Journal of Statistical Planning and Inference 29 (1991) 245–260
North-Holland

Bayes–Hermite quadrature

A. O'Hagan

Department of Mathematics, University of Nottingham, Nottingham NG7 2RD, UK

Received 18 June 1990; revised manuscript received 18 October 1990
Recommended by A.O. Dempster

Abstract: Bayesian quadrature treats the problem of numerical integration as one of statistical inference. A prior Gaussian process distribution is assumed for the integrand, observations arise from evaluating the integrand at selected points, and a posterior distribution is derived for the integrand and the integral. Methods are developed for quadrature in $\mathbb{R}^p$. A particular application is integrating the posterior density arising from some other Bayesian analysis.

Simulation results are presented, to show that the resulting Bayes–Hermite quadrature rules may perform better than the conventional Gauss–Hermite rules for this application. A key result is derived for product designs, which makes Bayesian quadrature practically useful for integrating in several dimensions. Although the method does not at present provide a solution to the more difficult problem of quadrature in high dimensions, it does seem to offer real improvements over existing methods in relatively low dimensions.

AMS Subject Classification: Primary 62C10; secondary 65D30.

Key words: Bayesian quadrature; numerical integration; Gaussian process; product rule; Gaussian quadrature.

1. Introduction

The problem of quadrature, or numerical integration, is simple and well known. We are interested in the value of an integral, such as

$$k = \int_A f(x)\,\mathrm{d}x,$$

but cannot obtain it analytically. A quadrature rule approximates k as a function of a number of specific values of $f(\cdot)$, say $f(x_1), f(x_2), \ldots, f(x_n)$.

The essence of the present paper is that quadrature is a statistical problem. There is an unknown quantity, k, about which we wish to make inference, using data $f(x_1), \ldots, f(x_n)$. A statistical formulation of the quadrature problem requires a model linking k to the $f(x_i)$s. In our approach $f(\cdot)$ is a random function. In the frequentist theory of inference, we might suppose that $f(\cdot)$ is drawn at random

from some population of functions. To assert a distribution for $f(\cdot)$ is then to describe the characteristics of the population. In a Bayesian approach, $f(\cdot)$ is random simply because it is numerically unknown. It has a known algebraic expression, but we do not know the numerical value of $f(x)$ for any x untill we actually calculate it. A distribution for $f(\cdot)$ then expresses the investigator's personal beliefs about it. We will adopt the Bayesian approach because of its simplicity. Specifically, we incorporate the observations $f(x_1), \ldots, f(x_n)$ by conditioning the prior distribution of $f(\cdot)$ on these n values which are now known. This is the posterior distribution of $f(\cdot)$, from which the posterior distribution of k is derived.

Diaconis (1988) traces the origins of this Bayesian approach to quadrature as far back as Poincaré (1896), and presents a useful review. More references are to be found in Sacks, Welch, Mitchell and Wynn (1989) and Ylvisaker (1987). The diversity of work in this area reflects to some extent the many varied uses of quadrature. Prior information about $f(\cdot)$ is very dependent on context, which results in different variations of the basic approach in different applications.

Our primary concern in this paper is to develop and apply Bayesian quadrature techniques in a specific context, which is itself strongly motivated by Bayesian statistics. Bayes' theorem asserts that the posterior density $f(\cdot)$ is proportional to the product of a prior density and a likelihood function, and a primary task of the Bayesian analysis will be to find its integral, the inverse of the proportionality constant. The integral will often not be determinable analytically, and quadrature techniques are necessary. There is another substantial literature on applying various non-Bayesian quadrature techniques to this fundamentally Bayesian problem. See Shaw (1988) and many references therein. To avoid confusion with our main subject, namely the Bayesian analysis of the quadrature problem, its application in integrating the posterior density arising in some other Bayesian problem will be refered to as 'the Bayesian application'.

It is important to stress that no single quadrature technique is appropriate to every kind of problem. Even within the restricted field of the Bayesian application, where $f(\cdot)$ is proportional to a density function in $\mathbb{R}^p$, different approaches are used depending upon the number of dimensions p. The major conventional technique for multiple integrals in up to about six or seven dimensions is to use cartesian products of one-dimensional rules. Specifically, Smith, Skene, Shaw, Naylor and Dransfield (1985) advocate products of Gauss–Hermite rules. Whereas product rules are practical in low dimensions, the number of function evaluations required rapidly escalates so that they become unrealistic to use in high dimensions. In the Bayesian application, a frequently used technique for integrating high dimensional posterior densities is Monte Carlo. When the integrand is a posterior density, it can generally be evaluated quickly and cheaply, but in contrast Sacks et al. (1989) describe other applications where evaluation of the function at a single point may require hours of time on a Cray supercomputer. They do no consider integration, but nevertheless use the same models as we present in this paper. The general technique of Bayesian quadrature is designed to make the fullest possible use of every function evaluation,

and so would be ideal when such evaluations are costly. However, we do not pursue those applications here.

There is one other quadrature technique in common use that recognises the statistical nature of the quadrature problem. This is the Monte Carlo method. However, O'Hagan (1987) criticises the underlying philosophy of Monte Carlo quadrature.

In Section 2 we present the general theory of Bayesian quadrature. Section 3 derives Bayes–Hermite quadrature as the Bayesian quadrature analogue of Gauss–Hermite rules for integration over $\mathbb{R}^p$. Simulation results suggest that Bayes–Hermite quadrature can be more accurate than Gauss–Hermite for the Bayesian application in one dimension. Section 4 is concerned with higher dimensional integrals. We present a general result on product rules, which facilitates the development of Bayes–Hermite product rules. Practical implementation of Bayesian quadrature is discussed in Section 5.

2. Bayesian quadrature

2.1. Model

Rewrite the basic integral as

$$k = \int_X f(x)\,\mathrm{d}G(x), \tag{2.1}$$

where $G(\cdot)$ is a measure over X. We could regard $k^{-1}f(\cdot)$ as a density with respect to the underlying measure $G(\cdot)$. We then formulate prior beliefs about $f(\cdot)$ via

$$f(x) = \boldsymbol{h}(x)^{\mathrm{T}}\boldsymbol{\beta} + e(x), \tag{2.2}$$

where $\boldsymbol{h}(\cdot)$ is a vector of q known functions of x (i.e. mapping X to $\mathbb{R}^q$), $\boldsymbol{\beta}$ is a vector of coefficients and $e(\cdot)$ represents departure of $f(\cdot)$ from the regression term. We suppose that our prior belief about $e(\cdot)$ is that it is a stationary, zero-mean, Gaussian process,

$$e(\cdot) \sim \mathrm{N}(0, v(\cdot,\cdot)), \tag{2.3}$$

where $v(\cdot,\cdot)$ is the covariance kernel

$$v(x,x') = \mathrm{Cov}(e(x), e(x')) = \sigma^2 c(\|x - x'\|). \tag{2.4}$$

Here $c(\cdot)$ is a monotone decreasing correlation function on $\mathbb{R}^+$ with $c(0) = 1$, and $\|x - x'\|$ denotes any distance measure on X. Then σ^2 is the variance of each $e(x)$.

We will assume that $c(\cdot)$ is known, but the question of unknown $c(\cdot)$ is considered briefly in Section 4.5. The variance σ^2 is unknown. Conditional on $\boldsymbol{\beta}$ and σ^2, the prior distribution of $f(\cdot)$ is a Gaussian process

$$f(\cdot) \mid \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(m(\cdot), v(\cdot,\cdot)), \tag{2.5}$$

where $m(x) = \boldsymbol{h}(x)^{\mathrm{T}}\boldsymbol{\beta}$. The prior model is completed by a prior distribution for $\boldsymbol{\beta}$ and σ^2. It would be simple to develop the theory with a proper, conjugate prior distribution at this stage, but prior information about $\boldsymbol{\beta}$ and σ^2 will typically be weak and difficult to elicit accurately. We therefore assume weak prior information, represented by the improper prior density

$$p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \tag{2.6}$$

The basic model (2.2) with $e(\cdot)$ defined by (2.3) and (2.4) is used in a similar context by Sacks et al. (1989), but its use in regression analysis goes back at least to Blight and Ott (1975). Essentially the same model underlies the method of kriging, see Cressie (1988), which is widely used by geologists. The localised regression model of O'Hagan (1978) is similar but incorporates $\boldsymbol{h}(\cdot)$ into $v(\cdot,\cdot)$, so that the errors are no longer stationary.

For greater generality, we replace (2.1) by

$$\boldsymbol{k} = \int_X \boldsymbol{r}(x) f(x)\, \mathrm{d}G(x), \tag{2.7}$$

where $\boldsymbol{r}(\cdot)$ is a known vector of p functions of x. Thus $\boldsymbol{k}$ is a p-vector whose i-th element is $\int_X r_i(x) f(x)\, \mathrm{d}G(x)$. We will continue to use the symbol k in the sense of (2.1), or of (2.7) in the case $\boldsymbol{r}(x) = (1)$.

2.2. *Posterior distributions*

We now obtain data comprising the value of $f(\cdot)$ at n 'design points' $x_1, x_2, \ldots, x_n$, yielding the observation vector

$$\boldsymbol{f} = (f(x_1), f(x_2), \ldots, f(x_n))^{\mathrm{T}}. \tag{2.8}$$

Posterior distributions are easily obtained. First, the posterior distribution of $f(\cdot)$ given $\boldsymbol{\beta}$ and σ^2 is the Gaussian process

$$f(\cdot) \mid \boldsymbol{f}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(m'(\cdot), v'(\cdot,\cdot)), \tag{2.9}$$

where

$$m'(x) = \boldsymbol{h}(x)^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{t}(x)^{\mathrm{T}}\boldsymbol{A}^{-1}(\boldsymbol{f} - \boldsymbol{H}\boldsymbol{\beta}), \tag{2.10}$$

$$\boldsymbol{t}(x) = \begin{pmatrix} c(\|x - x_1\|) \\ \vdots \\ c(\|x - x_n\|) \end{pmatrix}, \qquad \boldsymbol{H} = \begin{pmatrix} \boldsymbol{h}(x_1)^{\mathrm{T}} \\ \vdots \\ \boldsymbol{h}(x_n)^{\mathrm{T}} \end{pmatrix},$$

$$\boldsymbol{A} = \begin{pmatrix} 1 & c(\|x_1 - x_2\|) & \cdots & c(\|x_1 - x_n\|) \\ c(\|x_2 - x_1\|) & 1 & & c(\|x_2 - x_n\|) \\ \vdots & & \ddots & \\ c(\|x_n - x_1\|) & c(\|x_n - x_2\|) & & 1 \end{pmatrix},$$

$$v'(x, x') = \sigma^2\{c(\|x - x'\|) - \boldsymbol{t}(x)^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{t}(x')\}. \tag{2.11}$$

Posterior distributions of β and σ^2 derive simply from the fact that

$$f \mid \beta, \sigma^2 \sim \mathrm{N}(H\beta, \sigma^2 A),$$

and we find

$$\beta \mid f, \sigma^2 \sim \mathrm{N}(\hat{\beta}, \sigma^2 (H^{\mathrm{T}} A^{-1} H)^{-1}), \tag{2.12}$$

$$\sigma^2 \mid f \sim d\chi^{-2}_{n-q}, \tag{2.13}$$

where

$$\hat{\beta} = (H^{\mathrm{T}} A^{-1} H)^{-1} H^{\mathrm{T}} A^{-1} f$$

has a familiar generalised least squares form, and

$$d = f^{\mathrm{T}} \{A^{-1} - A^{-1} H (H^{\mathrm{T}} A^{-1} H)^{-1} H^{\mathrm{T}} A^{-1}\} f. \tag{2.14}$$

We have assumed that H has rank q. Otherwise the posterior distribution of β would be improper.

Now the integral (2.7) is a linear functional of $f(\cdot)$, and we immediately have from (2.9) that its posterior distribution given β and σ^2 is

$$k \mid f, \beta, \sigma^2 \sim \mathrm{N}(m', \sigma^2 U'), \tag{2.15}$$

where

$$m' = \int_X r(x) m'(x) \, \mathrm{d}G(x) = R\beta + TA^{-1}(f - H\beta), \tag{2.16}$$

$$R = \int_X r(x) h(x)^{\mathrm{T}} \, \mathrm{d}G(x), \tag{2.17}$$

$$T = \int_X r(x) t(x)^{\mathrm{T}} \, \mathrm{d}G(x), \tag{2.18}$$

$$U' = U - TA^{-1} T^{\mathrm{T}},$$

$$U = \int_X \int_X c(\|x - x'\|) r(x) r(x')^{\mathrm{T}} \, \mathrm{d}G(x) \, \mathrm{d}G(x'). \tag{2.19}$$

Notice that the corresponding prior distribution is

$$k \mid \beta, \sigma^2 \sim \mathrm{N}(R\beta, \sigma^2 U)$$

and we assume that both R and U exist, i.e. the integrals (2.17) and (2.19) converge. Otherwise there would be a non-zero prior probability that k itself did not exist. Combining (2.15) with (2.12) yields

$$k \mid f, \sigma^2 \sim \mathrm{N}(\hat{k}, \sigma^2 V), \tag{2.20}$$

with

$$\hat{k} = R\hat{\beta} + TA^{-1}(f - H\hat{\beta}), \tag{2.21}$$

$$V = U - TA^{-1} T^{\mathrm{T}} + (R - TA^{-1} H)(H^{\mathrm{T}} A^{-1} H)^{-1} (R - TA^{-1} H)^{\mathrm{T}}. \tag{2.22}$$

Finally, when we combine (2.20) with (2.13), the marginal posterior distribution of $\boldsymbol{k}$ is a multivariate t, which we express as

$$\boldsymbol{k} \mid f \sim t_{n-q}(\hat{\boldsymbol{k}}, d\boldsymbol{V}). \tag{2.23}$$

Its posterior mean and variance are

$$E(\boldsymbol{k} \mid f) = \hat{\boldsymbol{k}},$$
$$\mathrm{Var}(\boldsymbol{k} \mid f) = (n-q-2)^{-1} d\boldsymbol{V}. \tag{2.24}$$

provided $n > q+2$. Notice that the estimate $\hat{\boldsymbol{k}}$ is a linear function of the observation vector f. Specifically, $\hat{\boldsymbol{k}} = \boldsymbol{W}f$, where

$$\boldsymbol{W} = \boldsymbol{T}\boldsymbol{A}^{-1} + (\boldsymbol{R} - \boldsymbol{T}\boldsymbol{A}^{-1}\boldsymbol{H})(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{A}^{-1}. \tag{2.25}$$

2.3. Saturated designs

One special case of the estimate $\hat{\boldsymbol{k}}$ arises when the number of observations equals q. Then $\boldsymbol{H}$ is square and nonsingular, $\hat{\boldsymbol{\beta}} = \boldsymbol{H}^{-1}f$, and the posterior mean of $f(\cdot)$ which is given in general by

$$E(f(x) \mid f) = \boldsymbol{h}(x)^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \boldsymbol{t}(x)^{\mathrm{T}}\boldsymbol{A}^{-1}(f - \boldsymbol{H}\hat{\boldsymbol{\beta}}) \tag{2.26}$$

reduces to $\boldsymbol{h}(x)^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, the fitted regression line. Then the estimate of $\boldsymbol{k}$ is just the integral of $\boldsymbol{r}(x)$ times the fitted line. Most conventional deterministic quadrature rules (including Gaussian rules where $n < q$) may be derived in this way; see O'Hagan (1988). However, the data in these cases can provide no information about σ^2. The posterior distribution (2.23) has zero degrees of freedom, and is therefore improper. Our interest lies instead in the case $n > q$, where the Bayesian quadrature approach yields novel integration rules.

2.4. Optimal rules

The *design points* $x_1, x_2, \ldots, x_n$ may be chosen to optimise an appropriate criterion. One obvious set of criteria are the posterior variances of individual elements of $\boldsymbol{k}$, i.e. the diagonal elements of $\mathrm{Var}(\boldsymbol{k} \mid f)$. From (2.24),

$$\mathrm{Var}(\boldsymbol{k} \mid f) = \hat{\sigma}^2 \boldsymbol{V}, \tag{2.27}$$

where $\hat{\sigma}^2 = (n-q-2)^{-1}d$ is the posterior mean of σ^2. $\boldsymbol{V}$ is a matrix given in (2.22), depending on the design but not on the data. Before the data are observed, $\hat{\sigma}^2$ is unknown, and the design criterion must depend on the prior expectation of (2.27). It is easy to show that $E(d \mid \sigma^2) = n-q$, independent of the design, so the design criterion is to minimize the appropriate diagonal element of $\boldsymbol{V}$. In the case $\boldsymbol{r}(x) = (1)$, optimality reduces to minimizing v, the single element of $\boldsymbol{V} = (v)$.

3. Bayes-Hermite quadrature

3.1. Applying Bayesian quadrature

In the application of estimate (2.21) we must first evaluate $\boldsymbol{R}$ and $\boldsymbol{T}$, which are themselves defined as integrals in (2.17) and (2.18). Unless these can be done analytically in closed form, there is the danger of Bayesian quadrature degenerating into infinite regress. In practice, $G(\cdot)$ will generally be a standard, well-known probability distribution over X, and both $\boldsymbol{r}(\cdot)$ and $\boldsymbol{h}(\cdot)$ will comprise simple functions such as polynomials. It will therefore usually be straightforward to derive $\boldsymbol{R}$. However, the existence of a closed form for $\boldsymbol{T}$ will depend critically on the correlation function $c(\cdot)$. Another practical problem lies in the inverse of the matrix $\boldsymbol{A}$ in (2.21). $\boldsymbol{A}$ is an $n \times n$ matrix, and if its inverse is not known analytically then to invert it numerically is an order n^3 operation. If the number of design points is at all large, numerical inversion is not practical.

Notice, however, that to apply Gaussian quadrature methods does not require lengthy computations only because tables of design points and weights are readily available. The same can be provided for Bayesian quadrature. The Bayesian quadrature estimate is $\hat{\boldsymbol{k}} = \boldsymbol{W}\boldsymbol{f}$, where $\boldsymbol{W}$ is given by (2.25) and depends on the design points $x_1, x_2, \ldots, x_n$ at which $f(\cdot)$ is to be 'observed'. For given $c(\cdot)$, $\boldsymbol{h}(\cdot)$ and $\boldsymbol{r}(\cdot)$, and given design points, this matrix of weights need only be computed once and tabulated for future use. This would be done particularly for optimal designs. In Section 4.2 we tabulate some optimal Bayesian quadrature rules.

The real impact of the practical problems discussed above is in the degree of computation needed to find optimal designs. Unless we can find analytical or otherwise efficient ways of calculating $\boldsymbol{R}$, $\boldsymbol{T}$ and $\boldsymbol{A}^{-1}$, it may not be possible to derive optimal rules, particularly rules with many design points.

3.2. Bayes-Hermite formulation

We now consider a particular case in which the integrals $\boldsymbol{R}$ and $\boldsymbol{T}$ are available analytically. The quadrature problems for which Gauss-Hermite rules are advocated are those in which $X = \mathbb{R}^p$ with the density $f(\cdot)\,\mathrm{d}G(\cdot)$ being at least roughly approximated by a multivariate normal density, or by such a density multiplied by a polynomial. We develop here a Bayesian quadrature solution for this context, which we call Bayes-Hermite quadrature.

So let $X = \mathbb{R}^p$ and without loss of generality let $G(\cdot)$ be the standard p-dimensional normal distribution $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$:

$$\mathrm{d}G(\boldsymbol{x}) = (2\pi)^{-p/2} \exp(-\tfrac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})\,\mathrm{d}x. \tag{3.1}$$

We will suppose that the elements of $\boldsymbol{r}(\cdot)$ and $\boldsymbol{h}(\cdot)$ are any functions for which the integral (2.17) may be performed analytically. In particular, the strict analogue of Gauss-Hermite quadrature is obtained when elements of $\boldsymbol{h}(\cdot)$ are of the form

$\Pi_i x_i^{a_i}$. Finally, we propose that the correlation function also has a Gaussian form:

$$c(\|\boldsymbol{x}-\boldsymbol{x}'\|) = \exp\{-b(\boldsymbol{x}-\boldsymbol{x}')^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}')\}. \tag{3.2}$$

$$\therefore c(\|\boldsymbol{x}-\boldsymbol{x}'\|)\,\mathrm{d}G(\boldsymbol{x}) = \exp(-ab\boldsymbol{x}'^{\mathrm{T}}\boldsymbol{x}')\,\mathrm{d}G'(\boldsymbol{x}), \tag{3.3}$$

where $a=(1+2b)^{-1}$ and $G'(\cdot)$ is the multivariate normal distribution $\mathrm{N}(2ab\boldsymbol{x}', a\boldsymbol{I})$. Now (2.18) may also be performed analytically, and a second such operation also yields a closed form for (2.19).

Sacks et al. (1989) consider a more general class of correlation functions

$$c(\|\boldsymbol{x}-\boldsymbol{x}'\|) = \exp\left\{-b\sum_i |x_i - x_i'|^{\lambda}\right\},$$

where (3.2) is the case $\lambda=2$. It is possible also to evaluate $\boldsymbol{T}$ for the case $\lambda=1$, in terms of the univariate standard normal distribution function $\Phi(\cdot)$, but intermediate values of λ are not tractable. (Sacks et al. (1989) did not require $\boldsymbol{T}$ because their objective was not integration but interpolation.) Our choise of $\lambda=2$ reflects a belief in $f(\cdot)$ having a high degree of smoothness, which is appropriate in particular to the Bayesian application. The choice accords with the discussion of λ in Sacks et al. (1989) and is reinforced by some simulations which have found that $\lambda=2$ produces more accurate integration than $\lambda=1$. Further details are given in O'Hagan (1988).

3.3. *One-dimensional rules*

Consider the case of one-dimensional quadrature, $p=1$. Let $\boldsymbol{r}(x)=(1)$ and either $\boldsymbol{h}(x)=(1)$, $q=1$, or $\boldsymbol{h}(x)=(1, x, x^2)^{\mathrm{T}}$, $q=3$. Explicit formulae for Bayes-Hermite quadrature are now easily obtained. Optimal Bayes-Hermite rules (minimizing the variance v, as in Section 2.4) were derived for $n=3,4,5$ or 6 points, for $q=1$ or 3, and for various b. It is interesting to look at the three point designs, which have the form $(-x, 0, x)$. Table 1 shows values of x for various b and q.

Notice that as b decreases, corresponding to an increasingly smooth function $f(\cdot)$, the design points spread further out. It would appear that as $b\to 0$, both for $q=1$ and $q=3$, the rules are tending to the three point Gauss-Hermite rule which sets $x=\sqrt{3}=1.732$. Unfortunately, for small b the $\boldsymbol{A}$ matrix is very ill-conditioned, and numerically reliable figures could not be obtained for b less than 0.01. We con-

Table 1
Three point rules for varying b and q

q	b		
	1	0.1	0.01
1	1.152	1.599	1.716
3	1.369	1.645	1.722

jecture that for any q ($\leq n$), the n-point Bayes-Hermite rules tends to the n-point Gauss-Hermite rule as $b \to 0$.

The Bayes-Hermite rules are more conservative than Gauss-Hermite in the sense that they place design points nearer to the origin. The Gauss-Hermite rules are optimised for the case when $f(\cdot)$ is exactly a polynomial. The Bayesian quadrature model adds a random disturbance to $f(\cdot)$, and uncertainty about $f(x)$ for values of x close to the origin contributes most to uncertainty about k.

3.4. A random normal mixture

Any attempt to compare objectively the performance of different quadrature rules is doomed to failure. Quadrature rules are applied in many different contexts, and may perform very differently in each, so that comparisons made only in specific contexts can prove to be misleading. Nevertheless, the gradual accumulation of such experience is vital if users are to make good choices of rules in practice.

Motivated by our interest in the Bayesian application, we attempted to assess the performance of Bayes-Hermite rules in integrating the kind of density functions that typically arise in practical Bayesian statistics. Densities were generated randomly from the class of mixture distributions which can be written formally as

$$\gamma = (1-\alpha)\mathrm{N}(0,1) + \alpha \mathrm{N}(\mu, 0.3). \tag{3.4}$$

The two variables α and μ were given independent uniform distributions over (0.2, 0.6) and (0, 2) respectively. The density functions generated by this scheme are typical of those met with in Bayesian applications. Many are nearly normal in shape, but others show marked skewness or bimodality. With $g(\cdot)$ defined by (3.1) with $p=1$, the function $f(\cdot)$ is

$$f(x) = \gamma(x)/g(x) = 1 - \alpha + \alpha(0.3)^{-1/2} \exp[-\tfrac{1}{6}\{10(x-\mu)^2 - 3x^2\}].$$

Of course, (3.4) generates proper probability density functions, so that the true value of the basic integral is always

$$k = \int_{-\infty}^{\infty} f(x)\, \mathrm{d}G(x) = \int_{-\infty}^{\infty} \gamma(x)\, \mathrm{d}x = 1.$$

Table 2
Root mean squared errors for 5 point rules

q	b		
	1.0	0.5	0.1
1	0.0101	0.0086	0.0217
3	0.0237	0.0124	0.0231

Gauss-Hermite rules were compared with the Bayes-Hermite rules with $q=1$ or 3 and $b=1$, 0.5, or 0.1, for $n=3, 4, 5$ or 6 points, using 500 random selections from (3.4). The root mean squared error (RMSE) of $\hat{k}$ was calculated for each rule. Every Bayes-Hermite rule gave smaller RMSE than the Gauss-Hermite rule with the same number of points. Therefore, Bayes-Hermite integration seems to be superior to Gauss-Hermite over a range of values of q and b, for the kinds of integration problem represented by this simulation. For instance, Table 2 shows the RMSEs for the six five-point Bayes-Hermite rules. The RMSE for the five-point Gauss-Hermite rule was 0.0338.

Table 2 also shows a general finding in this simulation, that the rules which performed best were the Bayes-Hermite rules with $b=0.5$ and $q=1$.

4. Higher dimensions

4.1. Product designs

Higher-dimensional problems will demand increasing numbers of design points in order to achieve satisfactory quadrature. Inverting the resulting large A matrix numerically, makes it impractical to develop optimal designs in general. However, an important simplification arises with product designs. We first derive this result for general Bayesian quadrature.

Suppose that X is a Cartesian product $Y\times Z$, and we can write $x\in X$ as (y, z) with $y\in Y$ and $z\in Z$. Suppose also that the n design points $(x_1, x_2, \ldots, x_n)$ in X form a grid (i.e. cartesian product) of $n_Y n_Z$ points $x_i=(y_j, z_k)$, comprising all combinations of n_Y points $(y_1, y_2, \ldots, y_{n_Y})$ in Y and n_Z points $(z_1, z_2, \ldots, z_{n_Z})$ in Z. Suppose finally that the correlation function satisfies

$$c(\|(y,z)-(y',z')\|) = c_Y(\|y-y'\|_Y)c_Z(\|z-z'\|_Z). \tag{4.1}$$

Then we can write the correlation matrix of the design points as the Kronecker product $\boldsymbol{A}=\boldsymbol{A}_Y\otimes\boldsymbol{A}_Z$ of the two separate correlation matrices. For instance, the (i, j)th element of $\boldsymbol{A}_Y$ is $c_Y(\|y_i-y_j\|_Y)$. (We have assumed the sequence of design points $\{x_i\}$ to take points row-wise across the grid, so that the rows and columns of $\boldsymbol{A}$ will be properly arranged for this Kronecker product.) Then

$$\boldsymbol{A}^{-1} = \boldsymbol{A}_Y^{-1}\otimes\boldsymbol{A}_Z^{-1}. \tag{4.2}$$

This result makes it possible to invert the large $n\times n$ matrix $\boldsymbol{A}$ using only the $n_Y\times n_Y$ and $n_Z\times n_Z$ inversions of $\boldsymbol{A}_Y$ and $\boldsymbol{A}_Z$. If this is coupled with the ability to obtain $\boldsymbol{R}$ and $\boldsymbol{T}$ analytically, as in Bayes-Hermite formulations, it becomes possible to explore quadrature designs in higher dimensions. Designs may be obtained which are optimal within the class of product designs.

Calculation of $\boldsymbol{R}$ and $\boldsymbol{T}$ is simplified in general if we add some further reasonable assumptions. Let G be such that we can write

$$\int_X f(x)\,\mathrm{d}G(x) = \int_Y \left\{ \int_Z f(y,z)\,\mathrm{d}G_Z(z) \right\} \mathrm{d}G_Y(y). \tag{4.3}$$

That is, G is the product of independent measures G_Y and G_Z. Next assume that the q functions $\boldsymbol{h}(\cdot)$ comprise all the $q_Y q_Z$ products of q_Y functions of y and q_Z functions of z. We can then write (arranging the functions in $\boldsymbol{h}(\cdot)$ in the appropriate sequence)

$$\boldsymbol{H} = \boldsymbol{H}_Y \otimes \boldsymbol{H}_Z, \tag{4.4}$$

with $\boldsymbol{H}_Y$ and $\boldsymbol{H}_Z$ defined in the obvious way. Furthermore, the two Kronecker product forms are conformable. For instance

$$(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{H})^{-1} = (\boldsymbol{H}_Y^{\mathrm{T}}\boldsymbol{A}_Y^{-1}\boldsymbol{H}_Y)^{-1} \otimes (\boldsymbol{H}_Z^{\mathrm{T}}\boldsymbol{A}_Z^{-1}\boldsymbol{H}_Z)^{-1}.$$

We need $\boldsymbol{H}$ to have rank q for this inverse to exist, in which case $\boldsymbol{H}_Y$ has rank q_Y and $\boldsymbol{H}_Z$ has rank q_Z, and the two inverses on the right-hand side will also exist. In particular, $q_Y \leq n_Y$ and $q_Z \leq n_Z$.

Finally, suppose that the elements of $\boldsymbol{r}(\cdot)$ also comprise all the products of a set of functions of y and another set of functions of z. (Provided each element of $\boldsymbol{r}(\cdot)$ is such a product, we can always add further elements so that it comprises *all* products of the functions originally appearing.) Then we can arrange $\boldsymbol{r}(\cdot)$ and define $\boldsymbol{T}_Y$ and $\boldsymbol{T}_Z$ in the obvious way to achieve

$$\boldsymbol{T} = \boldsymbol{T}_Y \otimes \boldsymbol{T}_Z. \tag{4.5}$$

Putting (4.2), (4.4) and (4.5) into the results of Section 2 produces various other Kronecker product expressions. Of particular interest is (2.25) which becomes

$$\begin{aligned} \boldsymbol{W} = {} & (\boldsymbol{T}_Y \boldsymbol{A}_Y^{-1}) \otimes (\boldsymbol{T}_Z \boldsymbol{A}_Z^{-1}) \\ & + \{\boldsymbol{R}_Y (\boldsymbol{H}_Y^{\mathrm{T}}\boldsymbol{A}_Y^{-1}\boldsymbol{H}_Y)^{-1}\boldsymbol{H}_Y^{\mathrm{T}}\boldsymbol{A}_Y^{-1}\} \otimes \{\boldsymbol{R}_Z (\boldsymbol{H}_Z^{\mathrm{T}}\boldsymbol{A}_Z^{-1}\boldsymbol{H}_Z)^{-1}\boldsymbol{H}_Z^{\mathrm{T}}\boldsymbol{A}_Z^{-1}\} \\ & - \{\boldsymbol{T}_Y \boldsymbol{A}_Y^{-1}\boldsymbol{H}_Y (\boldsymbol{H}_Y^{\mathrm{T}}\boldsymbol{A}_Y^{-1}\boldsymbol{H}_Y)^{-1}\boldsymbol{H}_Y^{\mathrm{T}}\boldsymbol{A}_Y^{-1}\} \\ & \quad \otimes \{\boldsymbol{T}_Z \boldsymbol{A}_Z^{-1}\boldsymbol{H}_Z (\boldsymbol{H}_Z^{\mathrm{T}}\boldsymbol{A}_Z^{-1}\boldsymbol{H}_Z)^{-1}\boldsymbol{H}_Z^{\mathrm{T}}\boldsymbol{A}_Z^{-1}\}. \end{aligned} \tag{4.6}$$

In the case of a saturated design, i.e. $q = n$ which in turn implies $q_Y = n_Y$ and $q_Z = n_Z$, the first and third terms in (4.6) cancel and the second simplifies to

$$\boldsymbol{W} = \boldsymbol{R}\boldsymbol{H}^{-1} = (\boldsymbol{R}_Y \boldsymbol{H}_Y^{-1}) \otimes (\boldsymbol{R}_Z \boldsymbol{H}_Z^{-1}), \tag{4.7}$$

and $\boldsymbol{W}$ is itself a Kronecker product. This result corresponds to one of the standard approaches to multidimensional integrals in conventional quadrature theory. A product rule for integrating over X combines two separate rules for integrating over Y and Z. The design points consist of the cartesian product of the two sets of points for the component rules. Conventional rules invariably address the case $\boldsymbol{r}(x) = (1)$ (which automatically satisfies our assumption about $\boldsymbol{r}(\cdot)$ above), in which case $\boldsymbol{W}$ reduces to a vector of weights. In the conventional product rule the weight vector is the Kronecker product of the weight vectors for the two component rules. The

usual justification for such rules is that if each component rule integrates exactly a certain set of functions of y or z, then the product rule integrates exactly the set of all products of such functions. This corresponds to our assumption about $\boldsymbol{h}(\cdot)$ above. (4.7) expresses conventional product rules as special cases of Bayesian quadrature.

In Bayesian quadrature generally, however, we obtain this simple product rule form only with saturated designs. Otherwise, $\boldsymbol{W}$ is not a Kronecker product but a linear combination of three Kronecker products. Nevertheless, (4.6) allows Bayesian quadrature to be applied to multidimensional integrals using large numbers of points. The problem of inverting the large $n \times n$ matrix $\boldsymbol{A}$ is reduced by (4.2) to smaller integrals.

4.2. Bayes–Hermite product rules

Application to the Bayes–Hermite case is immediate. The correlation function (3.2) satisfies (4.1), and furthermore satisfies the more general expression

$$c(\|x-x'\|) = \prod_{i=1}^{p} c_i(|x_i - x_i'|) = \prod_{i=1}^{p} \exp\{-b(x_i - x_i')^2\}.$$

Similarly, the p-dimensional normal measure (3.1) not only satisfies (4.3), but is a product of p independent one-dimensional measures. We can therefore consider designs in $\mathbb{R}^p$ consisting of p-fold cartesian products of one-dimensional designs. The simplest rules to use in practice will be products of identical one-dimensional rules. Using an n_0-point one-dimensional rule in this way produces a p-dimensional *power rule* with $n = n_0^p$ points.

Consider for instance the optimal three-point design in one dimension for the case $b = 0.5$, $q = 1$. This sets points at $x = (-1.321, 0, 1.321)^{\mathrm{T}}$. The corresponding weight vector for the basic integral k_1 is $(0.2444, 0.5112, 0.2444)$. The product of this design with itself produces a nine-point design in two dimensions. Using (4.6) we obtain the appropriate weight vector for k_1. The four corner points, such as $(-1.321, -1.321)$, have weights 0.0624, whereas $0.2444^2 = 0.0597$. The center point $(0, 0)$ has weight 0.2595, whereas $0.5112^2 = 0.2613$. The other four points have weight 0.1227, compared with $0.2444 \times 0.5112 = 0.1249$.

However, this is not the optimal 3^2 rule. Taking the case $\boldsymbol{r}(x) = (1)$, and applying (4.2), (4.4) and (4.5) to (2.22) for a general power rule, the posterior variance of k is

$$\begin{aligned} v = u_0^p &- (t_0 A_0^{-1} t_0)^p + \{r_0 (H_0^{\mathrm{T}} A_0^{-1} H_0)^{-1} r_0\}^p \\ &- 2\{r_0^{\mathrm{T}} (H_0^{\mathrm{T}} A_0^{-1} H_0)^{-1} H_0^{\mathrm{T}} A_0^{-1} t_0\}^p \\ &+ \{t_0^{\mathrm{T}} A_0^{-1} H_0 (H_0^{\mathrm{T}} A_0^{-1} H_0)^{-1} H_0^{\mathrm{T}} A_0^{-1} t_0\}^p, \end{aligned} \tag{4.8}$$

where u_0, t_0, A_0, r_0 and H_0 are the appropriate quantities from the one-dimensional design. (t_0 and r_0 are the first rows of T_0 and R_0.)

Minimizing (4.8) shows that the optimal 3^2 design is actually the square of the

Table 3
Optimal power designs $(-x, 0, x)^p$ for $b = 0.5$, $q = 1$

p	1	2	3	4	5	6	7	8	9	10
x	1.321	1.334	1.342	1.347	1.350	1.351	1.351	1.351	1.350	1.348
p	15	25	50	100	200					
x	1.339	1.319	1.298	1.295	1.295					

design (−1.334, 0, 1.334). The difference is not great, and indeed the optimal 3^p designs for a range of p given in Table 3 show that it is not really necessary to tabulate a different three-point design to be used for each p. The p-th powers of a single design like (−1.345, 0, 1.345) will be adequate over all potentially useful p. The same has been found for the 4^p and 5^p designs.

Table 4 gives three-, four- and five-point designs recommended for general use in Bayes–Hermite integration, and particularly for the Bayesian application. The table gives all the data necessary to implement 3^p, 4^p or 5^p rules, for the basic integral k. Three weight vectors are given for each rule, corresponding to the three terms in (4.6). The first row is $\boldsymbol{T}_0\boldsymbol{A}_0^{-1}$, the second is $\boldsymbol{R}_0(\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{A}_0^{-1}\boldsymbol{H}_0)^{-1}\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{A}_0^{-1}$ and the third is $\boldsymbol{T}_0\boldsymbol{A}_0^{-1}\boldsymbol{H}_0(\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{A}_0^{-1}\boldsymbol{H}_0)^{-1}\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{A}_0^{-1}$, except that in each case the result is a vector because $\boldsymbol{r}(x) = (1)$ (so that both $\boldsymbol{T}_0$ and $\boldsymbol{R}_0$ are row vectors).

The rules tabulated here are sub-optimal for various reasons. First, we have approximated the slightly differing rules for different p (as in Table 3) by a single rule. Second, the rules are only optimal within a restricted class of designs. We have only searched among designs that place points symmetrically around the origin. It is possible that optimal rules do not always do this, although a few exhaustive searches

Table 4
Designs and weights for recommended power rules

$n_0 = 3$	Design	−1.345	0	1.345		
	Weights	0.234067	0.517635	0.234067		
		0.422807	0.154386	0.422807		
		0.416790	0.152189	0.416790		
$n_0 = 4$	Design	−1.780	−0.564	0.564	1.780	
	Weights	0.109864	0.388122	0.388122	0.109864	
		0.341081	0.158919	0.158919	0.341081	
		0.339707	0.158279	0.158279	0.339707	
$n_0 = 5$	Design	−2.167	−1.027	0	1.027	2.167
	Weights	0.048419	0.249079	0.403860	0.249079	0.048419
		0.327462	0.040800	0.263456	0.040800	0.327462
		0.327088	0.040753	0.263175	0.040753	0.327088

have not found such. However, it is certain that the optimal design with any fixed number of points in $\mathbb{R}^p$ will not be a product rule, and it is also the case that the optimal product rule will not typically be a power rule. These remarks are based on computations reported in O'Hagan (1989) and have the effect of emphasising the power of the Bayesian quadrature approach. The conventional method using Gauss–Hermite product rules employs powers of a common n_0-point rule for every p. A Bayes–Hermite rule which merely mimics this technique may be expected to provide improvements for every p in the same way as has been demonstrated for $p=1$. To do so requires no more computation than Gauss–Hermite and refers only to a short table such as Table 4 for its implementation. The above remarks, however, show that further improvements may be available. Rules that are optimal within the full class of product rules may be computed, will be just as simple to apply and will require only more extensive tables. If a means can be found to compute rules for large numbers of points with non-cartesian product configurations then further improvements may be achieved.

5. Applications and future research

5.1. Applying Bayes–Hermite quadrature

Application of Bayes–Hermite quadrature requires the specification of $\boldsymbol{r}(\cdot)$, $\boldsymbol{h}(\cdot)$ and b. Unless there are reasons to believe that $f(\cdot)$ approximates to the underlying $G(\cdot)$ times a regression model $\boldsymbol{h}(\cdot)^{\mathrm{T}}\boldsymbol{\beta}$ wherein $\boldsymbol{h}(\cdot)$ takes a *specific* form, the results of Section 3.4 suggest simply setting $\boldsymbol{h}(x)=(1)$. Unlike standard regression models, where adding regressor variables will always improve the fit, it seems better to allow the very general error term $e(\cdot)$ to smooth out the data than to add irrelevant regressor terms. The choice of b is more difficult. It is possible in principle to estimate b from the data, as is done by Sacks et al. (1989) in a different context. In the quadrature problem, their maximum likelihood estimation method would entail iteratively inverting $\boldsymbol{A}$ matrices numerically. A full Bayesian solution would require even heavier computation. We recommend $b=0.5$ for the Bayesian application. Suitable values for other problems will perhaps also be found through simulation.

Bayes–Hermite integration implies a choice of scale, through the simplifying assumption that $G(\cdot)$ is the standard normal distribution. The scale chosen for the simulated mixtures (3.4) was such that these distributions are not absurdly far from N(0, 1). We can expect poor results if we simulate distributions which are very far from normal, or whose mean and variance are far from 0 and 1 respectively. Two methods of choosing the scale can be mentioned. The Naylor–Smith iterative approach is described in Smith et al. (1985). Starting from an arbitrary scale, apply the quadrature rule (they use Gauss–Hermite) and estimate the mean and variance matrix of the distribution represented by $f(\cdot)\,\mathrm{d}G(\cdot)$. Use these to define a new rule,

and repeat until convergence. This procedure seems to work well in practice, although Shaw (1988) shows that convergence properties can be bad, even in nice-looking problems. Alternatively, when $f(\cdot)$ is differentiable we can approximate the mean by the mode of the density $f(\cdot)\,\mathrm{d}G(\cdot)$, and the variance by minus the inverse of the second derivative of $\log f(\cdot)\,\mathrm{d}G(\cdot)$ at the mode. The resulting scale could be refined by applying one or two iterations of the Naylor–Smith scheme.

In order to implement the Naylor–Smith iteration we require estimates of the mean μ_1 and variance μ_2 of the distribution being integrated. In one dimension we can let $\boldsymbol{r}(x) = (1, x, x^2)^{\mathrm{T}}$, so that $\boldsymbol{k} = (k_1, k_2, k_3)^{\mathrm{T}}$, $\mu_1 = k_2/k_1$ and $\mu_2 = (k_3/k_1) - (k_2/k_1)^2$. In higher dimensions $\boldsymbol{r}(\boldsymbol{x})$ must contain all terms x_i, x_i^2 and $x_i x_j$. Formulae for posterior inference about ratios of integrals are given in O'Hagan (1989).

5.2. *Future research*

There is much scope for further research in Bayesian quadrature. We have remarked that the real problems in quadrature lie in high dimensions, where product rules become impractical. It may be possible to identify more efficient patterns of points in high dimensions for which Bayesian quadrature is computationally feasible.

The choice of covariance structure in the Bayes–Hermite formulation is somewhat arbitrary. The relationship between the interpolant in such models and splines is well-established; see Kimeldorf and Wahba (1970). The literature in splines and elsewhere contains a variety of alternative formulations. A related point is that all such structures will have a parameter like b, representing the degree of smoothness. Although the Bayes–Hermite estimate $\hat{k}$ is accurate for a range of values of b, the posterior variance of k is typically very sensitive to b. The Bayesian quadrature method offers, in principle, not just an estimate $\hat{k}$ but a whole posterior distribution; in practice this will be of little value if it is too sensitive to prior assumptions.

In the Bayesian application, assuming a normal $G(\cdot)$ implies a strong prior belief that the tails of $f(\cdot)$ are thin. In practice, heavier tails are quite common. Work is in progress on adapting Bayesian quadrature to such problems.

Acknowledgements

The author thanks a referee for thought-provoking comments, and numerous colleagues, particularly at the University of Warwick where this work originated, for valuable discussions.

References

Blight, B.J.N. and L. Ott (1975). A Bayesian approach to model inadequacy for polynomial regression. *Biometrika* **62**, 79–88.

Cressie, N. (1988). Variogram. In: S. Kotz and N.L. Johnson, Eds., *Encyclopedia of Statistical Sciences, Vol. 9.* Wiley, New York, 489-491.

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

Kimeldorf, G. and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41**, 495-502.

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

O'Hagan, A. (1987). Monte Carlo is fundamentally unsound. *The Statistician* **36**, 247-249.

O'Hagan, A. (1988). Bayesian qyadrature. Warwick Statistics Research Report 159, University of Warwick, Coventry, UK.

O'Hagan, A. (1989). Integrating posterior densities by Bayesian quadrature. Warwick Statistics Research Report 177, University of Warwick, Coventry, UK.

Poincaré, H. (1896). *Calcul des Probabilités.* Georges Carré, Paris.

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

Shaw, J.E.H. (1988). Aspects of numerical integration and summarisation. In: J.M. Bernardo et al., Eds., *Bayesian Statistics 3.* Oxford University Press, Oxford, 411-428.

Smith, A.F.M., A.M. Skene, J.E.H. Shaw, J.C. Naylor and M. Dransfield (1985). The implementation of the Bayesian paradigm. *Comm. Statist. Ser. A* **14**, 1079-1102.

Ylvisaker, D. (1987). Prediction and design. *Ann. Statist.* **15**, 1-19.