Statistical Science
2019, Vol. 34, No. 1, 34–37
https://doi.org/10.1214/18-STS677
Main article: https://doi.org/10.1214/18-STS660
© Institute of Mathematical Statistics, 2019

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

Michael L. Stein and Ying Hung

We commend the authors for their serious effort to address some of the mathematical, conceptual and practical issues that arise when using probabilistic methods for evaluating integrals of deterministic functions and assessing the uncertainties in these methods. Applying Gaussian process models to deterministic but difficult to compute functions is, as this paper emphasizes, gaining increasing attention in the numerical analysis literature, but it has been actively pursued in the computer experiment literature since at least the seminal paper Sacks et al. (1989). In computer experiments, the interest is usually in interpolation or optimization of some complex deterministic function, rather than integration, but many of the issues raised in this work are also pertinent when interpolating. We would also point to Sacks and Ylvisaker (1970) as an early work that considers theoretical design issues when integrating Gaussian processes, although from the standpoint of assuming the unknown function really is a Gaussian process model with a known covariance structure.

The present paper takes the point of view, common in the approximation theory literature, that the unknown function lies in some specified RKHS. It then exploits the fact that this reproducing kernel can be viewed as the covariance function for a Gaussian process, making it possible to make Bayesian inferences based on this model. The problem with this approach, which has long been known but can still lead to confusion, is that if f is a realization of a Gaussian process with covariance function k, then its realizations are insufficiently smooth to be elements of the RKHS. This paper provides some theory and a number of examples showing that, despite this fundamental problem, the Bayesian inferences we get from the Gaussian process model may provide good point estimates and useful uncertainty assessments for the integral of f over some domain.

Michael Stein is Professor of Statistics, University of Chicago, Chicago, Illinois, USA (e-mail: stein@galton.uchicago.edu). Ying Hung is Assistant Professor of Statistics, Rutgers University, New Brunswick, New Jersey, USA (e-mail: yhung@stat.rutgers.edu).

The theory in this work focuses on posterior contraction to the true value of the integral. As the authors make clear, this is a very different notion than the posterior distribution providing accurate probability statements. Indeed, the paper also notes that there is no theory, asymptotic or otherwise, to support a claim that the posterior probability statements coming from Bayes theorem will have a valid probability interpretation when f is an element of the RKHS.

To clarify the issues, let us consider a simple example. Suppose we wish to integrate a function f on $[0,1]$ for which it is known $f(0) = f(1) = 0$. Brownian bridge on $[0,1]$ provides a Gaussian process model whose realizations satisfy this condition; its mean is 0 and its covariance function on $[0,1] \times [0,1]$ is $k(x,y) = \sigma^2(\min(x,y) - xy)$ for some $\sigma > 0$. For a Gaussian process Z with this covariance function observed at j/n for $j = 0, \ldots, n$, the optimal predictor of $\int_0^1 Z(x)\,dx$ is the trapezoidal rule and its RMSE can be shown to be $\sigma/(\sqrt{12}n)$. Realizations of this Gaussian process are nowhere differentiable with probability 1. In contrast, the elements of the RKHS with this kernel, which we can call $\mathcal{H}(k)$, are functions f that satisfy $f(0) = f(1) = 0$ and are absolutely continuous with an almost everywhere first derivative that is square integrable.

It is exceedingly implausible that a likelihood function or a solution of a deterministic differential equation of practical interest would be nowhere differentiable, so one should use caution when interpreting posterior probability statements based on this model. Lack of differentiability in these types of functions, when it occurs, tends to occur along lower-dimensional manifolds. For example, for $0 < s < t < 1$, suppose $f(x)$ equals 1 for $s < x < t$ and is 0 otherwise. This function shares some properties with Brownian bridge. First,

$$\lim_{n\to\infty} \sum_{j=1}^{n} \left\{ Z\left(\frac{j}{n}\right) - Z\left(\frac{j-1}{n}\right) \right\}^2 = \sigma^2 \tag{0.1}$$

almost surely and

$$\lim_{n\to\infty}\sum_{j=1}^{n}\left\{f\left(\frac{j}{n}\right)-f\left(\frac{j-1}{n}\right)\right\}^2=2, \tag{0.2}$$

so these limits are the same when $\sigma^2=2$. Note that the limit in (0.2) is 0 for any f in $\mathcal{H}(k)$. Second, the spectral representation for Brownian bridge closely resembles the Fourier series for f. Specifically,

$$Z(x)=\sum_{n=1}^{\infty}B_n\sin(\pi nx),$$

where the B_n's are independent $\mathrm{N}(0,2\sigma^2/(n\pi)^2)$, whereas

$$f(x)=\sum_{n=1}^{\infty}\frac{\cos(2\pi ns)-\cos(2\pi nst)}{\pi n}\sin(2\pi nx).$$

For Z, the standard deviations for the coefficients in the sine series decay like $1/n$ and, for f, the coefficients in the analogous sine series Fourier decay like $1/n$.

The integration errors for Z and f are also comparable. Specifically, the error of integration of the trapezoidal rule for f is $t-\lfloor nt\rfloor/n-(s-\lfloor ns\rfloor/n)$, which is $O(1/n)$, just like the RMSE for the integral of Z. But the correspondence is arguably much tighter than this. Specifically, first imagine that $S<T$ are the ordered values for two independent draws from a uniform distribution on $[0,1]$ and s and t are their realized values. Then straightforward calculation yields that the RMSE of the numerical integration is $1/(\sqrt{6}n)$, which is exactly what we get when $\sigma^2=2$ and (0.1) and (0.2) agree. However, the distribution of the error for f is not even asymptotically Gaussian, but instead follows a triangular distribution on $[-2/n,2/n]$ for all n sufficiently large. Note that if we define $f(x)=V$ on $S<x<T$ with $\mathrm{E}(V)=0$, $\mathrm{var}(V)=1$ and V, S and T independent, then $\mathrm{cov}(f(x),f(y))=2(\min(x,y)-xy)$, exactly the same as for Brownian bridge with $\sigma^2=2$.

Using either the famous result of Blackwell and Dubins on the merging of posterior distributions (Blackwell and Dubins, 1962) or by direct calculation, it is possible to show that if the two jump points are sampled from any density on $[0,1]^2$ that is bounded away from 0 and ∞, then the posterior distribution of the error is very nearly triangular on $[-2/n,2/n]$ in the sense that the variation distance between this distribution and the exact posterior tends to 0 as $n\to\infty$. We might interpret this result as saying that if the only aspects of f unknown were the jump points s and t with its value fixed at 1 between s and t, then any reasonable Bayesian model representing our uncertainty about their location would lead to asymptotically equivalent inferences about our uncertainty for the integral when using equally spaced design points. Furthermore, if we behave as if this f were a realization of Brownian bridge with σ^2 unknown and estimate it by $\hat{\sigma}^2=\sum_{j=1}^{n}\{f(\frac{j}{n})-f(\frac{j-1}{n})\}^2$, then (0.1) and (0.2) imply that $\hat{\sigma}/(\sqrt{12}n)$ gives an asymptotically valid estimate of the root mean squared integration error under the stochastic model for (S,T).

One might conclude that this example shows the appropriateness of treating deterministic functions as Gaussian processes even when that assumption is badly untrue. Perhaps, although the fact that f is not in the Sobolev space $\mathcal{H}_\alpha$ for any positive integer α means that none of the results in this paper apply to f. But there is a deeper problem in our view. Suppose one were faced with integrating a function on $[0,1]$ like f in practice that had jumps at unknown locations. Clearly there would be value in using an adaptive sampling strategy that added more observations between existing observations where a jump appears to occur. For the function f itself with jump points s and t unknown, an error rate exponentially decreasing in n would be attainable by using a divide and conquer approach to locating the two jump points. More generally, for a function on a region in d dimensions that is infinitely differentiable on most of its domain with limited differentiability on unknown lower dimensional manifolds, it would surely improve the integration to know the locations of these manifolds. In principle, one could attempt to accommodate such singularities through Gaussian process models by allowing nonstationarity in either the mean or the covariance function of the process.

A number of methods have been proposed to incorporate nonstationarity into Gaussian process models, including those that use spatial deformations and process convolutions. The idea behind the spatial deformation approach is to map nonstationary inputs in the original space into a dispersion space in which the process is stationary. Early work on this approach is due to Sampson and Guttorp (1992), although that work requires multiple realizations of the random process. The idea of process convolution is to construct nonstationary processes via the convolution of a family of independent stationary Gaussian processes. For example, Fuentes (2002) proposed a spatially-varying superposition of several independent stationary processes. These methods are usually computationally demanding. A faster alternative is to couple stationary processes with treed partitioning to account for the nonstationarity (Gramacy and Lee, 2008).

When the goal is to interpolate with estimated covariance parameters from a Gaussian process, selecting states by experimental design methods can be crucial. Zhu and Stein (2006) considered optimal design criteria that account for estimation uncertainty in both point and interval prediction using asymptotic approximations. In addition to regularly spaced sampling points, their resulting optimal designs often contain clusters of points. These clusters are important for capturing the local behavior of the process and it is this local behavior that determines what RKHS a function lies in. Although it is computationally infeasible to carry out an exhaustive search for the optimal design, some heuristic search algorithms have been proposed to solve this problem. Additionally, for moderately large sample sizes, the two-step algorithm proposed by Zhu and Stein (2006) appears to be effective in tackling the computational issue. It uses some of the sites to find the best design for prediction with known covariance parameters and then, conditional on these sites, uses the rest to find the best design for estimation of those covariance parameters. Discretization of the search space by Latin hypercube designs and the implementation of columnwise-pairwise exchanges can further enhance the search efficiency, especially for high-dimensional problems. Besides nonadaptive designs, a number of sequential sampling procedures have been proposed. For example, Gramacy and Lee (2009) draws ideas from active learning to construct an adaptive design and Joseph (2012) proposed a sequential strategy to add new points at the locations with maximum conditional prediction variance.

We found it a bit restrictive statistically to focus only on numerical uncertainty in the point estimate, such as the example given in Section 5.3. Inferences beyond point estimates, like uncertainty quantification of prediction intervals, can be of great importance in many applications. Note that an advantage of the MCMC is that it provides a direct way of sampling from the posterior, which is of greater interest than just the posterior mean.

In terms of the implementation of the methodology, we would like to mention the possibility of using conditional simulations of Gaussian processes as a tool for addressing some of the numerical issues related to Bayesian cubature, at least for stationary Gaussian processes in up to, say, three dimensions. Specifically, because exact simulation of a wide range of stationary Gaussian processes on a dense grid can be done efficiently using the discrete Fourier transform (Gneiting et al., 2006) and because conditional simulations of Gaussian processes can be readily done using point prediction and unconditional simulations (Chilès and Delfiner, 2012), we can move beyond models with a closed-from kernel mean as long as we are willing to approximate the integral by a dense sum. By carrying out multiple conditional simulations under a fitted Gaussian process model, we can use their average as an estimate of the posterior mean and their sample standard deviation as an estimate of the posterior standard deviation. A useful advantage of this conditional simulation approach is that it can be trivially applied when f is modeled as a known strictly monotonic pointwise transformation of a Gaussian process. For example, when integrating posterior distributions as in Sections 5.2 or 5.4, it may make more sense to treat the log-posterior as a Gaussian process than the posterior itself, in which case one can trivially exponentiate each conditional simulation of the Gaussian process and average these values to get a sample from the distribution of the integrated posterior given the available values of the posterior.

It is interesting that the probability statements in the examples based upon the Gaussian process models are often reasonably well calibrated. It is conceivable that there could be a theoretical justification for this finding, but this paper does not provide it and it may be difficult even to formulate a relevant asymptotic theory. Whenever uncertainty quantification is of interest, we consider it conceptually helpful to think about what a really smart Bayesian with lots of time and computing resources might do, and thus, find the theoretical results and practical examples in this paper worthy of careful consideration. However, we are skeptical that the solution to this problem will be found in deciding which RKHS to assume the function is an element. We would advocate instead taking the implications of the stochastic process model more seriously, which will often mean seeking models beyond stationary Gaussian processes.

REFERENCES

BLACKWELL, D. and DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Stat.* **33** 882–886. MR0149577

CHILÈS, J.-P. and DELFINER, P. (2012). *Geostatistics*: *Modeling Spatial Uncertainty*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2850475

FUENTES, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika* **89** 197–210. MR1888368

GNEITING, T., ŠEVČÍKOVÁ, H., PERCIVAL, D. B., SCHLATHER, M. and JIANG, Y. (2006). Fast and exact simulation of large Gaussian lattice systems in $\mathbb{R}^2$: Exploring the limits. *J. Comput. Graph. Statist.* **15** 483–501. MR2291260

GRAMACY, R. B. and LEE, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *J. Amer. Statist. Assoc.* **103** 1119–1130. MR2528830

GRAMACY, R. B. and LEE, H. K. H. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics* **51** 130–145. MR2668170

JOSEPH, V. R. (2012). Bayesian computation using design of experiments-based interpolation technique. *Technometrics* **54** 209–225. MR2967968

SACKS, J. and YLVISAKER, D. (1970). Statistical designs and integral approximation. In *Proc. Twelfth Biennial Sem. Canad. Math. Congr. on Time Series and Stochastic Processes*; *Convexity and Combinatorics* (*Vancouver, B.C.*, 1969) 115–136. Canad. Math. Congr., Montreal, QC. MR0277069

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SAMPSON, P. D. and GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.

ZHU, Z. and STEIN, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *J. Agric. Biol. Environ. Sci.* **11** 24–44.