Comment on "Probabilistic Integration: A Role in Statistical Computation?"

Fred J. Hickernell and R. Jagadeeswaran

Abstract. Probabilistic integration provides a criterion for stopping a simulation when a specified error tolerance is satisfied with high confidence. We comment on some of the modeling assumptions and implementation issues involved in designing an automatic Bayesian cubature.

Key words and phrases: Bayesian, fast algorithms, quasi-Monte Carlo.

1. WHEN TO STOP?

In highlighting the possibilities of probabilistic integration, the authors of Briol et al. (2018), henceforth abbreviated as PI, have suggested a useful stopping criterion for cubature. Numerical analysis provides an upper bound on the cubature error expressed as a product of the roughness of the integrand and the quality of our sampling scheme. For example, PI (5) quotes the error bound

$$(1.1) \quad |\hat{\Pi}[f] - \Pi[f]| \le \|f\|_{\mathcal{H}} \|\mu(\hat{\pi}) - \mu(\pi)\|_{\mathcal{H}},$$

where:

- the integrand, f, lies in a Hilbert space, $\mathcal{H}$,
- $\Pi[f]$ denotes the desired integral of f defined in terms of the probability measure π, and
- $\hat{\Pi}[f]$ denotes a cubature defined in terms of the discrete measure $\hat{\pi}$.

The *discrepancy* between π and $\hat{\pi}$ is defined as $\|\mu(\hat{\pi}) - \mu(\pi)\|_{\mathcal{H}}$. As the sample size, n, increases, a well-chosen sequence of discrete measures causes the discrepancy to tend to zero.

But, even if $\|\mu(\hat{\pi}) - \mu(\pi)\|_{\mathcal{H}}$ can be computed efficiently, one typically does not have a good estimate or bound on $\|f\|_{\mathcal{H}}$. Therefore, it is impractical to use

Fred J. Hickernell is Professor of Applied Mathematics and Vice Provost for Research, Illinois Institute of Technology, RE 208, 10 W. 32nd Street, Chicago, Illinois 60616, USA (e-mail: hickernell@iit.edu; URL: iit.edu/~hickernell). R. Jagadeeswaran is a Ph.D. student in Applied Mathematics, Illinois Institute of Technology, RE 208, 10 W. 32nd Street, Chicago, Illinois 60616, USA (e-mail: jrathin1@hawk.iit.edu).

(1.1) to determine an n satisfying the error criterion

$$(1.2) \quad |\hat{\Pi}[f] - \Pi[f]| \le \varepsilon,$$

where ε is the user-specified absolute error tolerance.

We believe that the practitioner would like an automatic cubature, that is, an algorithm with a stopping criterion that guarantees (1.2) (with high probability). Probabilistic integration, and in particular Bayesian cubature, as espoused in PI, fulfills that wish.

Bayesian cubature, as explained in PI, assumes that the integrand, f, may be modeled by a Gaussian stochastic process, $g \sim \mathcal{GP}(0, c)$, conditioned on g having the same values as f at the cubature nodes or states, $\{\boldsymbol{x}_i\}_{i=1}^n$. Thus, $\hat{\Pi}[g] = \hat{\Pi}[f]$. Furthermore, Bayesian cubature is designed to satisfy $\hat{\Pi}[g] = \mathbb{E}_n[\Pi[g]]$. Here, c is the covariance function (or kernel) for g. The definition of g allows us to construct credible intervals for the cubature error via Proposition 1 in PI, namely,

$$(1.3) \quad \mathbb{P}[|\hat{\Pi}[f] - \Pi[g]| \le 2.58\sqrt{\mathbb{V}_n[\Pi[g]]}] = 99\%,$$

$$(1.4) \quad \begin{aligned} &\mathbb{V}_n[\Pi[g]] \\ &= \Pi\Pi[c(\cdot, \cdot)] - \Pi[\boldsymbol{c}(\cdot, X)] C^{-1} \Pi[\boldsymbol{c}(X, \cdot)]. \end{aligned}$$

If the observed integrand, f, lies in the 99% middle of the sample space for g, and not in the 1% extreme, then increasing n until $2.58\sqrt{\mathbb{V}_n[\Pi[g]]}$ is no greater than ε ensures that (1.2) holds with 99% probability.

There are some practical obstacles to implementing this elegant recipe.

- How does one choose the covariance function c? While one may always choose the sample space large enough to include f, our use of the credible

interval as a stopping criterion assumes that f is not in the tails of the distribution $\mathcal{GP}(0, c)$. We discuss this question in the next section.

- The computational cost of computing $\mathbb{V}_n[\Pi[g]]$ involves matrix inversion, which requires $\mathcal{O}(n^3)$ operations in general. This typically takes much more time than the $\mathcal{O}(n)$ operations required to compute the cubature, $\hat{\Pi}[f]$, unless obtaining an integrand value is quite time-consuming. We discuss how to circumvent this problem by matching covariance functions and cubature nodes in Section 3.

There are commonalities and differences in the deterministic and Bayesian approaches to numerical integration. We discuss some of these in Section 4.

2. WHICH GAUSSIAN PROCESS?

As mentioned above, using a credible interval as a stopping criterion requires a careful choice of the covariance function, c. The width of the credible interval in (1.3) depends on $\mathbb{V}_n[\Pi[g]]$ given by (1.4). At first glance, nothing in (1.4) depends on the integrand data, $f = (f(x_i))_{i=1}^n$, although our intuition tells us that it should. The credible interval for the integral of $47f$ should be 47 times as wide as the credible interval for the integral of f.

When constructing the confidence interval for the mean of a scalar random variable, Y, from independent and identically distributed (IID) data, one must estimate the variance of Y by the sample variance. Analogously, when constructing the credible interval in (1.3) for the integral (mean) of a function, one must estimate the vertical scale factor inherent in the covariance function c.

We have recently explored Bayesian cubature as the basis for automatically selecting n to satisfy the error criterion (1.2) in Jagadeeswaran and Hickernell (2018), henceforth abbreviated as JH. As in Proposition 2 of PI, JH chooses the covariance function to take the form $c(x, x') = \lambda c_0(x, x'; \theta)$, where λ is the vertical scale factor, and the parameter θ determines the smoothness and other properties of the covariance function. An example of c_0 is the following (JH (36)):

$$
c_0(x, x'; \theta) = \prod_{l=1}^{d}[1 - (-1)^r \gamma B_{2r}(|x_l - x'_l|)],
$$

(2.1)
$$
\forall x, x' \in [0, 1]^d, \theta = (r, \gamma), r \in \mathbb{N}, \gamma > 0,
$$

where B_{2r} is the Bernoulli polynomial of degree $2r$. The smoothness of the covariance function increases with r. Covariance functions of this form appear in Hickernell (1996), Dick, Kuo and Sloan (2013).

Bernoulli polynomials are described in Chapter 24 of Olver et al. (2018).

To increase the possibility that our integrand f lies in the middle of the sample space, we also allow the Gaussian process g to have an arbitrary mean, m, so $g \sim \mathcal{GP}(m, \lambda c_0)$. One may imagine the situation where f represents an option payoff. Then, f is nonnegative and its mean is nonnegative. Assuming an improper prior on (m, λ), the posterior marginal for $\Pi[g]$ is a Student t-distribution with $n - 1$ degrees of freedom and with mean and variance both depending on the integrand data, f (JH (15–16)):

$$
\hat{\Pi}[f] = \mathbb{E}_n[\Pi[g]]
$$

(2.2a)
$$
= \left(\frac{(1 - \mathbf{1}^T C_0^{-1} \Pi[c_0(X, \cdot)]\mathbf{1}^T}{\mathbf{1}^T C_0^{-1} \mathbf{1}} \right.
$$
$$
\left. + \Pi[c_0(\cdot, X)] \right) C_0^{-1} f,
$$

$$
\mathbb{V}_n[\Pi[g]]
$$

$$
= \frac{1}{n-1} f^T \left(C_0^{-1} - \frac{C_0^{-1} \mathbf{1} \mathbf{1}^T C_0^{-1}}{\mathbf{1}^T C_0^{-1} \mathbf{1}} \right) f
$$

(2.2b)
$$
\times \left(\frac{(1 - \Pi[c_0(\cdot, X)] C_0^{-1} \mathbf{1})^2}{\mathbf{1}^T C_0^{-1} \mathbf{1}} \right.
$$
$$
+ \Pi\Pi[c_0(\cdot, \cdot)]
$$
$$
\left. - \Pi[c_0(\cdot, X)] C_0^{-1} \Pi[c_0(X, \cdot)] \right),
$$

(2.2c)
$$
\mathbb{P}[|\hat{\Pi}[f] - \Pi[g]| \le t_{n-1, 0.995} \sqrt{\mathbb{V}_n[\Pi[g]]}]
$$
$$
= 99\%.
$$

Here, $\mathbf{1}$ is a vector of ones, and $t_{n-1, 0.995}$ denotes the 99.5% quantile of the Student t-distribution with $n - 1$ degrees of freedom. For large n, $t_{n-1, 0.995} \approx 2.58$. The expressions in (2.2) are similar to the conclusion of PI, Proposition 2. The differences are due to the mean of the Gaussian process being left unspecified in JH, which reduces the degrees of freedom by one, and adds additional terms to the expressions for $\mathbb{E}_n[\Pi[g]]$ and $\mathbb{V}_n[\Pi[g]]$.

Hidden in the definition of c_0 is the parameter θ. One may place a discrete prior on θ, but this strikes us as rather arbitrary. Thus, in JH we advocate estimating θ by *empirical Bayes*, namely,

$$
\theta_{\text{EB}} = \underset{\theta}{\arg\min} \left\{ \log\left(f^T \left[C_0^{-1} - \frac{C_0^{-1} \mathbf{1} \mathbf{1}^T C_0^{-1}}{\mathbf{1}^T C_0^{-1} \mathbf{1}} \right] f \right) \right.
$$

(2.3)
$$
\left. + \frac{1}{n} \log(\det(C_0)) \right\}.
$$

JH also presents empirical Bayes as an alternative to assuming the improper prior on (m, λ), as discussed in PI Section 4.1.3. Under empirical Bayes, the posterior marginal for $\Pi[g]$ has the same mean as for the full Bayes approach, but a somewhat smaller variance.

JH also discusses the alternative of *generalized cross-validation* for estimating the correct covariance function from the integrand data, as alluded to in Section 4.1.2 of PI. The formulas for the Bayesian cubature and the credible interval width are significantly different than those for full Bayes.

3. SPEEDING UP THE COMPUTATION

Computing the estimate of θ in (2.3) and then the credible interval according to (2.2) involve matrix factorization and computing a matrix determinant, which requires as many as $\mathcal{O}(n^3)$ operations. On the other hand, the computational cost of obtaining the integrand data, f, is $\mathcal{O}(\$(f)n)$, where $\$(f)$ is the computational cost of a single integrand value.

If $\$(f)$ is extraordinarily large compared to the expected sample size n, then the cost of obtaining integrand data dominates, and the $\mathcal{O}(n^3)$ cost of matrix operations is unimportant. However, if $\$(f)$ is close to $\mathcal{O}(1)$, then the cost of matrix operations may make Bayesian cubature prohibitively costly.

JH presents a scenario where the cost of matrix operations may be reduced to $\mathcal{O}(n \log n)$ via fast transforms. The key is choosing covariance functions and cubature nodes that match. Let the matrix C_0 be decomposed in terms of its eigenvectors, which comprise the columns of V, and its eigenvalues, which comprise the diagonal elements of the diagonal matrix Λ:

$$C_0 = (C_1, \ldots, C_n) = \frac{1}{n} V \Lambda V^H, \quad V^H = nV^{-1},$$

$$V = (v_1, \ldots, v_n)^T = (V_1, \ldots, V_n).$$

Four assumptions are made regarding the covariance function, c_0, and the cubature nodes, $\{x\}_{i=1}^n$ (JH (25, 27)):

(3.1a) V may be identified analytically,

(3.1b) $v_1 = V_1 = \mathbf{1}$,

(3.1c) $\tilde{b} := V^H b$ requires only $\mathcal{O}(n \log(n))$ operations $\forall b$,

(3.1d) $\Pi[c_0(\cdot, x)] = 1 \quad \forall x$.

Here, $V^H b$ is called the *fast transform* of b because it takes fewer than the typical $\mathcal{O}(n^2)$ operations required for matrix-vector multiplication.

An example of matching covariance functions and cubature nodes is

- Shift-invariant covariance functions, c_0, which satisfy

$$c_0(x, x') = \mathring{c}_o(x - x' \bmod \mathbf{1}) \quad \forall x, x' \in [0, 1)^d,$$

for some $\mathring{c}_0$ with period 1 in each coordinate direction, and

- Shifted rank-1 integration lattice node sets, $\{x_i\}_{i=1}^n$, which satisfy

$$x, x', x'' \in \{x_i\}_{i=1}^n$$
$$\implies \quad x + x' - x'' \bmod \mathbf{1} \in \{x_i\}_{i=1}^n.$$

The covariance function in (2.1) is an example of a shift-invariant covariance function (Hickernell (1998)). Figure 1 (left) depicts a rank-1 integration node set Sloan and Joe (1994), Dick, Kuo and Sloan (2013). The reason that this family of covariance functions matches this family of cubature nodes and satisfies assumptions (3.1) is that the matrix C_0 is circulant and V may be written in terms of complex exponentials.

Under assumptions (3.1) one may express (2.2) and (2.3) in terms of the fast transforms of the integrand data and the first column of the matrix C_0 (JH Sections 3.2, 3.3):

(3.2a)
$$\tilde{f} := V^H f,$$
$$\ell = \text{diag}(\Lambda) = \tilde{C}_1 := V^H C_1,$$

(3.2b)
$$\hat{\Pi}[f] = \mathbb{E}_n[\Pi[g]] = \frac{1}{n} \sum_{i=1}^n f(x_i)$$
$$= \frac{\tilde{f}_1}{n} \quad \text{the sample average,}$$

(3.2c) $$\mathbb{V}_n[\Pi[g]] = \frac{1}{n(n-1)} \left(\frac{\ell_1}{n} - 1 \right) \sum_{i=2}^n \frac{|\tilde{f}_i|^2}{\ell_i},$$

(3.2d)
$$\theta_{\text{EB}} = \underset{\theta}{\text{argmin}} \left[\log\left(\sum_{i=2}^n \frac{|\tilde{y}_i|^2}{\ell_i} \right) + \frac{1}{n} \sum_{i=1}^n \log(\ell_i) \right].$$

Apart from the computations in (3.2a), which require $\mathcal{O}(n \log n)$ operations, all other calculations in (3.2) require only $\mathcal{O}(n)$ operations. The expression for $\mathbb{V}_n[\Pi[g]]$ excludes $i = 1$ in the sum because we allow g to have an arbitrary constant mean.
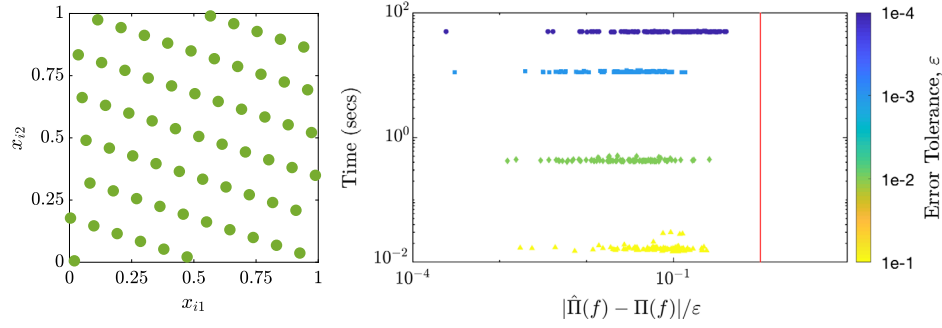
FIG. 1. *An example of shifted integration lattice nodes in two dimensions (left). The performance of Bayesian cubature for an option pricing example (right).*

Section 5 of JH presents several numerical experiments for Bayesian cubature with shift-invariant covariance functions and lattice nodes sets. We reproduce one such experiment in Figure 1 (right). The integrand is the payoff of an arithmetic mean Asian option. The covariance function is the one given by (2.1) with $r = 1$. A low degree of smoothness is chosen in view of the discontinuities in the partial derivatives of the integrand. The value of γ is determined by empirical Bayes as in (3.2d). The sample size, n, is increased in a sequence of powers of 2 until the stopping criterion implied by (2.2c), whose terms are computed quickly via (3.2), is satisfied. Four different values of the tolerance were tried, $\varepsilon = 0.1, 0.01, 0.001$, and 0.0001. The aim is for the cubature error, $|\Pi[f] - \hat{\Pi}[f]|$, to be no greater than, but not too much less than, the prescribed tolerance nearly all the time. In this experiment, the error tolerance is always met. As expected, the computation time increases as the tolerance decreases.

In this example, and others provided in Section 5 of JH, the cost of evaluating the integrand is modest, and so the cost of obtaining the needed integrand data, $\boldsymbol{f}$, is on the same order as the matrix-vector operations required to compute the credible interval. JH also provides examples of the empirical Bayes and generalized cross-validation approaches to determining the parameters inherent in the covariance function and to using credible intervals as stopping criteria for Bayesian cubature. All of these approaches are successful, which suggests that they should be explored over a larger range of examples.

4. BAYESIAN VERSUS DETERMINISTIC ANALYSIS

We return to the situation where the Gaussian process, g, has zero mean. Section 3.2 in PI sets the covariance function, c, identical to the reproducing kernel, k, of the Hilbert space containing the integrand, f,

for "aesthetic" reasons. While this makes the application of several results from numerical analysis of deterministic cubature more readily transferable to Bayesian cubature, we think that such a correspondence muddies the waters.

Suppose that $\{\phi_i\}_{i=1}^\infty$ is an orthonormal basis for the Hilbert space $\mathcal{H}$ with reproducing kernel k. Then

$$(4.1) \qquad k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^\infty \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x}').$$

Moreover, the norm of any $f = \sum_{i=1}^n a_i \phi_i \in \mathcal{H}$ is $\|f\|_{\mathcal{H}} = \|\boldsymbol{a}\|_2$. If $g = \sum_{i=1}^\infty A_i \phi_i$, with $A_i \overset{\text{IID}}{\sim} \mathcal{N}(0, 1)$, then $g \in \mathcal{GP}(0, c)$, where c is identical to k as defined in (4.1). While this may seem well and good, note that

$$\mathbb{E}(\|g\|_{\mathcal{H}}^2) = \mathbb{E}(\|A\|_2^2) = \|(1, 1, \ldots)\|_2^2 = \infty.$$

So setting c identical to k means that we are modeling an integrand with finite norm by a Gaussian process with an infinite expected squared norm. This seems counter-intuitive. Nevertheless, there are tantalizing mathematical similarities between the Bayesian and deterministic approaches to cubature.

When the optimal cubature weights are used, the deterministic error bound in (1.1) may be expressed as

$$(4.2a) \quad \hat{\Pi}[f] = \Pi[\hat{f}] = \Pi[\boldsymbol{k}(\cdot, X)]\boldsymbol{K}^{-1}\boldsymbol{f},$$

$$|\hat{\Pi}[f] - \Pi[f]|^2$$
$$(4.2b) \qquad \leq \|f - \hat{f}\|_{\mathcal{H}}^2 \{\Pi\Pi[k(\cdot, \cdot)]$$
$$- \Pi[\boldsymbol{k}(\cdot, X)]\boldsymbol{K}^{-1}\Pi[\boldsymbol{k}(X, \cdot)]\},$$

where $\hat{f}$ is the minimum Hilbert space norm interpolant of the integrand, f. The reason that $\|f\|_{\mathcal{H}}$ in (1.1) can be replaced by $\|f - \hat{f}\|_{\mathcal{H}}$ is that the cubature in (4.2a) integrates $\hat{f}$ exactly. Moreover, $f - \hat{f}$ is

orthogonal to $\hat{f}$ under the Hilbert space inner product. Also note that

$$(4.3) \qquad \|\hat{f}\|_{\mathcal{H}}^2 = \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}.$$

Compare the error bound in (4.2) to Proposition 2 of PI, which implies that

$$(4.4a) \quad \hat{\Pi}[f] = \mathbb{E}[\Pi[g]] = \Pi[\boldsymbol{c}_0(\cdot, X)] \boldsymbol{C}_0^{-1} \boldsymbol{f},$$

$$\mathbb{V}[\Pi[g]]$$

$$(4.4b) \quad = \frac{\boldsymbol{f}^T \boldsymbol{C}_0^{-1} \boldsymbol{f}}{n} \{ \Pi\Pi[c_0(\cdot, \cdot)]$$

$$- \Pi[\boldsymbol{c}_0(\cdot, X)] \boldsymbol{C}_0^{-1} \Pi[\boldsymbol{c}_0(X, \cdot)] \},$$

$$(4.4c) \quad \mathbb{P}[|\hat{\Pi}[f] - \Pi[f]|^2 \le t_{0.995,n}^2 \mathbb{V}[\Pi[g]]] = 99\%.$$

The formulas for the cubature $\hat{\Pi}[f]$ in both the deterministic and Bayesian senses are identical if k is identical to c_0. They are also independent of the vertical scale factor multiplying inherent in the definition of k or c_0, that is, the reproducing kernels k and $47k$ yield the same cubature rule. Likewise, the bound on the squared error in (4.2b) and the variance of the integral of the Gaussian process in (4.4c) both contain the common factor $\Pi\Pi[k(\cdot, \cdot)] - \Pi[\boldsymbol{k}(\cdot, X)] \boldsymbol{K}^{-1} \Pi[\boldsymbol{k}(X, \cdot)]$ if k is identical to c_0.

Matching the deterministic error bound in (4.2) to the Bayesian credible interval in (4.4) when k is identical to c_0 becomes possible if one applies (4.3) and asserts that

$$\|f - \hat{f}\|_{\mathcal{H}}^2 \le \frac{t_{0.995,n}^2 \|\hat{f}\|_{\mathcal{H}}^2}{n}$$

$$(4.5) \qquad = \frac{t_{0.995,n}^2 \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}}{n}.$$

Although this inequality is violated for some $f \in \mathcal{H}$, it holds for those $f \in \mathcal{H}$ that are well modeled by their minimum norm interpolants, $\hat{f}$. Thus, one can mimic Bayesian cubature via a deterministic cubature which assumes that the integrand satisfies inequality (4.5).

5. FURTHER MATTERS

A couple of matters deserve further investigation. How large a family of covariance functions must be considered for effective Bayesian cubature? A larger family increases the probability that the integrand in question lies in the middle of the space of Gaussian processes used to determine the stopping criterion. On the other hand, a larger family may require a more tedious choice of the underlying parameters $\boldsymbol{\theta}$.

The Bayesian approach to numerical integration assumes a Gaussian process. Do goodness-of-fit statistics confirm or discredit this assumption? How does the validity of this assumption affect the reliability of the proposed Bayesian automatic cubature? The alternative of Student t-processes has been suggested by Shah, Wilson and Ghahramani (2014).

Finally, probabilistic numerics—including Bayesian cubature—deserves further participation from statisticians, numerical analysts, and software developers alike. Statisticians and numerical analysts should become more conversant in each other's language and culture. Computational problems are better understood when one can look from multiple perspectives. Moreover, the algorithms that arise from probabilistic numerics should find their way into commonly used software libraries. Such libraries should be built using software engineering principles that are familiar to software developers, but perhaps not obvious to statisticians. Our Guaranteed Automatic Integration Library (GAIL) Choi et al. (2013–2017) is an example of such a library.

REFERENCES

BRIOL, F.-X., OATES, C. J., GRIOLAMI, M., OSBORNE, M. A. and SEJDINOVIC, D. (2019). Probabilistic integration: A role in statistical computation? *Statist. Sci.* **34** 1–22.

CHOI, S. C. T., DING, Y., HICKERNELL, F. J., JIANG, L., JIMÉNEZ RUGAMA, LL. A., LI, D., JAGADEESWARAN, R., TONG, X., ZHANG, K., ZHANG, Y. and ZHOU, X. (2013–2017). GAIL: Guaranteed Automatic Integration Library (Versions 1.0–2.2).

DICK, J., KUO, F. Y. and SLOAN, I. H. (2013). High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.* **22** 133–288. MR3038697

HICKERNELL, F. J. (1996). Quadrature error bounds with applications to lattice rules. *SIAM J. Numer. Anal.* **33** 1995–2016. MR1411860

HICKERNELL, F. J. (1998). Lattice rules: How well do they measure up? In *Random and Quasi-Random Point Sets. Lect. Notes Stat.* **138** 109–166. Springer, New York. MR1662841

JAGADEESWARAN, R. and HICKERNELL, F. J. (2018). Fast automatic Bayesian cubature using lattice sampling. Submitted for publication.

OLVER, F. W. J., LOZIER, D. W., BOISVERT, R. F., CLARK, C. W. and DALHUIS, A. B. O. (2018). Digital Library of Mathematical Functions.

SHAH, A., WILSON, A. G. and GHAHRAMANI, Z. (2014). Student-t processes as alternatives to Gaussian processes. In *Proceedings of the* 17*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 2014, *Reykjavik, Iceland, PLMR* (S. Kaski and J. Corander, eds.) **33** 877–885.

SLOAN, I. H. and JOE, S. (1994). *Lattice Methods for Multiple Integration. Oxford Science Publications*. Oxford University Press, New York. MR1442955