

Bayesian Nonparametrics for Complex Data - Concluding workshop

Department of Statistical Sciences, University of Padova, Room SC60

January 24, 2020

Project funded by



**SUPPORTING TALENT IN RESEARCH
@UNIVERSITY OF PADOVA**

Contents

Program	1
Abstracts	3
Bayesian nonparametric functional mixture modelling to uncover neurocardiovascular profiles in older Irish adults (<i>Bernardo Nipoti</i>)	3
Asymptotically exact variational Bayes for high-dimensional binary regression models (<i>Daniele Durante</i>)	3
Fast grouped posterior approximation for nonparametric density estimation (<i>Jan van Waaij</i>)	4
Graphical model selection for air quality time series (<i>Lucia Paci</i>)	4
Multiscale stick-breaking mixture models (<i>Marco Stefanucci</i>)	5
Expectation propagation for generalised quantile regression models (<i>Mauro Bernardi</i>)	5
Convex mixture regression for quantitative risk assessment controlling for confounders (<i>Arianna Falcioni</i>)	6
A convex mixture model for binomial regression (<i>Luisa Galtarossa</i>)	6
Bayesian models for environmental data analysis: a case study on wind speed in the Veneto area damaged by Vaia storm (<i>Sara Ceschin</i>)	7
Efficient posterior sampling for Bayesian log-linear models (<i>Laura D'Angelo</i>)	7
Composite mixture of loglinear model for multivariate categorical data (<i>Emanuele Aliverti</i>)	8
Importance conditional sampler for Pitman-Yor and GM-DDP mixtures (<i>Riccardo Corradin</i>)	8
Simultaneous transformation and rounding models for integer-valued data (<i>Antonio Canale</i>)	9
Author Index	11

Program

Invited speakers 1	Bernardo Nipoti	Bayesian nonparametric functional mixture modelling to uncover neurocardiovascular profiles in older Irish adults	9.00 - 10.30
	Daniele Durante	Asymptotically exact variational bayes for high-dimensional binary regression models	
	Jan van Waaij	Fast grouped posterior approximation for nonparametric density estimation	
Coffee break			10.30 - 11.00
Invited speakers 2	Lucia Paci	Graphical model selection for air quality time series	11.00 - 12.30
	Marco Stefanucci	Multiscale stick-breaking mixture models	
	Mauro Bernardi	Expectation propagation for generalised quantile regression models	
Lunch break			12.30 - 14.00
Junior session (talks by PhD and MSc students)	Arianna Falcioni	Convex mixture regression for quantitative risk assessment controlling for confounders	14.00 - 15.30
	Luisa Galtarossa	A convex mixture model for binomial regression	
	Sara Ceschin	Bayesian models for environmental data analysis: a case study on wind speed in the Veneto area damaged by Vaia storm	
	Laura D'Angelo	Efficient posterior sampling for Bayesian log-linear models	
Coffee break			15.30 - 16.00

Invited speakers 3	Emanuele Aliverti	Composite mixture of loglinear model for multivariate categorical data	16.00 - 17.30
	Riccardo Corradin	Importance conditional sampler for Pitman-Yor and GM-DDP mixtures	
	Antonio Canale	Simultaneous transformation and rounding models for integer-valued data	

Abstracts

Bayesian nonparametric functional mixture modelling to uncover neurocardiovascular profiles in older Irish adults

Bernardo Nipoti

University of Milano Bicocca

Invited
speakers 1
9.00 - 10.30

Cardiovascular ageing is one of the principal causes of physical and cognitive disability and mortality. Impaired blood pressure regulation with age is known to influence functional decline. As part of a larger project, which is investigating the relationship between ageing and its links to cardiovascular and neurocardiovascular functioning, we use a Bayesian nonparametric functional mixture model to uncover groups which might identify the major distinct neurocardiovascular profiles in older Irish adults aged 50+. Data were taken from The Irish Longitudinal Study on Ageing (TILDA) and consist of different measurements collected during the so called active-stand experiment.

Joint work with: Belinda Hernandez.

Asymptotically exact variational Bayes for high-dimensional binary regression models

Daniele Durante

Bocconi University

State-of-the-art methods for Bayesian inference on regression models with binary responses are impractical or inaccurate in high dimensions. To cover this gap we propose a novel variational approximation for the posterior distribution of the coefficients in high-dimensional probit regression. Our method leverages a representation with global and local variables but, unlike for classical mean-field assumptions, it avoids a fully factorized approximation, and instead assumes a factorization only for the local variables. We prove that the resulting variational approximation belongs to a tractable class of unified skew-normal distributions that preserves the skewness of the actual posterior and, unlike for state-of-the-art variational Bayes solutions, converges to the exact posterior as the number of predictors p increases. A scalable coordinate ascent variational algorithm is proposed to obtain the optimal parameters of the approximating densities. As we show with both theoretical results and an application to Alzheimer's data, such a routine requires a number of iterations converging to 1 as p

goes to infinite, and can easily scale to large p settings where expectation-propagation and state-of-the-art Markov chain Monte Carlo algorithms are impractical.

Joint work with: Augusto Fasano and Giacomo Zanella.

Fast grouped posterior approximation for nonparametric density estimation

Jan van Waaij
University of Padova

We propose a point estimator for the probability density function of an i.i.d. sample modelled via nonparametric mixtures of kernels. Our approach is inspired by Bayesian nonparametric mixtures and relies on approximating the full posterior expectation of the probability density function by means of an average of partial posterior expectations. The result can be computed exactly without resorting to MCMC. Finite sample performance of the methods are studied by means of simulated data.

Joint work with: Antonio Canale and Bernardo Nipoti.

Invited
speakers 2
11.00-12.30

Graphical model selection for air quality time series

Lucia Paci
Cattolica University of Milan

An objective Bayes approach based on graphical models is proposed for learning dependencies among multiple air quality time series within the framework of Vector Autoregressive (VAR) models. Using a fractional Bayes factor approach, we obtain the marginal likelihood in closed form and construct an MCMC algorithm for Bayesian graphical model determination with limited computational burden. We apply our method to study the interactions between multiple air pollutants over the municipality of Milan.

Joint work with: Guido Consonni.

Multiscale stick-breaking mixture models

Marco Stefanucci

University of Padova

We introduce a family of multiscale stick-breaking mixture models for Bayesian density estimation. The Bayesian nonparametric literature is dominated by single scale methods, exception made for Polya trees and allied approaches. Our proposal is based on the introduction of an infinitely-deep binary tree of random weights that grows according to a multiscale generalization of the stick-breaking process. The multiscale stick-breaking is paired with specific stochastic processes that generate sequences of parameters that induce stochastically ordered kernel functions. Properties of this family of multiscale stick-breaking mixtures are described. Focusing on a Gaussian specification, a Markov Chain Monte Carlo algorithm for posterior computation is introduced. The performance of the method is illustrated analyzing both synthetic and real data sets. The method is well-suited for data living in \mathbb{R} and is able to detect densities with varying degree of smoothing and local features.

Joint work with: Antonio Canale.

Expectation propagation for generalised quantile regression models

Mauro Bernardi

University of Padova

L_α -quantile regression models generalise quantiles ($\alpha = 1$) and expectiles ($\alpha = 2$) regression to account for the whole conditional distribution of the response variable. We introduce the L_α -quantile regression model and we present a new Bayesian estimation framework where regression parameters are learned by minimising the expected tilted check function. An approximated model evidence is obtained by employing the Expectation Propagation (EP) algorithm. The analytically intractable integration required by the parameters learning problem is solved by minimising the Kullback–Leibler divergence between the unnormalised posterior and a suitable approximating distribution usually belonging to the exponential family. We also provide some theoretical results concerning the consistency of the posterior distribution of the regression parameters under general priors. Moreover, the model selection problem is approached through an approximated Stochastic Search Variable Selection (SSVS–EP) algorithm based on the spike–and–slab prior. The effectiveness of the proposed model and parameter learning method are assessed on synthetic and real datasets.

Junior
session
14.00-15.30

Convex mixture regression for quantitative risk assessment controlling for confounders

Arianna Falcioni

University of Padova

We propose a convex mixture regression model for quantitative risk assessment that controls for the presence of possible confounders. The purpose is to estimate the risk related to the exposition to a given dose of possibly dangerous chemical agents on a continuous health measure. The model allows to incorporate one or more confounding variables in order to take into account the effect that these variables may have on health, regardless the exposure. We present a mixture of two densities at extreme doses, both with confounders dependent means, and express the conditional densities at each intermediate dose level via a convex combination of these extremal densities. A Bayesian approach is adopted and a Gibbs sampler for posterior inference is developed. The benefits of the method are outlined in simulations and applications to real data.

Joint work with: Antonio Canale.

A convex mixture model for binomial regression

Luisa Galtarossa

University of Padova

We introduce a convex mixture regression model to infer the effect of a potentially adverse exposure on a binary health outcome. Our construction assumes two extreme probabilities of observing a negative outcome at extreme doses, and relies on a convex combination of these extremal probabilities at each intermediate dose level. Inference is conducted by means of a Bayesian approach introducing a Gibbs sampler with closed-form full conditional posterior distributions.

Joint work with: Antonio Canale.

Bayesian models for environmental data analysis: a case study on wind speed in the Veneto area damaged by Vaia storm

Sara Ceschin
University of Padova

This study focuses on extreme values analysis of environmental phenomena. Classical extreme value theory describes the behaviour of extreme events through models based just on data that are maxima or over a threshold. Zorzetto, Canale and Marani (2019) introduced a Bayesian model which uses all the available information. Wind speed measured in North East Italy has been studied to understand the nature of Vaia storm. We focused on wind speed data which exhibit a clear bimodal behaviour, that we modeled by means of a mixture model approach. Simulation studies were conducted to assess the goodness of fit of the proposed models. Then the Bayesian model was tested to estimate maxima's distribution and return periods. The method is finally applied to the historical data leading to better understanding of the events of the Vaia storm.

Joint work with: Antonio Canale and Marco Marani.

Efficient posterior sampling for Bayesian log-linear models

Laura D'Angelo
University of Padova

Poisson log-linear models are among the most popular approaches for count regression. In the Bayesian context, however, there is a lack of specific computational tools to efficiently sample from the parameters' posterior distribution, and standard algorithms, as the random walk Metropolis-Hastings or the Hamiltonian Monte Carlo, are typically used. In this work we leverage the recent Polya-Gamma data augmentation scheme to propose an efficient Metropolis-Hastings proposal to simulate from the posterior distribution of log-linear models. The key idea is to exploit the Poisson approximation to the negative binomial and to adapt the Gibbs sampler of Polson et al. (2013) to a "pseudo" Gibbs sampler for log-linear models. This allows us to derive an efficient proposal distribution for a Metropolis-Hastings algorithm. Via simulation we show that the first order efficiency of our sampler is larger than that obtained with a uniform random walk proposal for every acceptance rate; and that its time per independent sample is competitive with that obtained using a Hamiltonian Monte Carlo approach.

Joint work with: Antonio Canale.

Invited
speakers 3
16.00-17.30

Composite mixture of loglinear model for multivariate categorical data

Emanuele Aliverti
University of Padova

Multivariate categorical data are routinely collected in many application areas. As the number of cells in the table grows exponentially with the number of variables, most cells will contain zero observations. This severe sparsity motivates appropriate statistical methodologies to reduce the number of free parameters, with penalized log-linear models and latent structure analysis being popular options. In this talk we propose a fundamentally new class of methods, which combines latent class analysis and log-linear to define a novel Bayesian methodology for characterizing interactions of multivariate categorical data. The method is used to investigate the relation among psychopathological profiles on a case study involving suicide attempts.

Importance conditional sampler for Pitman-Yor and GM-DDP mixtures

Riccardo Corradin
University of Milano Bicocca

Bayesian nonparametric mixtures are flexible models for density estimation and model based clustering. Within this family of models the Pitman-Yor mixtures (PYM) show a good balance between mathematical tractability and flexibility. Inference for this class of models is mainly performed by means of MCMC methods, which can be divided into marginal and conditional methods. We introduced in literature a new class of algorithm to perform estimation of PYM models, named importance conditional sampler (ICS), which, although conditional, is reminiscent of the Pólya urn marginal scheme and conveniently shares its degree of interpretability. The performance of the ICS is investigated and compared with commonly used competitors, by means of an extensive simulation study: our proposal has proved to be robust to the specification of the parameters characterising the distribution of the underlying process. We provided also an extension of the ICS strategy to the Griffiths-Milne dependent Dirichlet process mixture models, a family of models for partially exchangeable data, where we were able to analyze in a feasible time complex structures of dependence, induced by the use of multiple variables for grouping the data.

Joint work with: Antonio Canale and Bernardo Nipoti.

Simultaneous transformation and rounding models for integer-valued data

Antonio Canale
University of Padova

We propose a simple yet powerful framework for modeling integer-valued data, such as counts, scores, and rounded data. The data-generating process is defined by Simultaneously Transforming and Rounding (STAR) a continuous-valued process, which produces a flexible family of integer-valued distributions capable of modeling zero-inflation, bounded or censored data, and over- or underdispersion. The transformation is modeled as unknown for greater distributional flexibility, while the rounding operation ensures a coherent integer-valued data-generating process. An efficient MCMC algorithm is developed for posterior inference and provides a mechanism for adaptation of successful Bayesian models and algorithms for continuous data to the integer-valued data setting. Using the STAR framework, we design a new Bayesian Additive Regression Tree (BART) model for integer-valued data, which demonstrates impressive predictive distribution accuracy for both synthetic data and a large healthcare utilization dataset. For interpretable regression-based inference, we develop a STAR additive model, which offers greater flexibility and scalability than existing integer-valued models. The STAR additive model is applied to study the recent decline in Amazon river dolphins.

Joint work with: Daniel R. Kowal.

Author Index

Aliverti
Emanuele, 8

Bernardi
Mauro, 5

Canale
Antonio, 9

Ceschin
Sara, 7

Corradin
Riccardo, 8

D'Angelo
Laura, 7

Durante
Daniele, 3

Falcioni
Arianna, 6

Galtarossa
Luisa, 6

Nipoti
Bernardo, 3

Paci
Lucia, 4

Stefanucci
Marco, 5

van Waaij
Jan, 4