BAYESIAN STATISTICS 4, pp. 345–363
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, (Eds.)
© Oxford University Press, 1992

Some Bayesian Numerical Analysis

A. O'HAGAN
University of Nottingham, UK

SUMMARY

Bayesian approaches to interpolation, quadrature and optimisation are discussed, based on representing prior information about the function in question in terms of a Gaussian process. Emphasis is placed on how different methods are appropriate when the function is cheap or expensive to evaluate. A particular case of expensive functions is a regression function, where 'evaluation' consists of gaining observations (with the small added complication of measurement error).

Keywords: BAYES-HERMITE QUADRATURE; GAUSSIAN PROCESS; INTERPOLATION; NONPARAMETRIC REGRESSION; OPTIMISATION; QUADRATURE; SMOOTHING.

1. PRIORS FOR FUNCTIONS

Numerical techniques of interpolation, optimisation and quadrature have been studied for a very long time. They are all concerned with obtaining numerical approximations to specific properties of a given function — function values, locations of maxima and minima, integrals — that are not available exactly. In each case, the technique is based entirely on numerical evaluation of the function at a discrete set of points. The central theme of this paper is that all these problems are problems of inference. Inference is required about some unknown parameter — function value, integral, etc — based on data which are the function evaluations. Furthermore, of course, the mode of inference we shall use will be Bayesian. The formal structure is as follows.

The function in question is $f(\cdot)$, taking values $f(x)$ for $x \in \mathcal{X}$. We may or may not have an explicit expression for $f(\cdot)$, but in either case the numerical value $f(x)$ is unknown *a priori*. The prior distribution therefore consists of a distribution for the random function $f(\cdot)$. We then 'observe' the values $f = (f(x_1), \ldots, f(x_n))^T$ of the function at n discrete points in $\mathcal{X}$. The posterior distribution is simply the prior distribution conditioned on these values. Although in principle any prior distribution might be used, to reflect specific prior beliefs about $f(\cdot)$, as always it is the normal distribution that is the most tractable, and this paper will concentrate on techniques arising from representing prior information about $f(\cdot)$ in terms of a Gaussian process.

We shall assume the following hierarchical prior model. First,

$$f(\cdot)|\beta, \sigma^2 \sim N(h(\cdot)^T\beta, \sigma^2 v(\cdot, \cdot)). \tag{1}$$

That is, $f(\cdot)$ is a Gaussian process, conditional on hyperparameters β and σ^2, with mean $E(f(x)|\beta, \sigma^2) = h(x)^T\beta$ and covariance function $\operatorname{Cov}(f(x), f(x')|\beta, \sigma^2) = \sigma^2 v(x, x')$. The vector $h(\cdot)$ of q regressor functions $h(x) = (h_1(x), \ldots, h_q(x))^T$, and the covariance function $v(\cdot, \cdot)$ are known. The prior distribution of β and σ^2 is then given by

$$\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0^{-1}), \tag{2}$$

$$\sigma^2 \sim s_0\chi^{-2}_{a_0}. \tag{3}$$

The special case of an uninformative prior distribution on the hyperparameters, represented by $B_0 = 0$, $s_0 = a_0 = 0$, so that $p(\beta, \sigma^2) \propto \sigma^{-2}$, will often be appropriate.

This is about the most general structure that yields tractable posterior analysis. Some further generalisation can be achieved by letting a function of $f(\cdot)$ have this distribution; for instance, we could let $f(x) = f_1(x) + f_2(x)f_3(x)$, where $f_1(\cdot)$ and $f_2(\cdot)$ are known functions and $f_3(\cdot)$ has the distribution (1)–(3). The mathematics would be essentially unchanged.

Gaussian process priors for functions have been proposed several times in a variety of contexts, as will be apparent from the references throughout this paper.

2. INTERPOLATION AND SMOOTHING

2.1. *Cheap and Expensive Functions*

Another theme of this paper is that the most appropriate methods for a given problem depend on how costly it is to obtain each evaluation $f(x_i)$ of the function. In interpolation, for instance, it is implicit that the function is expensive to evaluate. Otherwise, if we wanted $f(x)$ we would just evaluate it exactly rather than approximate it by interpolation. Interpolation was a very important technique before the ready availability of computers. Then for most complex functions the only recourse was interpolation in a book of tables wherein every figure was the result of somebody's very laborious computation. Although this is not true for the commonly used functions today, there is still no shortage of functions whose evaluation is impractical except by extensive use of the most powerful computers. See Sacks, Welch, Mitchell, and Wynn (1989).

Optimisation and quadrature (numerical integration) are important even for cheap functions. When maxima/minima and integrals cannot be found analytically, numerical methods are essential. Particularly in high dimensional spaces, solving these problems will typically require very large numbers of function evaluations. Many methods then become impractical, even with cheap function evaluations.

2.2. *Interpolation*

Given the prior distribution (1)–(3) and data f, the posterior distribution is easily obtained. See for instance O'Hagan (1978). We find

$$f(\cdot)|\beta, \sigma^2, f \sim N(m(\cdot), \sigma^2 w(\cdot, \cdot)), \tag{4}$$

where

$$m(x) = h(x)^T\beta + t(x)^T A^{-1}(f - H\beta), \tag{5}$$

$$t(x) = \begin{pmatrix} v(x, x_1) \\ \vdots \\ v(x, x_n) \end{pmatrix}, \qquad H = \begin{pmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{pmatrix}, \tag{6}$$

$$A = \begin{pmatrix} v(x_1, x_1) & \dots & v(x_1, x_n) \\ \vdots & & \vdots \\ v(x_n, x_1) & \dots & v(x_n, x_n) \end{pmatrix}, \tag{7}$$

$$w(x, x') = v(x, x') - t(x)^T A^{-1} t(x'). \tag{8}$$

Also

$$\beta|\sigma^2, f \sim N(b, \sigma^2 B^{-1}), \tag{9}$$

where

$$b = B^{-1}(B_0 b_0 + H^T A^{-1} f), \tag{10}$$
$$B = B_0 + H^T A^{-1} H. \tag{11}$$

Finally,

$$\sigma^2 | f \sim s\chi_a^{-2},$$

where $a = a_0 + n - q$ and

$$s = s_0 + f^T\{A^{-1} - A^{-1}H(H^T A^{-1} H)^{-1} H^T A^{-1}\} f. \tag{12}$$

It follows that the posterior distribution of any particular $f(x)$ is t with degrees of freedom a, mean

$$m^*(x) = h(x)^T b + t(x)^T A^{-1}(f - Hb) \tag{13}$$

and variance $s\, w^*(x, x)/(a - 2)$, where

$$w^*(x, x') = w(x, x') + \{h(x) - H^T A^{-1} t(x)\}^T B^{-1}\{h(x') - H^T A^{-1} t(x')\}. \tag{14}$$

The natural interpolant is (13). In particular of course, at the observation points x_i we have $m^*(x_i) = f(x_i)$ and $w^*(x_i, x_i) = 0$.

We have already said that interpolation is only relevant for expensive functions. A related point is that when the function is very expensive it can be worth spending a great deal of effort to locate good *design* points $(x_1, \ldots, x_n)$. Sacks, Welch, Mitchell and Wynn (1989) develop designs to minimise the integral of the posterior variance

$$\int_{\mathcal{X}} w^*(x, x)\, \mathrm{d}x$$

over the space $\mathcal{X}$, and this is easily generalised as in O'Hagan (1978) to minimise a weighted average

$$\int_{\mathcal{X}} w^*(x, x)\, \mathrm{d}\Omega(x).$$

Sacks and Schiller (1988) advocate minimising the maximum of $w^*(x, x)$ over a subset of $\mathcal{X}$.

2.3. *The Prior Covariance Function*

The interpolant (13) has a simple and natural form. The first term is the fitted regression $h(x)^T b$ corresponding to the regression model $h(x)^T \beta$ assumed for the prior mean. The estimate of the coefficient vector, b, is a familiar Bayesian, weighted average of the prior mean b_0 and a generalised Least Squares estimate $\hat{\beta} = (H^T A^{-1} H)^{-1} H^T A^{-1} f$. In the case of a non-informative prior, b reduces to $\hat{\beta}$.

The second term in (13) is the interpolant of the residuals $f - Hb$ from this fitted line, and ensures that $m^*(x)$ interpolates the actual observations. This term takes the form

$$t(x)^T A^{-1}(f - Hb) = \sum_{i=1}^{n} k_i v(x, x_i), \tag{15}$$

where the k_i's are the elements of the vector $A^{-1}(f - Hb)$. Indeed, these are the unique coefficients that make the right hand side of (13) interpolate the observations. The interpolant

(13) will have properties of smoothness, such as differentiability, if and only if the prior covariance function $v(\cdot,\,\cdot)$ has those properties. This is an important guide to the form of covariance function that is appropriate. In particular, the one-dimensional Markov form

$$v(x,\,x') = \exp\{-b|x-x'|\} \tag{16}$$

used by Blight and Ott (1975) produces an interpolant (13) that is not differentiable at the observed points. The stationary Gaussian form for $x \in R^p$,

$$v(x,\,x') = \exp\{-b(x-x')^T V^{-1}(x-x')\} \tag{17}$$

is used by O'Hagan (1991) and Sacks, Welch, Mitchell and Wynn (1989), and produces smooth interpolation, infinitely differentiable everywhere. O'Hagan (1978) also uses (17), but in a rather different model from (1)–(3). Wahba (1978, 1983) chooses covariance functions that make the interpolant a spline function.

2.4. *Example*

An interesting example of interpolation is drawing contours. Consider the contours of the bivariate density function proportional to $\pi(\theta,\,\phi) = (1+\theta^2)^{-2}(1+\phi^2)^{-2}$, where θ and ϕ are independent t random variables with 3 degrees of freedom. It is easy to solve $\pi(\theta,\,\phi) = c$ for $\theta = 0$, or $\phi = 0$ or $\theta = \pm\phi$. This gives eight fixed points on any contour line. Figure 1 shows three contours, at $c = 0.4$, 0.03, 0.001, drawn by interpolating these eight points in each case. In order to do this, I set up the unknown function as a vector-valued $f(x)$ taking values in the plane, with x taking values in $[0,\,8)$. Then $f(0),\, f(1), \ldots, f(7)$ are the eight 'observed' points in sequence:

$$f(0) = (d_1,\,0), f(1) = (d_2,\,d_2), f(2) = (0,\,d_1), f(3) = (-d_2,\,d_2), \ldots, f(7) = (d_2,\,-d_2),$$

where

$$d_1 = (c^{-1/2} - 1)^{1/2}, \quad d_2 = (c^{-1/4} - 1)^{1/2}.$$

The preceding theory is easily generalised to vector $f(\cdot)$ (as in O'Hagan (1978)). The covariance function was defined to be

$$\mathrm{Cov}\,(f(x),\,f(x')) = \exp\{-b||x-x'||^2\}I_2,$$

with

$$||x-x'|| = \min\,((|x-x'|,\,8-|x-x'|)$$

reflecting the fact that the contour is a closed loop. The case $q = 0$ was used, removing the regression term and giving $f(\cdot)$ a zero prior mean.

The smoothing parameter b in the covariance function was chosen to give good fit of the interpolated contour to the true value, in the following way. The value of the density $\pi(m^*(1/2))$ was found for the interpolated value $x = 1/2$, and b was adjusted to make this close to the required value c. The resulting interpolated contours are extremely good. $\pi(m^*(x))$ was found for a large selection of x values and found to be within 2% of the true c. (This is for the outermost, and most difficult to interpolate, contour. Accuracy to within 0.001% was achieved on the innermost contour).

An alternative approach, transforming to polar coordinates and representing the contour as a scalar function of the angle, proved to be less accurate (and could anyway fail for a non-convex contour).

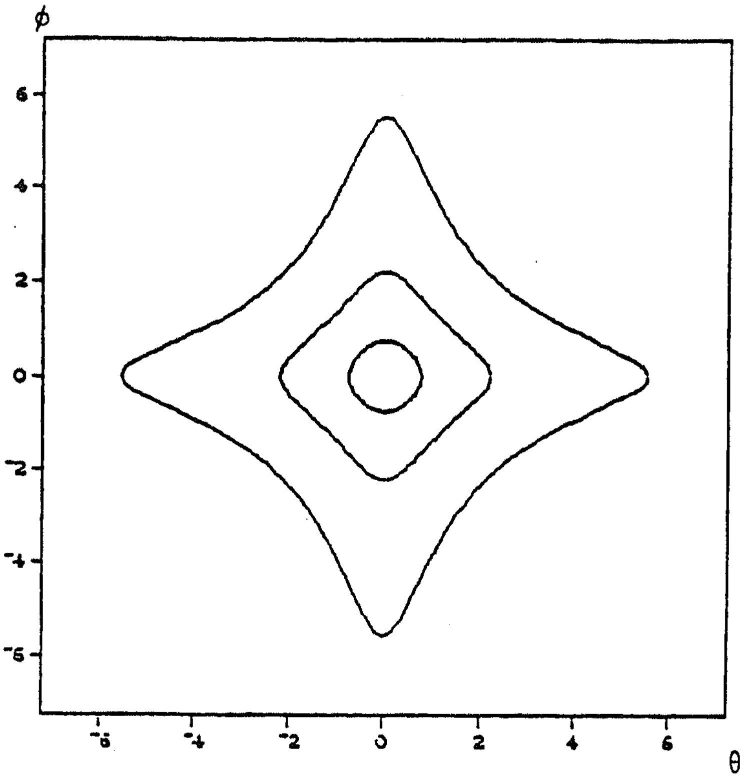


Figure 1. *Contours of product of independent t distributions.*

2.5. *Observing Derivatives*

There is a parallel theory for derivatives of $f(\cdot)$ itself. By considering the random variable $X_\delta = \delta^{-1}\{f(x+\delta) - f(x)\}$ and letting $\delta \to 0$ (for scalar x) it is easy to show that providing the random sequence X_δ converges, and providing that the indicated derivatives of $h(\cdot)$ and $v(\cdot,\, \cdot)$ exist, $f(\cdot)$ and all its derivatives are jointly normally distributed with

$$E(f^j(x)|\beta,\, \sigma^2) = h^j(x)^T\beta, \tag{18}$$

$$\operatorname{Cov}(f^j(x),\, f^i(x')|\beta, \sigma^2) = \sigma^2 v^{ji}(x,\, x'), \tag{19}$$

where

$$f^j(x) = \mathrm{d}^j f(x)/\mathrm{d}x^j, \quad h^j(x) = \mathrm{d}^j h(x)/\mathrm{d}x^j \quad \text{and} \quad v^{ji}(x,\, x') = \mathrm{d}^{i+j} v(x,\, x')/\mathrm{d}x^j \mathrm{d}x'^i.$$

Since the covariance function (16) is not differentiable, nor is $f(\cdot)$ in that case. On the other hand, it turns out that (17) is sufficiently smooth for $f(\cdot)$ to be infinitely differentiable everywhere with probability one (if $h(\cdot)$ is). This is another argument in favour of (17) rather than (16) as a prior covariance function. We are generally dealing with functions that we know to be differentiable, so that (16) simply does not represent that prior belief.

We can now consider making observations not of $f(\cdot)$ but of its derivatives. The idea is potentially useful in all problems. Although derivatives are often unavailable when $f(\cdot)$ is expensive, for cheap functions we can generally 'observe' them. It may in fact be even cheaper to observe $f'(x)$ when we are also observing $f(x)$. In the theory of Section 2.2 we need only replace the functions $v(\cdot,\, \cdot)$ and $h(\cdot)$ in (6) and (7) by appropriate derivatives, using (18) and (19). The result (13) will then not only pass through any observed values $f(x_i)$ but agree with any observed derivatives.

2.6. *Smoothing*

It is also simple to allow for observation error. If the function (or derivatives) cannot be evaluated, 'observed', exactly, the observation f can be represented as normally distributed with mean true values $(f(x_1), \ldots, f(x_n))$ and variance matrix $\sigma^2 V_f$, e.g. $V_f = v_f I_n$. Then we simply add V_f to A as defined in (7). This can occur with very expensive functions, but it can of course also be viewed in terms of a very standard statistical problem. In regression modelling, as an alternative to assuming that the regression function follows some simple parametrised form, we can let it be an arbitrary function $f(\cdot)$. The prior distribution (1)–(3) then says that the prior expectation of $f(\cdot)$ is a standard linear model $h(\cdot)^T\beta$, but the Gaussian process prior allows it to deviate smoothly from that form. This is the approach used in Blight and Ott (1975) and O'Hagan (1978). In the terms of this paper, statistical data are expensive function evaluations.

With A redefined to include the error V_f, (13) no longer passes through the observed points. Instead of interpolation, we have smoothing. If V_f is large, $m^*(x)$ is just the fitted linear model $h(x)^T b$.

2.7. *Implementation*

In implementation of these methods, the major practical difficulty is that of inverting the A matrix. Given n observations, this is an $n \times n$ matrix, and numerically inverting it is an order-n^3 operation. For sufficiently expensive functions, this will not be a problem, and anyway only needs to be done once, for any number of interpolations from the same data. Otherwise, if the function is only moderately expensive to evaluate and n is large, interpolation at any given point can be computed using only a more feasible number of nearest observations to that point.

Although A may be inverted analytically if we use the covariance function (16), so reducing the computation to order-n^2, we have rejected this function because of its poor interpolation properties. Analytic inversion of A is not possible with (17) or other realistic covariance functions. In two or more dimensions, however, O'Hagan (1991) shows that dramatic reduction of effort may be achieved if the observations lie on a rectangular grid. See also the discussion of tetrahedral designs in Section 3.3 below.

3. QUADRATURE

3.1. *Inference about an Integral*

Let

$$k = \int_{\mathcal{X}} f(x)\,\mathrm{d}M(x), \tag{20}$$

the integral of $f(\cdot)$ with respect to some measure $M(\cdot)$ on $\mathcal{X}$. This formulation is very general. For instance $M(\cdot)$ might be Lebesgue measure on $\mathcal{X} = R^p$ for vector x, so that k is just the (Lebesgue) integral of $f(\cdot)$. Or if $M(\cdot)$ is Lebesgue measure on $C \subset R^p$, and $M(B) = 0$ if $B \cap C = \phi$, then $k = \int_C f(x)\,\mathrm{d}x$.

The posterior distribution of k is given by (9), (12) and

$$k|f,\, \beta,\, \sigma^2 \sim N(m_k,\, \sigma^2 w_k) \tag{21}$$

where

$$m_k = \int_{\mathcal{X}} m(x)\,\mathrm{d}M(x), \tag{22}$$

$$w_k = \int_{\mathcal{X}^2} w(x,\, x')\,\mathrm{d}M(x)\,\mathrm{d}M(x'). \tag{23}$$

Its marginal posterior distribution is therefore t with degrees of freedom a and mean

$$m_k^* = \int_{\mathcal{X}} m^*(x)\,\mathrm{d}M(x) = h^{*T}b + t^{*T}A^{-1}(f - Hb)\,, \tag{24}$$

where

$$h^* = \int_{\mathcal{X}} h(x)\,\mathrm{d}M(x), \qquad t^* = \int_{\mathcal{X}} t(x)\,\mathrm{d}M(x). \tag{25}$$

This technique is called Bayesian quadrature by O'Hagan (1991), where it is explored in some detail. Earlier treatments are described in the review of Diaconis (1988).

3.2. *Bayesian Application*

Integration of complex, intractable, and often high-dimensional, posterior densities is a major interest in Bayesian statistics. Consider, therefore, two separate Bayesian analyses. One gives rise to a posterior density which we only know as proportional to $f(\cdot)$ (obtained as the product of prior density and likelihood). The second Bayesian analysis is the one considered in this paper, providing Bayesian inference about $f(\cdot)$, and in particular about integrals of $f(\cdot)$. The first analysis is only of interest here to provide the context for $f(\cdot)$. Specifically, $f(\cdot)$ is proportional to a density function, and given a reasonable amount of information may be expected to be smooth, and probably unimodal.

Suppose that $f(\theta)$ is proportional to a density function for a random vector $\theta \in R^p$ (the parameter vector in the first Bayesian analysis). Then we wish to estimate

$$\int r(\theta) f(\theta)\,\mathrm{d}G(\theta), \tag{26}$$

identifying $\mathrm{d}M(\cdot)$ in (22) with $r(\cdot)\mathrm{d}G(\cdot)$ and $\mathcal{X}$ with R^p. The interpretation of (26) is that the distribution of θ is represented as a density $f(\theta)$ with respect to a measure $G(\cdot)$ on R^p and we wish to find the expectation of $r(\theta)$. Or strictly, since the density of θ is only proportional to $f(\theta)$, the expectation of $r(\theta)$ is (26) divided by $\int f(\theta)\,\mathrm{d}G(\theta)$, which is itself the case $r(\theta) = 1$ of (26). Inference about ratios of integrals is discussed in O'Hagan (1989).

Choice of the measure $G(\cdot)$ is an important part of the prior specification. It would *not* be reasonable for a typical density in the usual sense, of a density with respect to Lebegue measure, to be represented by the prior distribution (1)–(3). However, if $G(\cdot)$ is a suitable multivariate normal probability measure, then it becomes reasonable to model $f(\cdot)$ in this way. O'Hagan (1991) deals specifically with 'Bayes-Hermite' quadrature, in which $G(\cdot)$ is multivariate normal (and without loss of generality is assumed to be the standard normal, $N(0, I)$) and the covariance function also takes the 'Gaussian' form (17). These choices have the effect of allowing the integrals (25) to be performed analytically. It is essential to be able to do this when $f(\cdot)$ is cheap to evaluate, as is typically the case when it is the posterior density in some other Bayesian analysis. Otherwise, the estimate (24) depends on integrals which are no easier to evaluate than the original problem (20).

3.3. *Multidimensional Designs*

Inversion of A was mentioned as a practical problem in Section 2.6 and is even more important when $f(\cdot)$ is cheap. To integrate a high-dimensional function inevitably requires the number of function evaluations, n, to be large. Although the evaluations may themselves be cheap, inversion of A requires order-n^3 computation and so completely dominates the exercise. In contrast the Monte Carlo method, described in the context of Bayesian applications by van Dijk and Kloek (1980, 1984) and Geweke (1989), processes n function evaluations in order-n computations. (An important new variant on Monte Carlo methods is given by West (1991).) Although Bayesian quadrature makes vastly more efficient use of the information in each function evaluation, Monte Carlo can make so many more function evaluations that it is generally a superior technique in high dimensions. It should also be mentioned here that recent applications of the 'Gibbs sampling' technique are proving even more effective than Monte Carlo for such cases; see Gelfand and Smith (1990), Gelfand, Hills, Racine-Poon and Smith (1990).

Nevertheless, it is of interest to develop Bayesian quadrature of multiple integrals, for two reasons. First, for sufficiently expensive functions it is essential to make maximum use of function evaluations, for which purpose Bayesian quadrature is the best available technique. Second, by suitable choice of design points $x_1, \ldots, x_n$ it is possible either to make inversion of A analytically feasible or to reduce it to a much smaller problem. An example of the latter is the development of cartesian product designs in O'Hagan (1991), which exploits a Kronecker product form for A. We present here an even simpler design where A can be inverted analytically.

Suppose that the points $x_1, \ldots, x_{p+1}$ lie on a regular simplex in R^p. That is, the distance is the same between all pairs of points. Then with covariance function (17), A takes the intraclass form, where all diagonal elements are equal and all off-diagonal elements are equal. This matrix is trivial to invert and the following results are then easy to derive. We assume the Bayes-Hermite form, combining (17) with letting $G(\cdot)$ be the $N(0, I_p)$ measure. We also let $h(x) = 1$, so that the regression part of (1) consists just of an unknown mean β. Then let $(x_1, \ldots, x_{p+1})$ be a simplex centred on the origin, with all points at a distance d from the origin, so that the distance between pairs of points is $d\{2(p+1)/p\}^{\frac{1}{2}}$. Then the posterior variance of the simple integral $\int f(\theta)\,\mathrm{d}G(\theta)$ is minimised by setting

$$d^2 = p(2+p)(1+2b)\ln(1+2b)\{2b(p+2+4b+4pb)\}^{-1}. \tag{27}$$

Setting $b = 1$ and evaluating (27) for various p we find the optimal tetrahedral designs shown in Table 1.

p	1	2	3	5	10	100
d	0.6704	0.9077	1.0849	1.3640	1.8792	5.7636

Table 1. *Distance d from origin of points in optimal tetrahedral design in p dimensions.*

As $p \to \infty$, $d^2/p \to (1+2b)\ln(1+2b)\{2b(1+4b)\}^{-1}$. Therefore in the case $b = 1$ shown in Table 1, d is asymptotically $0.5741\,p^{\frac{1}{2}}$.

Such designs are admittedly of little practical importance. They may provide a very cursory and quick exploration of a multidimensional function, but $p+1$ points in p dimensions is far too few for adequate integration. Conversely, the product designs require of the order of m^p observations, which is impractical when p is large. Good accuracy should be achievable with far fewer observations. Much work remains to be done on Bayesian quadrature in many dimensions, but there are some interesting points already to consider. For instance, the optimal tetrahedral designs expand in distance from the origin proportionally to $\sqrt{p}$ as p increases, and this is also observed in optimal product designs; see O'Hagan (1991).

Some idea of the likely patterns of good designs may be found from identifying small optimal designs. Optimal designs of $n = 3$, 4 and 5 points in $p = 2$ dimensions have been found for the model assumed in Table 1. The three point tetrahedral design in Table 1 proves to be the optimal three point design. The optimal four point design might be expected to be a square, but it is not. Nor is it a rectangle (a 2 × 2 product design), but a skew rhombus. The optimal five point design is a kind of flattened pentagon. These designs are shown in Figure 2. Note that the model is spherically symmetrical, so rotations of these designs around the origin are equally good.

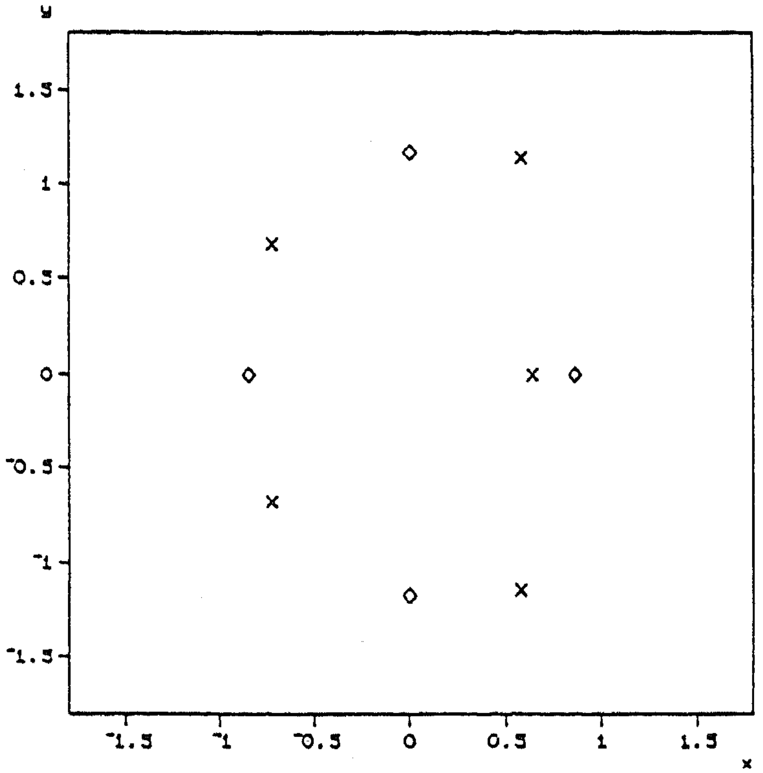


Figure 2. *Optimal* 4(= ◇) *and* 5(= ×) *point designs in two dimensions.*

These findings suggest that optimal designs will not in general have any natural geometric form.

3.4. *The Value of Observing Derivatives*

The possibility of using observations of derivatives of the function in a quadrature problem does not seem to have been explored before. Yet the development in Section 2.5 above makes it clear that to do so adds no extra complexity to the analysis. I can report here only the results of some simple computations in one dimension, taking the Bayes-Hermite context again.

The optimal n point design, at least for small n, comprises only direct observations of the function, even when each observation is allowed to be either of $f(x_i)$ or its first derivative $f'(x_i)$. If this optimal n point design is used, observing $f'(x_i)$ *in addition* to $f(x_i)$ at all those points adds *no* information about the integral. (I would like to have a theoretical proof of this rather interesting empirical result.) All of this seems to suggest that nothing is to be gained from using derivatives (at least the first derivatives), but their use can be valuable in a sequential context. Suppose that the optimal two point design has been used, and we now decide to take an extra observation to improve the posterior variance of the integral. Then the best third observation is of a first derivative, rather than of the function itself. The same is true of adding one or two further points to the optimal three point design.

Further investigation is needed of the use of derivatives in quadrature.

4. OPTIMISATION

4.1. *Optimising the Interpolant*

Optimisation is the problem of finding a minimum (or maximum) of $f(\cdot)$. Numerical analysts have constructed powerful algorithms for optimising functions in many dimensions, and the important feature of such algorithms is their sequential nature. Based on some or all of the function evaluations made up to the current iteration, the algorithm determines one or more new points to be evaluated in the next iteration, with the objective of homing in on the minimum.

It is of particular interest to consider how the Bayesian approach may improve efficiency for expensive functions. Suppose that n function evaluations have been made, and the posterior mean of the function is now the interpolant (13). With no further function evaluations of the expensive $f(\cdot)$, we can now optimise instead the cheap function $m^*(\cdot)$. (Even for large n, we need only invert A once for any number of evaluations of $m^*(\cdot)$, so this is genuinely a cheap function.) Various search strategies could now be devised to determine the next evaluation of $f(\cdot)$ required at each iteration, but the simplest is just to set x_{n+1} to argmin $m^*(\cdot)$. Here is a simple example of the effect of this strategy.

The function was set at $f(x) = -x - 2 \ln x$. The covariance function (17) was used with $b = 0.1$. (Similar results are achieved for a range of b values.) $f(x)$ was initially evaluated at $x_1 = 0.1$, $x_2 = 2.5$, $x_3 = 4.9$. The algorithm subsequently produced $x_4 = 3.173$, $x_5 = 2.563$, $x_6 = 2.244$, $x_7 = 2.069$, $x_8 = 2.00744$, $x_9 = 2.000103$.

Other similar examples of this technique, albeit only in one dimension, have generally produced rapid convergence once a good spread of points has been established. Indeed, convergence is often as fast as the Newton–Raphson algorithm, which requires both first and second derivatives to be available (and so makes two evaluations at each step). In the case of expensive functions, derivatives will very often be unavailable, and standard optimisation methods using only function evaluations are very much slower. This Bayesian technique therefore promises to be highly effective in such cases.

4.2. *Response Surfaces*

Section 2.6 introduced the idea of a regression function as an expensive function. Observations can only be obtained by experimentation, which is typically costly compared with computation time, and are then subject also to experimental error. An important problem in industrial statistics is to optimise a regression function. The techniques of response surface methodology are also sequential, therefore. See Box and Draper (1987). Optimising a fitted non-parametric regression using the smoothing techniques of Section 2.6 could undoubtedly contribute to that methodology.

ACKNOWLEDGEMENTS

I would like to thank John Kent for assistance with the theory of derivatives of Gaussian processes, and Arieh Epstein for discussions on optimisation.

REFERENCES

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* New York: Wiley.

Diaconis, P. (1988). Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV* **1** (S. S. Gupta and J. Berger, eds.), New York: Springer, 163–175.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1340.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. B* **40**, 1–42, (with discussion).

O'Hagan, A. (1989). Integrating posterior densities by Bayesian quadrature. *Tech. Rep.* **159**, University of Warwick, Coventry, UK.

O'Hagan, A. (1991). Bayes-Hermite quadrature. *J. Statist. Planning and Inference* , (to appear).

Sacks, J. and Schiller, S. (1988). Spatial designs. *Statistical Decision Theory and Related Topics IV* **2**. (S. S. Gupta and J. Berger, eds.), New York: Springer, 385–399.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci* **4**, 409–435.

van Dijk, H. K. and Kloek, T. (1980). Further experience in Bayesian analysis using Monte Carlo integration. *J. Econometrics* **14**, 307–328.

van Dijk, H. K. and Kloek, T. (1984). Experiments with some alternatives for simple importance sampling in Monte Carlo integration. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 511–530, (with discussion).

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. B* **40**, 364–372.

Wahba, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *J. Roy. Statist. Soc. B* **45**, 133–150.

West, M. (1991). Bayesian computations: Monte Carlo density estimation. *J. Roy. Statist. Soc. B* , (to appear).

DISCUSSION

L. M. BERLINER (*Ohio State University, USA*)

Professor O'Hagan presents us a paper with two goals. The first goal is to review efforts in Bayesian numerical analysis based on Gaussian random function prior specifications. The second goal is to present some interesting applications of such specifications of current interest. Professor O'Hagan succeeds on both accounts; We should be grateful for this well-written and informative paper.

The first point for discussion I wish to raise involves the role of subjective Bayesian thinking in the use of Gaussian random field priors. In the second paragraph of the paper, O'Hagan writes "... this paper will concentrate on techniques arising from representing prior information about $f(\cdot)$ in terms of a Gaussian process". My question is "What prior information is actually represented?" Specifically, consider (1). One can readily ascribe meaning to the specification of the mean function of the process. However, I believe the specification of the covariance function $v(\cdot,\cdot)$ is a difficult problem. While we may find some comfort in being told that v controls the degree of smoothness of realizations of the process, I do not believe we have a clear picture of what typical realizations of a Gaussian process actually look like for various v. The motivation of my concern is my perception of the view users of these methods seem to take concerning results. Suppose we observe f at the design points without error. The posterior mean can be drawn as an interpolant of the data and, because of the advertised advantage of the Bayesian approach to function estimation, we can also associate posterior variances with our function estimates. The success of the approach may be well and good, but as a skeptical discussant, I suggest that the posterior mean may typically be smoother than typical realizations of the posterior or prior process. (This phenomenon may be amplified in the realistic, hierarchical model suggested by O'Hagan in that the posterior process involves t-distribution tails.) Has the posterior mean computation actually been averaged over realizations, which may or not seem reasonable, to produce a "nice" function? If "almost all" realizations of the process are unbelievable, what solace can we really take in the ability to find posterior variances?

To potentially help to alleviate the concerns raised above, consider the following:

(i) I think intuition can be obtained by thorough examination of the covariance function v. In one dimension, consider the common covariance function (see (17))

$$v(x,x') = \exp\{-b(x-x')^2\}.$$

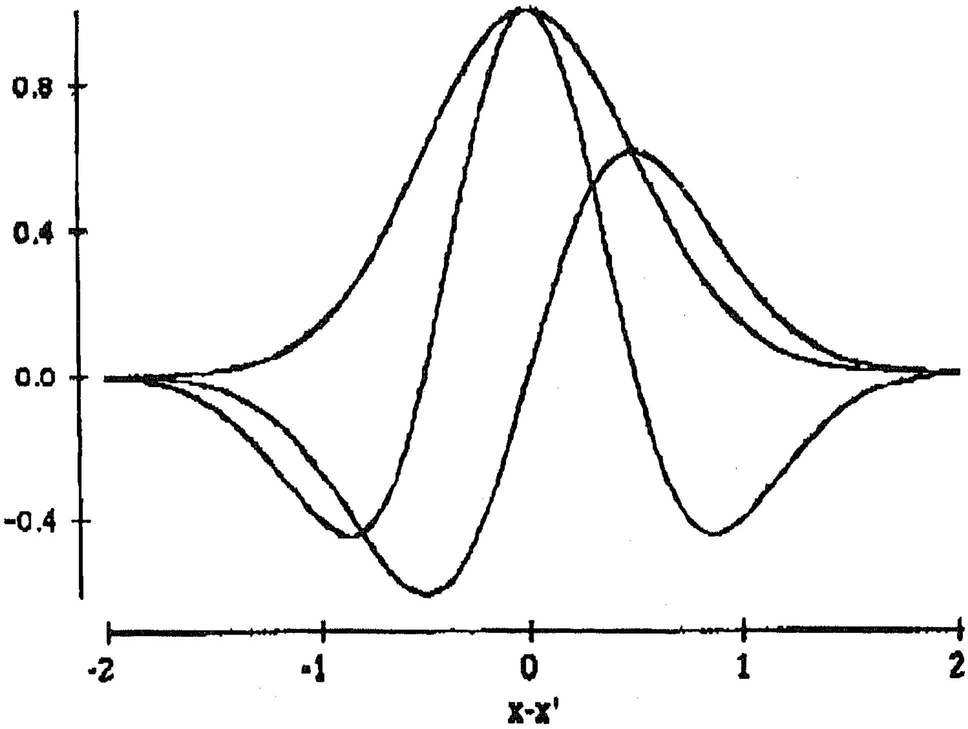
This function is very well understood by statisticians, and, hence, we may have a reasonable "feeling" for the implied covariance on f. Perhaps we can learn a little more by investigating the implied (see (18) and (19)) functions (let f^1 denote the first derivate of f)

$$v^{0,1} = \mathrm{Cov}\,(f(x), f^1(x')) = 2b(x-x')\exp\{-b(x-x')^2\}$$

and

$$v^{1,1} = \mathrm{Cov}\,(f^1(x), f^1(x')) = 2b[1-2b(x-x')^2]\exp\{-b(x-x')^2\}$$

For $b=2$, the three corresponding correlation functions are graphed as functions of $(x-x')$ below. Of course, both v and $v^{1,1}$ are even functions about zero, while $v^{0,1}$ is an odd function. Note that $f(x)$ and $f^1(x)$ are uncorrelated. The absolute value of the correlation between $f(x)$ and $f^1(x')$ grows as $|x-x'|$ does to a maximum of $e^{-1/2}$ at $|x-x'| = (2b)^{-1/2}$.



(ii) I think analysis beyond the reporting of the posterior means and variances should become a standard part of applied work in this area. For example, when possible, analysts might consider providing several "approximate" realizations from the appropriate posterior. If the realizations look ridiculous, the analysis may be questionable. Such realizations could be obtained as simulated random functions from the posterior. One possible method involves "brute force"; namely, we might approximate the process as an appropriate multivariate normal distribution. More sophisticated simulation techniques deserve investigation.

(iii) Depending on the application, the following suggestion may be useful. It may be desirable to try to model the real prior information available through the mean, or parametric hierarchical models for the mean, of the process. After exhausting prior information, the rest of the uncertainty is modelled as a Gaussian random process.

In the balance of this discussion, I will focus on the applications O'Hagan discusses. The main ongoing work described involves Bayesian interpolation, integration, and optimization based on Gaussian random fields. The primary contention of the author is that the analysis of such problems via Gaussian random process methods is appropriate when the function f of interest is extremely expensive to compute. The cost may involve time and/or money associated with experiments either in the "field" or on the computer. For such problems the notion of attempting to use data in an efficient manner as emphasized by O'Hagan is quite sensible.

In Section 3, O'Hagan presents discussion of Bayesian quadrature in which the integrand is modeled as a realization of a Gaussian process. The analysis seems reasonable, though some care should be given to rigor. For example, the estimate given in (22) need not exist if, as suggested as an example by the author, M is Lebesgue measure on R^p. Also, I doubt that the methods described by the author in the context of numerical Bayesian integration are, typically, serious competitors to sampling/simulation based methods. Forgive the repetition, but, I must ask when a Gaussian random process model is actually sensible for the likelihood-prior product? Also, sampling based methods such as Gibbs' Sampling enjoy the advantage of producing a sample from the posterior distribution, estimates of various marginal posteriors, as well as estimates of various posterior expectations. These comments apply to day-to-day Bayesian work. However, in defense of the current paper, it is difficult for me to imagine performing a Gibbs' Sampling analysis for a problem in which a *single* (efficiently programmed) computation of the function of interest requires several hours of supercomputer time.

O'Hagan presents optimization problems as another potential application of Gaussian random function analyses. While I agree that optimization problems may be a rich arena for the application of these methods, I think O'Hagan's discussion in this context falls short of what it should have been. First, the example analyzed is far too simple to be taken seriously as a test case. Though not exploited in the current paper, the extreme complexity functions often of interest in global optimization may lend credence to Gaussian random processes as priors. Also, there are various Bayesian and pseudo-Bayesian methods not mentioned by the author. Substantial, and directly pertinent, work based on Gaussian processes has been done in a series of contributions by J. Mockus (1989). Other work on optimization with a Bayesian flavor include Laarhoven (1988) and Laud, Berliner, and Goel (1989).

I was somewhat disappointed in the discussion of applications. I wish O'Hagan had told us more about the role of real Bayesian thinking for Gaussian processes in areas in which such methods are almost becoming routine. (See Wahba, 1989, for references.) For example, the computation of the posterior mean of a Gaussian random field is known as Kriging in the geostatistics literature; in some oceanography circles, a special case is known as "objective analysis". (What a great name!) Bayesian thinking has much to offer in these applications. For more evidence, see the exchange between Cressie (1990) and Wahba (1990).

In conclusion, I have emphasized that, in principle, in each application of the methods discussed, one should ask to what degree the Gaussian random field prior actually reflects prior beliefs. While I believe this question is fundamental, I am not a fool; I also believe in using procedures which seem to work even if a firm footing for the procedures is not available. Statisticians and other mathematical modelers must surely take such a view in order to actually do, rather than just talk about, something in practice. However, this operational view produces answers which must always be treated with a degree of suspicion commensurate with the level of belief in the model.

B. BETRÒ (*CNR-IAMI, Italy*)

In (continuous) optimization, we must distinguish between local problems, in which the function to be optimized (objective function) is assumed to be unimodal, and global problems, in which the global optimum is sought in absence of guaranteed unimodality. In the first case, a quadratic model of the function is adequate in a neighborhood of the extremum, and efficient methods exploit this local structure. It is hard to believe that the introduction of a stochastic model would improve the situation. Considering the example given in the paper, minimization of the function $x - 2\log x$ in the interval $[0, 5]$ by subroutine E04ABF in the NAG library gives, after 9 function evaluations, 1.999999995 as an estimate of the minimum. Notice that E04ABF does not require evaluation of derivatives.

The fact with stochastic processes, in particular with Gaussian processes, is that their sample paths are far from being unimodal. Therefore they are better suited for global, rather than local, optimization problems. Unfortunately, as it is clear from (15), the optimization of the interpolant still requires the solution of a global optimization problem which, although dealing with a function cheaper than the original one, is not necessarily a simpler one. In one dimension the interpolant is possibly unimodal between the evaluation points, so that global optimization is affordable, but in the multidimensional case the situation obviously becomes dramatically worse.

There is a certain amount of literature on the use of Bayesian methods for global optimization. Other approaches, different from those modelling the objective function by a stochastic process, have been proposed. For a recent survey see Betrò (1991).

D. E. MYERS (*University of Arizona, USA*)

In examining a new derivation it is important to ask whether a new interpolator is obtained, or, if not, whether it provides new properties or whether it is more general than other developments. As given in the paper the interpolator is

$$m^*(x) = \Sigma k_i v(x, x_i) + \Sigma b_j h_j(x). \qquad (eq.13)$$

Using the uninformative prior on the hyperparameters wherein $B_0 = 0$ and using the notation in the paper, the coefficients in m^* are obtained as the solution of

$$\begin{bmatrix} A & H \\ H^T & 0 \end{bmatrix} \begin{bmatrix} k \\ b \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \qquad (M1)$$

k is the vector of k_i's. Under weaker conditions than used in the paper the interpolator can be rewritten as

$$m^*(x) = \Sigma \lambda_i(x) f(x_i) \qquad (M2)$$

where

$$\begin{bmatrix} A & H \\ H^T & 0 \end{bmatrix} \begin{bmatrix} \lambda_x \\ \mu_x \end{bmatrix} = \begin{bmatrix} t(x) \\ h(x) \end{bmatrix}. \qquad (M3)$$

The system (M3) can also be directly obtained by requiring that (M2) be unbiased and have minimum error variance, which is then given by $w^*(x, x)$. Eq. 13 is known in the numerical analysis literature as the radial basis function interpolator, Micchelli (1986). Both the thin plate and the smoothing spline are special cases. The form given by (M2) is known as the kriging estimator in the geostatistical literature. The equivalence of eq. 13 and M2 is well-known, Myers (1988). However either form is more general than indicated by the specific Bayesian analysis given in the paper. For example, conditional positive definiteness (with respect to the $h_{j,s}$) of the kernel function v is sufficient to ensure the invertability of the coefficient matrix in either (M1) or (M3) and hence to uniquely determine an interpolator. The assumption of normality precludes the use of conditional positive definite kernels. The prior should instead focus on specifying: i.) the type of generalized covariance, ii.) parameters in the generalized covariance, iii.) the functions $h_j, j = 0, \ldots, p$. Normality is neither needed nor particularly desirable, the prior on σ^2 is unimportant since the value of σ^2 has no effect on the interpolated values, only on the error variance. Note that the error variance does not depend on the data values nor do the coefficients in (M2). One of the advantanges of the use of (M2) is that it leads easily to the estimation of any linear functional of f, one simply applies that linear functional to the entries on the right hand side of (M3). This formulation also shows why one would minimize w^* in order to design a sampling plan. Sample plan design has received extensive consideration in the geostatistical literature, see Barnes (1989). Finally either form of m^* easily generalizes to the vector valued case as noted in Myers (1991).

J. PILZ (*Bergakademie Freiberg, Germany*)

I would like to comment on some issues raised in Section 2 on interpolation and smoothing. First, let me draw attention to geostatistical prediction techniques that have become known under the heading "kriging". From a statistical viewpoint, the kriging predictor is merely the best linear unbiased predictor of a random function (field) defined on some spatial domain. The natural interpolant (13) exactly coincides with the Bayesian kriging predictor introduced by Kitanidis (1986) and Omre (1987). The well-known (universal) kriging predictor comes out as a limiting case when $B_0 = 0$, i.e., when specifying a noninformative prior for (β, σ^2).

Pilz (1991b) developed a robust Bayesian version of (13) which only requires approximate knowledge of the first and second order moments of β: Assume

$$b_0 \in \mathcal{B} \quad \text{and} \quad B_0 \geq \bar{B}$$

where $\mathcal{B}$ is some subset of R^q supposed to be symmetric around some center point $b_1 \in R^q$ and $\bar{B}$ is some given positive definite matrix such that $B_0 - \bar{B}$ is positive semidefinite, i.e., the prior covariance matrix of $\beta|\sigma^2$ is bounded from above by $\sigma^2 \bar{B}^{-1}$. Then a robust Bayesian interpolant minimizing the maximum possible Bayes risk (posterior variance) is given by (13) with b replaced by

$$b^* = B^* \left((\bar{B}^{-1} + B_1)^{-1} b_1 + H^T A^{-1} f \right), B^* = \left((\bar{B}^{-1} + B_1)^{-1} + H^T A^{-1} H \right)^{-1}$$

where B_1 maximizes the trace functional

$$G(B_p) = \operatorname{tr} \left(H^T A^{-1} H + (\bar{B}^{-1} + B_p)^{-1} \right)^{-1}$$

defined on the set of all centered moment matrices $B_p = \int_{\mathcal{B}} (t - b_1) \times (t - b_1)^T P(dt)$ generated by the class of probability measures P over $\mathcal{B}$. Obviously, b^* is the three-stage hierarchical Bayes linear estimator of β, where, at the third stage, the hyperparameter $E\beta$ has mean b_1

and (least favourable) covariance matrix B_1. For further details and explicit form solutions for B_1 see Pilz (1991a), Sections 6.3, 15.5 and 17, where the case of an inadequate model for the mean function $h(\cdot)^T\beta$ is also treated.

In the geostatistics literature, the unknown covariance structure of the random function is inferred from the variogram $E(f(x) - f(x'))^2$. Under the assumption of covariance stationarity, the covariance function and the variogram are equivalent tools. For interpolation or smoothing it is very essential to capture the behaviour of the variogram near the origin. In this respect, a warning about the use of the Gaussian covariance function (17) is in order, even if one deals with smooth phenomena. Stein (1989) shows that slight misspecifications of this covariance function can have dramatic effects on prediction, he also proposes a smoooth alternative.

A full Bayesian approach to interpolation and smoothing would also require some prior modeling for the parameters of the covariance function, e.g., for the decay rate b in (16) or (17). Unfortunately, covariance parameters are not easy to handle in a Bayesian framework. For the case, however, that the covariance function admits a parametric linear model

$$v(x,x') = \theta_1 k_1(x-x') + \cdots + \theta_m k_m(x-x'), m \leq 3$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is an unknown vector of variance components and $k_1, \ldots, k_m$ are given correlation functions, Bayes invariant quadratic estimators for v (invariant w.r.t. translations $f + H\beta$ for all $\beta \in R^q$) have been derived by Stuchlý (1989), see also Pilz (1990).

A. H. SEHEULT (*University of Durham, UK*) and
J. A. SMITH (*University of Durham, UK*)

Professor O'Hagan discusses several problems in numerical analysis from a Bayesian perspective, concentrating on estimation and design issues in interpolation, integration and optimisation. A related problem is that of locating a zero, one which we considered in our poster presentation at the Meeting. Our model explicitly includes the location of the zero as random quantity, and the "slope" of the function is given a stationary covariance structure, leading to a nonstationary function process similar (but different) to that in O'Hagan (1978). A one-step-ahead sequential procedure selects the next design point to minimise the mean squared error from zero of the linear Bayes predictor of the function. The variance contribution to the mean squared error inhibits the next design point from being the zero of the predictor (the bias contribution) by not letting it wander too far from previous design points, and is in contrast to the optimisation method suggested by Professor O'Hagan where succesive points are chosen to optimise the predictor. However, in a series of papers, Schagen (1979, 1980a, 1980b, 1984) develops a sequential optimisation criterion which explores the function before gradually homing in on the optimum. This is achieved by incorporating (after each function evaluation) an estimate of the posterior probability of missing a "bump" in part of the function. Schagen also includes a method for estimating the "smoothing parameter" in the covariance function.

The distinction between expensive and cheap functions is important. Effective design will depend crucially on careful elicitation of prior information, especially for expensive functions. Moreover, fixed designs are questionable in this context; but if they are used experimenters should be free to modify them after contrasting observation with prediction. Also questionable in this context is slavish adherence to the widely-held, narrow view of Bayesian inference expounded by Lindley (1992) that "... the conditional distribution ... is a complete description of your uncertainty ... after the data have been seen". Expensive functions are often computer codes for complicated models of complex phenomena, and after each run of the code—each function evaluation—unanticipated relevant information often becomes available, and in such cases it is perhaps better to regard the conditional distribution as one input to posterior beliefs.

Finally, we believe that considerable care should be taken when attempting to compare the relative merits of Bayesian and classical methods of numerical analysis, especially when the former are developed in a context-free manner. The strength of Bayesian thinking is surely to regard and treat each application as unique: the danger of fixed methodology is that it encourages release from the burden of thinking!

REPLY TO THE DISCUSSION

The discussants have raised a number of issues, giving me helpful suggestions and a number of references, for which I am very grateful. As I tried to make clear, many if not all of the techniques I presented have been developed before, independently and often in quite different guises, by workers in several disparate fields. To those references I should add the work of Upsdell (1985), who derives a substantial amount of related theory.

Berliner, Myers and Pilz all draw attention to the work in geostatistics on methods known there as 'kriging'. Kriging developed as a classical 'best linear estimator' technique, as explained by Pilz, and as Myers shows, the interpolant (13) can be seen just as a solution of a set of linear equations. He thereby links it to work of numerical analysts on splines and radial basis functions. The Bayesian nature of these theories has also been demonstrated within those fields.

Myers points out that normality is not necessary to derive the interpolant. The robust formulation given by Pilz emphasises this fact, and (13) can be derived even more simply as a Bayes linear estimator as advocated by Goldstein (1988). The work of Seheult and Smith is also based on Bayes linear estimation. Myers further points out that the interpolant is independent of the variance parameter σ^2, and so the prior distribution I assume for σ^2 is irrelevant. Of course, posterior variances *will* depend on σ^2 and its prior distribution, and more work is needed on how to make use of the full posterior distribution rather than just its mean.

This brings in comments by Berliner and Betrò on the nature of sample paths. Sample paths of the posterior distribution will indeed be less well behaved than the posterior mean, and will indicate when it is reasonable to assume a Gaussian process prior distribution. Seheult and Smith stress the need to think carefully about prior information. They are right, of course, but to specify a prior distribution for a whole function in a realistic amount of time and effort necessitates compromise. Any prior distribution can only be an approximation to true prior beliefs, and sensitivity of posterior inference to that approximation should be looked for. It is clear that some posterior inferences will be sensitive to the Gaussian process prior assumption. To compute the posterior distribution of the number of local maxima (or modes), for instance, would not be sensible, since the Gaussian process is likely to produce 'bumpy' realisations.

Berliner makes a number of suggestions about assessing the realism of my prior modelling. I am not sure of the usefulness of his plots of covariance functions of derivatives. Their general form seems reasonable to me, and I cannot imagine having strong opinions about features like the separation $x - x^*$ at which the correlation between $f(x)$ and its first derivative $f'(x^*)$ is maximised. But I can see that some qualitative conclusions are important, such as that whenever $f'(x)$ exists it will be independent of $f(x)$.

Berliner makes useful comments on my quadrature ideas. In particular, he again reminds us of the need for care in modelling. The realism of the Gaussian process assumption depends on the underlying measure with respect to which I define the integral. For instance, the Bayes-Hermite formulation means that beliefs about the ratio of the density function being integrated to the approximating normal density, are represented as a stationary Gaussian process. If the tails of the density are heavier than the normal, this would not be at all realistic. More work is needed on prior formulations appropriate to a variety of contexts. It is worth saying that this is a strength of the Bayesian approach. Traditional numerical analysis takes account of context only in an anecdotal and unsystematic way. Careful attention to representing prior beliefs that truly reflect the real context, will guarantee posterior inferences that are appropriate to that context.

Betrò and Berliner comment on optimisation. Both rightly criticise my too-simple example, and both give references to several other Bayesian approaches to optimisation which I will need to study. The work of Seheult and Smith on finding a zero of a function is closely related. Their search strategy which does not initially home in too quickly, but spreads early points in a way that gives good information about the form of the function, sounds very interesting.

It seems that no discussion at Peñíscola was complete without a mention of Gibbs sampling, and Berliner obliges us here. It nicely reinforced my point about expensive and cheap functions. Sampling methods do seem to be the best we have for integrating high-dimensional *cheap* functions. Their inefficient use of function evaluations then matters less than their efficient coverage of the parameter' space. We do not yet know how to deploy points deterministically in high dimensions in efficient, sparse ways and such that inference from them is practical for cheap functions. But Gibbs sampling is inappropriate for expensive functions, and I believe that research on methods for expensive functions will in the long run allow us to dispense with randomisation completely.

ADDITIONAL REFERENCES IN THE DISCUSSION

Barnes, R. (1989). A partial history of spatial sampling design. *The Geostatistics Newsletter* **3**, 10–12.

Betrò, B. (1991). Bayesian methods in global optimization. *Journal of Global Optimization* **1**, 1–14.

Cressie, N. (1990). Letters to the Editor: Reply to comment on Cressie. *Amer. Statist.* **44**, 256–258.

Goldstein, M. (1988). Adjusting belief structures. *J. Roy. Statist. Soc. B* **50**, 133–154.

Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research* **22**, 499–507.

Laarhoven, P. J. M. van (1988). *Theoretical and Computational Aspects of Simulated Annealing*. Amsterdam: Centrum voor Wiskunde en Informatica.

Laud, P., Berliner, L. M. and Goel, P. K. (1989). A stochastic probing algorithm for global optimization. *Computing Science and Statistics. Proccedings of the 21st Symposium on the Interface* (K. Berk *et al.* , eds.). Alexandria, Virginia: ASA.

Lindley, D. V. (1992). Is our view of Bayesian Statistics too narrow?. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 1–15, (with discussion).

Micchelli, C. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation* **2**, 11–22.

Mockus, J. (1989). *Bayesian Approach to Global Optimization.* Dordrecht: Kluwer Academic.

Myers, D. E. (1988). Interpolation with positive definite functions. *Sciences de la Terre* **28**, 252–265.

Myers, D. E. (1991). An alternative Bayesian formulation for spatial interpolation. (Unpublished *Tech. Rep.*).

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. B* **40**, 1–42, (with discussion).

Omre, H. (1987). Bayesian kriging – merging observations and qualified guesses in kriging. *Math. Geology* **19**, 25–39.

Pilz, J. (1990). Bayes estimation of variograms and Bayesian collocation. *TUB Dokumentation Heft* **51**, Vol. II. 565–576.

Pilz, J. (1991a). *Bayesian Estimation and Experimental Design in Linear Regression Models.* New York: Wiley.

Pilz, J. (1991b). Robust Bayes linear prediction of regionalized variables. (Unpublished *Tech. Rep.*).

Schagen, I. P. (1979). Interpolation in two dimensions — a new technique. *J. Inst. Maths. Applics* **23**, 53–59.

Schagen, I. P. (1980a). Stochastic interpolating functions — applications in optimization. *J. Inst. Maths. Applics* **26**, 93–101.

Schagen, I. P. (1980b). The use of stochastic processes in interpolation and approximation. *Int. J. Comput. Math. B* **8**, 63–76.

Schagen, I. P. (1984). Sequential exploration of unknown multi-dimensional functions as an aid to optimization. *IMA J. Num. Anal.* **4**, 337–346.

Smith, J. A. and Seheult, A. H. (1991). Linear Bayes methods for locating zeros of deterministic functions. (Unpublished *Tech. Rep.*).

Stein, M. L. (1989). The loss of efficiency in kriging prediction caused by misspecifications of the covariance structure. *Geostatistics* **1** (M. Armstrong, ed.). Dordrecht: Kluwer Academic, 273–282.

Stuchlý, J. (1989). Bayes unbiased estimation in a model with three variance components. *Aplikace Mat.* **34**, 375–386.

Upsdell, M. P. (1985). *Bayesian Inference for Functions*. Ph.D. Thesis, University of Nottingham.

Wahba, G. (1990). Letters to the Editor: Comment on Cressie. *Amer. Statist.* **44**, 255–256.