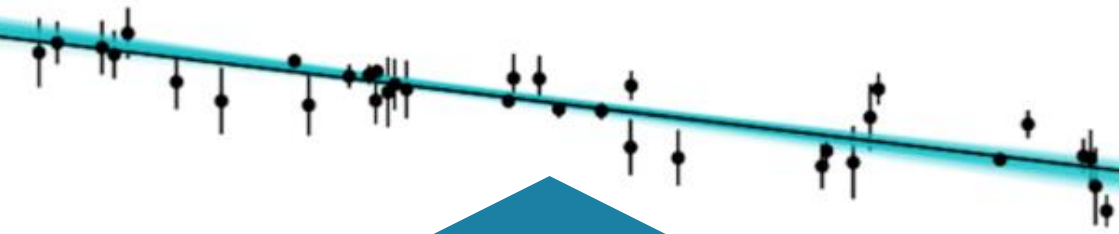
UNIVERSITY OF SURREY



TOWARDS A CROSS-RESEARCH PLATFORM FOR HOSTING BAYESIAN DATA-FITTING TOOLS

ONLINE EVENT
WORKSHOP PROGRAMME

14 JULY 2021

OUR SPONSOR

IAS | Institute of Advanced Studies

The Institute of Advanced Studies (IAS) at the University of Surrey sponsors workshops and Fellowships at the 'cutting edge' of science, engineering, social science and the humanities. Through this scheme the Institute fosters interdisciplinary collaborations and encourages a flow of international scholars to visit, enjoy their stay at Surrey and leave behind excellent ideas and innovations.

ias.surrey.ac.uk

INTRODUCTION

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.

This free virtual one-day workshop will bring together 30+ stakeholders in both research and industry to facilitate knowledge exchange in the quantification of uncertainty in complex data-fitting problems and methods for Bayesian sampling and optimization. There will be a combination of pedagogical talks and research talks in two sessions:

1. Complex data-fitting problems and uncertainty analysis

2. Bayesian sampling and optimization

Three keynote speakers will give the pedagogical talks:

1. Introduction to uncertainty and sequential data assimilation
Dr Naratip Santitissadeekorn from the University of Surrey

2. The essentials of Markov Chain Monte Carlo
Dr Alex Shestopaloff from The Alan Turing Institute and Queen Mary University

3. An Introduction to Nested Sampling
Dr Josh Speagle from Harvard University

The day will end with a discussion session that will cover many of the topics that arise during the workshop. Its aim will be to pave the first steps towards creating a cross-research platform for testing a wide range of Bayesian samplers on various complex data - fitting problems.

The speakers are all invited to submit 2-page papers to the Conference Proceedings.

Organisers:

Dr Dr Payel Das, Department of Physics

Dr Alex Shestopaloff, The Alan Turing Institute

Prof David Lloyd, Department of Mathematics

PROGRAMME

WEDNESDAY 14TH JULY

(UK Time)

SESSION: **Complex data-fitting problems and uncertainty analysis**

09.00 - 09.20	Welcome and opening statement (Payel Das)
09.20 - 10.00	**KEYNOTE:** Introduction to uncertainty and sequential data assimilation (Naratip Santitissadeekorn)
10.00 - 10.20	Transfer learning-based surrogate assisted evolutionary bi-objective optimization for objectives with different evaluation times (Xilu Wang)
10.20 - 10.40	Nuclear physics in a machine learning era (Alessandro Pastore)
10.40 - 11.00	Coffee break and Networking
11.00 - 11.20	Surrogate assisted calibration of computational fluid dynamics models (Tao Chen)
11.20 - 11.40	Bayesian calibration of building energy models for uncertainty analysis through test cells monitoring (C.M. Calama-González)
11.40 - 12.20	KEYNOTE: The essentials of Markov Chain Monte Carlo (Alex Shestopaloff)
12.20 - 13.20	Lunch Break

Continued ▶

(UK Time)

SESSION: **Bayesian sampling and optimization**

13.20 - 13.40	Fitting noisy data with noisy models (Denis Erkal)
13.40 - 14.00	Using MCMC methods for fitting data: from failing to learning (Joaquín García de la Cruz)
14.00 - 14.20	Parallel, black-box and gradient-free inference (Minas Karamanis)
14.20 - 14.40	Coffee break and Networking
14.40 - 15.20	KEYNOTE: An Introduction to Nested Sampling (Josh Speagle)
15.20 - 15.40	Using adaptive Hamiltonian Monte Carlo for training artificial neural networks (Payel Das)
15.40 - 16.00	Applications of MCMC Bayesian sampling methods (Linghan Li)
16.00 - 17.00	Discussion

ABSTRACTS AND BIOS

Follows Programme order

Uncertainty of an influence network driven by count data

Naratip Santitissadeekorn, University of Surrey, UK

It is usually difficult to make an accurate prediction of complex processes with deterministic models. In many situations, it is more useful to consider a probability of certain events. Stochastic models of complex processes incorporate some element of uncertainty to account for lack of knowledge about important physical parameters, random variability of model correct model. This talk will discuss the basic idea of uncertainty from the Bayesian viewpoint and how we might propagate such an uncertainty in light of observed data.

Bio:

Naratip obtained his PhD at Clarkson University in 2008 and then took up a postdoctoral fellowship at the University of New South Wales in 2008-2011 followed by a postdoctoral fellowship at the University of North Carolina in 2011-2014. In 2014, he was appointed to a position of lecturer in Data Assimilation at the University of Surrey and was promoted to a Senior lecturer in September 2020.

Transfer Learning Based Surrogate Assisted Evolutionary Bi-objective Optimization for Objectives with Different Evaluation Times

Xilu Wang, University of Surrey, UK

The Gaussian process has been widely used as a surrogate model for optimizing expensive multi-objective problems due to its ability for providing predictions with uncertainty. Little attention has been paid to more general and realistic optimization scenarios where different objectives are evaluated by different computer simulations or physical experiments with different time complexities (latencies) and only a very limited number of function evaluations is allowed for the slow objective. We propose a transfer learning scheme within a surrogate-assisted evolutionary algorithm framework to augment the training data for the surrogate for the slow objective function by transferring knowledge from the fast one. Specifically, a hybrid domain adaptation method aligning the second-order statistics and marginal distributions across domains is introduced. A Gaussian process (GP) model based co-training method is adopted to predict the value of the slow objective and the uncertainty provided by GPs is used to select the augmented synthetic training data, thereby enhancing the approximation quality of the GP of the slow objective. Our experimental results demonstrate that the proposed algorithm outperforms existing surrogate and non-surrogate-assisted delay-handling methods on a range of bi-objective optimization problems. The approach is also more robust to varying levels of latency and correlation between the objectives.

Bio:

Xilu Wang is a PhD student from the Computer Science Department at the University of Surrey. Her current research interests include evolutionary & transfer optimization and machine learning. Related research topics are surrogate-assisted evolutionary algorithms, gaussian process and Bayesian optimization, and multi-objective optimization

Continued ►

Nuclear Physics in a machine learning era

Alessandro Pastore, University of York, UK

With a few lines of Python code, it is possible to train a neural network over a data-set and then use it to make extrapolations. These techniques are routinely used in particle physics and condensed matter and only recently some pioneering work has been done to apply them to the nuclear physics case. Given the incredible potential of these techniques, is it still necessary to invest time to build complex nuclear models to do the same thing? Why not directly use machine learning algorithms to analyse the data?
In my talk, I will present some applications of machine learning techniques to the case of nuclear masses: by using either neural networks [1] or Gaussian Process Emulators [2] I will show how to use these algorithms to reproduce this particular observable. In particular I will consider the case of a neural network to reproduce nuclear masses without any underlying model and with a model to improve performances. This procedure may be very helpful: in the short term, it will help us detect possible trends in the data and eventually perform reliable predictions in nearby regions of the nuclear chart; in the long term, by interpreting the algorithms, we may learn what is the missing physics in the nuclear model we currently use.

[1] Pastore, A., & Carnini, M. (2021). Extrapolating from neural network models: a cautionary tale. Journal of Physics G: Nuclear and Particle Physics
[2]Shelley, M., & Pastore, A. (2021). A new mass model for nuclear astrophysics: crossing 200 keV accuracy. Universe, 7(5), 131

Bio:

Alessandro has been a lecturer at York University since 2015. He has a PhD in nuclear Physics at Milano in 2008 and undertook postdoc positions at Jyvaskyla University (2009-2010), Lyon University (2010-2012) Bruxelles University (2013-2015), and CEA - Paris (2015-2015). He currently researches the use of Nuclear Energy Density Functional (NEDF) theory to study properties of atomic nuclei.

Surrogate assisted calibration of computational fluid dynamics models

Tao Chen, University of Surrey, UK

Computational fluid dynamics (CFD) is a simulation technique widely used in chemical and process engineering applications. However, computation has become a bottleneck when calibration of CFD models with experimental data (also known as model parameter estimation) is needed. In this research, the kriging meta-modelling approach (also termed Gaussian process) was coupled with expected improvement (EI) to address this challenge. A new EI measure was developed for the sum of squared errors (SSE) which conforms to a generalised chi-square distribution and hence existing normal distribution-based EI measures are not applicable. The new EI measure is to suggest the CFD model parameter to simulate with, hence minimising SSE and improving match between simulation and experiments. The usefulness of the developed method was demonstrated through a case study of a single-phase flow in both a straight-type and a convergent-divergent-type annular jet pump, where a single model parameter was calibrated with experimental data. This talk is based on a journal article we previously published in the AIChE Journal.

Bio:

Tao Chen received the B.Eng. (2000) and M.Eng. (2002) degrees both from Department of Automation, Tsinghua University, China, and the PhD (Chemical Engineering) from Newcastle University in 2006, UK. Prior to joining University of Surrey in 2011 as Lecturer, he worked as Research Associate at Newcastle University (2006-2007), then Assistant Professor at Nanyang Technological University, Singapore (2008-2010). He is a computational and data scientist with 15+ years research experience in mathematical and statistical modelling, and for solving problems in industries like pharmaceutical, healthcare, cosmetic, food, water processing, etc. He has provided R&D consultancy to leading organisations in the cosmetic, consumer goods and steelmaking sectors. He currently leads a research group of 6 doctoral and 2 postdoctoral researchers and have experience in lecturing to small (ca. 10 students) and large (270+) classes, at both undergraduate and postgraduate levels.

Continued ▶

Bayesian calibration of building energy models for uncertainty analysis through test cells monitoring

C.M. Calama-González, University of Seville, Spain

The improvement of energy efficiency of existing buildings is key for meeting 2030 and 2050 energy and CO2 emission targets. Thus, building simulation tools play a crucial role in evaluating the performance of energy retrofit actions, not only at present, but also under future climate scenarios. A Bayesian calibration approach, combined with sensitivity analysis, is applied to reduce the discrepancies between measured and simulated hourly indoor air temperatures. Calibration is applied to a test cell case study developed using the EnergyPlus building simulation software. Several scenarios are evaluated to determine how different variables may impact the calibration process: orientations, activation of mechanical ventilation, different blind aperture levels, etc. Uncertainties associated with model inputs (fixed parameters in the energy model), model discrepancies due to physical limitations of the building energy model (simplifications when compared to the real performance of the building), errors in field observations and noisy measurements were also accounted for. Even though uncalibrated models were within the uncertainty ranges specified by the ASHARE Guidelines, pre-calibration simulation outputs over-predicted measurements up to 3.2 ÂºC. After calibration, the average maximum temperature difference was reduced to 0.68 ÂºC, improving the results by almost 80%. Thus, these techniques are proven to improve the level of agreement between on-site measurements and simulated outputs. Besides, the implementation of this methodology is useful for calibrating and validating indoor hourly temperatures and, consequently, provide adequate results for thermal comfort assessment.

Bio:

C.M. Calama-González is a PhD student in the Architecture Programme of the University of Seville (Spain). She is involved in the research project “Parametric Optimization of Double Skin Facades in the Mediterranean Climate to Improve Energy Efficiency Under Climate Change Scenarios”, funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund. Her research interests are large database analysis, building energy modelling, numeric optimisation, building thermal and energy assessment, climate change and building retrofit.

Continued ▶

The essentials of Markov Chain Monte Carlo

Alex Shestopaloff, Queen Mary University of London, UK

In my talk, I will start by introducing the basics of Markov Chain Monte Carlo (MCMC) methods. I will start with the Metropolis and Gibbs samplers, and the proceed to the Hamiltonian Monte Carlo sampler. A key focus will be to describe the domains of applicability of each of these sampling methods and the difficulties they encounter when applied. I will then describe strategies to overcome some of the difficulties encountered with basic versions of these methods. I will also touch on output diagnostics, and determining when a sampler is working as desired. Finally, I will consider the case of sampling where the likelihood of a model is expensive to compute and how MCMC can be used in this situation. The latter case may be of interest in astronomy applications.

Bio:

Alex Shestopaloff is a Lecturer in Statistics at the School of Mathematical Sciences, Queen Mary University of London and a Fellow of the Alan Turing Institute. He is interested in statistical computing, in particular, Markov Chain Monte Carlo methods for performing Bayesian inference for complex stochastic models. Before joining QMUL he was a Research Fellow at the Alan Turing Institute and completed his PhD in Statistics at the University of Toronto in 2016.

Fitting noisy data with noisy models

Denis Erkal, University of Surrey, UK

In astrophysics, we sometimes need to compare noisy data with noisy models. I will present one such scenario in which this arises when the model has shot noise since it is an N-body simulation consisting of a discrete set of particles. I will briefly explain the astrophysical motivation of the problem. I will then explain how we fit these models to the data using an MCMC as long as the shot noise in the model is sufficiently small and discuss some pitfalls. I will also present a toy model which has the same characteristics to better explore this problem.

Bio:

Denis is a lecturer in the astrophysics research group at the University of Surrey. He was previously a postdoc at the Institute of Astronomy at the University of Cambridge. He received his PhD from the University of Chicago in 2013. His work has mainly focussed on understanding on how the Milky Way was built by the accretion of many smaller systems. In particular, he is interested in tidal streams which form as globular clusters or dwarf galaxies are disrupted by the tides of the Milky Way. These streams roughly follow orbits and are excellent tracers of the potential of our Galaxy.

Using MCMC methods for fitting data: from failing to learning.

Joaquín García de la Cruz, Liverpool John Moores University, UK

MCMC methods are a great tool for fitting data because they explore the whole parameter space and, more importantly, they are able to deliver uncertainties. Uncertainties are crucial because they show how reliable the fit is. When the fit looks reasonable and the uncertainties are not very high, you can claim that you were able to describe your data successfully. However, what happens when your fits do not look so good; either because the values are unrealistic in the context of the system you are studying, or because the uncertainties are too high for the fits to be reliable? Is it the data or the approach to fitting the data the cause of failing? In this talk, I will talk about my personal experience using MCMC methods during the course of my PhD. Firstly, I will show through different examples how the physical interpretation of the fitted values and their uncertainties was crucial in solving problems in my research. Sometimes, failing to fit the data --especially due to high uncertainties-- led me to new insights about system I was struggling to understand, even taking that project into a whole new direction. Secondly, I will talk about situations where I still struggle fit the data, and I will share my insights as to why.

Bio:

Joaquín García de la Cruz is a last year PhD student working on galactic archaeology using simulations of galaxies in their cosmological context. His interests are in the domain of galactic evolution, galaxy stellar populations, and Milky Way science. His funding program (STFC CDT) has a special focus on data science, which has allowed him to receive some training in different Big Data fields, as well as a 6 months long placement in a private company working with their data science team. So, both Astrophysics and data science interest JoaquÃ-n equally, and he would like to learn as much as possible from both fields.

Continued ►

Parallel, black-box and gradient-free inference

Minas Karamanis, University of Edinburgh, UK

Slice Sampling has emerged as a powerful Markov Chain Monte Carlo algorithm that adapts to the characteristics of the target distribution with minimal hand-tuning. However, Slice Sampling's performance is highly sensitive to the user-specified initial length scale hyperparameter and the method generally struggles with poorly scaled or strongly correlated distributions. To this end, we introduce Ensemble Slice Sampling (ESS) and its Python implementation, zeus, a new class of algorithms that bypasses such difficulties by adaptively tuning the initial length scale and utilising an ensemble of parallel walkers in order to efficiently handle strong correlations between parameters. These affine-invariant algorithms are trivial to construct, require no hand-tuning, and can easily be implemented in parallel computing environments. Empirical tests show that Ensemble Slice Sampling can improve efficiency by more than an order of magnitude compared to conventional MCMC methods on a broad range of highly correlated target distributions. In cases of strongly multimodal target distributions, Ensemble Slice Sampling can sample efficiently even in high dimensions. We argue that the parallel, black-box and gradient-free nature of the method renders it ideal for use in scientific fields such as physics, astrophysics and cosmology which are dominated by a wide variety of computationally expensive and non-differentiable models.

Bio:

Minas Karamanis is a 3rd year PhD student at the University of Edinburgh. His main interests lie in the fields of cosmology, astrostatistics, Bayesian inference and machine learning. His work focuses on the development of novel statistical methods that can handle the challenges posed by modern cosmological and astronomical analyses.

An Introduction to Nested Sampling

Joshua Shen Speagle, University of Toronto, Canada

Quantifying model uncertainty and performing model selection within a Bayesian framework is becoming an ever-larger part of scientific analysis both within and outside of astronomy. I will present a brief introduction to Nested Sampling, a complementary framework to Markov Chain Monte Carlo approaches that is designed to estimate marginal likelihoods (i.e. Bayesian evidences) and posterior distributions, outline some of their pros and cons, and briefly discuss more recent extensions such as Dynamic Nested Sampling. I will also briefly highlight `dynesty`, an open-source Python package designed to make it easy for researchers to applying Nested Sampling approaches to various "black box" likelihoods present in their work.

Bio:

Josh is a Banting & Dunlap Postdoctoral Fellow jointly hosted between the Department of Statistical Sciences, David A. Dunlap Department of Astronomy & Astrophysics, and the Dunlap Institute for Astronomy & Astrophysics at the University of Toronto and supervised by Prof. Gwen Eadie. He earned his PhD from Harvard University under the joint supervision of Profs. Doug Finkbeiner and Charlie Conroy along with Profs. Daniel Eisenstein and Alyssa Goodman. He develops methods and analyzes large datasets to understand how galaxies like our own Milky Way form, behave, and evolve. This work lies in the interdisciplinary fields of astrostatistics and data science at the intersections of statistics, astronomy, and computer science. He is particularly interested in developing techniques to jointly model observations from separate (but often complementary) datasets such as imaging, spectroscopy, and time series.

Continued ▶

Using adaptive Hamiltonian Monte Carlo for training artificial neural networks

Payel Das, University of Surrey, UK

A fundamental parameter that is difficult and time consuming to estimate in Milky Way astronomy is the age of a star. Recent work has shown that for a particular class of stars, the age can be estimated empirically from a number of relatively easily measured observables. One way of replicating the relation between these observables and age is by constructing a artificial neural network using a training set. However, a simple feedforward neural network with 9 inputs, 4 outputs, and 10 neurons in a single hidden layer between the input and output layers has 114 parameters that are likely highly correlated. Deriving posterior distributions for such a high number of parameters can be challenging using Bayesian samplers. Here we present the application of an adaptive Hamiltonian Monte Carlo sampler called NUTS (No U-Turn sampler) in the Python PyMC3 package in determining posterior distributions for the parameters of the neural network, and posterior predictive distributions for ages given new observables.

Bio:

Payel is currently a UKRI Future Leaders Fellow in the Astrophysics research group at Surrey, working on the GLEAM project. She has a PhD in Astrophysics from the Max Planck Institute for Extraterrestrial Physics in Germany. She then left Astrophysics for a few years and explored how homes can be optimally designed for energy efficiency and comfort at UCL before returning to Astrophysics as a PDRA at the University of Oxford. Her research interests lie in unveiling the evolutionary histories and dark matter contents of nearby galaxies through forensic studies of the chemical and dynamical properties of the stars within them. She uses an interdisciplinary approach, combining physics-based equilibrium dynamical models and chemical evolution models with statistics-based machine learning and Bayesian tools to efficiently interpret big datasets.

Applications of MCMC Bayesian sampling methods

Linghan Li, University of Surrey, UK

Nowadays, the volume of science or engineering data has increased substantially, and a variety of models have been developed to help to understand the observations. Markov chain Monte Carlo (MCMC) has been established as the standard procedure of inferring these model parameters subject to the available data in a Bayesian framework. Real systems such as interacting galaxies require complex models and these models are computationally prohibitive. The goal of this project is to provide a flexible platform for connecting a range of efficient algorithms for any user-defined circumstances. It will also serve as a testbed for assessing new state-of-the-art model-fitting algorithms. The most commonly used MCMC methods are variants of the Metropolis-Hastings (MH) algorithm. At the beginning of this project and in this article, the standard MH-MCMC algorithm together with affine-invariant ensemble MCMC, which has dominated astronomical analysis over the past few decades, has been tested to reveal the performance of each sampler for the problems with known solutions. The Hamiltonian Monte Carlo algorithm was also tested and it shows in which circumstance that it outperforms the other two..

Bio:

Linghan Li is currently a Post-Doctoral Research Associate (PDRA) working with Payel, Denis, Justin, and Alex (from Alan Turning Insitute) at the University of Surrey. His current project is about parameter optimization for disequilibrium models, with the aim to produce a plug-and-play open-source Python package that enables users to find the most efficient Bayesian sampler for their 'disequilibrium modelling' problem. Linghan is new to the Bayesian Data-fitting research field; he has an academic background with a Ph.D. in ocean wave energy extraction from the University of Southampton, UK, and a Bachelor's degree in ship science which he studied at Harbin Engineering University, China.

Continued ►

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

University of Surrey
Guildford, GU2 7XH, UK

surrey.ac.uk