Bayesian Analysis (2016) **11**, Number 4, pp. 1285–1293

Contributed Discussion on Article by Chkrebtii, Campbell, Calderhead, and Girolami*

Comment[1] by François-Xavier Briol[2,3], Jon Cockayne[4], and Onur Teymur[5]

Abstract. We commend the authors for an exciting paper which provides a strong contribution to the emerging field of probabilistic numerics. Below, we discuss aspects of prior modelling for differential equations which will need to be considered thoroughly in future work.

Keywords: probabilistic numerics, uncertainty quantification, numerical analysis.

Introduction

The majority of probabilistic numerics (PN) solvers, including the present paper, take a Bayesian viewpoint and hence require several modelling choices including prior specification. As with any inference problem, there exists a trade-off between representing prior beliefs and choosing a prior which is convenient and/or readily interpretable mathematically. We believe that the consequences of these assumptions are often discussed in too little detail and therefore highlight below several issues to consider.

Computational Complexity

Of interest was the discussion into reduction of the computational complexity by exploiting compactly supported covariance function. The authors note in Section 3.2 that while such a choice will yield a method involving inversion of a sparse matrix, this is not explored further – though this will have an effect on the rate of convergence of the estimator. We believe that a study of the extent of this effect is of some importance, as there is a clear trade-off here between steps desired to achieve a required tolerance, and the computational cost of each step.

*Main article DOI: 10.1214/16-BA1017.

[1] This work was completed within the Probabilistic Numerics working group which is part of the 2016–2017 SAMSI programme on Optimization.

[2] FXB was supported by EPSRC [EP/L016710/1].

[3] Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom, f-x.briol@warwick.ac.uk

[4] Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom, j.cockayne@warwick.ac.uk

[5] Department of Mathematics, Imperial College London, London, SW7 2AZ, United Kingdom, o@teymur.uk

DOI: 10.1214/16-BA1017A

Tractability

One issue is the intractability of the joint conditional predictive probability distribution in Section 2.1 which depends on analytically convolving covariance functions. This is not possible except for a few simple kernels; relying upon such construction therefore significantly restricts the range of priors available.

With this in mind, we note that differentiating kernels is often easier than integrating them. Unless there is a specific reason to model u_t, it may therefore be more convenient to define a kernel for u and differentiate it to obtain a kernel for u_t.

Another interesting point is that this trade-off is also encountered in Bayesian Quadrature, a PN method for integration. A table of kernels which can be integrated analytically is provided in Briol et al. (2015) and may be of interest to users of the present methodology.

Boundary Conditions

An important point for partial differential equations is how best to make use of boundary information. The authors observe that, for ordinary differential equations, it is simple to encode the initial condition in the prior, but generalising this to partial differential equations is significantly more challenging owing to the fact that the boundaries will now typically be a manifold of dimension larger than zero.

Significant work in this area includes that of Owhadi (2015) and Cockayne et al. (2016) which select covariances based on Green's functions, though the computations involved are challenging and such closed-form conditioning is narrowly applicable as a result. In general Owhadi and Scovel (2015) shows that conditioning over the entire boundary is well-defined from a mathematical perspective, provided the boundary operator is linear. We would be interested in whether this can be generalised in a tractable way so that, for example, we can define a prior over those functions which satisfy the boundary conditions exactly.

Relationship to Known Integrators

A desideratum (although not always a requirement) for a probabilistic method is that the estimate given by some readily-calculated statistic of the posterior distribution corresponds to the output of a classical numerical solver. The advantage here is that the theory of such solvers is highly developed and certain properties – convergence, stability, etc. – can potentially be inherited. This method does not, to the best of our knowledge, satisfy this property. However, more recent work which builds on this work includes a general construction provided by Conrad et al. (2016) and careful choice of the kernel within a similar framework has subsequently been shown to correspond to Runge–Kutta methods of order less than four (Schober et al., 2014) and linear multi-step methods of arbitrary order (Teymur et al., 2016).

Conclusion

Once again, we would like to congratulate the authors of this paper which provided foundational work upon which many subsequent PN methods have been built. We hope to have highlighted some of the important issues relating to the development of priors which will hopefully influence future work in this area.

References

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2015). "Probabilistic Integration: A Role for Statisticians in Numerical Analysis?" arXiv:1512.00933. 1286

Cockayne, J., Oates, C. J., Sullivan, T., and Girolami, M. (2016). "Probabilistic Meshless Methods for Partial Differential Equations and Bayesian Inverse Problems." arXiv:1605.07811. 1286

Conrad, P., Girolami, M., Särkkä, S., Stuart, A., and Zygalakis, K. (2016). "Statistical Analysis of Differential Equations: Introducing Probability Measures on Numerical Solutions." *Statistics and Computing*. doi: `http://dx.doi.org/10.1007/s11222-016-9671-0`. 1286

Owhadi, H. (2015). "Bayesian Numerical Homogenization." *SIAM Multiscale Modeling & Simulation*, 13(3): 818–828. MR3369060. doi: `http://dx.doi.org/10.1137/140974596`. 1286

Owhadi, H. and Scovel, C. (2015). "Conditioning Gaussian Measure on Hilbert Space." arXiv:1506.04208. 1286

Schober, M., Duvenaud, D., and Hennig, P. (2014). "Probabilistic ODE Solvers with Runge–Kutta Means." In *Advances in Neural Information Processing Systems*, 739–747. 1286

Teymur, O., Zygalakis, K., and Calderhead, B. (2016). "Probabilistic Linear Multistep Methods." In *Advances in Neural Information Processing Systems*. arXiv:1610.08417. 1286

Comment by William Weimin Yoo[1]

Abstract. We begin by introducing the main ideas of the paper, and we give a brief description of the method proposed. Next, we discuss an alternative approach based on B-spline expansion, and lastly we make some comments on the method's convergence rate.

Keywords: differential equation, discretization uncertainty, B-splines, tensor product B-splines, convergence rate.

I would like to congratulate the authors for such an interesting research. The Bayesian method with the probabilistic solver introduced is highly innovative and practical. The various examples presented in the paper show the wide applicability of the proposed method. However, I do find the title a bit of a misnomer, as I initially thought that the authors are constructing credible sets for the fixed but unknown solution u^* of the differential equation.

The inverse problem that the authors are trying to solve, in its most basic form is this: Suppose you have observations $Y = Au + \varepsilon$, where ε is some normal errors and u follows $u_t = f(t, u, \theta)$. Here, A is a known transformation from the state space u to the observation space Y, u_t is the first order derivative with respect to its argument t, f is the known form of the differential equation, and θ's are the equation's parameters. The method proposed consists of two steps, with one nested within the other. First, solve for u probabilistically to obtain a discretized solution at some grid points. Then we embed these discretized version of u in a Bayesian hierarchical framework to estimate θ. To model discretization uncertainty associated with using only u evaluated at grid points, the authors endow priors based on Gaussian process jointly on u and u_t, where the covariance function is constructed by convolving kernels.

There is an alternative and perhaps a conceptually easier way to achieve the same result. We can first represent u by a B-spline series, i.e., $u(t) = \sum_{j=1}^{J} \vartheta_j B_{j,q}(t)$ with $B_{j,q}(\cdot)$ denoting the jth B-spline of order q, and we endow the coefficients ϑ_j's with normal priors. Here, the number of basis J plays the role of $1/\lambda$, where λ is the length-scale parameter defined in the paper. It turns out that the first derivative of this u is another B-spline series $u_t(t) = \sum_{j=1}^{J-1} \vartheta_j^{(1)} B_{j,q-1}(t)$ where $\vartheta_j^{(1)}$ is some weighted first order finite difference of the ϑ_j's ((4.23) of Schumaker (2007)). Therefore, u and u_t are jointly normal and their associated covariance matrices are banded due the support separation property of B-splines. To enforce the given initial condition, we can condition the joint prior (u, u_t) on $u^*(0)$.

Moreover, this approach can be generalized to the partial differential equation case, where we take tensor product of B-splines to model both the spatial and temporal components, i.e., $u(x,t) = \sum_{j_1=1}^{J} \sum_{j_2=1}^{J} \theta_{j_1,j_2} B_{j_1,q}(x) B_{j_2,q}(t)$. As in the univariate case, partial derivatives of tensor product B-splines will be another tensor-product B-splines ((3.2) of Yoo and Ghosal (2016)). Hence we will obtain the same Gaussian process prior

[1] Mathematical Institute, Leiden University, The Netherlands, yooweimin0203@gmail.com

for u and all its mixed partial derivatives if we endow normal priors on the coefficients. As before, we enjoy some simplification in computing the covariance matrices because they are banded.

In addition, I would like to comment on the effect of grid point distribution on the convergence rate of the proposed algorithm. Intuitively, one would expect that the grid points should be chosen roughly uniformly across the domain $[0, L]$. Suppose we choose grid points $\{t_1, t_2, \ldots, t_N\}$ and we further assume that they are quasi-uniform, i.e., $h/\min_i(t_i - t_{i-1}) \leq C$ for some constant $C > 0$ with $h = \max_i(t_i - t_{i-1})$. In other words, the max grid increment is of the same order as the min grid increment. Then it follows that h is of the order of $1/N$ and by Theorem 1 of Chkrebtii et al. (2016), the rate of convergence is $O(N^{-1})$. Therefore for quasi-uniform grids (which includes uniform discrete grids), increasing the number of grid points will result in more accurate solution.

The paper under discussion Chkrebtii et al. (2016) makes significant contribution to the new field of probabilistic numerics. I have learnt a great deal by reading this paper, and I hope that there will be more papers in uncertainty quantification for differential equation models in the future.

References

Chkrebtii, O. A., Campbell, D. A., Calderhead, B., and Girolami, M. A. (2016). "Bayesian solution uncertainty quantification for differential equations." *Bayesian Analysis*, Advance publication. doi: http://dx.doi.org/10.1214/16-BA1017. 1289

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, New York, third edition. MR2348176. doi: http://dx.doi.org/10.1017/CBO9780511618994. 1288

Yoo, W. W. and Ghosal, S. (2016). "Supremum norm posterior contraction and credible sets for nonparametric multivariate regression." *Annals of Statistics*, 44(3): 1069–1102. MR3485954. doi: http://dx.doi.org/10.1214/15-AOS1398. 1288

Comment by Jon Cockayne[1]

I would like to thank the authors for their interesting and very clearly presented paper discussing probabilistic solvers for ordinary differential equations (ODEs) and partial differential equations (PDEs).

1 Nature of the Uncertainty Quantification

I am particularly interested in the nature of the uncertainty quantification provided over the forward model. Considering the ODE

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t) = f(t, u) \ ,$$

we note that Skilling (1992) advocates construction of a probabilistic model for the vector field $f(t, u)$, the uncertainty of which is then propagated to the solution u itself. This is "Bayesian" in that all evaluations of f are incorporated into the estimate of u.

Conversely in this work it seems that there is an inconsistency in the posterior distributions obtained. To consider a simply toy example, suppose we wish to solve the linear ODE

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t) = f(t),$$

where f is independent of u, and the problem is thus linear. For a Bayesian treatment of this problem, we endow u with a prior and update it based on evaluations of the vector field $f(t)$ at different t_i, $i = 1, \dots, N$, where $t_i > t_{i-1}$. If we suppose $u_1 \overset{d}{=} u|(t_1, f(t_1))$ and $u_2 \overset{d}{=} u|(t_1, f(t_1)), (t_2, f(t_2))$ then we do not expect $u_1(t_1)$ is equal in distribution to $u_2(t_1)$, as a result of having obtained more information about the vector field in u_2 which would have an impact on our belief about the distribution $u_2(t_1)$.

However in the present work, it is impossible for u_1 to depend upon $f(t_2)$; that is, our new beliefs about f at t_n cannot have any impact on the distribution of u_m for $m < n$. Thus we appear to have imposed a filtration on the σ-algebra of the probability space which is not inherent to the problem. As a result the posterior distributions cannot be regarded as a full Bayesian update, which I believe this casts some doubt on the "Bayesian" nature of the uncertainty quantification provided in the Skilling (1992) sense, as well as on the information efficiency of the method.

The work is similar to the recently published work of Kersting and Hennig (2016) and Schober et al. (2016), in that the uncertainty is generated by a methodology similar to "filtering" in the data assimilation literature; the full Bayesian posterior would be

[1] Department of Statistics, University of Warwick, Coventry, CV4 7AL, j.cockayne@warwick.ac.uk

given by solution of the correspond "smoothing" problem. I would be interested to see whether this can be incorporated into the present work.

2 Treatment of Partial Differential Equations

The treatment of evolutionary PDEs is also of interest, in light of recent developments of probabilistic meshless methods (PMM) for PDEs (Cockayne et al., 2016). In Section 5.3 I was interested to see the reduction of the Navier Stokes PDE to a large system of ODEs. It is an interesting point for probabilistic numerics, that many problems can be formulated by multiple equivalent numerical schemes; one wonders how the solution obtained by solving this system of ODEs would compare to direct solution of a PDE system, and how consistent the posterior measures generated would be.

Similarly, in Section 5.4 the authors have solved the heat equation by a "forward in time, continuous in space" formulation; if I understand this correctly, we treat the spatial component by a Gaussian process model and discretise the temporal component using the methods of this paper. In light of my comments on the provided uncertainty quantification and considering that this evolutionary system is linear, I wonder how this solution would compare to the fully Bayesian solution provided by the PMM.

References

Cockayne, J., Oates, C. J., Sullivan, T., and Girolami, M. (2016). "Probabilistic Meshless Methods for Partial Differential Equations and Bayesian Inverse Problems." arXiv:1605.07811. 1291

Kersting, H. and Hennig, P. (2016). "Active Uncertainty Calibration in Bayesian ODE Solvers." In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, 309–318. AUAI Press. 1290

Schober, M., Särkkä, S., and Hennig, P. (2016). "A Probabilistic Model for the Numerical Solution of Initial Value Problems." arXiv:1610.05261. 1290

Skilling, J. (1992). Bayesian Solution of Ordinary Differential Equations." In *Maximum Entropy and Bayesian Methods*, 23–37. Dordrecht: Springer Netherlands. 1290

Comment by Michael Schober[1] and Philipp Hennig[2]

Abstract. We welcome the paper by Chkrebtii et al. which provides a thorough analysis of Gaussian process ordinary differential equation (ODE) solvers and their applications in inverse problems. We present some remarks on the interaction between modelling requirements and computational cost.

1 Introduction

An increasing number of manuscripts on probabilistic numerics in general and probabilistic ODE solvers in particular highlight the rising interest and importance in this area of research. The manuscript by Chkrebtii et al. (2016) presents a principled approach to solve a variety of problems related to numerical approximation of differential equations. Their level of detail combined with the open access to all code should be a standard for all researchers working on practical methods.

2 On Sufficient Kernel Conditions

One of the paper's main contribution is a set of sufficient conditions for $\mathcal{O}(h)$ convergence. Among those requirements is that the covariance function has to decay with the step size, $\lambda = \mathcal{O}(h)$. As the authors point out, this has the side-effect that the resulting Gram matrices are sparse, essentially banded matrices; and this structure could be used to achieve low computational cost through approximations.

It is interesting in this context to consider the structure of classic methods for the solution of initial value problems. They are constructed "the other way round", with explicitly linear algorithmic structure, which is then designed to achieve high convergence order. Recall the two main families of numerical methods for initial value problems, Runge–Kutta methods (1) and linear multistep methods (2):

$$x_{n+1} = x_n + h\sum_{i=0}^{k} b_i Y_{n,i}, \qquad Y_{n,j} = f(t_n + hc_j, x_n + h\sum_{i=0}^{j-1} w_{ji}Y_{n,i}), \tag{1}$$

$$\sum_{i=-1}^{q} \alpha_i x_{n-i} = h\sum_{i=-1}^{q} \beta_i f_{n-i}, \qquad f_{n-i} = f(t_{n-i}, x_{n-i}). \tag{2}$$

One can think of these as three sets of ingredients: problem dependent *dynamic* memory contents $[(x_n, Y_{n,j})], (x_n, f_n)]$, a step size h, and step size agnostic *static* method parameters $[(c_j, w_{ji}, b_i), (\alpha_i, \beta_i)]$. Similarly, $\lambda = \mathcal{O}(h)$ leads to almost independent inference problems on each subinterval $[t_n, t_{n+1}]$. Recently, Schober et al. (2016) have proposed a Gauss–Markov process prior

$$\mathrm{d}\begin{pmatrix} X \\ X' \\ X'' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} X \\ X' \\ X'' \end{pmatrix}\mathrm{d}t + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} d\omega, \tag{3}$$

[1] MPI for Intelligent Systems, Tübingen, Germany, mschober@tue.mpg.de
[2] MPI for Intelligent Systems, Tübingen, Germany, ph@tue.mpg.de

which directly encodes the linear computational cost requirement. The authors have also been able to show convergence rate of $\mathcal{O}(h^3)$.

3 Probabilistic ODE Solvers with Reproducible Output

Chkrebtii et al. (2016) achieve accurate uncertainty quantification through sampling. This approach of injecting randomness into the problem to reduce dependence between error and estimate is elegant in its formal simplicity. Its downside is a comparably high computational cost, since Monte Carlo estimates only converge at stochastic rate. In the context of numerical methods, computational efficiency is particularly crucial. Kersting and Hennig (2016) recently proposed an alternative based on Bayesian quadrature, which suggests that calibrated uncertainty can also be achieved in an entirely deterministic fashion. Combined with recent convergence results on Bayesian quadrature (Briol et al., 2015), it might be interesting to see whether theoretical guarantees can be found for uncertainty quantification in this similar setting.

References

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2015). "Probabilistic Integration: A Role for Statisticians in Numerical Analysis?" arXiv:1512.00933 [stat.ML]. 1293

Chkrebtii, O. A., Campbell, D. A., Girolami, M. A., and Calderhead, B. (2016). "Bayesian Solution Uncertainty Quantification for Differential Equations." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1017. 1292, 1293

Kersting, H. P. and Hennig, P. (2016). "Active Uncertainty Calibration in Bayesian ODE Solvers." In Janzing and Ihlers (eds.), *Uncertainty in Artificial Intelligence (UAI)*, volume 32. 1293

Schober, M., Särkkä, S., and Hennig, P. (2016). "A Probabilistic Model for the Numerical Solution of Initial Value Problems." arXiv:1610.05261. 1292