BISP12 ONLINE WORKSHOP

The Twelfth Workshop on Bayesian Inference in Stochastic Processes

27-28 May 2021

PROGRAM – DAY 2

28 May 2021

Virtual poster presentations

Virtual posters are available at the link: bisp12.imati.cnr.it/virtual_poster.php

Chair: Elisa Varini, CNR-IMATI, Italy

CEST time

13.00-13.50

Poster 9 - **Gaussian Process with multiple degrees of freedom**
Izhar Asael Alonzo Matamoros, Aki Vehtari

Poster 10 - **Bayesian Variable Selection for Gaussian Copula regression models**
Angelos Alexopoulos, Leonardo Bottolo

Poster 11 - **Gaussian process modelling including derivatives in the framework of computational soil and earth structure simulations**
Carmen van Meegen, Carla Henning, Swetlana Herbrandt, Katja Ickstadt, Tim Ricken

Poster 12 - **Lithological Tomography with the Correlated Pseudo-Marginal Method**
Lea Friedli, Niklas Linde, David Ginsbourger, Arnaud Doucet

Poster 13 - **A Hierarchical Model of Nonhomogeneous Poisson Processes for Twitter Retweets**
Clement Lee, Darren Wilkinson

Poster 14 - **Estimating the effectiveness of permanent price reductions for competing products using multivariate Bayesian structural time series models**
Fiammetta Menchetti, Iavor Bojinov

Poster 15 - **Bayesian approach to illness-death models for assessing the progression of recurrent hip fractures**
Fran Llopis-Cardona, Carmen Armero, Gabriel Sanfélix-Gimeno

Poster 16 - **Hierarchical models for Bayesian inference of brain activity from Magneto/Electro-encephalography data**
Alessandro Viani, Gianvittorio Luria, Harald Bornfleth, Alberto Sorrentino

Poster 17 - **Bayesian Time-Varying Tensor Vector Autoregressive Models for Dynamic Effective Connectivity**
Wei Zhang, Michele Guindani, Sonia Petrone, Ivor Cribben

Invited talks

Chair: Mike Wiper, Universidad Carlos III, Spain

14.00-14.40 Talk 6 - **Bayesian Network Inference With Uncertain Evidence and Parameters**
Paul Pao-Yen Wu, Fabrizio Ruggeri, Kerrie Mengersen
Discussant: David Banks, Duke University, USA

14.45-15.25 Talk 7 - **Bayesian forecasting dynamic models under attacks**
Roi Naveiro
Discussant: Mike West, Duke University, USA

15.25-15.45 Break

Chair: Refik Soyer, George Washington University, USA

15.45-16.25 Talk 8 - **Quasi-stationary Monte Carlo methods via stochastic approximation**
Andi Q. Wang, Murray Pollock, Gareth Roberts, David Steinsaltz
Discussant: Giacomo Zanella, Università Bocconi, Italy

16.30-17.10 Talk 9 - **Compositions of discrete random probabilities for inference on multiple samples**
Giovanni Rebaudo, Augusto Fasano, Antonio Lijoi, Igor Prünster
Discussant: Bernardo Nipoti, Università di Milano-Bicocca, Italy

Chair: Antonio Pievatolo, CNR-IMATI, Italy

17.15-17.55 Talk 10 - **Bayesian Conditional Auto-Regressive LASSO Models to Learn Sparse Networks with Predictors**
Yunyi Shen, Claudia Solis-Lemus
Discussant: Nicholas Polson, University of Chicago, USA

18.00-18.05 **Closing**

Poster 9 – 28 May 2021 | 13:00-13:50 CEST

Gaussian Process with multiple degrees of freedom

Izhar Asael Alonzo Matamoros [1], Aki Vehtari [1]

(1) Aalto University, Finland.

asael_am@hotmail.com; aki.vehtari@aalto.fi

Abstract: Gaussian process regression is a popular tool for Bayesian modeling because of its flexibility for imposing non-parametric priors over functions. However, one major limitation of this model is its non-robustness, so that its predictive accuracy is compromised in the presence of outlier observations. Although several solutions using heavy-tailed distributions such as student-t processes or student-t distribution for the likelihood could be applied, that could still not suffice since it is reasonable to assume that the behavior of the tails differs depending on every observation. For this situation, assuming that every variable has its own degree of freedom should lead to a modeling improvement. In this work, we introduce the multiple degrees of freedom Gaussian process model obtained restructuring the process' covariance matrix such that its posterior follows a Generalized student-t distribution. This distribution can only be described through its hierarchical structure assigns a student-t distribution with different degrees of freedom to every observation. We derive the model's hierarchical structure, demonstrate its benefits with examples, and finally use the method in a real data application.

Keywords: Gaussian Process; hierarchical structures; generalized student-t distribution; heavy-tailed distribution.

Poster 10 – 28 May 2021 | 13:00-13:50 CEST

Bayesian Variable Selection for Gaussian Copula regression models

Angelos Alexopoulos (1), *Leonardo Bottolo* (2,3,4)

(1) MRC Biostatistics Unit, University of Cambridge, (2) Department of Medical Genetics, University of Cambridge, (3) The Alan Turing Institute, London, (4) MRC Biostatistics Unit, University of Cambridge.

angelos@mrc-bsu.cam.ac.uk

Abstract: We develop a novel Bayesian method to select important predictors in regression models with multiple responses of diverse types. In particular, a sparse Gaussian copula regression model is used to account for the multivariate dependencies between any combination of discrete and/or continuous responses and their association with a set of predictors. We utilize the parameter expansion for data augmentation strategy to construct a Markov chain Monte Carlo algorithm for the estimation of the parameters and the latent variables of the model. Based on a centered parametrization of the Gaussian latent variables, we design an efficient proposal distribution to update jointly the latent binary vectors of important predictors and the corresponding non-zero regression coefficients. The proposed strategy is tested on simulated data and applied to two real data sets in which the responses consist of low-intensity counts, binary, ordinal and continuous variables.

Keywords: Gaussian copula; Mixed data; Multiple-response regression models; Sparse covariance matrix; Variable selection.

Poster 11 – 28 May 2021 | 13:00-13:50 CEST

Gaussian process modelling including derivatives in the framework of computational soil and earth structure simulations

Carmen van Meegen [(1)], Carla Henning [(2)], Swetlana Herbrandt [(1)], Katja Ickstadt [(1)], Tim Ricken [(2)]

(1) Chair of Mathematical Statistics with Applications in Biometrics, TU Dortmund University, Germany, (2) Institute of Mechanics, Structural Analysis, and Dynamics, University of Stuttgart, Germany.

fvanmeegen@statistik.tu-dortmund.de; herbrandt@statistik.tu-dortmund.de; ickstadtg@statistik.tu-dortmund.de; fcarla.henning@isd.uni-stuttgart.de; tim.rickeng@isd.uni-stuttgart.de

Abstract: Improving safety concepts in geotechnical applications is an aim in our research project on quantification and assessment of polymorphic uncertainties in computational simulations of earth structures within the DFG priority program SPP 1886 on polymorphic uncertainty quantification. Therefore, a better understanding of the nature and influence of the uncertainties in the material properties is required. We propose a gradient-enhanced surrogate model, i.e. Gaussian process, with a subsequent sensitivity analysis to examine the importance and type of the influence of a material parameter. We illustrate this method using a numerical soil and earth structure simulation.
Computational simulations of reliable earth structure assessment imply a very high complexity of the soil structure on different length scales, the load bearing mechanism and the residual stress state due to individual load histories. To describe the strongly coupled and transient solid-fluid response behaviour, the theory of porous media (TPM) is used and solved via the finite element method[1,2]. Since these computational simulations are usually expensive in terms of their long run time, multiple evaluations of the TPM model for probabilistic analyses are not feasible. Hence, in the present contribution, the focus lies on the adaption of a suitable surrogate model as a basis for sensitivity analysis[3]. For this purpose, Gaussian processes are introduced as surrogates that provide a framework for the use of exploratory statistical methods, e.g. (Bayesian) sensitivity analysis. Since the model's tangential information of the underlying physical model can be obtained with minimal additional effort during each individual evaluations of the simulation, it is suitable to extend the Gaussian process by the gradient[4]. This has two advantages: The fit of the surrogate model is improved and the (TPM) model runs required for reliable estimation are reduced.

Keywords: Gaussian Process; Gradient-Enhanced Kriging; (Bayesian) Sensitivity Analysis.

References

[1] De Boer, R., Theory of Porous Media, Springer, New York (2008).

[2] Ehlers,W., and Bluhm, J., Foundations of multiphasic and porous materials, Porous Media: Theory, Experiments and Numerical Applications, Springer, New York (2002), 3–86.

[3] Oakley, J. E., and O'Hagan, A. (2004), Probabilistic sensitivity analysis of complex models: A Bayesian approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66, 751–769.

[4] Lookwood, B. A., and Anitescu, M. (2012), Gradient-Enhanced Universal Kriging for Uncertainty Propagation, Nuclear Science and Engineering, 170:2, 168–195.

Poster 12 – 28 May 2021 | 13:00-13:50 CEST

Lithological Tomography with the Correlated Pseudo-Marginal Method

Lea Friedli [(1)], Niklas Linde [(1)], David Ginsbourger [(2)], Arnaud Doucet [(3)]

(1) Institute of Earth Sciences, University of Lausanne, Switzerland, (2) Institute of Mathematical Statistics and Actuarial Science and Oeschger Center for Climate Change Research, University of Bern, Switzerland, (3) Department of Statistics, Oxford University, United Kingdom.

lea.friedli@unil.ch; niklas.linde@unil.ch; david.ginsbourger@stat.unibe.ch; doucet@stats.ox.ac.uk

Abstract: We consider an inverse problem in which hydrogeological parameters are inferred from geophysical data, and the intermediate geophysical properties are treated as latent variables. The petrophysical relationship linking the hydrogeological and geophysical properties is generally non-linear and includes significant scatter. In this setting, it is often necessary to estimate the intractable likelihood of observing the geophysical data given the hydrogeological parameters. The Pseudo-Marginal method relies on an unbiased approximation of this likelihood based on Monte-Carlo averaging over samples from the petrophysical relationship, thereby, ensuring that the scattered nature of the petrophysical relationship is taken into account. To reduce the number of required Monte Carlo samples while increasing the efficiency of the resulting Metropolis-Hastings scheme, we ensure low-variance approximations of the likelihood ratio by correlating the samples used in two subsequent iterations of the Markov chain. In geophysical settings with large data sets exhibiting low noise levels, it is essential to rely on importance sampling distributions that well represent the posterior random field of petrophysical scatter around the current hydrogeological property field. For comparison purposes, we use an inversion ignoring the petrophysical uncertainty and a method inferring the posterior distribution of the latent variables along with that of the hydrogeological target variables. We evaluate the performance of the various methods by using a synthetic example in which we invert for porosity, using crosshole ground-penetrating radar (GPR) travel times as geophysical data. We use a (50 × 50)-dimensional pixel-based parameterization of the multi-Gaussian porosity field with known statistical parameters, resulting in a parameter space of high dimension. To obtain an efficient Metropolis-Hastings regardless, we apply a multi-chain algorithm with adaptive proposal scheme and, to avoid dependencies on the discretization of the target space and to prevent too narrow exploration of the posterior due to the adaptation of the proposal scheme, we adjust the proposal scheme such that it preserves the prior distribution. We find that the correlated pseudo-marginal with the prior-preserving proposal scheme outperforms the other approaches by enhancing the exploration of the posterior distribution, reducing the mean Kullback-Leibler divergence to the analytical posterior by up to 99%.

Keywords: Hydrogeophysical inversion, Latent variable model, Intractable Likelihood function, Pseudo-Marginal method.

Poster 13 – 28 May 2021 | 13:00-13:50 CEST

A Hierarchical Model of Nonhomogeneous Poisson Processes for Twitter Retweets

Clement Lee [(1)], *Darren Wilkinson* [(2)]

(1) Lancaster University, UK, (2) Newcastle University, UK

clement.lee@lancaster.ac.uk

Abstract: We present a hierarchical model of nonhomogeneous Poisson processes (NHPP) for information diffusion on online social media, in particular Twitter retweets. The retweets of each original tweet are modelled by a NHPP, for which the intensity function is a product of time-decaying components and another component that depends on the follower count of the original tweet author. The latter allows us to explain or predict the ultimate retweet count by a network centrality-related covariate. The inference algorithm enables the Bayes factor to be computed, to facilitate model selection. Finally, the model is applied to the retweet datasets of two hashtags.

Keywords: Poisson process; Information diffusion; Markov chain Monte Carlo; Model selection.

Poster 14 – 28 May 2021 | 13:00-13:50 CEST

Estimating the effectiveness of permanent price reductions for competing products using multivariate Bayesian structural time series models

Fiammetta Menchetti [(1)], Iavor Bojinov [(2)]

(1) University of Florence, Italy, (2) Harvard Business School, USA.

fiammetta.menchetti@unifi.it; ibojinov@hbs.edu

Abstract: The Florence branch of an Italian supermarket chain recently implemented a strategy that permanently lowered the price of numerous store brands in several product categories. To quantify the impact of such a policy change, researchers often use synthetic control methods for estimating causal effects when a subset of units receive a single persistent treatment, and the rest are unaffected by the change. In our applications, however, competitor brands not assigned to treatment are likely impacted by the intervention because of substitution effects; more broadly, this type of interference occurs whenever the treatment assignment of one unit affects the outcome of another. This paper extends the synthetic control methods to accommodate partial interference, allowing interference within predefined groups but not between them. Focusing on a class of causal estimands that capture the effect both on the treated and control units, we develop a multivariate Bayesian structural time series model for generating synthetic controls that would have occurred in the absence of an intervention enabling us to estimate our novel effects. In a simulation study, we explore our Bayesian procedure's empirical properties and show that it achieves good frequentists coverage even when the model is misspecified. We use our new methodology to make causal statements about the impact on sales of the affected store brands and their direct competitors. Our proposed approach is implemented in the CausalMBSTS R package.

Keywords: Causal Inference; Partial Interference; Synthetic Controls; Bayesian Structural Time Series.

Poster 15 – 28 May 2021 | 13:00-13:50 CEST

Bayesian approach to illness-death models for assessing the progression of recurrent hip fractures

Fran Llopis-Cardona [(1)], Carmen Armero [(2)], Gabriel Sanfélix-Gimeno [(1)]

(1) Health Services Research Unit, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Spain, (2) Department of Statistics and Operational Research, Universitat de València, Spain.

llopis_fracar@gva.es

Abstract: Multi-state models are a broad class of stochastic models that focus on trajectories in which individuals are allowed to move over time among different states. Many models for survival analysis can be also defined as stochastic processes. Approaching multi-event survival problems via stochastic processes provides richer and complex models that include dependencies on past events or transitions. A particular multi-state model, the so-called illness-death model, is of special interest when dealing in competing risks scenarios with a unidirectional communication between the competing events. Transition probabilities between states are modelled from transition intensities, which are the hazard rate functions within the survival framework. We applied a Bayesian illness-death model to a real-world study of recurrent hip fracture (the PREV2FO cohort). We retrospectively followed 34 491 patients aged 65 and over who were discharged after a hospitalization due to an osteoporotic hip fracture between 2008-2015. We considered two possible events after the first hip fracture: recurrent fracture and death. The illness-death model also allowed to consider a transition from refracture to death. We assume Markovian transitions from the initial state to disease and death. Nevertheless, progression from refracture to death is modelled by a semi-Markov process, which implies dependence on the time spent in the fracture state. Transition times are modelled through transition intensities via Cox proportional hazard models. The subsequent posterior distribution was approximated via MCMC methods. Posterior distributions for each transition probability were estimated. Posterior probabilities of refracture were higher for women than for men, whilst men showed higher mortality rates. Age was observed to increase every risk, the risk of refracture, death without refracture and death after refracture. As a result, lower free-event probabilities were estimated among older patients.

Keywords: multi-state models, semi-Markov process, transition intensity, transition probability.

Poster 16 – 28 May 2021 | 13:00-13:50 CEST

Hierarchical models for Bayesian inference of brain activity from Magneto/Electro-encephalography data

Alessandro Viani [(1)], Gianvittorio Luria [(1)], Harald Bornfleth [(2)], Alberto Sorrentino [(1)]

(1) Dipartimento di Matematica, Università di Genova, Italy, (2) BESA GmbH, Gräfelfing, Germany.

viani@dima.unige.it; luria@dima.unige.it; harald.bornfleth@besa.de; sorrentino@dima.unige.it

Abstract: We present a powerful generalization of a previously described model and algorithm for bayesian inference in hierarcical models making usage of Rao-Blackwellized Sequential Monte Carlo (SMC). We discuss and analyze the consequences, in a full Bayesian approach, related with the introduction of a specific hyperprior on the standard deviation of a set of conditionally linear Gaussian (CLG) variables.
The main result is the prospect to remove the limitations imposed on the model by the choice of Gaussian priors without losing the well known benefits related with CLG models. The proposed generalization is applied in the contest of the estimation of brain activity from magneto/electro-encephalographic (M/EEG) data, providing numerical results obtained on a set of synthetic datasets for M/EEG and an experimental dataset from MNE-Python package. The comparison with the previously described algorithm shows how the approximation to the posterior distribution using the proposed generalization remains extremely stable under a large scale of values for the hyperparameter, while the same analysis managed on the classical algorithm provides divergent results, invalidated by a wrong choice of the parameter.
Concluding, obtained results show that the generalization virtually removes the dependence on the hyperparameter, whose estimation in real applications is often non trivial. Moreover in the analyzed application it allows to truly reconstruct multiple dipoles and complex configurations at the same time guaranteeing uncertainty quantification.

Keywords: Bayesian inverse problems; hierarchical models; hyperprior; M/EEG; conditionally Gaussian models; sequential Monte Carlo.

Poster 17 – 28 May 2021 | 13:00-13:50 CEST

Bayesian Time-Varying Tensor Vector Autoregressive Models for Dynamic Effective Connectivity

Wei Zhang (1), Michele Guindani (2), Sonia Petrone (1), Ivor Cribben (3)

(1) Bocconi University, Italy, (2) University of California, Irvine, USA, (3) University of Alberta, Canada.

wei.zhang@unibocconi.it; mguindan@uci.edu; sonia.petrone@unibocconi.it; cribben@ualberta.ca

Abstract: Recent developments in functional magnetic resonance imaging (fMRI) investigate how some brain regions directly influence the activity of other regions of the brain *dynamically* throughout the course of an experiment, i.e. the so-called dynamic effective connectivity. Time-varying vector autoregressive (TV-VAR) models have been employed to draw inferences for this purpose, but they are very computationally intensive, since the number of parameters to be estimated increases quadratically with the number of time-series. In this paper, we propose a computationally efficient time-varying Bayesian VAR approach for modeling high-dimensional time series. The proposed framework employs a tensor decomposition for the VAR coefficient matrices at different lags. Dynamically varying connectivity patterns are captured by assuming that at any given time the VAR coefficient matrices are obtained as a mixture of only an active subset of components in the tensor decomposition. Latent binary time-series select the active components at each time via a convenient Ising prior specification.
Sparsity-inducing priors are employed to allow for global-local shrinkage of the coefficients, to determine automatically the rank of the tensor decomposition and to guide the selection of the lags of the auto-regression. The proposed prior structure encourages sparsity in the tensor structure and allows to ascertain model complexity through the posterior distribution. We show the performances of our model formulation via simulation studies and data from a real fMRI study involving a book reading experiment.

Keywords: Time-varying vector autoregressive models; Tensor factorization; Brain effective connectivity.

Invited talk 6 – 28 May 2021 | 14:00-14:40 CEST

Bayesian Network Inference With Uncertain Evidence and Parameters

Paul Pao-Yen Wu [1,3,4], Fabrizio Ruggeri [2,3], Kerrie Mengersen [1,3,4]

(1) Queensland University of Technology, Australia, (2) CNR-IMATI, Italy (3) ARC Centre of Excellence in Mathematical and Statistical Frontiers (ACEMS), Australia (4) Centre for Data Science (CDS), Australia

p.wu@qut.edu.au; fabrizio@mi.imati.cnr.it; k.mengersen@qut.edu.au

Abstract: Bayesian Networks (BNs) have been widely applied across a wide range of domains including environmental, medical, information technology domains, and as expert systems. However, although the method represents relationships between factors in a network via conditional probabilities, the parameters of the model are inherently deterministic. In addition, uncertainty associated with observations that can be used for model learning and inference, referred to as evidence, is not captured to the extent typical of Bayesian statistics. Typically classified as hard, virtual and soft evidence, existing approaches tend to represent evidence as single observations in the context of inference. In the case of virtual and soft evidence, uncertainty about the observation of a variable is represented conceptually as a discrete or continuous distribution, independent of other variables. However, uncertainty can stem from many sources including measurement error, uncertainty around model parameters and uncertainty in expert knowledge. Computation of full posterior distributions that incorporate this uncertainty in a Bayesian manner can address many of these challenges and enable better communication for decision support.

We discuss the challenges of full Bayesian inference in a BN with potentially multiple observations in an arbitrary network of factors. The likelihood rapidly becomes intractable with even comparatively simple BNs. In addition, there is a key trade-off between computational efficiency and tractability, a key benefit of BN methodology, and rigorous treatment of uncertainty. One potential approach is to update conditional probabilities given observations as is done with BN learning, of which the most common method is to use Expectation Maximisation (EM). This can be done in a more computationally efficient manner by combining traditional BN inference with Gibbs sampling based on Markov blankets of the network. We demonstrate the utility of this preliminary approach on a number of case studies and discuss future extensions for inference under uncertainty for simple to complex systems modelled using BNs.

Keywords: Bayesian Network; evidence; measurement error; Gibbs.

Discussant: David Banks, Duke University, USA

Invited talk 7 – 28 May 2021 | 14:45-15:25 CEST

Bayesian forecasting dynamic models under attacks

Roi Naveiro [1]

(1) Institute of Mathematical Sciences (ICMAT-CSIC), Spain.

roi.naveiro@icmat.es

Abstract: The last decade has seen the rise of Adversarial Machine Learning (AML). AML studies how an intelligent adversary can manipulate data to fool inference engines, and how to protect those systems against such attacks. The ultimate goal of AML is to produce inference systems that are robust against data manipulation attacks that an adversary may produce. Thus, having relevant models of how an adversary might modify input data to an inference system is key to guarantee protection against adversarial attacks.
In the last few years, several attacking models to classification and regression systems has been released. Probably, the most popular ones target computer vision. However, relevant data manipulation attacks have appeared in other fields including fake news detection and spam detection. Much less attention has been paid to attacks to time series forecasting models. In this talk, we introduce a decision analysis based attacking strategy that could be utilized against Bayesian dynamic forecasting models. We illustrate the proposed attacking strategy using a simulation study. Finally, we provide several directions of future work.

Keywords: Bayesian forecasting; Adversarial machine learning; Poisoning attacks; Bayesian model monitoring.

Discussant: Mike West, Duke University, USA

Invited talk 8 – 28 May 2021 | 15:45-16:25 CEST

Quasi-stationary Monte Carlo methods via stochastic approximation

Andi Q. Wang [(1)], *Murray Pollock* [(2)], *Gareth Roberts* [(3)], *David Steinsaltz* [(4)]

(1) University of Bristol, UK, (2) Newcastle University, UK, (3) University of Warwick, UK and (4) University of Oxford, UK

andi.wang@bristol.ac.uk; murray.pollock@newcastle.ac.uk; Gareth.O.Roberts@warwick.ac.uk;steinsal@stats.ox.ac.uk

Abstract: Quasi-stationary distributions describe the long-term behaviour of an absorbed stochastic process, conditioned on non-absorption. Such objects arise naturally in the study of population dynamics, but have very recently been used to construct scalable MCMC methods to perform exact Bayesian inference on tall datasets. I will review this recent development, and then describe an alternative methodological approach and subsequent extensions, based on stochastic approximation techniques to simulating quasi-stationary distributions.

Keywords: Monte Carlo methods, quasi-stationary distribution, diffusion, stochastic approximation.

Discussant: Giacomo Zanella, Università Bocconi, Italy

Invited talk 9 – 28 May 2021 | 16:30-17:10 CEST

Compositions of discrete random probabilities for inference on multiple samples

Giovanni Rebaudo[(1)], Augusto Fasano[(2)], Antonio Lijoi [(3)], Igor Prünster [(3)]

(1) University of Texas at Austin, USA (2) University of Turin, Italy, (3) Bocconi University, Italy

rebaudo.giovanni@gmail.com

Abstract: Bayesian hierarchical models have proved to be an effective tool when observations are from different populations or studies, since they naturally allow borrowing information across groups, while allowing the subjects in the same group to share the same unknown distribution. We consider models induced by compositions of discrete random probabilities measures, most notably Pitman-Yor processes. Such compositions are well suited to account for both clustering of populations and clustering of observations. We identify an analytical expression of the distribution of the induced random partition, and this allows us to gain a deeper insight into the theoretical properties of the model while deriving predictive distributions and urn schemes. The proposed models can be used as a building block for addressing problems of density estimation, prediction with species sampling data and testing of distributional homogeneity. The theoretical results further lead us to devise novel MCMC sampling schemes whose effectiveness will be discussed through illustrative examples involving simulated and real data.

Keywords: Bayesian Nonparametrics; Species Sampling Processes; Partial Exchangeability; Random Partitions.

Discussant: Bernardo Nipoti, Università di Milano-Bicocca, Italy

Invited talk 10 – 28 May 2021 | 17:15-17:55 CEST

Bayesian Conditional Auto-Regressive LASSO Models to Learn Sparse Networks with Predictors

Yunyi Shen [(1)], Claudia Solis-Lemus [(1)].

(1) University of Wisconsin-Madison, USA

yshen99@wisc.edu; solislemus@wisc.edu

Abstract: We introduce a novel Bayesian conditional auto-regressive (CAR) LASSO model to infer a sparse network structure with nodes for responses and for predictors with applications to microbiome research. Directed edges between a predictor and a response represent conditional links, and undirected edges among responses represent correlations. Specifically, our model can estimate a microbial network that represents the dependence structure of a multivariate response (e.g. abundances of microbes) while simultaneously estimating the effect of a set of predictors that influence the network (e.g. diet, weather, experimental treatments). In addition, our method produces a sparse interpretable graph via LASSO penalization which can become adaptive so that different shrinkage can be applied to different edges. Furthermore, our model is able to equally handle small and big data and is computationally inexpensive through an efficient Gibbs sampling algorithm. With hierarchical structure, we extend the model to binary, counting and compositional responses by adding an appropriate sampling distribution to the core Normal model.

Keywords: Linear regression; Compositional data; Interaction network; Graphical model.

Discussant: Nicholas Polson, University of Chicago, USA