Towards a cross-research platform for hosting Bayesian data-fitting tools

Workshop report

Wednesday 14th July 2021

The event was a one-day virtual workshop, held over Zoom on Wednesday 14th July 2021. The event was co-organized by Dr Payel Das (Surrey Department of Physics), Dr David Lloyd (Surrey Department of Mathematics), and Dr Alex Shestopaloff (Queen Mary University of London and Alan Turing Institute). The event was attended by 46 delegates from industry and a wide range of research disciplines including Astrophysics, Mathematics, Astrodynamics, Building Physics, Nuclear Physics, Computer Science, and Engineering. The workshop consisted of a mixture of pedagogical keynotes, research talks, and a discussion session.

Symposium aims

Scientists and engineers enhance their understanding of the world by fitting models to data. The model parameters are however inherently uncertain due to observational errors in the data, and structural uncertainties in the model. Bayesian sampling methods offer an approach to quantifying the uncertainty in model parameters by inferring the full posterior probability distribution of the model as a function of its parameters. In static systems, the models are usually quick to run for each set of parameters. Real systems such as interacting galaxies, the weather, or flu epidemics often change rapidly with time, requiring more complex models. Applying Bayesian methods to these problems can be computationally prohibitive. The aim of the workshop was to bring together stakeholders in both research and industry to facilitate knowledge exchange in complex data-fitting problems and methods for Bayesian sampling and optimization.

Event themes

Pedagogical and contributed talks
The day was split into two sessions:

1. *Complex data-fitting problems and uncertainty analysis*: this focussed on establishing the diversity of complex data-fitting problems that researchers encounter.
2. *Bayesian sampling and optimization*: this focussed on establishing the landscape of the various Bayesian sampling procedures that are available.

The two sessions were interspersed with three pedagogical talks to help set the scene for the interdisciplinary audience. Our first keynote speaker was Dr Naratip Santitissadeekorn from the University of Surrey, who introduced the concept of quantifying uncertainty, and Bayes' rule. The second speaker giving a pedagogical talk was Dr Alex Shestopaloff from Queen Mary University of London, who introduced the basics of Markov Chain Monte Carlo (MCMC) methods. The final pedagogical talk was given by Dr Josh Speagle from Harvard University who introduced nested sampling, a complementary framework to MCMC that is designed to estimate marginal likelihoods (i.e. Bayesian evidences) and posterior distributions.

Key themes that arose from the pedagogical and contributed talks were:

1. *The varied landscape of complex data-fitting problems*: The complexity in fitting models to data appears to arise due to i) existence of multiple objective functions (e.g. Dr Xilu Wang discussed a general problem where one objective function is much faster to calculate than the other), ii) computationally-demanding objective functions (Dr Tao Chen and Carmen Calama Gonzalez spoke about expensive CFD and building physics models in engineering), and iii) noisy models (Dr Denis Erkal spoke about fitting astrophysical data with noisy N-body simulations).
2. *The use of a Gaussian process for surrogate-assisted data-fitting:* In the case of computationally-prohibitive objective functions, several researchers in engineering used a combination of sensitivity analyses to reduce the dimensionality of the parameter search (Carmen Calama Gonzalez uses a Morris sensitivity analysis to reduce the number of parameters to vary in her building physics model) and Gaussian Processes to replace the model (Dr Xilu Wang, Dr Tao Chen). A Gaussian Process was also used by Dr Alessandro Pastore to better predict observed nucleus masses.
3. *The range of MCMC methods available for Bayesian model fitting*: There are a significant number of MCMC methods available to carry out a Bayesian model-fitting analysis. Dr Alex Shestopaloff introduced Metropolis-Hastings and Gibbs sampling, which uses a single walker to explore the posterior hypervolume spanned by the model parameters. Minas Karamanis introduced emcee, a Python implementation of Goodman & Weare's Affine Invariant MCMC ensemble sampler, which uses an ensemble of walkers to explore the posterior distribution. He also discussed a new highly efficient sampler implemented in Python that combines the ensemble walker approach, with 'slice sampling'. Dr Alex Shestopaloff and Dr Payel Das spoke about Hamiltonian Monte Carlo (HMC), which borrows methods from Hamiltonian Mechanics in Physics to create a particularly efficient sampler for high-dimensional problems.
4. *Nested sampling as a Bayesian alternative to MCMC*: Although most of the talks focussed on the MCMC approach to Bayesian model fitting, Dr Josh Speagle's pedagogical talk discussed a complementary approach called nested sampling. This approach tries to best approximate the Bayesian evidence integral, which involves the integral of the posterior distribution over the parameter hypervolume. In doing so however, it creates samples from the posterior distribution.
5. *Challenges in applying Bayesian sampling methods*: The main challenges in applying Bayesian methods are i) their computing requirement, ii) choosing the most appropriate method, iii) selecting the best tuning parameters. Engineers often use Gaussian Process as a surrogate for expensive models to speed up the model-fitting process. Dr Linghan Li compared various Bayesian methods on the a range of test problems and found that nested sampling gave the best overall performance, HMC was particularly suited to high-dimensional problems like artificial neural networks, and the default settings for emcee were only good for low to intermediate-dimensionality problems. He also suggests optimal choices for tuning parameters.

Discussion session

The event closed with a discussion guided by questions that had been collected from delegates during the weeks preceding the workshop. The questions focussed on three primary issues that arose during the workshop: 1) what are the barriers to adopting Bayesian statistics in data-fitting problems? 2) how do we choose the best Bayesian sampler for the problem? 3) how do we deal with time-consuming and noisy models? During the talks and discussion, it became clear that there are two key barriers to parallel progress in applying Bayesian methods across research disciplines: 1) ba terminology barrier between mathematicians, physicists, engineers, and industry that needs to be addressed, and 2) a lack of research exchange.

Next steps – Outcomes

The speakers contributed either full short papers or abstracts to a conference proceedings, which has been compiled. The next steps are focussed on removing the barrier of parallel progress in applying Bayesian methods across research disciplines:

1. With the new network, we will build a new lexicon that enables mathematicians, engineers, physicists, and industry to communicate.
2. We will set up a Wiki page on Bayesian sampling methods that draws on the themes identified here.
3. We will organize future events to continue research exchange between different communities.

Videos from the event are available to watch on Youtube.

Acknowledgements

We gratefully acknowledge the support of our sponsors: the Institute of Advanced Studies at the University of Surrey. We would like to thank Mirela Dumic and Vicki Blamey from IAS for all their support and advice throughout the process. We would also like to thank Denis Erkal who co-chaired the sessions with Payel Das.