdot\}$ is a kernel function, typically a pdf symmetric about the origin, $x_k: k = 1, \ldots, s$ is an independent random sample from a distribution with pdf $f(\cdot)$ and $h \equiv h_s$ is the smoothing constant.

Asymptotic expressions for the mean and variance of the estimated log-likelihood (3.1), using the kernel estimator (3.2), can be obtained from Taylor series approximations (cf. Bartlett, 1963). We assume that $f(\cdot)$ is twice differentiable on the real line and that $\delta(u) > 0$ for all u. Then,

$$E[\ln \hat{f}(y)] = \ln f(y) + \frac{1}{2} h^2 \{f''(y)/f(y)\} \int u^2 \delta(u)\, du + o(h^2)$$

and

$$\mathrm{Var}\{\ln \hat{f}(y)\}\} = \{shf(y)\}^{-1} \int \{\delta(u)\}^2 du + o\{(sh)^{-1}\}.$$

From the additive property of expectation, we deduce that

$$E[L^*(\boldsymbol{\theta})] = L(\boldsymbol{\theta}) + \frac{1}{2} h^2 \int u^2 \delta(u)\, du \sum_{i=1}^{n} \{f''(y_i)/f(y_i)\} + o(h^2).$$

This does not offer any obvious opportunities to correct for bias, since $f''(\cdot)$ is difficult to estimate in practice. However, if $f''(\cdot)$ is continuous on the real line, $E\{f''(Y)/f(Y)\} = 0$, so that for large n we might reasonably expect the leading term in the bias of $L^*(\boldsymbol{\theta})$ to be small relative to $L(\boldsymbol{\theta})$. The variance of $L^*(\boldsymbol{\theta})$ can similarly be shown to be $O\{(sh)^{-1}\}$. It follows that $L^*(\boldsymbol{\theta})$ is consistent for $L(\boldsymbol{\theta})$ if $h \to 0$, $sh \to \infty$ as $s \to \infty$, and that the variance and squared bias balance asymptotically if $h = O(s^{-0.2})$, a familiar result. An application of Silverman's (1978a) fundamental decomposition theorem shows that $L^*(\boldsymbol{\theta})$ is asymptotically Normally distributed.

3.3. *Implementation of the Kernel Method*

Our implementation of (3.2) reflects the context in which we are working, and we do not suggest that it is universally appropriate. In particular, and in contrast to the use of kernel estimators in exploratory data analysis, (*a*) in an informal sense, we expect the underlying density to be smooth; (*b*) we are not concerned about possible outliers amongst the x_k; (*c*) we are applying the estimator not to a fixed sample size s, but rather to a sequence of increasing values of s with, ultimately, s of the order of several thousand.

There is general agreement that (3.2) can give good results using almost any symmetric, unimodal kernel function $\delta(\cdot)$; see, for example, Epanechnikov (1969) and Silverman (1978b). Primarily because of its computational simplicity, we have used a parabolic kernel,

$$\delta(u) = \begin{cases} 0.75(1 - u^2): & -1 \leqslant u \leqslant 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

An immediate consequence is that $\hat{f}(y_i)$ may be zero, in which case $L^*(\boldsymbol{\theta})$ is undefined. This may

indicate either that the model is grossly wrong, or that there is at least one outlier amongst the y_i, and we see no need to disguise its occurrence by using a strictly positive kernel. On the other hand, Silverman (1982) has recently given an algorithm for the evaluation of (3.2) via a Fast Fourier-transform, for which a Normal kernel function is particularly convenient.

Two difficulties remain to be resolved:

(i) a way of handling distributions with restricted support.

(ii) a choice for the value of the smoothing constant h, and its dependence on the simulation size s.

The most common manifestation of (i) is when the data are necessarily non-negative. In this context, Boneva *et al.* (1971) recommend reflecting in the origin that part of $\hat{f}(\cdot)$ which refers to the negative half-line, whilst Copas and Fryer (1980) work with logarithms of the data. Our preferred solution is to use the data $\mathbf{y}$ to estimate by maximum likelihood a power transformation to approximate Normality (Box and Cox, 1964), and to work with a reflected kernel estimator on this transformed scale. The reflection induces the constraint $\hat{f}'(0) = 0$, but its effect is often negligible on the transformed scale. There is of course no need to back-transform the estimate for computation of $L(\boldsymbol{\theta})$.

Difficulty (ii) has attracted considerable attention. See, for example, Habbema *et al.* (1974), Silverman (1978b), Gratton (1979), Davis (1981) and Bowman (1982). If the transformed x_k are exactly Normally distributed and the kernel function (3.3) is used, the value of h which asymptotically minimizes the mean integrated square error (MISE) on the transformed scale is

$$h = 2.35\,\sigma s^{-0.2}, \tag{3.4}$$

where $\sigma \equiv \sigma(\boldsymbol{\theta})$ is the standard deviation and can be estimated in the obvious way. Whilst the power transformation based on the y_i cannot guarantee approximate Normality of the x_k throughout the parameter space, it should be reasonably effective in the vicinity of the maximum likelihood estimate of $\boldsymbol{\theta}$.

One further consideration is the systematic tendency of kernel estimators to inflate the variance of the underlying distribution. To eliminate this effect, we calculate $\hat{f}(\cdot)$ not from the simulated values of x_k but from

$$x_k^* = \bar{x} + (x_k - \bar{x})\sqrt{(1 - h^2 v/\sigma^2)}: \quad k = 1, 2, \ldots, s, \tag{3.5}$$

where $v = \int u^2 \delta(u)\,du$. This device was suggested by Fryer (1976) as a bias-reduction technique and by Efron (1979) and Silverman (1981) as a means of simulating samples from a smooth distribution whose mean and variance match the sample mean and variance of the x_k.

Empirical investigations with simulated samples from Normal mixtures and log-gamma distributions suggest firstly that (3.5) can give a worthwhile reduction in MISE even for moderately skewed densities, and secondly that a value of h somewhat larger than that implied by (3.4) is desirable, in part as a consequence of using (3.5) but also because this improves the estimation of $\log\{f(\cdot)\}$ in the tails. Clearly, both (3.4) and (3.5) are inappropriate for multimodal densities and, in particular, (3.4) will then tend to give too large a value of h. With this qualification, we arrive at the following method of implementation:

(i) use the data $\mathbf{y}$ to estimate a power transformation to approximate Normality;

(ii) for each $\boldsymbol{\theta}$, simulate $\mathbf{x}$ on this transformed scale and shrink towards $\bar{x}$ using (3.5) with $v = 0.2$ to correspond to the kernel function (3.3);

(iii) set $h = 2.7\,\hat{\sigma} s^{-0.2}$ and evaluate (3.2) using (3.3), but with $\mathbf{x}^*$ in place of $\mathbf{x}$.

3.4. *An Example*

Fig. 1 showed the true log-likelihood function $L(\theta)$ for an independent random sample of size $n = 15$ from an exponential distribution with parameter $\theta = 1$, together with a seven-point simple moving average of the estimated log-likelihood $L^*(\theta)$ obtained as described in Section 3.3. Fig. 3 reproduces this information, together with two further estimated log-likelihoods based on a reflected kernel estimate on the untransformed scale and on a transformation to approximate

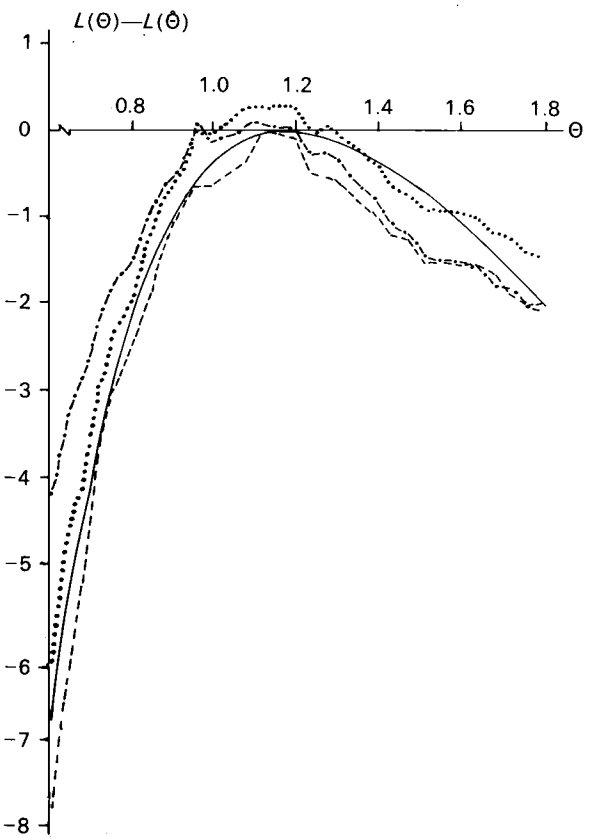

Fig. 3. Estimated log-likelihood functions for exponential example. - - -, Reflection; -.-.-, transformation; . . . . ., transformation plus shrinkage; ——, true $L(\theta)$.

Normality but without the shrinkage. Note in particular that omission of the shrinkage shifts the estimated log-likelihood to the left, consistent with the variance inflation effect noted in Section 3.3. To provide a more objective comparison amongst the three methods of estimating the log-likelihood, we repeated the experiment 100 times and evaluated the empirical mean square of $(\theta^* - \hat{\theta})/\hat{\theta}$, where θ^* denotes the value which maximizes the seven-point simple moving average of $L^*(\theta)$ again elevated at $\theta = 0.60\ (0.02)\ 1.80$. The results in Table 1 suggest that there is little

TABLE 1

Empirical relative mean square error of θ^* for $\hat{\theta}$

Type of kernel estimator	$MS\{(\theta^* - \hat{\theta})/\hat{\theta}\}$	*Standard error*
Reflection	0.0084	0.0016
Transformation	0.0071	0.0009
Transformation plus shrinkage	0.0035	0.0007

to choose between the reflection and transformation methods, but that the shrinkage towards the sample mean approximately halves the mean square error of θ^* for $\hat{\theta}$.

4. ESTIMATION OF $\boldsymbol{\theta}$ IN THE MULTI-DIMENSIONAL CASE

For a set of data $\mathbf{y}$, let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimator for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ the estimator produced by numerical maximization of the estimated log-likelihood, $L^*(\boldsymbol{\theta})$. The statistical properties of $\boldsymbol{\theta}^*$ under repeated sampling of $\mathbf{y}$ differ from those of $\hat{\boldsymbol{\theta}}$ because of the random variation in the evaluation of $L^*(\boldsymbol{\theta})$. However, this additional source of random variation can be controlled through the number of simulations performed. Standard results on the asymptotic variance matrix of $\hat{\boldsymbol{\theta}}$ rely on a quadratic approximation to the log-likelihood, $L^*(\boldsymbol{\theta})$, which we would wish to recover from the estimated surface, $L^*(\boldsymbol{\theta})$. We therefore require an algorithm for function maximization which can tolerate random error in function evaluations, which can ensure that at each stage in the maximization sufficiently many simulations are performed, and which furnishes a quadratic approximation to $L(\boldsymbol{\theta})$ near $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. This section describes the development of an algorithm and illustrates its use on data from standard distributions. Further details are given by Gratton (1981, Ch. 4).

4.1. *Development of the SQ Algorithm*

Dixon (1972) and Murray (1972) review a large number of methods for maximization of a function of several variables. Most methods assume that the function can be evaluated effectively without error so that, for example, derivatives can be evaluated numerically if not analytically. Nelder and Mead (1965) describe a simplex method, subsequently referred to as NM, which does not use information on derivatives and is therefore potentially useful even when function evaluation is subject to random error.

The problem of function maximization in the presence of random noise is embraced by the theory of stochastic approximation (Wasan, 1969). The classic stochastic approximation algorithm is Blum's (1954) multidimensional extension of a method proposed by Kiefer and Wolfowitz (1952). This is a gradient method which uses observed differences between function values to estimate the partial derivatives of the objective function. The theoretical properties of the Kiefer-Wolfowitz algorithm have been thoroughly explored (see, for example, Wasan, 1969, Ch. 3 and Ch. 5, pp. 88-91) but the use of noise-corrupted function values to estimate partial derivatives seems a delicate operation. Springer (1969) presents a direct search algorithm which is similar in essence to NM and which shows up favourably in empirical comparisons with a version of the Kiefer-Wolfowitz algorithm. Chambers (1977) reviews various methods of function maximization and concludes that direct search methods "are particularly appropriate for functions which are not smooth".

We therefore adopt NM as the starting point for the development of our algorithm. A FORTRAN program is given by O'Neill (1971); subsequent remarks by Chambers and Ertel (1974), Benyon (1976) and Hill (1978) should be noted. In p dimensions, NM requires the user to specify an initial guess $\boldsymbol{\theta}_0$ for $\hat{\boldsymbol{\theta}}$ and p initial step-lengths δ_i. The algorithm then constructs p points $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \delta_i \mathbf{u}_i$, where $\mathbf{u}_i$ is a unit-length vector in the direction of the ith co-ordinate axis, and the $p+1$ points $\boldsymbol{\theta}_i: i = 0, 1, \ldots, p$ form the first "current simplex". The point corresponding to the lowest of the $p+1$ values $L(\boldsymbol{\theta}_i)$ in the current simplex is then replaced by a new point chosen in an adaptive manner. This process continues until the sample variance, v_b say, of the $p+1$ values $L(\boldsymbol{\theta}_i)$ is less than a specified constant, at which stage the simplex has typically contracted to a small neighbourhood of $\hat{\boldsymbol{\theta}}$.

A difficulty with NM when used with noise-corrupted function values $L^*(\boldsymbol{\theta})$ is that it tends to over-contract towards a point at which a spuriously high value of $L^*(\boldsymbol{\theta})$ has been observed. We need to ensure that if $L(\boldsymbol{\theta}_1) > L(\boldsymbol{\theta}_2)$ then $L^*(\boldsymbol{\theta}_1) > L^*(\boldsymbol{\theta}_2)$ with high probability. To achieve this we incorporate the following criterion for periodically increasing the simulation size s so that the simulation-induced $\mathrm{Var}\{L^*(\boldsymbol{\theta})\}$ is always small, relative to v_b. For each point $\boldsymbol{\theta}_i$ in the current simplex, let d_i denote the difference between the estimated log-likelihoods calculated from two half-samples of $\frac{1}{2}s$ simulations with $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. Then, $v_w = (\Sigma_{i=1}^{p+1} d_i^2)/\{4(p+1)\}$ is an estimate of $\mathrm{Var}\{L^*(\boldsymbol{\theta})\}$, pooled over the $p+1$ points $\boldsymbol{\theta} = \boldsymbol{\theta}_i$.

If $v_b < cv_w$, we increase s by a factor of 4. We found that testing this criterion with $c = 6$ after every five function evaluations, and whenever a contraction of the whole simplex is indicated, gave good results. The precise value of c is rather arbitrary. A larger value would increase the probability of correctly identifying the point in the current simplex corresponding to the lowest of the $L(\boldsymbol{\theta}_i)$, but at the cost of a more rapid increase in s.

In our experience, this typically gives a good point estimate of $\boldsymbol{\theta}$, but it does not provide direct information on the curvature of $L(\boldsymbol{\theta})$ near its maximum. The second phase of the *SQ* algorithm therefore consists of a sequence of quadratic fitting experiments using central composite rotatable designs as described by Cochran and Cox (1957, Ch. 8). The first such design is centred on the vertex of the final simplex corresponding to the current estimate of the maximizing value of $\boldsymbol{\theta}$, the principal axes of the design are in the direction of the remaining vertices and points on the axes are set at four times the distance from the centre of the corresponding vertices of the simplex.

On successful termination of this second phase, the point estimate $\boldsymbol{\theta}^*$ is the value which maximizes the fitted quadratic. The approximate variance of $\boldsymbol{\theta}^*$ is obtained as

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\theta}^*) &= \mathrm{Var}_{\mathbf{Y}}\{E_{\mathbf{X}}(\boldsymbol{\theta}^* \mid \mathbf{Y})\} + E_{\mathbf{Y}}\{\mathrm{Var}_{\mathbf{X}}(\boldsymbol{\theta}^* \mid \mathbf{Y})\} \\ &\simeq \mathrm{Var}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}}) + \mathrm{Var}_{\mathbf{X}}(\boldsymbol{\theta}^* \mid \mathbf{y}). \end{aligned} \tag{4.1}$$

The first term on the right side of (4.1) is estimated from the parameter values, $\hat{\boldsymbol{\beta}}$ say, of the fitted quadratic surface whilst the second, which represents the simulation-induced variance in the point estimate $\boldsymbol{\theta}^*$, can be estimated from the estimated variance matrix of $\hat{\boldsymbol{\beta}}$ by the delta technique, since $\boldsymbol{\theta}^*$ is itself an explicit function of $\hat{\boldsymbol{\beta}}$.

Termination of this second phase requires the following five criteria to be satisfied, using the standard tests of sums of squares described in Cochran and Cox (1957, Ch. 8) and taking "significant" to mean at the conventional 5 per cent level:

(i) a maximum is predicted;
(ii) the sum of squares for lack of fit to the quadratic surface is not significant;
(iii) the sum of squares for the quadratic terms is significant;
(iv) the sum of squares for the linear terms is not significant;
(v) each term in the estimated $\mathrm{Var}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})$ matrix is at least 100 times as large as the equivalent term in the $\mathrm{Var}_{\mathbf{X}}(\boldsymbol{\theta}^* \mid \mathbf{y})$ matrix.

Failure to meet the above criteria invokes one or more of the following modifications to the current design, as indicated in Table 2:

(a) increase the simulation size;
(b) increase inter-point distances in the design;
(c) decrease inter-point distances in the design;
(d) if a maximum is predicted, reshape the design in accordance with current estimates of the quadratic terms;
(e) if a maximum is not predicted, re-shape the design by increasing interpoint distances along any axis corresponding to a one-dimensional slice with a predicted minimum.

The strategy shown in Table 2 has been arrived at by a combination of heuristic argument and practical experimentation, and is not claimed to be optimal.

4.2. *Empirical Assessment of the SQ Algorithm*

In order to investigate the performance of the *SQ* algorithm, we first generated data $\mathbf{y}$ as the expected order statistics of an independent random sample of size $n = 25$ from the standard Normal distribution, yielding maximum likelihood estimates $\hat{\mu} = 0.0000$ and $\hat{\sigma}^2 = 0.9159$. Keeping these data fixed, we then estimated μ and σ^2 from 50 replications of the *SQ* algorithm, using randomly determined starting values for the parameters and the random number sequence. The initial simulation size was $s = 50$.

Table 3a gives the results of this simulation experiment. Because the data are fixed, our interest

TABLE 2
Strategy for modification of experimental design

Termination criteria satisfactory (√) or unsatisfactory (×)					*Recommended modifications (√)*				
(i)	(ii)	(iii)	(iv)	(v)	(a)	(b)	(c)	(d)	(e)
√	√	√	√	×	√	.	.	.	.
√	√	√	×	.	.	.	.	√	.
√	√	×	√	.	√	√	.	.	.
√	√	×	×	.	.	√	.	.	.
√	×	√	√	.	√	.	√	√	.
√	×	√	×	.	.	.	√	√	.
√	×	×	.	.	√	.	√	.	.
×	√	√	√	.	.	.	.	.	√
×	√	√	×	.	√	√	.	.	√
×	.	×	√	.	.	√	.	.	.
×	.	×	×	.	√	√	.	.	.
×	×	√	√	.	.	.	√	.	√
×	×	√	×	.	√	.	√	.	√

is in the ability of the *SQ* algorithm to recover the properties of $\hat{\boldsymbol{\theta}}$ conditional on $\mathbf{y}$, and the "true" values for the elements of Var $(\hat{\boldsymbol{\theta}})$ referred to in Table 3a are therefore based on the observed Fisher information matrix. For each replicate of the *SQ* algorithm, the estimated values for Var $(\hat{\boldsymbol{\theta}})$ are calculated from the curvature of the fitted quadratic surface. The results suggest that the *SQ* algorithm produces approximately unbiased estimators of both $\hat{\boldsymbol{\theta}}$ and the elements of Var $(\hat{\boldsymbol{\theta}})$, conditional on $\mathbf{y}$; the biggest discrepancy is the difference between $\hat{\sigma}^2$ and the mean of its implicit model estimate of about five times the empirical standard error. Note also that the figures 0.00013 and 0.00053 in the final column of Table 3a, which effectively represent $\mathrm{Var}_{\mathbf{X}}(\boldsymbol{\theta}^* \mid \mathbf{y})$, are of a smaller order of magnitude than the corresponding Var $(\hat{\boldsymbol{\theta}})$ as a direct

TABLE 3
Empirical assessment of the SQ algorithm

(a) *Normal*

Quantity	*True value*	*Implicit model estimates (50 replicates) Mean (standard error)*	*Variance*
$\hat{\mu}$	0.0000	−0.0024 (0.0016)	0.00013
$\hat{\sigma}^2$	0.9159	0.9332 (0.0032)	0.00053
Var $(\hat{\mu})$	0.0366	0.0405 (0.0016)	0.00015
Cov $(\hat{\mu}, \hat{\sigma}^2)$	0.0000	−0.0001 (0.0017)	0.00016
Var $(\hat{\sigma}^2)$	0.0671	0.0670 (0.0023)	0.00026

(b) *Weibull*

Quantity	*True value*	*Implicit model estimates (50 replicates) Mean (standard error)*	*Variance*
$\hat{\lambda}$	0.9657	0.9740 (0.0022)	0.00024
$\hat{k}$	1.0852	1.0892 (0.0018)	0.00016
Var $(\hat{\lambda})$	0.0422	0.0472 (0.0021)	0.00022
Cov $(\hat{\lambda}, \hat{k})$	−0.0119	−0.0133 (0.0013)	0.00009
Var $(\hat{k})$	0.0287	0.0299 (0.0019)	0.00019

consequence of our incorporating the stopping criterion (v), above.

The Normal distribution is self-serving for the implementation of the kernel method described in Section 3.3 and to some extent for the quadratic fitting phase of the *SQ* algorithm. In a second experiment, we therefore generated data **y** as the expected order statistics of an independent random sample of size $n = 25$ from a Weibull distribution with distribution function $F(y) = 1 - \exp(-\lambda y^k)$ and $\lambda = k = 1$. A Box-Cox analysis yielded a transformation to the power 0.22 to convert the data to approximate Normality, the success of which was confirmed by a Normal probability plot. The remainder of the experiment continued as for the Normal example. The results, given in Table 3b, are encouragingly similar in nature, although the average time to convergence of the algorithm was greater by a factor of about two, reflecting the greater difficulty of obtaining a good quadratic approximation to the log-likelihood.

5. RESPONSE OF FLOUR-BEETLES TO POISON

Hewlett (1974) describes an experiment in which male and female flour-beetles (*Tribolium castaneum*) are sprayed with insecticide (pyrethrins B) in Risella 17 oil. Hewlett's data, here reproduced as Table 4, show the resulting frequency distribution of times from dosage to death and overall proportions of deaths for a series of eight experiments involving male or female beetles and four different concentrations of insecticide. The beetles were fed during the experiment to eliminate control mortality. Hewlett uses analysis of variance and probit analysis (Finney, 1971) on transformed times of death and overall mortality levels, respectively, but says that this: "is not fully satisfactory. What is required is some single process of computation for estimating simultaneously the parameters of distributions of tolerances and times to death".

TABLE 4

Numbers of male (M) and female (F) flour-beetles dying in successive time intervals following spraying with insecticide at four different concentrations (reproduced from Hewlett, 1974)

Time interval (days)	*Concentration (mg/cm² deposit)* 0.20		0.32		0.50		0.80	
	M	*F*	*M*	*F*	*M*	*F*	*M*	*F*
0–1	3	0	7	1	5	0	4	2
1–2	11	2	10	5	8	4	10	7
2–3	10	4	11	11	11	6	8	15
3–4	7	8	16	10	15	6	14	9
4–5	4	9	3	5	4	3	8	3
5–6	3	3	2	1	2	1	2	4
6–7	2	0	1	0	1	1	1	1
7–8	1	0	0	1	1	4	0	1
8–9	0	0	0	0	0	0	0	0
9–10	0	0	0	0	0	0	1	1
10–11	0	0	0	0	0	0	0	0
11–12	1	0	0	0	0	1	0	0
12–13	1	0	0	0	0	1	0	0
No. of survivors	101	126	19	47	7	17	2	4
Total no. treated	144	152	69	81	54	44	50	47

Puri and Senturia (1972) criticize the classical probit model for data of this type, and propose a family of stochastic models with the following structure. The notional amount of insecticide in an insect's system as a function of time, t, is a stochastic process $\{Z(t)\}$ with $Z(0) \equiv z_0$, a notional initial dose (here, concentration). The insecticide is eliminated from the insect's system in a series of releases of independent random amounts V_i at random times T_i; thus, if $N(t)$ is the number of $T_i \leqslant t$, then

$$Z(t) = \max\left\{0, z_0 - \sum_{i=1}^{N(t)} V_i\right\}.$$

Finally, the hazard to the insect is a specified function $h(z, t)$ such that $h(z, t)\delta t$ is the probability of death in $(t, t + \delta t)$, conditional on survival to t and $Z(t) = z$. For the special case in which $N(t)$ has a Poisson distribution with mean αt, the V_i are exponentially distributed with parameter β and $h(z, t) = \gamma z$, Puri and Senturia obtain a complicated but numerically tractable expression for $F(y) = P\{\text{death at or before time } y\}$; note that $F(\infty) \leqslant 1$ denotes the overall mortality level. We refer to this as the basic PS model.

We first use maximum likelihood estimation to fit the basic PS model to Hewlett's data, fitting separate values of α, β and γ to the eight columns of Table 4. The fit is generally poor, because the model can only generate reverse-J-shaped distributions of survival time whereas all eight combinations of sex and z_0 suggest unimodal, positively skewed distributions. Table 5 compares observed and expected frequencies for male beetles with $z_0 = 0.2$ mg/cm^2. Formal application of Pearson's X^2 gives an attained significance level of about 6 per cent.

TABLE 5

Observed and expected frequencies for survival times of male flour-beetles after dosage with pyrethrins at concentration $z_0 = 0.2$ mg/cm^2

Time interval (days)	*Observed frequencies*		*Expected frequencies for*		
			Basic PS	*Extended* PS	*Extended* PS ($\beta = \gamma = 0$)
0–1	3		9.6	4.0	3.4
1–2	11		7.9	8.5	7.4
2–3	10		6.4	9.4	8.4
3–4	7		5.1	7.8	7.3
4–5	4		3.9	5.7	5.5
5–6	3	5	5.3	6.0	6.3
6–7	2				
7–8	1				
8–9	0				
9–10	0	3	4.5	2.7	3.4
10–11	0				
11–12	1				
12–13	1				
No response	101		101.3	99.9	102.3
Total	144		144.0	144.0	144.0
χ^2			9.01	1.82	2.86
d.f.			4	3	5
P-value			0.06	0.61	0.72

The simplest explanation of an initially increasing hazard is that a fixed level of insecticide might progressively weaken the insect. We therefore consider an "extended PS" model for which $h(z, t) = (\gamma + \rho t)z$. In the terminology of Section 1, this defines an implicit statistical model and we use the methods of Sections 3 and 4 to fit it. We follow Hewlett in using the same transformation to the power 0.309 for all eight sets of data; analyses using separate transformations gave substantially the same results. Since the individual times to death are not available, we estimate the probabilities corresponding to the frequency distributions in Hewlett's Table 4 as

$$p_j^* = \int_{a_{j-1}}^{a_j} \hat{f}(y)\, dy,$$

where $\hat{f}(\cdot)$ is estimated from ungrouped simulated realizations of the model and the a_j refer to the class boundaries in Table 4, but on the transformed scale. Because the grouping in Table 4 is relatively fine, we prefer this to the multinomial approach described in Section 2. Table 5 includes estimates of the expected frequencies, based on $s = 5000$ simulations, for this extended PS model which now gives an acceptable fit; with $s = 5000$, the simulation-induced random variation has a negligible effect on the X^2 statistic. Table 6 shows the eight sets of point estimates $\boldsymbol{\theta}^*$. In obtaining some of these we experienced considerable difficulty with the quadratic fitting phase of the algorithm. The explanation for this lies in the estimated values of β^*. Since the expected amount of each release is β^{-1}, our point estimates imply that in most cases, the first release exceeds z_0 with high probability. Plots of slices through the estimated likelihood surface confirm that $L(\boldsymbol{\theta})$ is virtually constant as β approaches zero.

TABLE 6

Point estimates for the extended PS *model*

	z_0	α^*	$1/\beta^*$	γ^*	ρ^*
Male	0.20	0.49	0.19	0.08	0.31
	0.32	0.13	9.09	0.10	0.56
	0.50	0.08	1.22	0.16	0.36
	0.80	0.11	0.33	0.12	0.26
Female	0.20	0.80	2.63	0.48	0.31
	0.32	0.36	2.17	0.29	0.28
	0.50	0.13	5.56	0.09	0.10
	0.80	0.06	1.03	0.10	0.19

Setting $\beta = 0$ has the simple physical interpretation that all the insecticide is eliminated in a single release. With this simplification, analytical progress is again possible. Let $g(t)$ be the pdf of time to release. For given z_0 and hazard $h(z, t)$, the probability of death at or before time y is

$$F(y) = 1 - \int_0^y \exp\left\{-\int_0^t h(z_0, u)\,du\right\} g(t)\,dt - \exp\left\{-\int_0^y h(z_0, u)\,du\right\} \int_y^\infty g(t)\,dt.$$

In particular, if we retain the forms for $h(\cdot)$ and $g(\cdot)$ assumed in the extended PS model, this reduces to

$$F(y) = 1 - \alpha \int_0^y \exp\left[-\{(\alpha + \gamma z_0)\,t + \tfrac{1}{2}\rho z_0 t^2\}\right] dt - \exp\left[-\{(\alpha + \gamma z_0)y + \tfrac{1}{2}\rho z_0 y^2\}\right]. \quad (5.1)$$

Maximum likelihood estimates of γ in this model turn out to be close to, and certainly not significantly different from, zero for all eight sets of data. Our final model therefore takes the form of (5.1), but with $\gamma = 0$. Table 5 includes expected frequencies corresponding to the data for male beetles and $z_0 = 0.2$ mg/cm^2. Maximum likelihood estimates, standard errors and X^2 goodness-of-fit statistics for all eight sets of data are given in Table 7; the model fits well, with the single exception of female beetles, $z_0 = 0.2$.

If the model is a reasonable representation of the underlying biology, the parameters should vary systematically, if at all, with sex and z_0. A plot of $\ln \hat{\alpha}$ against z_0, Fig. 4a, shows a clear, approximately linear trend, with $\hat{\alpha}$ consistently larger for females than for males. The corresponding plot of $\ln \hat{\rho}$ against z_0, Fig. 4b, shows $\hat{\rho}$ consistently smaller for females than for males but no simple relationship with z_0. For both males and females, likelihood ratio tests formally reject the hypothesis of constant ρ against the unrestricted alternative, with attained significance levels of less than 0.1 per cent and about 0.3 per cent for males and females respectively. A physical interpretation of the complete set of parameter estimates is that the average time to

TABLE 7

Point estimates, standard errors and χ^2 goodness-of-fit statistics for the extended PS *model with $\beta = \gamma = 0$*

	z_0	$\hat{\alpha}$ (*standard error*)	$\hat{\rho}$ (*standard error*)	X^2 (*P-value*)
Male	0.20	0.286 (0.036)	0.284 (0.056)	2.86 (0.722)
	0.32	0.116 (0.027)	0.589 (0.091)	6.45 (0.168)
	0.50	0.047 (0.018)	0.345 (0.053)	4.60 (0.467)
	0.80	0.013 (0.009)	0.197 (0.029)	3.86 (0.570)
Female	0.20	0.364 (0.051)	0.203 (0.058)	16.43 (0.001)
	0.32	0.250 (0.040)	0.336 (0.070)	5.62 (0.132)
	0.50	0.099 (0.025)	0.107 (0.024)	3.67 (0.300)
	0.80	0.027 (0.014)	0.175 (0.028)	4.91 (0.179)

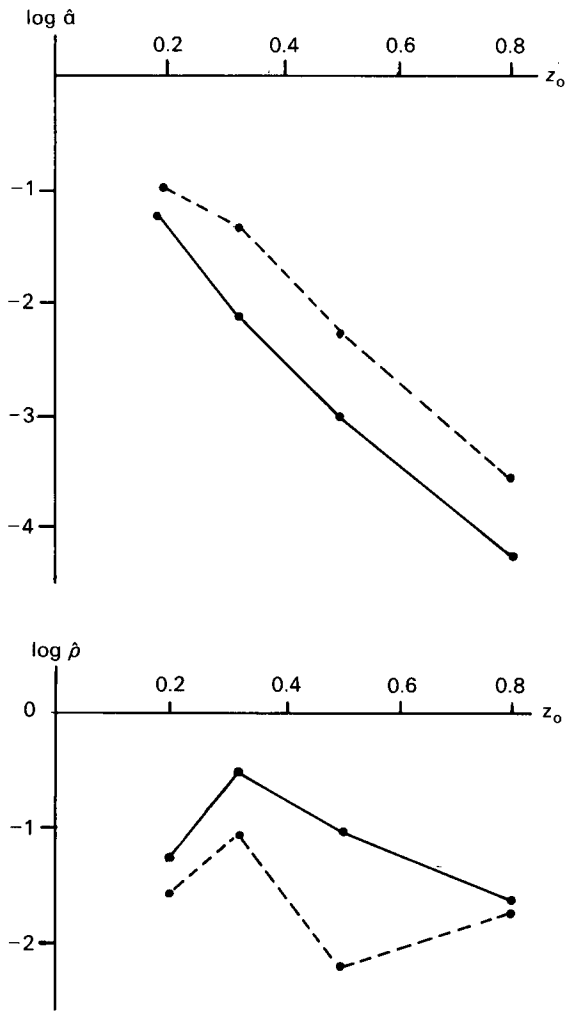

Fig. 4. Relationship between parameter estimates and concentration. ——, Males; - - -, females.

release increases with z_0 and is longer for males than for females, and that for a given value of z_0, males are at a higher risk than females.

6. SPATIAL PATTERN OF DISPLACED AMACRINE CELLS IN THE RETINA OF A RABBIT

Fig. 5 shows the locations of 152 displaced amacrine cells in an area of approximately 1070 x 600 μ in the retinal ganglion cell layer of a rabbit. These cells process "light-on" information in the eye; a second, apparently independent, pattern of cells which deal with "light-off" information is not shown. A quantitative description of the evident regularity of such patterns is of interest, because of its implications for development and retinal sampling efficiency (Hughes, 1981).

Pairwise interaction point processes (Ripley, 1981, pp. 166–167) provide a useful class of models for regular spatial point patterns, but are notoriously intractable. These models assume that the likelihood of n points at specified locations in a given region is proportional to

$$\prod_{i<j} h(t_{ij}), \tag{6.1}$$

where t_{ij} is the distance between the ith and jth points and $h(\cdot)$ is a suitably parameterized "interaction function".

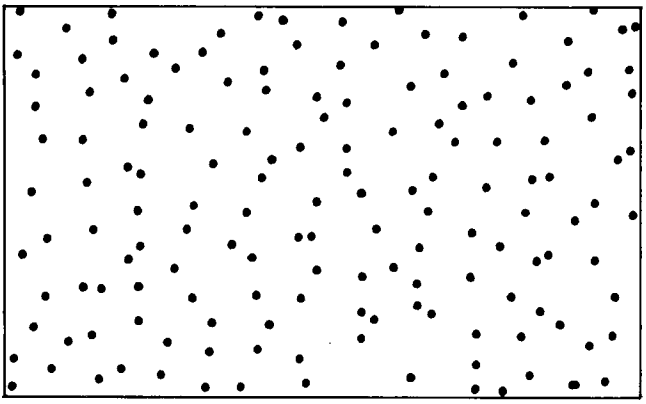

Fig. 5. Locations of 152 displaced amacrine cells within a 1070 X 600 μ rectangular region of rabbit retinal ganglion cell layer.

In constructing a parametric model for the cell data, inspection of Fig. 5 suggests setting $h(t) = 0$ for $t < \delta$, i.e. no two points (cell centres) can be separated by less than a distance δ. A further, natural condition to impose on $h(\cdot)$ is that $h(t) = 1$ for $t > \rho$, i.e. points interact only over a finite *range* ρ. It is then a short step to a three-parameter class of models with

$$h(t) = \begin{cases} 0: & t < \delta, \\ \{(t-\delta)/(\rho-\delta)\}^{\beta}: & \delta \leqslant t \leqslant \rho, \\ 1: & t > \rho. \end{cases} \tag{6.2}$$

Note that (6.2) embraces a one-parameter class of "simple inhibition" or "hard-core" models when $\delta = \rho$, but that this class turns out to be inadequate to describe the cell data.

Likelihood inference for a general pairwise-interaction process is precluded by the inaccessibility of the normalizing constant for (6.1), although some progress is being made in this direction using techniques imported from statistical mechanics (Ogata and Tanemura, 1981).

A more *ad hoc*, but generally practicable, approach is described in Diggle (1983, Ch. 5). In the present context, this operates as follows.

For parameter estimation we use the simplex phase of the *SQ* algorithm to minimize the quantity

$$d_s(\boldsymbol{\theta}) = \int_0^{t_0} \left[\{\hat{K}(t)\}^c - \left\{ s^{-1} \sum_{j=1}^{s} \hat{K}_j(t; \boldsymbol{\theta}) \right\}^c \right]^2 dt. \tag{6.3}$$

In (6.3), $\hat{K}(t)$ and $\hat{K}_j(t; \boldsymbol{\theta})$ denote Ripley's (1981, p. 159) estimator for the reduced second moment measure of a stationary point process, evaluated for the data and a simulation of the model, respectively. In addition, t_0 and c are "tuning constants". We used $t_0 = 1{\cdot}50\,\mu$ and $c = \frac{1}{2}$, and obtained point estimates $\boldsymbol{\theta}^* = (\delta^*, \beta^*, \rho^*) = (19, 1.67, 76)$. This implies considerably stronger spatial regularity than would be imposed by the physical dimensions of the cells, each of which is approximately circular with a diameter of about $10\,\mu$.

For an informal assessment of goodness-of-fit, Fig. 6 shows two graphical comparisons between the data and 19 simulations of the fitted model. In each, a summary description of the data is plotted as ordinate against the mean of the 19 simulations as abscissa, together with the upper and lower extremes from the 19 simulations to give an indication of sampling fluctuations under the model. In Fig. 6a, the "summary description" is the empirical distribution function (EDF) of distances from each of the n points of the process to its nearest neighbouring point, whilst Fig. 6b shows the EDF of distances from each of $m \cong n$ sampling origins in a regular grid arrangement to the nearest point of the process. Each plot suggests a good fit, although there seems to be no very satisfactory way of testing the fit, whilst formally allowing for the effects of parameter estimation. However, faith in the statistical adequacy of the model is strengthened by the use of two complementary summary descriptions, neither of which has a direct bearing on the method of parameter estimation.

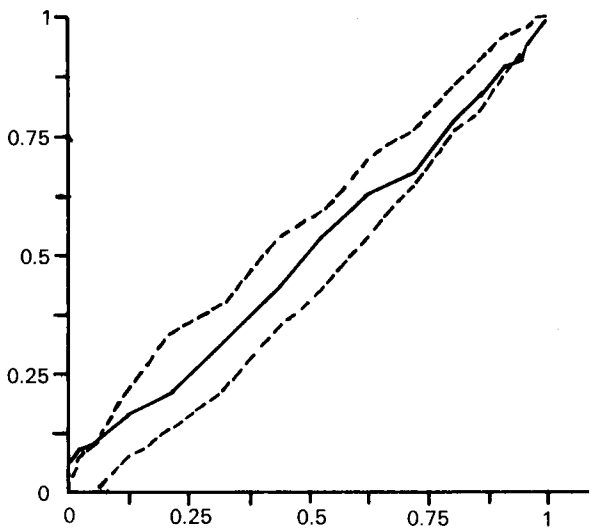

Fig. 6a

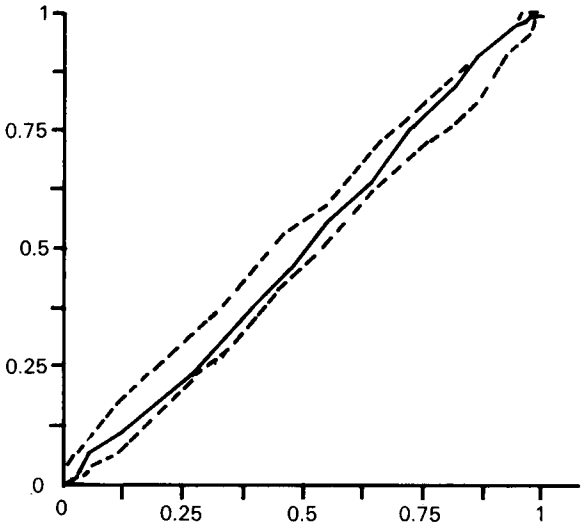

Fig. 6b

Fig. 6. Goodness-of-fit assessment for the pairwise interaction model. ——, Data; - - - -, upper and lower extremes from 19 simulations of model. (a) EDF of distances from points to nearest neighbouring points. (b) EDF of distances from sampling origins to nearest points.

7. CONCLUDING REMARKS

A simulation-based methodology for parameter estimation extends the range of models which can be fitted to a set of data. An immediate consequence is that a scientifically plausible implicit statistical model need not be ruled out on the grounds of mathematical intractability. Conversely, there is of course no value in fitting an arbitrary implicit model in favour of a simpler prescribed model unless the former has a scientific basis.

Our analysis of the extended PS model in Section 5 showed that the *SQ* algorithm should not be applied rigidly. Like any optimization algorithm it can encounter difficulties and, particularly in view of the presence of random noise in each function evaluation, these are best investigated by direct examination of the log-likelihood surface. In our example, this suggested a reduction from four to three parameters. In other instances, a transformation of the parameter space may be more appropriate in order to improve the quadratic approximation to $L(\boldsymbol{\theta})$.

Section 5 also illustrated the difficulty of fitting over-parameterized implicit statistical models, since any near-singularity in the parameter space will make optimization by quadratic approximation very problematical. The consequent acceptance of poorer quality estimates in turn has adverse implications for hypothesis tests based on estimated log-likelihoods. Although $L^*(\boldsymbol{\theta})$ is approximately unbiased for $L(\boldsymbol{\theta})$, the definition of $\hat{\boldsymbol{\theta}}$ necessitates that $L(\hat{\boldsymbol{\theta}}) \geqslant L(\boldsymbol{\theta}^*)$, so that $L^*(\boldsymbol{\theta}^*)$ is negatively biased for $L(\hat{\theta})$, and the bias will increase as poorer quality estimates $\boldsymbol{\theta}^*$ are accepted. This suggests that tests based on estimated log-likelihood ratios will tend to be conservative.

Our estimated log-likelihoods are a practical proposition only for models which generate independent realisations of low-dimensional random variables. However, the example of Section 6 showed that the simplex phase of the *SQ* algorithm can be useful for the analysis of more complicated stochastic structures by *ad hoc* methods. The quadratic fitting phase of the algorithm is primarily used to give good estimates of the curvature of the objective function in the vicinity of a maximum whose approximate location has already been identified in the simplex phase. The curvature estimates give approximate standard errors for parameter estimates if the

objective function is a log-likelihood, but are of limited value otherwise.

Most of the computing was carried out on an IBM 370/168 mainframe programmed in FORTRAN or, occasionally, APL. This simply reflects what was available to us. Because the work makes heavy demands on CPU time, but is relatively modest in its storage and software requirements, it would be well suited to implementation on a mini-computer (the results in Section 6 were obtained using a VAX 11/750), and might even be within the capacity of a modern, dedicated micro. To give some indication of what is involved, the mean time for the 50 replicates of the Weibull experiment reported in Section 4.2 was about 590 CPU seconds.

ACKNOWLEDGEMENTS

The paper was completed while P. J. D. was a Visiting Research Scientist at CSIRO Division of Mathematics and Statistics, Canberra.

R. J .G. received financial support from the Science and Engineering Research Council.

REFERENCES

Barnard, G. A. (1963) Contribution to the discussion of Prof. Bartlett's paper. *J. R. Statist. Soc.* B, **25**, 294.

Bartlett, M. S. (1963) Statistical estimation of density functions. *Sankhyā* (A), **25**, 245–254.

Bean, S. J. and Tsokos, C. P. (1980) Developments in non-parametric density estimation. *Int. Statist. Rev.*, **28**, 267–287.

Benyon, P. R. (1976) Remark AS R15. Function minimization using a simplex procedure. *Appl. Statist.*, **25**, 97.

Besag, J. and Diggle, P. J. (1977) Simple Monte Carlo tests for spatial pattern. *Appl. Statist.*, **26**, 327–333.

Blum, J. R. (1954) Multidimensional stochastic approximation method. *Ann. Math. Statist.*, **25**, 737–744.

Boneva, L. I., Kendall, D. G. and Stefanov, I. (1971) Spline transformations: three new diagnostic aids for the statistical data-analyst (with discussion). *J. R. Statist. Soc.* B, **33**, 1–70.

Bowman, A. (1982) A comparative study of some kernel-based non-parametric density estimators. Research Report, Manchester–Sheffield School of Probability and Statistics.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Buckland, S. T. (1980) A modified analysis of the Jolly–Seber capture–recapture model. *Biometrics*, **36**, 419–435.

Chambers, J. M. (1977) *Computational Methods for Data Analysis.* New York: Wiley.

Chambers, J. M. and Ertel, J. E. (1974) Remark AS R11. A remark on Algorithm AS 47 "Function minimization using a simplex procedure". *Appl. Statist.*, **23**, 250–251.

Cochran, W. G. and Cox, G. M. (1957) *Experimental Designs.* New York: Wiley.

Copas, J. B. and Fryer, M. J. (1980) Density estimation and suicide risks in psychiatric treatment. *J. R. Statist. Soc.* A, **143**, 167–176.

Davis, K. B. (1981) Estimation of the scaling parameter for a kernel-type density estimate. *J. Amer. Statist. Ass.*, **76**, 632–636.

Diggle, P. J. (1983) *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.

Dixon, L. C. W. (1972) *Nonlinear Optimisation.* London: English University Press.

Efron, B. (1979) Bootstrap methods – another look at the jack-knife. *Ann. Statist.*, 7, 1–26.

Epanechnikov, V. A. (1969) Non-parametric estimation of a multivariate probability density. *Theor. Prob. Appl.*, **14**, 153–158.

Finney, D. J. (1971) *Probit Analysis*, 3rd ed. Cambridge: University Press.

Fryer, M. J. (1976). Some errors associated with the non-parametric estimation of density functions. *J. Inst. Math. Appl.*, **18**, 371–380.

Fryer, M. J. (1977) A review of some non-parametric methods of density estimation. *J. Inst. Math. Appl.*, **20**, 335–354.

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

Gratton, R. J. (1979) Generalizing the test graph method for estimating functions of an unknown density. *Biometrika*, **66**, 663–667.

——— (1981) Unpublished Ph.D. Thesis. University of Newcastle upon Tyne.

Habbema, J. D. F., Hermans, J. and Van Der Broek, K. (1974) A stepwise discriminant analysis programme using density estimation. In *Compstat, 1974* (G Bruckman, ed.), pp. 101–110. Vienna: Physica Verlag.

Hewlett, P. S. (1974) Time from dosage to death in beetles, *Tribolium castaneum*, treated with pyrethrins or DDT, and its bearing on dose-mortality relations. *J. Stored Prod. Res.*, **10**, 27–41.

Hill, I. D. (1978) Remark AS R28. A remark on Algorithm AS 47: Function minimization using simplex procedure. *Appl. Statist.*, **27**, 380–382.

Hoel, D. G. and Mitchell, T. J. (1971) The simulation, fitting and testing of a stochastic cellular proliferation model. *Biometrics*, **27**, 191–199.
Hope, A. C. A. (1968) A simplified Monte Carlo significance test procedure. *J. R. Statist. Soc.* B, **30**, 582–598.
Hughes, A. (1981) Cat retina and the sampling theorem: the relation of transient and sustained brisk-unit cut-off frequency to α and β-mode cell density. *Exp. Brain Res.*, **42**, 196–202.
Kiefer, J. and Wolfowitz, J. (1952) Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, **23**, 463–466.
Kronmal, R. A. and Tarter, M. E. (1968) The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Ass.*, **30**, 925–952.
Kubitschek, H. E. (1962) Normal distribution of cell generation rate. *Exptl. Cell Res.*, **26**, 39–50.
Loftsgaarden, D. O. and Quesenberry, C. P. (1965) A non-parametric estimate of a multivariate density function. *Ann. Math. Statist.*, **36**, 1049–1051.
Marriott, F. H. C. (1979) Barnard's Monte Carlo tests: how many simulations? *Appl. Statist.*, **28**, 75–77.
Murray, W. (ed.) (1972) *Numerical Methods for Unconstrained Optimisation*. London: Academic Press.
Nelder, J. A. and Mead, R. (1965) A simplex method for function minimisation. *Comp. J.*, **7**, 303–313.
Ogata, Y. and Tanemura, M. (1981) Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.*, **33**, 315–338.
O'Neill, R. (1971) Algorithm AS 47. Function minimization using a simplex procedure. *Appl. Statist.*, **20**, 338–345.
Plackett, R. L. (1977) Methods of inference for latent models. Unpublished.
Puri, P. S. and Senturia, J. (1972) On a mathematical theory of quantal response assays. *Proc. 6th Berk. Symp. Math. Stat.*, **4**, 231–247.
Ripley, B. D. (1981) *Spatial Statistics*. Chichester: Wiley.
Rosenblatt, M. (1956) Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.
Ross, G. J. S. (1972) Stochastic model-fitting by evolutionary operation. In *Mathematical Models in Ecology* (Jeffers, J. R. N., ed.), pp. 297–308. Oxford: Blackwell Scientific Publications.
Silverman, B. W. (1978a) Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.*, **6**, 177–184.
——— (1978b) Choosing the window width when estimating a density. *Biometrika*, **65**, 1–11.
——— (1981) Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* B, **43**, 97–99.
——— (1982) Algorithm AS 176. Kernel density estimation using the Fast Fourier transform. *Appl. Statist.*, **31**, 93–97.
Springer, B. G. F. (1969) Numerical optimisation in the presence of random variability: the single factor case. *Biometrika*, **56**, 65–74.
Tapia, R. A. and Thompson, J. R. (1978) *Non-parametric Probability Density Estimation*. Baltimore: John Hopkins University Press.
Titterington, D. M. (1980) A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.
Wasan, M. T. (1969) *Stochastic Approximation*. Cambridge: University Press
Wegman, E. J. (1972a) Non-parametric probability density estimation: I. A summary of available methods. *Technometrics*, **14**, 513–546.
——— (1972b) Non-parametric probability density estimation: II. A comparison of density estimation methods. *J. Statist. Comp. Simul.*, **1**, 225–245.

DISCUSSION OF THE PAPER BY DR DIGGLE AND DR GRATTON

Dr B. W. Silverman (University of Bath): The subject of tonight's paper is of fundamental importance in statistics. Far too much statistical modelling is "method-oriented" rather than "problem-oriented", in that the models are chosen for their statistical tractability rather than their appropriateness to the real process being studied. The work of Diggle and Gratton is an important step in the development of a methodology whereby any "mechanism-defined" model can be fitted to observed data.

There are, of course, very many areas of application where implicit statistical models arise. For example, *compartmental models* form a wide class of models for the analysis of systems in biology, medicine and pharmacology. See, for further discussion and references, Rodda *et al.* (1975) and Norton *et al.* (1980). The basic idea of compartmental models, somewhat related to the flour-beetle example, is that the system of interest, for instance the human body, is made up of several compartments. The flow of some material (such as a drug) between the various compartments is modelled by a system of differential equations which may have random components. Except in very special cases, the statistical structure of the observed responses will