

MASTERCLASS IN BAYESIAN STATISTICS

CIRM, 22 - 26 OCTOBER, 2018

COURSES

Nicolas Chopin

An introduction to Sequential Monte Carlo

Abstract: This course will give a gentle introduction to SMC (Sequential Monte Carlo algorithms):

- motivation: state-space (hidden Markov) models, sequential analysis of such models; non-sequential problems that may be tackled using SMC.
 - Formalism: Markov kernels, Feynman-Kac distributions.
 - Monte Carlo tricks: importance sampling and resampling
 - standard particle filters: bootstrap, guided, auxiliary
 - maximum likelihood estimation of state-space models
 - Bayesian estimation of these models: PMCMC, SMC².
-

Kerrie Mengersen

Bayesian modelling

Abstract: This tutorial will be a beginner's introduction to Bayesian statistical modelling and analysis. Simple models and computational tools will be described, followed by a discussion about implementing these approaches in practice. A range of case studies will be presented and possible solutions proposed, followed by an open discussion about other ways that these problems could be tackled.

Christian P. Robert

Bayesian computational methods

Abstract: This is a short introduction to the many directions of current research in Bayesian computational statistics, from accelerating MCMC algorithms, to using partly deterministic Markov processes like the bouncy particle and the zigzag samplers, to approximating the target or the proposal distributions in such methods. The main illustration focuses on the evaluation of normalising constants and ratios of normalising constants.

Håvard Rue***Bayesian computing with INLA***

Abstract: This talk focuses on the estimation of the distribution of unobserved nodes in large random graphs from the observation of very few edges. These graphs naturally model tournaments involving a large number of players (the nodes) where the ability to win of each player is unknown. The players are only partially observed through discrete valued scores (edges) describing the results of contests between players. In this very sparse setting, we present the first nonasymptotic risk bounds for maximum likelihood estimators (MLE) of the unknown distribution of the nodes. The proof relies on the construction of a graphical model encoding conditional dependencies that is extremely efficient to study n -regular graphs obtained using a round-robin scheduling. This graphical model allows to prove geometric loss of memory properties and deduce the asymptotic behavior of the likelihood function. Following a classical construction in learning theory, the asymptotic likelihood is used to define a measure of performance for the MLE. Risk bounds for the MLE are finally obtained by subgaussian deviation results derived from concentration inequalities for Markov chains applied to our graphical model.

Aki Vehtari***Model assessment, selection and averaging***

Abstract: The tutorial covers cross-validation, and projection predictive approaches for model assessment, selection and inference after model selection and Pseudo-BMA and Bayesian stacking for model averaging. The talk is accompanied with R notebooks using `rstanarm`, `bayesplot`, `loo`, and `projpred` packages.

SPEAKERS

Simon Barthelmé

Correcting variational approximations

Abstract: Variational methods can be used to obtain simple approximations to complicated distributions, and in a Bayesian setting this enables (computationally) cheap inference and prediction. Unfortunately, simple approximations can sometimes be too simple and the results may be inaccurate. Instead of giving up on variational inference entirely, an alternative is to find ways of correcting a variational approximation. In this talk I'll survey some correction methods, from importance sampling to perturbation series.

Marie-Pierre Etienne

Sequential Monte Carlo smoother for partially observed stochastic differential equation

Abstract: The recent advances in biologging provide small and cheap devices to follow the movement of wild animals and arise new ecological questions regarding the link between animals and their environment. Most statistical models for analysing such data rely on State Space Modelling and use discrete time Markov Models. However those approaches don't provide any solution to estimation a functional depending on the whole trajectory, even the unobserved part. Considering time continuous model such as stochastic differential equations (SDE) circumvents the issue of irregular time step but the inference of such models might be difficult. Euler Maruyama approximation is the most frequent proposed solution but we highlight that this method might not be quiet inaccurate with typical frequency acquisition of movement ecology. Furthermore, depending on the technology, the relocation error might not be neglected and the relocations have to be considered as noisy observations of a hidden continuous process. We extent the PaRIS algorithm and propose a sequential Mote Carlo algorithm to approximate smoothed additive functionals of diffusion processes observed with noise.

Chris Holmes

Bayesian learning at scale with approximate models

Abstract: Bayesian learning from data is predicated on the likelihood function, $f(x|\theta)$, being true, in that for some parameter setting $\theta_0 \in \Theta$ we have $x \sim F_{\theta_0}(x)$. In reality all models are false. If the data is simple and small, and the models are sufficiently rich, then the consequences of model misspecification may not be severe. Increasingly however data is being captured at scale, both in terms of the number of observations as well as the diversity of data modalities. If Bayesian inference is to remain at the forefront of data-science then we will need to develop the theory and computational methods that accommodate and respect the approximate nature of scalable models.

In this talk I will discuss some our research in this field motivated by large-scale health applications involving high-dimensional multivariate measurements on 100,000s of subjects.

Adam M. Johansen

Asymptotic Genealogies of Sequential Monte Carlo Algorithms

Abstract: We consider a class of interacting particle system which covers many of the sequential Monte Carlo (SMC) methods, which are widely-used in applied statistics and cognate disciplines across a range of domains.

We consider the genealogical tree embedded into such particle systems by the resampling operation, and identify conditions, as well as an appropriate time-scaling, under which it converges to the Kingman coalescent in the infinite system size limit in the sense of finite-dimensional distributions. This makes the plethora of distributional results known for the Kingman coalescent available for analysis of SMC algorithms, which we illustrate by characterising the limiting mean and variance of the tree height, as well as of the total branch length of the tree. It also greatly improves the tractability of genealogies of SMC methods, which are known to be closely connected to the performance of these algorithms. The conditions which we require to prove convergence are strong, but we demonstrate by simulation that they do not appear to be necessary. This is joint work with Jere Koskela, Paul Jenkins and Dario Spano.

Sylvain Le Corff

Nonparametric maximum likelihood estimation for large random graphs with latent data

Abstract: This talk focuses on the estimation of the distribution of unobserved nodes in large random graphs from the observation of very few edges. These graphs naturally model tournaments involving a large number of players (the nodes) where the ability to win of each player is unknown. The players are only partially observed through discrete valued scores (edges) describing the results of contests between players. In this very sparse setting, we present the first nonasymptotic risk bounds for maximum likelihood estimators (MLE) of the unknown distribution of the nodes. The proof relies on the construction of a graphical model encoding conditional dependencies that is extremely efficient to study n -regular graphs obtained using a round-robin scheduling. This graphical model allows to prove geometric loss of memory properties and deduce the asymptotic behavior of the likelihood function. Following a classical construction in learning theory, the asymptotic likelihood is used to define a measure of performance for the MLE. Risk bounds for the MLE are finally obtained by subgaussian deviation results derived from concentration inequalities for Markov chains applied to our graphical model.

Bruno Nicenboim

Cognitive models of memory processes in sentence comprehension: A case study using Bayesian hierarchical modeling

Abstract: Understanding a sentence relies on storing phrases in our memory as we read or hear them, and retrieving only the right phrase at the right moment to create dependencies between phrases. In the sentence shown below, for example, when we reach the verb “called out” a linguistic dependency must be completed between this verb and “the Mouse” in order to understand the main assertion of the sentence.

“At last the Mouse, who seemed to be a person of authority among them, called out[...]”
— Lewis Carroll, Alice’s Adventures in Wonderland

Bayesian modeling allows us to investigate the underlying cognitive process that leads to a successful dependency completion by implementing different theories as computational models, fit the models to data, and compare them. I will use **NicenboimVasishth2017Models** as a case study to exemplify this process.

NicenboimVasishth2017Models disentangle the predictions of two influential models of sentence comprehension: the activation-based model (**LewisVasishth2005**) and the direct-access model (**McElree2000**). These two models assume that constraints on working memory influence the speed and accuracy of dependency resolution processes. Although the models make virtually identical predictions when comparing only mean reading times (or comprehension accuracy), the underlying process assumed is very different in these.

The two models are implemented as hierarchical Bayesian models in the probabilistic programming language Stan (**Stan2017**) and fit to reading times and accuracy simultaneously. Posterior predictive checks and model comparison using k-fold cross validation (**VehtariEtAl2017**) shows that the direct-access model provides a better fit to the data than the activation-based model. The comparison also reveals why the activation-based model fails and how it could be improved. More generally, this work leverages the capabilities of the Bayesian framework for flexibly developing computational models of competing theories, and demonstrates how these models can be compared to provide novel insights.

Sebastian Reich

Interacting particle approximations for state and parameter estimation

Abstract: We will summarise a unifying mathematical framework and algorithmic approaches for state and parameter estimation of partially observed diffusion processes, which arise from a combined McKean-Vlasov and optimal transportation perspective on the associated filtering and smoothing problem. The popular ensemble Kalman filter as well as more recent developments such as the feedback particle filter or the ensemble transform particle filter including their smoother extensions can be seen as particular algorithms arising from such a framework.

Adeline Samson

Stochastic models in biology and statistical challenges

Abstract: Stochastic models in biology are of various types: point processes, discrete time models, continuous time models, models with latent variables. The estimation of their parameters faces different challenges: multidimensional process, hidden coordinates, partial observations. In this talk, I will present some statistical methods that have been proposed in this context.

Eric-Jan Wagenmakers

Bayesian Inference with JASP

Abstract: JASP is an open-source statistical software package with an attractive graphical user-interface (jasp-stats.org). For many statistical practitioners, JASP helps unlock methodologies that would have otherwise remained inaccessible. I will illustrate the basic functionality of JASP, how it can be used for teaching and research, and how it can be extended using modules. One of the primary goals of JASP is to remove the technical obstacles that have so far prevented empirical researchers from taking full advantage of the benefits that the Bayesian paradigm affords. Consequently, JASP prominently features state-of-the-art Bayesian procedures for popular analysis scenarios involving t-tests, ANOVA, and regression.

Giacomo Zanella

Scalable Importance Tempering and Bayesian Variable Selection

Abstract: We propose a Monte Carlo algorithm to sample from high-dimensional probability distributions that combines Markov chain Monte Carlo and importance sampling. We provide a careful theoretical analysis, including guarantees on robustness to high-dimensionality, explicit comparison with standard schemes and illustrations of the potential improvements in efficiency. Simple and concrete intuition is provided for when the novel scheme is expected to outperform standard schemes. When applied to Bayesian variable selection problems, the novel algorithm is orders of magnitude more efficient than available alternative sampling schemes and allows to perform fast and reliable fully Bayesian inferences with tens of thousand regressors.

PRACTICAL TUTORIALS

Julien Stoehr and Guillaume Kon Kam King

Good practice in R: code, packages and R Markdown

Abstract: As nicely written by Hadley Wickham R guru, "Good coding style is like correct punctuation: you can manage without it, but it sure makes things easier to read." In this tutorial, we will review some general good practice to write 1) R code, 2) R packages and 3) R Markdown and knitr documents. In a first part, we will present some ideas of the tidyverse style guide (following <http://style.tidyverse.org/>) including the styler package (interactively restyle selected text) and the lintr package (automated style checks). The second part will be devoted to writing R packages, which is at the basis of reproducibility. We will learn how to use the Roxygen package to (quite automatically) turn personal code into reusable code (following <http://r-pkgs.had.co.nz/>). The last part of the talk explains how to put together text and chunks of code to produce nice formatted output. This may be useful to produce teaching documents, software paper, vignettes for R package just to name a few. We will focus on R Markdown and knitr through simple examples. Make sure you have the latest versions of RStudio, styler, lintr, Roxygen and knitr packages installed.

Aki Vehtari

The Stan software

Abstract: The tutorial covers cross-validation, and projection predictive approaches for model assessment, selection and inference after model selection and Pseudo-BMA and Bayesian stacking for model averaging. The talk is accompanied with R notebooks using rstanarm, bayesplot, loo, and projpred packages.