CONFERENCE

Bayesians Statistics in the Big Data Era (event 1912)

Dates: 26-30 November 2018

Place: CIRM (Marseille Luminy, France)

DESCRIPTION

Bayesian methods are now firmly established in the fields of Statistics and Machine Learning and are being increasingly applied to "Big Data". However, there are still gaps in knowledge about the theory, methodology, computation and application of Bayesian methods in this context.
This conference will bring together an international and interdisciplinary group of researchers and practitioners to share insights, research, challenges and opportunities in developing and using Bayesian statistics in the Big Data era. The anticipated outcomes include: knowledge transfer, new collaborative networks, new research directions and new statistical tools to address challenging problems in the real world.

Themes

- Emerging Theory & Methods: Bayesian methodology for big data modelling and analysis
- Enabling Computation: Bayesian computation for big data
- New Insights: Applications of Bayesian analysis using big data
- Program

The program will include oral presentations and substantial time for discussion after each presentation, as well as poster presentations with specific sessions for presentation and discussion.
The program will also include dedicated time for research discussion, collaboration and networking. Sessions on topics such as career pathways and mentoring will be scheduled for postgraduates and early career researchers.

SCIENTIFIC COMMITTEE

- Pierre Pudlo (Aix-Marseille Université)
- Christian P. Robert (Université Paris-Dauphine)
- Kerrie Mengersen (Queensland University of Technology)

ORGANIZING COMMITTEE

- Jean-Marc Freyermuth (Aix-Marseille Université)
- Jean-Michel Marin (Université de Montpellier)
- Kerrie Mengersen (QUT Brisbane)
- Denys Pommeret (Aix-Marseille Université)
- Pierre Pudlo (Aix-Marseille Université)

KEYNOTES

- Sudipto Banerjee (UCLA)
- Amy Herring (Duke)
- Sylvia Frühwirth-Schnatter (WU Vienna
- Peter Mueller (UT Austin):

SPEAKERS

- Pierre Alquier (ENSAE ParisTech)
- Louis Aslett (Durham University)
- Tamara Broderick (MIT)
- Noel Cressie (University of Wollongong)
- Marco Cuturi (ENSAE, Paris)
- David B. Dunson (Duke University)
- Gregor Kastner (WU Vienna)
- Gary Koop (University of Strathclyde, Glasgow)
- Antonio Lijoi (Bocconi University)
- Jean-Michel Marin (Université de Montpellier)
- Antonietta Mira (Università della Svizzera italiana and University of Insubria)
- Igor Prünster (Bocconi University)
- Stéphane Robin (AgroParisTech)
- Heejung Shim (University of Melbourne)
- Rebecca Steorts (Duke University)
- Darren Wilkinson (Newcastle University)

ORAL PRESENTATIONS

- Atanu Bhattacharjee (Tata Memorial Centre)
- Marta Crispino (INRIA Grenoble)
- Christel Faes (Hasselt University Belgium)
- Logan Graham (University of Oxford)
- Zitong Li (University of Melbourne)
- Benoit Liquet (Université de Pau et des Pays de L'Adour)
- Jia Liu (University of Helsinki)
- Reza Mohammadi (University of Amsterdam)
- Ahihiko Nishimura (University of California - Los Angeles)
- Monica Patriche (University of Bucharest)
- Gajendra Vishwakarma (Indian Institute of Technology Dhanbad)

Bayesian Statistics in the Big Data Era: 26-30 November 2018

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00 AM	Workshop (event 1911) Presentations, Conclusion	Peter Müller	Sylvia Fruhwirth-Schnatter	Sudipto Banerjee	Stephane Robin
9:30 AM					Benoit Liquet
10:00 AM		Antonio Lijoi	Gary Koop	Jia Liu	Antonietta Mira
10:30 AM		Break	Break	Break	Break
11:00 AM		Igor Pruenster	Monica Patriche	David Dunson	Gajendra Vishwakarma
11:30 AM		Akihiko Nishimura	Christian Robert	Atanu Bhattacharjee	Jean-Michel Marin
12:00 PM	Conference Registration	Pierre Pudlo	Gregor Kastner	Noel Cressie	Future-Think Presentations
12:30 PM	12:30 - 2.00 Lunch	12:30 - 2.00 Lunch	12:30 - 2.00 Lunch	12:30 - 2.00 Lunch	Conference Conclusion and Lunch
2:00 PM	Amy Herring	Break	Free afternoon	Break	
3:00 PM	Christel Faes	Heejung Shim		Marco Cuturi	
3:30 PM	Louis Aslett	Zitong Li		Reza Mohammadi	
4:00 PM	Coffee break	Coffee break		Coffee break	
4:30 PM	Marta Crispino	Pierre Alquier		Tamara Broderick	
5:00 PM	Rebecca Steorts (V)	Logan Graham		Darren Wilkinson	
5:30 PM	"Future-Think" Themed Discussion	"Future-Think" Themed Discussion		"Future-Think" Themed Discussion	
6:30 PM	Break	Break		Break	
7:30 PM	7:30 Dinner	7:30 Dinner	7:30 Dinner	7:30 Dinner	
8:30 PM - 10:00 PM	Welcome evening	Posters			

https://www.chairejeanmorlet.com/2018-2-mengersen-pudlo-1912.html

SPONSORS

Conference Program: Monday 26th November 2018, 14:00 – 18:30

Presenter	Title
Amy Herring Duke University	*Centered Partition Processes: Lumping versus Splitting in Sparse Health Data*
Christel Faes Hasselt University Belgium	*Accounting for residential history in disease mapping*
Louis Aslett Durham University	*Privacy and Security in Pooled Bayesian Inference*
Marta Crispino INRIA Grenoble	*Bayesian preference learning*
Rebecca Steorts, Duke University	*Video*

Presenter: **Amy Herring**
Centered Partition Processes: Lumping versus Splitting in Sparse Health Data
E-mail: amy.herring@duke.edu
Authors: Sally Paganin, University of Padova, Amy H. Herring, Duke University, Andrew F. Olshan,
Affiliation: The University of North Carolina at Chapel Hill, and David B. Dunson, Duke University

In many health studies, interest often lies in assessing health effects on a large set of outcomes or specific outcome subtypes, which may be sparsely observed, even in big data settings. For example, while the overall prevalence of birth defects is not low, the vast heterogeneity in types of congenital malformations leads to challenges in estimation for sparse groups. However, lumping small groups together to facilitate estimation is often controversial and may have limited scientific support.

There is a very rich literature proposing Bayesian approaches for clustering starting with a prior probability distribution on partitions. Most approaches assume exchangeability, leading to simple representations in terms of Exchangeable Partition Probability Functions (EPPF). Gibbs-type priors encompass a broad class of such cases, including Dirichlet and Pitman-Yor processes. Even though there have been some proposals to relax the exchangeability assumption, allowing covariate-dependence and partial exchangeability, limited consideration has been given on how to include concrete prior knowledge on the partition. We wish to cluster birth defects into groups to facilitate estimation, and we have prior knowledge of an initial clustering provided by experts. As a general approach for including such prior knowledge,
we propose a Centered Partition (CP) process that modifies the EPPF to favor partitions close to an initial one. Some properties of the CP prior are described, a general algorithm for posterior computation is developed, and we illustrate the methodology through simulation examples and an application to the motivating epidemiology study of birth defects.

Presenter: **Christel Faes**
Accounting for residential history in disease mapping
E-mail: christel.faes@uhasselt.be
Authors: Christel Faes
Affiliation: Hasselt University Belgium

Mesothelioma is a rare cancer caused by exposure to asbestos. The period from exposure to development of the disease varies between 20 to 40 years. Individuals living in nearby areas of asbestos factories have higher risk to develop mesothelioma cancer. Standard disease mapping methods make use of counts at locations of diagnosis or death.
However, for diseases with long latency periods, the history of residential locations could be of greater interest. We investigate a method that takes into account the residential history of patients. A spatial multiple membership model is proposed, using pancreatic cancer as a control disease. Results show the impact of the residential mobility on the geographical risk estimation in Belgium, as well as the importance of acknowledging for the latency period of a disease.

Presenter: **Louis Aslett**
Privacy and Security in Pooled Bayesian Inference
E-mail: louis.aslett@durham.ac.uk
Authors: Louis Aslett
Affiliation: Durham University

The growth of data sets in Bayesian analyses brings with it concerns surrounding privacy and security, both for the raw data during model fitting and for the potential leaking of sensitive information via the fitted model. This talk will present recent developments combining different privacy and security methodologies to provide protection in the setting of multiple parties wanting to pool their data to produce a Bayesian model fitted on the combined data, considering security and privacy in both the inference and post inference phases.

Presenter: **Marta Crispino**
Bayesian preference learning
E-mail: marta.crispino@]inria.fr
Authors: Marta Crispino, Mistis team,
Affiliation: INRIA Grenoble, Rhône-Alpes, https://sites.google.com/site/crispinostat/home

Preference data occur when assessors express comparative opinions about a set of items, by rating, ranking, pair comparing, liking or clicking. The purpose of preference learning is to (i) infer on the shared consensus preference of a group of users; or (ii) estimate for each user her individual ranking of the items, when the user indicates only incomplete preferences. We develop a new computationally tractable method for Bayesian inference in the Mallows model for preference learning that performs inference on the consensus and individual rankings, also when based on partial rankings, such as top-k items and pairwise comparisons, even in the particular case of non-transitive patterns in the data. We develop approximate stochastic algorithms that allow a fully probabilistic analysis, leading to coherent quantifications of uncertainties. Our method makes probabilistic predictions on the class membership of assessors based on their ranking of just some items, and predicts missing individual preferences, as needed in many applications.

Presenter: **Rebecca Steorts**
Via Video
E-mail: beka@stat.duke.edu
Authors: Rebecca Steorts
Affiliation: Duke University

Conference Program: Tuesday 27th November 2018, 9:00 – 18:30

Presenter	Title
Peter Müller UT Austin	*Scalable Bayesian Nonparametric Clustering and Classification*
Antonio Lijoi Bocconi University	*Nonparametric priors for covariate-dependent data*
Igor Prünster Bocconi University	*Hierarchies of discrete random probabilities*
Akihiko Nishimura, University of California	*Computational advances in "large n and large p" sparse Bayesian regression for binary and survival outcomes*
Pierre Pudlo Aix-Marseille Université	*Approximation Bayesian Computation and Model Choice*
Heejung Shim, University of Melbourne	*Bayesian multi-scale Poisson models for analyses of high-throughput sequencing data in genomics*
Zitong Li, University of Melbourne	*Bayesian non-parametric regression for analyzing time course quantitative genetic data*
Pierre Alquier ENSAE, Paris Tech	*Informed Sub-Sampling MCMC: Approximate Bayesian Inference for Large Datasets*
Logan Graham, University of Oxford	*Causality in Modern Machine Learning: A Review*

Presenter: **Peter Müller**
Scalable Bayesian Nonparametric Clustering and Classification
E-mail: pmueller@math.utexas.edu
Authors: Peter Müller
Affiliation: UT Austin

We develop a scalable multi-step Monte Carlo algorithm for inference under a large class of nonparametric Bayesian models for clustering and classification. We discuss two strategies. One is based on a consensus Monte Carlo approach. It splits the data into shards and then combines subset posteriors to recover joint inference. We propose several alternative algorithms that are suitable to address the computational challenge of full posterior inference for a random partition in a large data set, and a variation to process big data. A second strategy exploits predictive recursion to build up posterior inference for the complete data. The methods are applicable for a wide range of Bayesian nonparametric mixture models. We apply the proposed schemes to inference for a large data base of gene-gene interactions extracted from the online search tool ``Zodiac'', and to inference for electronic health records (EHR) using a product partition model with regression on covariates. Under simulated and benchmark data sets the proposed methods compare favorably with other clustering algorithms, including k-means, DP-means, DBSCAN, SUGS, streaming variational Bayes (SVB) and an EM algorithm.

Presenter: **Antonio Lijoi**
Nonparametric priors for covariate-dependent data
E-mail: antonio.lijoi@unibocconi.it
Authors: Antonio Lijoi
Affiliation: Bocconi University

In the last few years there has been a growing interest in Bayesian nonparametric priors suited for modeling data, or latent variables, that display forms of dependence more general than exchangeability. The talk will discuss a model based on discrete random measures that define a class of nonparametric priors for inference with covariate-dependent data.
Predictive distributions and posterior characterizations will be presented, along with algorithms that allow to determine approximate Bayesian inferences of interest. The discussion will be completed by illustrations with simulated and real data.

Presenter: **Igor Prünster**
Hierarchies of discrete random probabilities
E-mail: igor@unibocconi.it
Authors: Igor Prünster
Affiliation: Bocconi University

Hierarchies of discrete probability measures define nonparametric priors that have become popular in several applied areas. This is due to them naturally representing multiple heterogeneous populations and allowing for "sharing of information" across multiple samples.
A complete distributional picture of hierarchical normalized random measures (encompassing the hierarchical Dirichlet and Pitman–Yor processes) is given. Characterizations of the (partially exchangeable) partition structure, including the distribution and the asymptotics of the number of clusters, and of the posterior structure are provided. Moreover, based on a suitable finite-dimensional approximation of such priors, a conditional Gibbs sampling algorithm is devised, which allows for the analysis of large-scale datasets.

Presenter: **Akihiko Nishimura**
Computational advances in "large n and large p" sparse Bayesian regression for binary and survival outcomes
E-mail: akihiko4@g.ucla.edu
Authors: Akihiko Nishimura
Affiliation: University of California

In a modern observational study based on healthcare databases, the number of observations is typically in the order of 10^5 ~ 10^6 and that of the predictors in the order of 10^4 ~ 10^5. Despite the large sample size, the data rarely provide enough information to reliably estimate such a large number of parameters. Sparse regression provides a potential solution to this problem. There is a rich literature on desirable theoretical properties of the Bayesian approach based on shrinkage prior. On the other hand, the development of scalable methods for the required posterior computation has largely been limited to the p >> n case with continuous outcomes. In this talk, I will discuss the recently developed computational techniques in the "large n and large p" setting for binary and survival time outcomes. We apply our algorithm to a large-scale observational study with n = 72,489 and p = 22,175, designed to assess the relative risk of intracranial hemorrhage from two alternative blood anti-coagulants. Our algorithm demonstrates an order of magnitude speed-up in the posterior computation.

Presenter: **Pierre Pudlo**
Approximation Bayesian Computation and Model Choice
E-mail: pierre.PUDLO@univ-amu.fr
Authors: Pierre Pudlo
Affiliation: Aix-Marseille Université

Presenter: **Heejung Shim**

Bayesian multi-scale Poisson models for analyses of high-throughput sequencing data in genomics

E-mail: heejung.shim@unimelb.edu.au
Authors: Heejung Shim
Affiliation: University of Melbourne

In genomics, there are many applications which involve estimating the differences in molecular phenotypes between multiple groups of samples. The development of cheap high-throughput sequencing technologies with experiment protocols has increased the use of high-throughput sequencing data as measurements of molecular phenotypes (e.g., RNA-seq for gene expression, ChIP-seq for transcription factor binding, ATAC-seq for chromatin openness, CASE-seq for transcription start site usage). The technologies measure the data of diverse types at unprecedented scales and provide high-resolution measurements on how molecular phenotypes vary along the whole genome in each sample. However, typical analyses fail to exploit the full potential of these high-resolution measurements, instead aggregating the data at coarser resolutions, such as genes, or windows of fixed length.

We will present methods that fully exploit the high-resolution data as well as model the count nature of the sequence data directly. Specifically, we assume that the data for each sample follow inhomogeneous Poisson processes with spatially structured underlying intensity function. Then, we adapt Bayesian multi-scale models for inhomogeneous Poisson processes to develop methods for estimating differences in the underlying intensity function between multiple groups of samples. Applying new methods to ATAC-seq data to estimate differences in chromatin openness, we will demonstrate the potential of the proposed methods over a simple window-based approach. We will also briefly discuss applications of the proposed approaches for analyses of different types of high-throughput sequencing data, such as RNA-seq, and Hi-C data.

Presenter: **Zitong Li**

Bayesian non-parametric regression for analyzing time course quantitative genetic data

E-mail: zitong.li1@unimelb.edu.au
Authors: Zitong Li[1]*, Jarno Vanhadalo[2], Mikko Sillanpää[3]
Affiliation:
[1]Melbourne Integrative Genomics and School of Mathematics and Statistics, University of Melbourne, Australia
[2]Department of Mathematics and Statistics and Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland
[3]Department of Mathematical Sciences and Biocenter Oulu, University of Oulu, Oulu, Finland

In life science, studying genetic regulation of dynamic biological process such as development and growth using time course data is of primary interest to many applied scientists. Recent advances in high throughput genotyping and phenotyping techniques can cost effectively generate large scale of genotype and repeated phenotype data sets. Identifying essential biomarkers explaining quantitative trait variation from such high dimensional data is a very challenge statistical inference problem. Therefore, there is an urgent need to develop efficient statistical and computational methods for analyzing time course genetic data.

We propose a Bayesian Gaussian process (GP) regression approach for quantitative trait locus (QTL) mapping of longitudinal traits. The GP priors are used to model the continuously varying coefficients which describe how the effects of molecular markers on the quantitative trait are changing over time. A flexible Mátern covariance function was specified in GPs to induce the smoothness in temporally varying coefficients. The hyper-parameters in GPs determining the optimal degree of smoothness were estimated by an efficient empirical Bayes algorithm. Furthermore, stepwise procedure is developed to search through the model space in terms of genetic variants, and a Bayesian model posterior probability was used as a stopping rule to focus on only a small set of putative QTL. We also discuss an interesting connection between GP and some existing semi-parametric approaches such as penalized B splines and wavelets. On a simulated and three real data sets, our GP approach demonstrates great flexibility on modelling different type of phenotypic trajectories with low computational demand, and high statistical power to identify QTL. Software is available as a MATLAB package 'GPQTLmapping', and they can be downloaded from GitHub.

Presenter: **Pierre Alquier**
Informed Sub-Sampling MCMC: Approximate Bayesian Inference for Large Datasets
E-mail: pierre.alquier@ensae.fr
Authors: Pierre Alquier, Florian Maire and Nial Friel
Affiliation: ENSAE, Paris Tech

In this talk I will introduce a framework for speeding up Bayesian inference conducted in presence of large datasets. We design a Markov chain whose transition kernel uses an (unknown) fraction of (fixed size) of the available data that is randomly refreshed throughout the algorithm. Inspired by the Approximate Bayesian Computation (ABC) literature, the subsampling process is guided by the fidelity to the observed data, as measured by summary statistics. The resulting algorithm, Informed Sub-Sampling MCMC (ISS-MCMC), is a generic and flexible approach which, contrary to existing scalable methodologies, preserves the simplicity of the Metropolis-Hastings algorithm. Even though exactness is lost, i.e. the chain distribution approximates the posterior, we study and quantify theoretically this bias and show on a diverse set of examples that it yields excellent performances when the computational budget is limited. If available and cheap to compute, we show that setting the summary statistics as the maximum likelihood estimator is supported by theoretical arguments.

Presenter: **Logan Graham**
Causality in Modern Machine Learning: A Review
E-mail: logan@robots.ox.ac.uk
Authors: Logan Graham
Affiliation: University of Oxford

While causality has attracted considerable research recently by the machine learning community, little is understood about the state of integrating causality in modern machine learning. Most work has focused on using machine learning to improve causal inference. However, there is considerable opportunity for the inverse: we review the how tools from causal inference have been shown to improve the learning, efficiency, and generalization of machine learning approaches to machine learning problems. We discuss the sets of tools and ideas that researchers can incorporate in their work, and speculate on high-impact areas of future research to accelerate the integration of causality in machine learning.

Conference Program: Wednesday 28th November 2018, 09:00 – 12:30

Presenter	Title
Sylvia Fruhwirth-Schnatter WU, Vienna	*Data Mining through Markov Chain Mixtures with Applications in Labor Economics and Marketing*
Gary Koop University of Strathclyde, Glasgow	*Composite Likelihood Methods for Large Bayesian VARs with Stochastic Volatility*
Monica Patriche University of Bucharest	*Equilibrium existence for Bayesian generalized games in choice form and applications*
Christian Robert, Université Paris-Dauphine	*Inference in generative models using the Wasserstein distance*
Gregor Kastner WU, Vienna	*Bayesian Inference in Many Dimensions: Examples from Macroeconomics and Finance*

Presenter: **Sylvia Fruhwirth-Schnatter**

Data Mining through Markov chain mixtures with applications in labor economics and marketing

E-mail: sylvia.fruehwirth-schnatter@wu.ac.at
Authors: Sylvia Fruhwirth-Schnatter
Affiliation: WU, Vienna

Data mining methods based on finite mixture models are quite common in many areas of applied science, such as marketing, to segment data and to identify subgroups with specific features. Recent work shows that these methods are also useful in micro econometrics to analyze the behavior of workers in labor markets. Since these data are typically available as time series with discrete states, clustering kernels based on Markov chains with group-specific transition matrices are applied to capture both persistence in the individual time series as well as cross-sectional unobserved heterogeneity. Markov chains clustering has been applied to data from the Austrian labor market, (a) to understanding the effect of labor market entry conditions on long-run career developments for male workers (Frühwirth-Schnatter et al., 2012), (b) to study mothers' long-run career patterns after first birth (Frühwirth-Schnatter et al., 2016), and (c) to study the effects of a plant closure on future career developments for male worker (Frühwirth-Schnatter et al., 2018). To capture non-stationary effects for the later study, time-inhomogeneous Markov chains based on time-varying group specific transition matrices are introduced as clustering kernels. For all applications, a mixture-of-experts formulation helps to understand which workers are likely to belong to a particular group. Finally, it will be shown that Markov chain clustering is also useful in a business application in marketing and helps to identify loyal consumers within a customer relationship management (CRM) program.

References

Frühwirth-Schnatter, S., Stefan Pittner, S., Weber, A. and Winter-Ebmer, R. (2018): Analysing Plant Closure Effects Using Time-Varying Mixture-of-Experts Markov Chain Clustering. Annals of Applied Statistics, forthcoming.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A. and Winter-Ebmer, R. (2016): Mothers' long-run career patterns after first birth. Journal of the Royal Statistical Society, Series A, 179, 707-725.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A. and Winter-Ebmer, R. (2012): Labor Market Entry and Earnings Dynamics: Bayesian Inference Using Mixtures-of-Experts Markov Chain Clustering. Journal of Applied Econometrics, 27, 1116–-1137.

Presenter: **Gary Kopp**
Composite Likelihood Methods for Large Bayesian VARs with Stochastic Volatility
E-mail: gary.koop@strath.ac.uk
Authors: Gary Kopp, Joined work with Joshua C.C. Chan, Eric Eisenstat and Chenghan Hou
Affiliation: University of Strathclyde, Glasgow

Adding multivariate stochastic volatility of a áexible form to large Vector Autoregressions (VARs) involving over a hundred variables has proved challenging due to computational considerations and over-parameterization concerns. The existing literature either works with homoskedastic models or smaller models with restrictive forms for the stochastic volatility.
In this paper, we develop composite likelihood methods for large VARs with multivariate stochastic volatility. These involve estimating large numbers of parsimonious sub-models and then taking a weighted average across these sub-models. We discuss various schemes for choosing the weights. In our empirical work involving VARs of up to 196 variables, we show that composite likelihood methods have similar properties to existing alternatives used with small data sets in that they estimate the multivariate stochastic volatility in a áexible and realistic manner and they forecast comparably. In very high dimensional VARs, they are computationally feasible where other approaches involving stochastic volatility are not and produce superior forecasts than natural conjugate prior homoscedastic VARs.

Presenter: **Monica Patriche**
Equilibrium existence for Bayesian generalized games in choice form and applications
E-mail: monica.patriche@yahoo.com
Authors: Monica Patriche
Affiliation: University of Bucharest

We propose a new definition of a stochastic game, in the spirit of the competitive economy: the Bayesian generalized game in choice form. This game is characterized by constraint correspondences and a Bayesian choice profile under restrictions, expressing the choices of agents, depending on the set of nature states in the world. Our model generalizes, in a Bayesian setting, the ones introduced by Ferrara and Stefanescu in [1] and by the author in [2]. This work is also a continuation of the r

Presenter: **Christian Robert**
Inference in generative models using the Wasserstein distance
E-mail: xian@ceremade.dauphine.fr
Authors: Christian Robert; joint work with Espen Bernton, Pierre Jacob and Mathieu Gerber
Affiliation: Université Paris-Dauphine

A growing range of generative statistical models are such the numerical evaluation of their likelihood functions is intractable. Approximate Bayesian computation and indirect inference have become popular approaches to overcome this issue, simulating synthetic data given parameters and comparing summaries of these simulations with the corresponding observed values. We propose to avoid these summaries and the ensuing loss of information through the use of Wasserstein distances between empirical distributions of observed and synthetic data. We describe how the approach can be used in the setting of dependent data such as time series, and how approximations of the Wasserstein distance allow the method to scale to large data sets. In particular, we propose a new approximation to the optimal assignment problem using the Hilbert space-filling curve. We provide an in-depth theoretical study, including consistency in the number of simulated data sets for a fixed number of observations and posterior concentration rates. The approach is illustrated with various examples, including a multivariate g-and-k distribution, a toggle switch model from systems biology, a queueing model, and a Lévy-driven stochastic volatility model.

slides: https://www.slideshare.net/xianblog/inference-in-generative-models-using-the-wasserstein-distance-ini

Presenter: **Gregor Kastner**
Bayesian Inference in Many Dimensions: Examples from Macroeconomics and Finance
E-mail: gregor.kastner@wu.ac.at
Authors: Gregor Kastner
Affiliation: WU, Vienna

Statistical inference for dynamic models in high dimensions often comes along with a huge amount of parameters that need to be estimated. Thus, to handle the curse of dimensionality, suitable regularization methods are of prime importance, and efficient computational tools are required to make practical estimation feasible. In this talk, we exemplify how these two principles can be implemented for models of importance in macroeconomics and finance. First, we discuss a Bayesian vector autoregressive (VAR) model with time-varying contemporaneous correlations that is capable of handling vast dimensional information sets. Second, we propose a straightforward algorithm to carry out inference in large dynamic regression settings with mixture innovation components for each coefficient in the system.

Presenter	Title
Sudipto Banerjee UCLA	*High-Dimensional Bayesian Geostatistics*
Jia Liu University of Helsinki	*Bayesian model-based spatiotemporal survey design for log-Gaussian cox process*
David Dunson Duke University	*Generalized Bayes for robust and scalable inferences from high-dimensional data*
Atanu Bhattacharjee, Tata Memorial Centre	*Time-Course Data Prediction for Repeatedly Measured Gene Expression*
Noel Cressie University of Wollongong	*Inference for Spatio-Temporal Changes of Arctic Sea Ice*
Marco Cuturi ENSAE, Paris	*Regularized Optimal Transport*
Reza Mohammadi University of Amsterdam	*High-dimensional Bayesian inference for Graphical Models with Application to Brain Connectivity*
Tamara Broderick CSAIL, MIT	*Automated Scalable Bayesian Inference via Data Summarization*
Darren Wilkinson Newcastle University	*A Compositional Approach to Scalable Bayesian Computation and Probabilistic Programming*

Presenter: **Sudipto Banerjee**
High-Dimensional Bayesian Geostatistics
E-mail: sudipto@ucla.edu
Authors: Sudipto Banerjee
Affiliation: UCLA

With the growing capabilities of Geographic Information Systems (GIS) and user-friendly software, statisticians today routinely encounter geographically referenced data containing observations from a large number of spatial locations and time points. Over the last decade, hierarchical spatiotemporal process models have become widely deployed statistical tools for researchers to better understand the complex nature of spatial and temporal variability. However, fitting hierarchical spatiotemporal models often involves expensive matrix computations with complexity increasing in cubic order for the number of spatial locations and temporal points. This renders such models unfeasible for large data sets. I will present a focused review of two methods for constructing well-defined highly scalable spatiotemporal stochastic processes. Both these processes can be used as ``priors" for spatiotemporal random fields. The first approach constructs a low-rank process operating on a lower-dimensional subspace. The second approach constructs a Nearest-Neighbor Gaussian Process (NNGP) that ensures sparse precision matrices for its finite realizations. Both processes can be exploited as a scalable prior embedded within a rich hierarchical modeling framework to deliver full Bayesian inference. These approaches can be described as model-based solutions for big spatiotemporal datasets. The models ensure that the algorithmic complexity has n floating point operations (flops), where n is the number of spatial locations (per iteration). We compare these methods and provide some insight into their methodological underpinnings.

Keywords: Bayesian modeling; Directed Acyclic Graphs; Gaussian Processes; Low-rank models; Scalable models; Spatial stochastic processes

Presenter: **Jia Liu**
Bayesian model-based spatiotemporal survey design for log-Gaussian cox process
E-mail: jia.liu@helsinki.fi
Authors: **Jia Liu**
Affiliation: University of Helsinki

In geostatistics, the design for data collection is central for accurate prediction and parameter inference. One important class of geostatistical models is log-Gaussian Cox process (LGCP) which is used extensively, for example, in ecology. However, there are no formal analyses on optimal designs for LGCP models in spatial or spatiotemporal domains. In this work, we compare traditional balanced and uniform random designs in situations where analyst has prior information on intensity function of LGCP. Moreover, we propose a new spatially balanced rejection sampling design which directs sampling to locations that are a priori expected to provide most information. Designs are evaluated using the average predictive variance loss function and the Kullback-Leibler divergence between prior and posterior for the LGCP intensity function. Our results show that the proposed rejection sampling method outperforms traditional balanced and uniform random sampling designs. We perform also a case study applying our new sampling design to plan a survey for species distribution modeling on larval areas of two commercially important fish stocks on Finnish coastal areas. The case study results show that rejection sampling designs give considerable benefit compared to traditional designs. Results show also that best performing designs may vary considerably between target species.

Keywords: Bayesian inference; Kullback-Leibler information; log Gaussian Cox process; spatiotemporal; species distribution; survey design.

Presenter: **David Dunson**
Generalized Bayes for robust and scalable inferences from high-dimensional data
E-mail: dunson@duke.edu
Authors: David Dunson
Affiliation: Duke University

In this talk I describe some recent ideas on scaling up Bayesian inferences for massive datasets, which involve huge sample sizes and/or very high-dimensional data. I propose new classes of scalable MCMC algorithms based on biased subsampling and multiscale representations; these approaches are designed to converge to a target distribution corresponding to an exact posterior distribution. Often it is useful to instead employ approximations to speed up computation and achieve more robust inferences in big data settings. We propose some notions of "generalized Bayes" inferences that are useful in this regard, and illustrate these approaches through several biomedical applications involving very high-dimensional (but not necessarily large sample size) data.

Presenter: **Atanu Bhattacharjee**,
Time-Course Data Prediction for Repeatedly Measured Gene Expression
E-mail: atanustat@gmail.com
Authors: Atanu Bhattacharjee[1], Gajendra K. Vishwakarma[2]
Affiliation:
[1]Centre for Cancer Epidemiology,The Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Tata Memorial Centre,Navi Mumbai-410210.Tata Memorial Centre
[2] Department of Applied Mathematics, Indian Institute of Technology(ISM)-Dhanbad, India

Variability in time course gene expression data is natural phenom-ena. The intention of this work is to predict the future time point data through observed sample data point. The Bayesian inference is carried to serve the objective. A total of 6 replicates 3 time point's data of 218 genes expression is adopted to illustrate the method. The estimates are found consistent with HPD interval to predict the future time point gene expression value. This proposed method can be adopted in other gene expression data setup to predict the future time course data.

Keywords Crossover Trial _ Bayesian algorithm _ Multiple Imputation _Three Arms _ Three Periods

Presenter: **Noel Cressie**
Inference for Spatio-Temporal Changes of Arctic Sea Ice
E-mail: ncressie@uow.edu.au
Authors: Noel Cressie[1], Dr Bohai Zhang[2]
Affiliation:
[1]National Institute for Applied Statistics Research Australia (NIASRA), Building 39C, University of Wollongong, NSW 2522, Australia
[2]School of Statistics and Data Science, Nankai University, China.

Arctic sea-ice extent has been of considerable interest to scientists in recent years, mainly due to its decreasing trend over the past 20 years. In this talk, I propose a hierarchical spatio-temporal generalized linear model (GLM) for binary Arctic-sea-ice data, where data dependencies are introduced through a latent, dynamic, spatio-temporal mixed-effects model. By using a fixed number of spatial basis functions, the resulting model achieves both dimension reduction and non-stationarity for spatial fields at different time points. An EM algorithm is used to estimate model parameters, and an MCMC algorithm is developed to obtain the predictive distribution of the latent spatio-temporal process. The methodology is applied to spatial, binary, Arctic-sea-ice data for each September over the past 20 years, and several posterior summaries are computed to detect changes of Arctic sea-ice cover. The fully Bayesian version is under development and will be discussed.

Keywords: Dynamic temporal dependence; EM algorithm; GLM for binary data

Presenter: **Marco Cuturi**
Regularized Optimal Transport
E-mail: marco.cuturi@ensae.fr
Authors: Marco Cuturi
Affiliation: ENSAE, Paris

I will present in this talk a review of recent numerical advances in the field of optimal transport. These advances arise from a suitable regularization of the original linear programming (Kantorovich) formulation of the optimal transport problem. That reformulation yields an efficient numerical scheme that can be run on parallel architectures and which can be easily automatically differentiated. I will described a few applications of these advances.

Presenter: **Reza Mohammadi**
High-dimensional Bayesian inference for Graphical Models with Application to Brain Connectivity
E-mail: a.mohammadi@uva.nl
Authors: Reza Mohammadi
Affiliation: University of Amsterdam

In graphical models, Bayesian frameworks provide a straightforward tool, explicitly incorporating underlying graph uncertainty. In principle, the Bayesian approaches are based on averaging the posterior distributions of the quantity of interest, weighted by their posterior graph probabilities. However, Bayesian inference has not been used in practice for high-dimensional graphical models, because computing the posterior graph probabilities is hard and the number of possible graph models is very large. In this talk, we discuss the computational problems related to Bayesian structure learning and we offer several solutions to cope the high-dimensionality problems. We apply our method to high-dimensional fMRI data from brain connectivity studies to show its empirical usefulness. In addition, we have implemented our method in the R packages BDgraph and ssgraph which are available online at CRAN.

Presenter: **Tamara Broderick**

Automated Scalable Bayesian Inference via Data Summarization

E-mail: tbroderick@csail.mit.edu
Authors: Tamara Broderick
Affiliation: CSAIL, MIT

Bayesian methods are attractive for analyzing large-scale data due to in part to their coherent uncertainty quantification, ability to model complex phenomena, and ease of incorporating expert information. Many standard Bayesian inference algorithms are often computationally expensive, however, so their direct application to large datasets can be difficult or infeasible. Other standard algorithms sacrifice accuracy in the pursuit of scalability. We take a new approach.

Namely, we leverage the insight that data often exhibit approximate redundancies to instead obtain a weighted subset of the data (called a "coreset") that is much smaller than the original dataset. We can then use this small coreset as input to existing Bayesian inference algorithms without modification. We provide theoretical guarantees on the size and approximation quality of the coreset. In particular, we show that our method provides geometric decay in posterior approximation error as a function of coreset size. We validate on both synthetic and real datasets, demonstrating that our method reduces posterior approximation error by orders of magnitude relative to uniform random subsampling.

Presenter: **Darren Wilkinson**

A Compositional Approach to Scalable Bayesian Computation and Probabilistic Programming

E-mail: darren.wilkinson@ncl.ac.uk
Authors: Darren Wilkinson
Affiliation: Newcastle University

In the Big Data era, some kind of hierarchical "divide and conquer" approach seems necessary for the development of genuinely scalable Bayesian models and algorithms, where (solutions to) sub-problems are combined to obtain (solutions to) the full problem of interest. It is therefore unfortunate that statistical models and algorithms are not usually formulated in a composable way, and that the programming languages typically used for scientific and statistical computing fail to naturally support the composition of models, data and computation. The mathematical subject of category theory is in many ways the study of composition, and provides significant insight into the development of more compositional models of computation. Functional programming languages which are strongly influenced by category theory turn out to be much better suited to the development of scalable statistical models and algorithms than the imperative programming languages more commonly used. Expressing algorithms in a functional/categorical way is not only more elegant, concise and less error-prone, but provides numerous more tangible scalability benefits, such as automatic parallelisation and distribution of computation. Categorical concepts such as monoids, functors, monads and comonads turn out to be useful for formulating (Monte Carlo based) Bayesian inferential algorithms in a composable way. Further, probability monads form the foundation for the development of flexible and compositional probabilistic programming languages.

Conference Program: Friday 30th November 2018, 09:00 – 12:30

Presenter	Title
Stéphane Robin AgroParisTech	*Shortened Bridge Sampler: Using Deterministic Approximations to Accelerate SMC for Posterior Sampling*
Benoit Liquet Université de Pau et des Pays de L'Adour	*Bayesian Variable Selection Regression Of Multivariate Responses For Group Data*
Antonietta Mira Università della Svizzera italiana and University of Insubria	*Bayesian dimensionality reduction via the identifications of the data intrinsic dimensions*
Gajendra Vishwakarma, Indian Institute of Technology Dhanbad	*A Bayesian Approach for Dynamic Treatment Regimes in Presence of Competing Risk Analysis*
Jean-Michel Marin Université de Montpellier	*Local tree methods for classification*

Presenter: **Stéphane Robin**
Shortened Bridge Sampler: Using Deterministic Approximations to Accelerate SMC for Posterior Sampling
E-mail: robin@agroparistech.fr
Authors: Stéphane Robin, Sophie Donnet
Affiliation: AgroParisTech

The Stochastic Block-Model (SBM) has become one of the most popular tools for analysing the topology of interaction networks. It can be generalized to account for the effect of covariates on the intensity of the links between edges. Because of the intricate dependency structure, the statistical inference of SBM models raises a series of issues and most approaches rely on approximations, with few theoretical guarantees.

On the other hand, Sequential Monte Carlo (SMC) has become a standard tool for Bayesian inference of complex models. This approach can be computationally demanding, especially when initialized from the prior distribution.

We focus on a weighted version of the SBM model (including covariates). We propose a bridge sampling scheme starting from a deterministic approximation of the posterior distribution and targeting the true one. The resulting Shortened Bridge Sampler (SBS) relies on a sequence of distributions that is determined in an adaptive way.

We use of the proposed approach to analyse the organization of some ecological networks.

Presenter: **Benoit Liquet**
Bayesian Variable Selection Regression Of Multivariate Responses For Group Data
E-mail: benoit.liquet@univ-pau.fr
Authors: Benoit Liquet
Affiliation: Université de Pau et des Pays de L'Adour

We propose two multivariate extensions of the Bayesian group lasso for variable selection and estimation for data with high dimensional predictors and multi-dimensional response variables. The methods utilize spike and slab priors to yield solutions which are sparse at either a group level or both a group and individual feature level. The incorporation of group structure in a predictor matrix is a key factor in obtaining better estimators and identifying associations between multiple responses.

Presenter: **Antonietta Mira,** Università della Svizzera italiana and University of Insubria
Bayesian dimensionality reduction via the identifications of the data intrinsic dimensions
E-mail: antonietta.mira@usi.ch
Authors: Michele Allegra, Francesco Denti, Elena Facco, Alessandro Laio, Michele Guindani, Antonietta Mira
Affiliation: Università della Svizzera italana Università della Svizzera italiana; University of Insubria

Even if they are defined on a space with a large dimension, data points usually lie onto a hypersurface, or manifold, with a much smaller intrinsic dimension (ID). The recent TWO-NN method (Facco et al., 2017, Scientific Report), allows estimating the ID when all points lie onto a single manifold.
TWO-NN only assumes that the density of points is approximately constant in a small neighborhood around each point. Under this hypothesis, the ratio of the distances of a point from its first and second neighbour follows a Pareto distribution that depends parametrically only on the ID, allowing for an immediate estimation of the latter. We extend the TWO-NN model to the case in which the data lie onto several manifolds with different ID. While the idea behind the extension is simple (the Pareto is replaced by a mixture of K Pareto distributions), a non-trivial Bayesian scheme is required for estimating the model and assigning each point to the correct manifold. Applying this method, which we dub Hidalgo (heterogeneous intrinsic dimension algorithm), we uncover a surprising ID variability in several real-world datasets. Hidalgo obtains remarkable results, but its main limitation consists in fixing a priori the number of component in the mixture. To adopt a fully Bayesian approach, a possible extension would be the specification of a prior distribution for the parameter K. Instead, we employ a flexible Bayesian Nonparametric approach and model the data as an infinite mixture of Pareto distributions using a Dirichlet Process Mixture Model. The approach allows to evaluate the uncertainty relative to the number of mixture components. Since the posterior distribution has no closed form, we employ the Slice Sampler algorithm for posterior inference. From preliminary analyses performed on simulated data, the model provides promising results.

Presenter: **Gajendra Vishwakarma**
A Bayesian Approach for Dynamic Treatment Regimes in Presence of Competing Risk Analysis
E-mail: vishwagk@rediffmail.com vishwagk@iitism.ac.in
Authors: Gajendra K. Vishwakarma
Affiliation: Department of Applied Mathematics, Indian Institute of Technology Dhanbad

A sequencing rule is considered to formulate the dynamic treatment regime (DTR). However, this sequence rule is based on clinical relevance, prior evidence about the best performing therapy and as per requirement to treat a patient in a specific scenario. The challenge occurred when the objective of a study offers a concluding remark about best effective therapy among all possible combinations of treatment management schedule treated with a sequence rule. The time-to-event data analysis is the only available method to figure out the best effective treatment in the context of oncology research. However, the presences of competing risk event of death in time-to-event data is unavoidable and it becomes more challenging to take a decision about the best effective treatment strategy. We developed the statistical methodology to handle the competing risk -time-to-event data analysis in DTR. The analysis is performed with the Bayesian approach to obtain the best effective treatment strategy. We introduce the OpenBUGS function, which provides the comparison and estimation of different treatment sequences in time-to-event competing risk data analysis adopting the newly proposed statistical approach. This developed method is handy to boost up the personalized medicine in oncology setup through supportive decision rule.

Keywords: Dynamic Treatment Regime, Competing Risk, Sequential rule, Bayesian, Personalized Medicine.

Presenter: **Jean-Michel Marin**
Local tree methods for classification
E-mail: jean-michel.marin@umontpellier.fr
Authors: Jean-Michel Marin
Affiliation: Université de Montpellier

For intractable likelihood models, the idea that underlines a very large class of methods is to learn from simulations coming from generative models. Naturally, to learn from these simulations, machine learning methods are increasingly used. That is typically the case in the Bayesian paradigm where random forests strategies have prooven to have good empirical properties either for model choice questions or parameter inference problems. Approximate Bayesian Computation strategies via random forests perform a global learning and the number of simulations from the generative model needs to be quite large. To decrease that number, something crucial in the big data context, local learning can be more efficient. In this talk, we discuss about local tree methods.

Conference Posters:

Erlis Ruli ruli@stat.unipd.it

Objective model selection with proper scoring rules and improper priors

The Bayes factor (BF) is the standard model selection tool in the Bayesian framework. However, it is well known that the BF tends to be sensitive to the prior distributions of the models under comparison and therefore requires careful elicitation. Furthermore, the BF cannot be used with objective improper priors, because of the dependence of the marginal likelihood on the arbitrary scaling constants of the model prior densities. Recently, Dawid and Musio (2015-2017) propose to solve this problem by replacing the marginal log-likelihood by an homogeneous proper scoring rule, which is insensitive to the scaling constants. We apply and study this methodology in the context of continuous exponential families. A couple of examples will be provided

Farzana Jahan f.jahan@hdr.qut.edu.au and Kerrie Mengersen

Bayesian Empirical Likelihood Spatial Model applying Leroux Structure

Bayesian Empirical Likelihood (BEL) methods have been applied to spatial data analysis for small area estimation (SAE) by extending the Fay-Herriot (FH) model [1] in recent times. The present study is an attempt to develop a Bayesian semiparamtric model for spatial data analysis utilising the popular Leroux prior for spatial dependence.

Hoang Nguyen hoang.nguyen@uc3m.es

Variational Inference for high dimensional structured factor copulas

Factor copula models have been recently proposed for describing the joint distribution of a large number of variables in terms of a few common latent factors. In this paper, we employ a Bayesian procedure to make fast inferences for multi-factor and structured factor copulas. To deal with the high dimensional structure, we apply a variational inference (VI) algorithm to estimate the different specifications of factor copula models.
Compared to the Markov chain Monte Carlo (MCMC) approach, the variational approximation is much faster and could handle a sizeable problem in a few seconds. Another issue of factor copula models is that the bivariate copula functions connecting the variables are unknown in high dimensions. We derive an automatic procedure to recover the hidden dependence structure. By taking advantages of the posterior modes of the latent variables, we select the bivariate copula functions based on minimizing Bayesian information criterion (BIC). The simulation studies in different contexts show that the procedure of bivariate copula selection could be at least 80% accuracy in comparison to the true generated copula model. We illustrate our proposed procedure with high dimensional real dataset.

Julyan Arbel julyan.arbel@inria.fr

Bayesian neural networks increasingly sparsify their units with depth

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust of the paper is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularisation scheme at the level of the units, ranging from convex penalties for the first two layers - weight decay for the first and Lasso for the second - to non convex penalties for deeper layers. Thus, despite weight decay does not allow for the weights to be set exactly to zero, sparse solutions tend to be selected for the units from the second layer onward. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.
https://arxiv.org/abs/1810.05193

Khuyen Le lekhuyen.maths@gmail.com

Connected component selection for Linear Discriminant Analysis in high dimension and applications to medical imaging

We consider the problem of classifying a normally distributed vector x 2 Rp into one of K groups of p-dimensional normal distribution when the variable number p is much larger than the observation number N. In such a case, Bayesian classi_ers such as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) usually do not be applied. Under the sparsity assumption of the inverse covariance matrix (a.k.a the precision matrix), we propose an adaptation of the LDA, namely Adapted-LDA, for high dimension classi_cation by including a sparse precision matrix estimation over all populations. This estimate is obtained by solving a Graphical Least Absolute Shrinkage and Selection Operator (GLASSO) problem. Moreover, of interest is the case of block diagonal precision matrix, its blocks are represented by connected components in the graphical model. Each connected component in the graph has its own capacity for discriminating data. We propose a method for computing the discriminant capacity of each connected component, then we select the connected components whose discriminant capacity are the largest. Both the Adapted-LDA and the connected component selection method are applied to real data, where we classify patients of Alzheimer's Diseases from the Healthy Control people. We show that our method, Adapted-LDA, outperforms competing procedures. Besides, the connected component selection method not only reduces significantly the misclassification of the Adapted-LDA as well as other classification methods, but it also reduces the computational time of these methods.

Keywords: LDA, FDA, Sparse Precision matrix Estimation (SPE), GLASSO, ADMM, connected component selection, PET.

Maxime Rischard, mrischard@g.harvard.edu Pierre Jacob, Natesh Pillai

Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation.

Posterior distributions often feature intractable normalizing constants, called marginal likelihoods or evidence, that are useful for model comparison via Bayes factors. This has motivated a number of methods for estimating ratios of normalizing constants in statistics. In computational physics the logarithm of these ratios correspond to free energy differences. Combining unbiased Markov chain Monte Carlo estimators with path sampling, also called thermodynamic integration, we propose new unbiased estimators of the logarithm of ratios of normalizing constants. As a by-product, we propose unbiased estimators of the Bayesian cross-validation criterion. The proposed estimators are consistent, asymptotically Normal and can easily benefit from parallel processing devices.

Oluwole, K. Oyebamiji[1] and David Leslie[1]
[1]Mathematics and Statistics, Lancaster University, LA1 4YF, UK.
o.oyebamiji@lancaster.ac.uk; d.leslie@lancaster.ac.uk

Bayesian optimal weighting scheme for combining simulation ensemble for global climate projection

To quantify uncertainties associated with climate projections, simulations derived from multiple climate modelling centres are being widely used for the assessment. The typical approach for combining this data is to average the model results known as unweighted multi-model projection. The multi-model ensemble averaging is a biased approach because models are likely to perform differently. We develop an optimal weighting scheme that uses the Bayesian framework for integrating global ozone observations provided by the Tropospheric Ozone Assessment Report (TOAR) with an ensemble of simulation obtained from the Chemistry-Climate Model Initiative (CCMI) partners. A kriging-based approach was employed to spatially interpolate the coarse field observation onto a fine resolution grid. We achieved computational efficiency by clustering and averaging the interpolated high-resolution data into eight different regions of the world. We do not average the model outputs over the year to capture the temporal dependence, so the data fusion is dynamically weighted. The model ensembles are intelligently weighted using the eigendecomposition of similarity matrices derived from multi-output models. We then used a combination of the multivariate dynamic linear model with Gaussian process regression to compute the temporally-varying weight for each spatial region by matching the multiple independent observations against the weighted model outputs of the past. The estimated weights are then used to scale future projections of global ozone distribution. Our technique accounts for the fact that different models are likely better for particular regions or conditions.

Keywords: Eigendecomposition, Optimal weighting scheme, Dynamic linear model, Bayesian model, MCMC, Ozone

Paul-Marie Grollemund paul-marie.grollemund@umontpellier.fr

Elicitation of Experts' Knowledge for Functional Linear Regression

We present an approach to elicit experts' knowledge about the Bliss model, which is a parsimonious Bayesian Functional Linear Regression model. We derive an informative prior from elicited information and we define weights to tune prior information contribution on the estimators.

Victor Penha victoraspenha@gmail.com

Do parasites affect bird plumage coloration? An analysis on plumage reflectance and carotenoid plumage deposition

Birds are highly visual animals, and their plumage coloration are important factors related to communication and predation risk avoidance. Plumage coloration can be determined by pigments, such as carotenoids and melanin. Carotenoids are also used for other physiological roles, stimulating the immune system and being stored as fat within tissues. Carotenoid-based plumage coloration mainly come from the diet, and because feather coloration is directly related to the amount of carotenoids an individual's ingest, carotenoid-based plumage is an honest signal of an individual's ability to forage and compete for carotenoid-containing resources. Several factors can change the balance of physiological trade-offs associated to carotenoid investment, including pathogens, such as haemosporidian parasites. Therefore, the goal of the present study was to evaluate the association between the carotenoid-based plumage as well as the body condition of birds and malaria parasites. We captured individuals from several species of Passeriformes in an area of *sensu strictu* Cerrado, in Uberlandia, Minas Gerais, Brazil. We used 10 mist-nets to capture individuals and collected a small blood sample from the brachial vein, to prepare blood smears and then use it to a nested PCR procedure to screen individuals for malaria parasites presence. A small number of feathers were also collected from each individual in order to analyze plumage coloration variables, such as achieved saturation, carotenoid chroma, hue and the maximum reflectance at the ultraviolet spectrum. Morphometric measurements were also taken in order to calculate the scale mass index. Only the achieved saturation and carotenoid chroma was negatively explained by the presence of parasites. The achieved saturation was also negatively explained by the amount of parasites one individual had. The scaled mass index and the achieved saturation along with the hue were negatively corelated. Energetic imbalances may have individuals trade carotenoid to plumage and the immune system to cope with parasites. These results are important since studies have shown that females can discriminate differences in saturation and prefer males with higher concentrations of carotenoids, showing how important parasites may be for sexual selection in natural bird populations.