Bayesian Inference in Stochastic Processes

BISP-11

Book of Abstracts

Madrid, June 12 - 14, 2019

Real Academia de Ciencias, Valverde 22

https://www.methaodos.org/congresos-methaodos/index.php/bisp11/bisp11

Real Academia de Ciencias Exactas, Físicas y Naturales
AXA Research Fund
ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

Plenary Talks

Bayesian dimensionality reduction via identifications of data intrinsic dimensions

Antonietta Mira [(1)], *Michale Allegra* [(2)], *Elena Facco* [(2)], *Alessandro Laio* [(2)], *Francesco Denti* [(3)], *Michale Guindani* [(4)]

(1) Università della Svizzera italiana and Insubria University, (2) Sissa (International School for Advanced Studies), (3) University of Milan – Bicocca, (4) University of California, Irvine

antonietta.mira@usi.ch

Abstract: 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. Let us consider the data as a configuration of a Poisson Process (PP) with an intensity proportional to the true density. TWO-NN makes only a weak assumption: locally, on the scale of the second nearest neighbor, the original PP can be well approximated by a homogeneous one. In other words, we assume that in a small neighborhood around each point the intensity of the underlying PP is approximately constant. Under this hypothesis, the ratio of the distances of a point from its first and second neighbor follows a Pareto distribution that depends parametrically only on the ID, allowing for an immediate estimation of the latter.

We have extended 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, 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. The Hidalgo model obtains remarkable results, but its limitation consists in fixing a priori the number of components in the mixture. We then let $K \to \infty$, adopting a Bayesian Nonparametric approach: we model the data with a Dirichlet Process Mixture Model as an infinite mixture of Pareto distributions. This approach can consider the uncertainty relative to the number of mixture components. Since the

posterior distribution has no closed form, to sample from it we rely on the Slice Sampler algorithm. From preliminary analyses performed on simulated data, the model provides promising results. In particular, the method reveals the ability to detect the ID of MCMC outputs, which can be used as an additional diagnostic method to test the identifiability of the entire set of parameters.

Key Words: Dimensionality reduction, Hierarchical model, Slice Sampler

Non-Gaussian spatial and spatio-temporal processes

Alexandra M. Schmidt.

Department of Epidemiology, Biostatistics and Occupational Health. McGill University

alexandra.schmidt@mcgill.ca

Abstract: In the analysis of most spatial and spatio-temporal processes in environmental studies, observations present skewed distributions, with a heavy right or left tail. Usually, a single transformation of the data is used to approximate normality, and stationary Gaussian processes are assumed to model the transformed data. Spatial interpolation and/or temporal prediction are routinely performed by transforming the predictions back to the original scale. The choice of a distribution for the data is key for spatial interpolation and temporal prediction. In this talk, I will start discussing the advantages and disadvantages of using a single transformation to model such processes. Then I will discuss some recent advances in the modeling of non-Gaussian spatial and spatio-temporal processes.

Keywords: Covariance function; Stationarity; Spatial interpolation; Temporal prediction

Replication or exploration? Sequential design for stochastic simulation experiments

Mickael Binois [(1)], *Jiangeng Huang* [(2)], *Robert B Gramacy* [(2)], *Mike Ludkovski* [(3)]

(1) Argonne National Laboratory, (2) Virginia Tech, (3) University of California, Santa Barbara

rbg@vt.edu

Abstract: We investigate the merits of replication, and provide methods that search for optimal designs (including replicates), in the context of noisy computer simulation experiments. We first show that replication offers the potential to be beneficial from both design and computational perspectives, in the context of Gaussian process surrogate modeling. We then develop a lookahead based sequential design scheme that can determine if a new run should be at an existing input location (i.e., replicate) or at a new one (explore). When paired with a newly developed heteroskedastic Gaussian process model, our dynamic design scheme facilitates learning of signal and noise relationships which can vary throughout the input space. We show that it does so efficiently, on both computational and statistical grounds. In addition to illustrative synthetic examples, we demonstrate performance on two challenging real-data simulation experiments, from inventory management and epidemiology.

Keywords: computer experiment, Gaussian process, surrogate model, input-dependent noise, replicated observations, lookahead

Real Academia de Ciencias Exactas, Físicas y Naturales
AXA Research Fund
ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

Dimensionality Reduction for Stochastic Processes

Vadim Sokolov [1], *Nick Polson*[2], *Laura Schutz* [1]

(1) George Mason University, Department of Systems Engineering and Operations Research, (2) University of Chicago

vsokolov@gmu.edu

Abstract: Bayesian algorithms such as Markov Chain Monte Carlo or Bayesian optimization can quickly become computationally prohibitive or even infeasible for high dimensional problems. In many applications, however, the underlying dynamics of a stochastic process typically can be represented in a lower dimensional space. We will review existing linear and nonlinear dimensionality reduction methods, such as Laplacian eigenmaps and restricted Boltzmann machines. Further, we will present some new results for nonlinear dimensionality techniques based on deep learning models. We will demonstrate our approach in the context of Bayesian optimization algorithms applied to a stochastic process defined by a complex agent-based model. Finally, we discuss directions for future research.

Keywords: dimensionality reduction, deep learning, Bayesian optimization

From infinity to here: a Bayesian nonparametric perspective of finite mixture models

Raffaele Argiento [(1)], *María de Iorio*[(2)]

(1) University of Torino and Collegio Carlo Alberto, (2) YaleNUS College and Department of Statistical Science, University College London

raffaele.argiento@unito.it

Abstract: Modelling via finite mixtures is one of the most fruitful Bayesian approach, particularly useful for clustering when there is unobserved heterogeneity in the data. The most popular algorithm under this approach is the reversible jump MCMC that can be nontrivial to design, especially in high-dimensional spaces. We will show how nonparametric methods can be transferred into the parametric framework. We first introduce a class of almost sure finite discrete random probability measures obtained by normalization of finite point processes. Then, we use the new class as mixing measure of a mixture model and derive its posterior characterization. The resulting class encompasses the popular finite Dirichlet mixture model. In order to compute posterior statistics, we propose an alternative to the reversible jump: borrowing notation from the nonparametric Bayesian literature, we set up a conditional MCMC algorithm based on the posterior characterization of the unnormalized point process. The flexibility of the model and the performances of our algorithm are illustrated on simulated and real data.

Keywords: Mixture models, Bayesian nonparametric, Model-based Clustering

The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions

Tamara Broderick

Computer Science and AI, MIT

tbroderick@mit.edu

Abstract: Discovering interaction effects on a response of interest is a fundamental problem faced in biology, medicine, economics, and many other scientific disciplines. In theory, Bayesian methods for discovering pairwise interactions enjoy many benefits such as coherent uncertainty quantification, the ability to incorporate background knowledge, and desirable shrinkage properties. In practice, however, Bayesian methods are often computationally intractable for even moderate-dimensional problems. Our key insight is that many hierarchical models of practical interest admit a particular Gaussian process (GP) representation; the GP allows us to capture the posterior with a vector of O(p) kernel hyper-parameters rather than O(p^2) interactions and main effects. With the implicit representation, we can run Markov chain Monte Carlo (MCMC) over model hyper-parameters in time and memory linear in p per iteration. We focus on sparsity-inducing models and show on datasets with a variety of covariate behaviors that our method: (1) reduces runtime by orders of magnitude over naive applications of MCMC, (2) provides lower Type I and Type II error relative to state-of-the-art LASSO-based approaches, and (3) offers improved computational scaling in high dimensions relative to existing Bayesian and LASSO-based approaches.

Invited talks

Distributed Bayesian Inference For Massive Scale Spatial/Spatio-temporal Data

Rajarshi Guhaniyogi [(1)], *Cheng Li* [(2)], *Terrance Savitsky* [(3)], *Sanvesh Srivastava* [(4)],

(1) University of California Santa Cruz, (2) National University of Singapore, (3) U.S. Bureau of Labor Statistics, (4) University of Iowa

rguhaniy@ucsc.edu

Abstract: Flexible hierarchical Bayesian modeling of massive data is challenging due to poorly scaling computations in large sample size settings. This talk will focus on spatial/spatio-temporal process models for analyzing geostatistical data, which typically entail computations that become prohibitive as the number of spatial locations and/or time points becomes large. We propose a three-step divide-and-conquer strategy within the Bayesian paradigm to achieve massive scalability for any spatial/spatio-temporal process model. We partition the data into a large number of subsets, apply a readily available Bayesian spatial/spatio-temporal process model on every subset in parallel, and optimally combine the posterior distributions estimated across all the subsets into a pseudo-posterior distribution that conditions on the entire data. The combined pseudo posterior distribution is used for predicting the responses at arbitrary locations and for performing posterior inference on the model parameters and the residual spatial surface. This approach offers significant advantages in applications where the entire data are or can be stored on multiple machines. Under the standard theoretical setup, we show that if the number of subsets is not too large, then the Bayes risk of estimating the true residual spatial surface using the pseudo posterior distribution decays to zero at a nearly optimal rate. A variety of simulations and a geostatistical analysis of the Pacific Ocean sea surface temperature data validate our theoretical results.

Keywords: Gaussian processes, Wasserstein Barycenter, Posterior Contraction Rate, Large Spatial/spatio-temporal data, Distributed Inference

Bayesian inference for large-scale phylogenetic multivariate probit models

Zhenyu Zhang [(1)], *Akihiko Nishimura,* [(1)] *Philippe Lemey* [(2)], *Marc A. Suchard* [(1)]

(1) University of California, Los Angeles, USA, (2) Rega Institute, KU Leuven, Belgium

zyz606@ucla.edu

Abstract: Inferring correlation among biological traits is an important topic in evolutionary biology and sheds insight into underlying disease mechanisms. Multivariate Brownian diffusion along a random evolutionary tree regularly models continuous-valued traits. Most measured traits are, however, binary. We jointly model the correlation between continuous and binary traits by projecting binary dimensions into a continuous latent space, giving rise a phylogenetic multivariate probit model. Inference under this model remains problematic with increasing sample size, requiring sampling from a high-dimensional truncated Gaussian distribution. Best current approaches scale quadratically in sample size. To overcome this barrier, we develop a bouncy particle sampler that exploits novel dynamic programming over the tree to yield samples in linear computational order. In an application, we estimate correlation between 28 HIV virulence and immunological epitope traits across 535 viruses that necessitates repeated sampling from a 14,980-dimensional truncated Gaussian. Importantly, our inference framework suits a large class of covariance structures beyond comparative biology that we characterize.

Keywords: Phylogenetics, Mixed Outcomes Types, Piecewise-Deterministic Markov Chain Monte Carlo, Dynamic Programming

Bayesian inference for leverage stochastic volatility models

Hoang Nguyen [(1)], Roberto Casarin[(2)], M. Concepción Ausín[(1, 3)], Pedro Galeano[(1, 3)]

(1) Universidad Carlos III de Madrid, Spain, (2) University Ca' Foscari of Venice, Italy, (3) UC3M-BS Institute of Financial Big Data (IFiBiD), Spain

hoang.nguyen@uc3m.es

Abstract: Stochastic volatility models are commonly used to explain the evolution of financial time series. The recent studies have concentrated on leverage effects where shocks to returns and innovations of volatilities are negatively correlated. A bivariate normal distribution has been proposed in the literature that might be restrictive due to the assumption of univariate Gaussian for both shocks and innovations. In this study, we model the leverage effect through a bivariate copula function, hence allow for more flexible univariate marginals and asymmetric dependences. We compare the current advanced algorithms in Bayesian inference such as Sequential Monte Carlo (SMC) and Sequential Variational Inference (SVI). The SVI approach is not only accurate but also reduces the computational burden. We illustrate the proposed models with different financial data sets.

Keywords: leverage stochastic volatility, Sequential Monte Carlo, Sequential Variational Inference

Variational Inference for Diffusion Proceses

Thomas Ryder, Andrew Golightly, A. Stephen McGough, Dennis Prangle

Newcastle University

t.ryder2@newcastle.ac.uk

Abstract: Parameter inference for stochastic differential equations can be challenging due the intractability of the underlying diffusion process. Working with the Euler Maruyama discretisation for the diffusion, we use advances in variational inference to jointly learn the parameters and the diffusion paths. Using a standard mean-field variational approximation of the parameter posterior, we investigate recurrent neural networks and other methods to approximate the posterior for the diffusion paths conditional on the parameters. These approaches learn how to bridge between observations in a very similar way to a conditioned diffusion. The resulting black-box inference method can be applied to any SDE system with light tuning requirements.

Keywords: Variational Inference, Diffusion Processes, Machine Learning

Threatened Markov Decision Processes

Víctor Gallego, Roi Naveiro

Institute of Mathematical Sciences (ICMAT-CSIC), Spain

roi.naveiro@icmat.es

Abstract: Markov Decision Processes (MDPs) are a long-standing tool for modeling decision making within stochastic processes where some variables are under the control of the decision maker. In In the context of MDPs, the main objective is to maximize the expectation of the discounted utility. In Reinforcement Learning (RL) settings, Q-learning is the dominant computational approach for solving MDPs.

Recently, due to the rise of machine learning applications in security settings, a new paradigm is being faced in MDPs: how do we model decision making when there is an adversary that interferes with the reward generating process? Clearly, a new framework that allows reasoning about opponents within MDPs becomes necessary.

In this paper, we introduce Threatened Markov Decision Processes (TMDPs), an extension of MDPs that incorporates the possibility of explicitly modeling opponents interfering with the reward generating process. In addition, we provide a modification of the standard Q-learning rule to solve TMDPs in RL settings.

We propose two approaches to make inference about the opponent decision making process: the first one is the Bayesian version of the standard fictitious play in Game Theory; the second is a level-k thinking scheme. Finally, we propose a Bayesian approach to learn about the opponent type when there is uncertainty about it, as is the case in most security settings.

Empirical evidence is given supporting the proposed framework and showing its benefits in adversarial RL settings.

Keywords: Markov Decision Processes, Bayesian Inference, Agent modeling, Security Games

An Adversarial Risk Analysis Approach for Differential Games: A Botnet Defense Model

Jorge González-Ortega, Antonio Gómez-Corral

Institute of Mathematical Sciences (ICMAT-CSIC), Spain

jorge.gonzalez@icmat.es

Abstract: We consider Differential Games (DGs) corresponding to conflict situations in which players choose strategies over time. In this context, typical applications include defense, counter-terrorism and finance. Attempts to solve DGs have focused mostly on Game Theory (GT) based on variants of Nash equilibria, but this is not satisfactory in many of the above applications since beliefs and preferences of adversaries will not be readily available, frequently violating GT common knowledge assumptions. Alternatively, Adversarial Risk Analysis (ARA) supports one of the players (defender), minimizing her subjective expected costs, treating the adversaries' (attackers') decisions as random variables. To do so, ARA models the attackers' decision-making problems and, under assumptions about their rationality, tries to assess their probabilities and utilities, acknowledging uncertainty, to predict their (random) optimal actions. Our approach is illustrated through a botnet defense example with an underlying SIS epidemic model (Bensoussan et al., 2010), in which we compare both the GT and ARA solutions.

Keywords: Non-cooperative games, Decision analysis, Nonzero-sum games, Cybersecurity.

Understanding Priors in Bayesian Neural Networks at the Unit Level

Julyan Arbel [(1)], *Mariia Vladimirova* [(1)], *Jakob Verbeek* [(1)], *Pablo Mesejo* [(2)]

(1) Inria Grenoble Rhône-Alpes, France, (2) University of Granada, Spain

julyan.arbel@inria.fr

Abstract: We investigate deep Bayesian neural networks with Gaussian weight priors and a class of ReLU-like nonlinearities. Bayesian neural networks with Gaussian priors are well known to induce an L^2, weight decay, regularization. Our results characterize a more intricate regularization effect at the level of the unit activations. Our main result establishes that the induced prior distribution on the units before and after activation becomes increasingly heavy-tailed with the depth of the layer. We show that first layer units are Gaussian, second layer units are sub-exponential, and units in deeper layers are characterized by sub-Weibull distributions. Our results provide new theoretical insight on deep Bayesian neural networks, which we corroborate with experimental simulation results on convolutional networks.

Keywords: Bayesian neural network, sparsity, heavy-tailed distribution

Tajima heterochronous coalescent.

Lorenzo Cappello [1], Amandine Vebe [2], Julia A. Palacios[1]

(1) Stanford University, USA, (2) École Polytechnique, France

cappello@stanford.edu

Abstract: The observed variation in molecular samples modeled with Kingman's n-coalescent allows to infer evolutionary parameters such as past population dynamics. In this framework, such a variation is modeled as a mutation process superimposed on a stochastic genealogy sampled from the Kingman n-coalescent. However, the state space of Kingman's genealogies grows superexponentially; thus inference is computationally unfeasible already for small sample sizes. An alternative to Kingman coalescent has been proposed in the literature, the Tajima n-coalescent, which relies on a coarser genealogical resolution, reducing the state space substantially. Such process does not accommodate samples collected at different times, a situation that in applications is both real (e.g. ancient DNA, influenza viruses) and desirable since it reduces the variance of the estimate. In order to fill this gap, we introduce a new process, called Tajima heterochronous n-coalescent, define the exact likelihood of observed mutations given a Tajima's genealogies, and present a Bayesian nonparametric procedure to infer past population size. We propose a new sampler to explore the space of Tajima's genealogies. We compare our procedure with state-of-art algorithms.

Keywords: Gaussian processes, Coalescent, Effective population size.

New Inference for Spatio-temporal Point Processes

Erin M. Schliep [(1)], *Alan E. Gelfand* [(2)]

(1) University of Missouri, (2) Duke University

schliepe@missouri.edu

Abstract. We expand inference for stochastic point processes defined over space and time by introducing the notion of velocity. For a realization of a stochastic point process, we obtain the instantaneous gradient of the intensity surface in time and space for a given location and time. The ratio of these two gradients provides a velocity in that direction at that location and time, i.e., speed of change in intensity in that direction. This velocity can be interpreted in terms of speed of change in chance for an event. We apply our methodology to a spatio-temporal point pattern of theft events in San Francisco in 2012. We specify a log Gaussian Cox process model to explain the events and use a sparse Gaussian process model approximation to expedite the demanding computation for model fitting and gradient calculation. We estimate the velocity of the point pattern, where the magnitude and direction of the minimum velocity provides the slowest rate and direction of movement required to maintain constant chance for an event.

Keywords: Directional derivative processes; Log Gaussian Cox process; Markov chain Monte Carlo; Nearest neighbor Gaussian process.

Monte Carlo prediction intervals for dynamic queueing systems

Anthony Ebert

Queensland University of Technology

ac.ebert@qut.edu.au

Abstract. Dynamic queueing systems model queues where demand varies strongly with time, such as airport terminals. The complexity of these models implies that posterior inference generally requires a likelihood-free approach. Based on posterior samples from approximate Bayesian computation, we develop decision support tools for managing dynamic queueing networks. We demonstrate these tools with real data from an airport terminal combined with a realistic scenario of a delayed flight.

Keywords: Dynamic queues, likelihood-free methods, decision support.

Non-Gaussian Translation Processes in Stochastic Differential Equation and Spatio-temporal Models

Robert Richardson

Brigham Young University, USA

richardson@stat.byu.edu

Abstract: Gaussianity and linearity are two of the most common assumptions made to simplify modeling of stochastic processes that evolve over time. When both are present it can severely complicate parameter estimation and inference of existing modeling approaches. A non-Gaussian translation process, which is essentially a Gaussian copula process used in reverse, can be used to adjust Brownian motion in a stochastic differential equation (SDE) or more generally, any marginal distribution specification of any stochastic process. Through a simple correction factor, a non-Gaussian SDE can be converted into a non-linear SDE, including one that is already non-linear, resulting in a simplified model. Similarly, this allows for more flexibility in non-Gaussian spatial and spatio-temporal models and spatial GLMs.

Modelling the Proliferation of Terrorist Activity

Gentry White, Michael D. Porter, Fabrizio Ruggeri

Queensland University of Technology, University of Virginia, CNR IMATI

gentry.white@qut.edu.au

Abstract: In this paper we present a Bayesian hierarchical model for a convolution of non-homgeneous Poisson point process and a self-exciting negative-binomial point process where the excitation is a function of the number of fatalities in each event. This convolution represents two distinct mechanisms for the proliferation of terrorist activity over time. The hierarchical structure of the resulting MCMC scheme allows for the likelihood of this convolution to be evaluated as a single Poisson process by defining the negative-binomial process as conditional Poisson point process with a gamma-distribution for its expectation. This model is applied to data from Iraq, Afghanistan and Israel over the period from 2000-2016.

Keywords: Self-exciting Point Process, Terrorism, Contagion, Diffusion

Contributed talks

The semi-Markov beta-Stacy process

Andrea Arfé [(1)], Stefano Peluso [(2)], Pietro Muliere (presenter) [(1)]

(1) Università Bocconi, Milan, Italy, (2) Università Cattolica di Milano, Italy

pietro.muliere@unibocconi.it

Abstract: We introduce the semi-Markov beta-Stacy process (SMBS), a novel stochastic process useful for the Bayesian non-parametric analysis of semi-Markov processes. We show that the SMBS process is conjugate with respect to data generate from the observation of a semi-Markov process for a finite time, a property which greatly simplifies the task of performing probabilistic predictions. We show that the predictive distributions associated to the SMBS process can be characterized as a reinforced random walk on a specific system of urn.

A Bayesian Hedonic Regression Model for Art Price

Enrique ter Horst

Universidad de Los Andes, Colombia

ea.terhorst@uniandes.edu.co

Abstract: Hedonic pricing models are based on the premise that the prices of marketed goods are related to their attributes. The traditional OLS regression applied to hedonic pricing models assumes that, when using time series, the estimated coefficients with respect to each of the attributes remain constant. This does not have to be the case. We propose a Bayesian dynamic estimation of the hedonic regression model in which the estimated coefficients are time-varying and apply it to art prices. We find that, using a sample of 27,124 paintings sold at auction by 63 Pop artists (2001-2013), the estimated coefficients from the dynamic regression model fluctuate noticeably through time, and also that certain types of artworks, which might be regarded as "safer", declined in price by less than less safer paintings during the financial crisis (2008-2009). We also estimated a Pop-Art price index, finding that, in the semesters prior to the crisis, Pop-Art prices increased much faster than what an OLS estimation would have suggested.

Adversarial α-divergence Minimization for Bayesian Approximate Inference

Simón Rodríguez [1], *Daniel Hernández Lobato* [2]

(1) Instituto de Ciencias Matemáticas, (2) Universidad Autónoma de Madrid

simon.rodriguez@icmat.es, daniel.hernandez@uam.es

Abstract: In the Bayesian approach to machine learning a task of critical importance is doing inference about the potential values of the latent variables of the model given the observed data. Typically, this task involves computing a posterior distribution for the latent variables using Bayes' rule. A difficulty is, however, that computing this posterior distribution is intractable in most practical situations. Therefore, one has to resort to approximate inference methods. A popular technique for this task is variational inference (VI). VI is based on minimizing the Kullback-Leibler (KL) divergence between an approximate distribution q and the target posterior. By solving this optimization problem, q is enforced to look similar to the posterior. Notwithstanding, a limitation of VI is that it is restricted to work with parametric distributions q, which may suffer from strong approximation bias. A simple way to increase the flexibility of q is to let this distribution be obtained as a non-linear parametric transformation (i.e., a multi-layer perceptron) applied to a source of white Gaussian noise. If the non-linearity is complex enough, almost any distribution can be represented like this. The problem is that, although sampling from such a distribution is relatively easy, it will lack a closed form expression for the resulting density. This prevents the use of VI, which needs to evaluate the log-ratio between q and the prior over the latent variables p. To solve this problem, a technique recently proposed is adversarial variational Bayes (AVB). In AVB the log-ratio between the approximate and the prior distribution is estimated as the optimal decision boundary of a classifier that discriminates between samples of these two distributions. In this talk, we propose to extend AVB to minimize, instead of the KL-divergence, the α-divergence, in an approximate way, where $\alpha \in (0, 1]$ is a parameter of the divergence. The α-divergence is a generalization of the KL-divergence. In particular, when $\alpha \to 0$, the α-divergence tends to the KLdivergence optimized by VI. By contrast, if $\alpha = 1$, the α-divergence is the reversed KL-divergence, i.e., the KL-divergence between the exact

posterior and q. This reversed form is approximately optimized by other algorithms for approximate inference such as expectation propagation. The method proposed can in consequence be seen as a generalization of AVB that allows to optimize a more general class of divergences, resulting in flexible approximate distributions q with different properties. We have evaluated such a method in the context of Bayesian neural networks. The experiments carried out (involving several regression problems extracted from the UCI repository) show that one can obtain better prediction results than those of AVB and VI, in terms of the mean squared error and the test log-likelihood, by using intermediate values of α.

A Gaussian Process Model for Multi-class Classification with Noisy Input Measurements

Bryan Zaldivar [1, 2], Cesar Garrido Merchán [1], Carlos Villacampa-Calvo Universidad [1] *Daniel Hernández Lobato*[1]

(1) Universidad Autónoma de Madrid, (2) Instituto de Física Teórica bryan.zaldivarm@uam.es, eduardo.garrido@uam.es, carlos.villacampa@uam.es, daniel.hernandez@uam.es

Abstract: In scientific experiments, measurements always come with associated uncertainties. Consequently, when performing classification tasks on the corresponding datasets one should deal with noise in a statistically consistent manner, not only in the output space (as usual) but also in the input space. Importantly, the level of noise associated with the input space is sometimes known beforehand, which means that it can be readily incorporated into the modeling process of the observed data. To our knowledge, input noise has not been properly addressed in the machine learning literature so far in the context of multi-class classification. In this talk, we describe a principled approach to account for input noise in this type of learning problems. With this goal, we describe a multi-class Gaussian process (GP) classifier in which the observed attributes may have been contaminated with additive Gaussian noise. For this, we introduce in our model extra latent variables that represent the actual noiseless input measurements. A difficulty, however, is that exact Bayesian inference in such a model is intractable and one has to resort to approximate techniques in practice. We describe an efficient algorithm based on variational inference that can be used to approximate the required computations. This algorithm is scalable which means that it allows to address very large datasets. We plan to compare this model with standard methods for multi-class GP classification in synthetic and real-world experiments, the latter involving data coming from an open problem in the context of astrophysics, which motivated the present work.

Approximate Inference in Deep Gaussian Processes by Minimizing Alpha Divergences

Carlos Villacampa-Calvo, Gonzalo Hernández Muñoz, Daniel Hernández Lobato

Universidad Autónoma de Madrid

carlos.villacampa@uam.es, gonzalo.hernandez@uam.es, daniel.hernandez@uam.es

Abstract: Deep Gaussian processes (DGPs) are a multi-layer hierarchical generalization of Gaussian processes (GPs) that have been shown to provide better generalization performance and better uncertainty estimation of the target values. Notwithstanding, exact inference is infeasible in DGPs, even for regression problems, since the predictive distribution in the second layer and above is no longer Gaussian due to the randomness of the inputs coming from the previous layer. Therefore, in practice, when using these models, one has to resort to approximate inference techniques. Most works in the literature address this problem by using Variational Inference (VI) as the underlying approximate algorithm, while a few authors have considered the use of Expectation Propagation (EP). VI and EP (in an approximate way) target different versions of the Kullback-Leibler (KL) divergence (i.e., the direct and the reversed version of this divergence) to enforce that the approximate distribution looks similar to the exact posterior distribution. As an alternative, in this paper we propose to approximately minimize the α-divergence for approximate inference in DGPs. The α-divergence can be seen as a generalization of the KL divergence and by changing the α parameter one can interpolate between the KL divergence used in VI (i.e., $\alpha \to 0$) and the KL divergence used in EP (i.e., $\alpha = 1$). With the goal described, we propose a modified version of Power Expectation Propagation (PEP), where we approximate the output distribution at each layer of the DGPs using Monte Carlo samples. The proposed method generalizes VI and EP and allows to interpolate between these two techniques simply by varying the α parameter. Preliminary experiments on 8 regression UCI datasets show that we can obtain better results for DGPs in terms of the test log-likelihood if we choose higher values of α, i.e., $\alpha \approx 1$. On the other hand, if we are interested in minimizing the mean squared error of the DGP we show that it is better to consider intermediate values of α, i.e., values close to 0.5.

Bayesian inference for controlled branching processes: ABC methodology.

Miguel González, Rodrigo Martínez, Carmen Minuesa, Inés M. del Puerto

University of Extremadura, Spain

mvelasco@unex.es

Abstract: Controlled branching processes are stochastic growth population models in which the number of individuals with reproductive capacity in each generation is controlled by a random control function. The behaviour of these processes is strongly related to their main parameters. In practice, these values are unknown and their estimation is necessary. The purpose of this work is to examine the Approximate Bayesian Computation (ABC) methods and to propose appropriate summary statistics for them in the context of these processes. This methodology enables to approximate the posterior distributions of the parameters of interest satisfactorily without explicit likelihood calculations and under a minimal set of assumptions. In particular the tolerance rejection algorithm, the sequential Monte Carlo ABC algorithm, and a post-sampling correction method are provided. The accuracy of the proposed methods are illustrated by the way of a simulated example developed with the statistical software R.

Acknowledgements: This research has been supported by the Ministerio de Economía y Competividad (grant MTM2015-70522-P), the Junta de Extremadura (grants IB16099 and GR18103) and the Fondo Europeo de Desarrollo Regional.

Approximate Bayesian inference using the cross-entropy method

Dennis Prangle

Newcastle University

dennis.prangle@newcastle.ac.uk

Abstract: For complex models, exact Bayesian inference or near-exact Monte Carlo methods can be infeasibly computationally expensive. Approximate methods are a common alternative. One popular method is variational inference (VI), which uses optimisation to select an approximate posterior from a family of candidate distributions. The optimisation objective is the reverse Kullback-Leibler (KL) divergence (measuring information lost when the posterior is used in place of the approximation). This choice is computationally convenient but produces results with an undesirable property: they tend to be over-concentrated compared to the correct posterior.

Using the forward KL divergence would be preferable. It tends to produce under-concentrated approximate posteriors, which can be corrected using importance sampling or other Monte Carlo methods. However this objective function is not computationally tractable using standard VI methods.

This work describes how optimisation of the forward KL divergence is possible using the cross-entropy method, a method for rare-event simulation. Adapting recent advances in VI to this setting provides an effective algorithm for approximate Bayesian inference. The method is illustrated for complex models in both likelihood-based and likelihood-free settings.

Keyworks: Variational inference, approximate inference, ABC

Semiparametric Bayesian Forecasting with an Application to Stochastic Volatility

Martina Danielova Zahariev[(1)] *and Fabian Goesslin* [(2)]

(1) Erasmus University Rotterdam, (2) Private industry

zaharieva@eur.ese.nl

Abstract: We propose a new and highly flexible Bayesian sampling algorithm for nonlinear state space models under nonparametric distributions. The estimation framework combines a particle filtering and smoothing algorithm for the latent process with a Dirichlet process mixture model for the error term of the observable variables. In particular, we overcome the problem of constraining the models by transformations or the need for conjugate distributions. We use the Chinese restaurant representation of the Dirichlet process mixture, which allows for a parsimonious and generally applicable sampling algorithm. Thus, our estimation algorithm combines a pseudo marginal Metropolis Hastings scheme with a marginalized hierarchical semiparametric model. We test our approach for several nested model specifications using simulated data and provide density forecasts. Furthermore, we carry out a real data example using S&P 500 returns.

Keywords: Bayesian Nonparametrics, Particle Filtering, Stochastic Volatility, MCMC, Forecasting

Precision Aggregated Local Models

Adam M. Edwards and Robert B. Gramacy

Abstract: Gaussian Process (GP) models have long been used as a flexible method for nonparametric regression. Despite their accuracy, they are infeasible for larger data sets due to the scaling of the computational burden (O N3) and the storage requirements (O N2). Typical methods to combat intractability have focused on splitting the larger GP into smaller problems that can be treated as functionally independent either by partitioning the domain space [Rushdi et al. (2016)], or partitioning the data themselves into discrete sets [Tresp (2000)]. These methods get around the functional dependence of the individual GP models they use by assigning no weight to most of the models during prediction, or exploiting global models learning theoretically the same surface respectively, however each solution has its drawbacks. Partition models maintain accuracy, but lose continuity at the boundaries [Park et al. (2011), Park and Huang (2016)]. Typical averaged models, on the other hand, maintain absolute continuity in both the mean and variance surface while oversmoothing the function as a whole. Using Local Approximate Gaussian Processes (LAGP) [Gramacy and Apley (2013)] as a method to build local experts, the Precision Aggregated Local Models (PALM) approach bridges the gap between these two approaches to create a locally accurate global model that maintains absolute continuity. We address the functional dependence of individual GP models, creating a unified surface with drastically reduced computational burden O N2 C + NC n 3 C , and storage requirement O N2 C + NC n 2 C , where NC , nc N.

References

Gramacy, R. B. and Apley, D. W. (2013). “Local Gaussian process approximation for large computer experiments.” arXiv:1303.0383 [stat]. ArXiv: 1303.0383.

Park, C. and Huang, J. Z. (2016). “Efficient Computation of Gaussian Process Regression for Large Spatial Data Sets by Patching Local Gaussian Processes.” Journal of Machine Learning Research, 17, 174, 1–29.

Park, C., Huang, J. Z., and Ding, Y. (2011). “Domain Decomposition Approach for Fast Gaussian Process Regression of Large Spatial Data Sets.” Journal of Machine Learning Research, 12, May, 1697–1728.

Rushdi, A., P. Swiler, L., T. Phipps, E., D’Elia, M., and Ebeida, M. (2016). “VPS: Voronoi Piecewise Surrogate Models for High-Dimensional Data Fitting.” International Journal for Uncertainty Quantification, 7.

Tresp, V. (2000). “A Bayesian Committee Machine.” Neural Computation, 12, 11, 2719–27

On combining design criteria by window selection in Bayes-Parzen training of binary discriminative classifiers

Marcelino Lázaro and Aníbal R. Figueiras-Vidal

Signal Theory and Communications, Department University Carlos III of Madrid

mlazaro@tsc.uc3m.es

Abstract: Trainable transformation discriminative machines, such as Multi-Layer Perceptrons, are effective for solving classification problems. In binary cases, these architectures provide a unidimensional variable z, which serves to construct a nonlinear output o(z) whose sign determines de decision. Besides using nonlinear activations, it is possible to model the probability density of z under each hypothesis by means of (unidimensional) Parzen windows. Then, forcing z=0 as the classification threshold, the Bayes risk can be estimated and its minimization is an alternative for designing the classifier [1]. Note that this is a sort of combination between generative and discriminative principles. The above formulation allows to easily select the window shapes for imposing different characteristics to the classification results. For example, Dirac delta windows lead to (intractable) error count minimization, step (improper) windows to the original Perceptron Rule algorithm, and double-sided windows produce forms of separation margins. Obviously, there is the possibility of selecting windows that combine several of these characteristics, an easy and attractive way to create new classification algorithms. In this contribution, we will present a parametric family of windows which permits a combination of maximal margin, Perceptron Rule, and error count to train the classification machine. Some experiments will show the practical advantages of such flexible formulations. We will close our presentation suggesting other interesting possibilities for selecting windows.

Keywords: Machine learning, Bayes risk, Parzen windows

[1] M. Lázaro, M. H. Hayes, A. R. Figueiras-Vidal. “Training neural network classifiers through Bayes risk minimization applying unidimensional Parzen windows”, Pattern Recognition 77, 204-215, 2018.

This work has been partly supported by grant MacroADOBE (TEC2015-67719-P, MINECO / FEDER / UE)

Probabilistic programming for stochastic process models

Victor Gallego

ICMAT-CSIC

victor.gallego@icmat.es

Abstract: Probabilistic programming offers powerful tools for Bayesian inference, prediction and decision making queries. A probabilistic program defines a joint distribution that typically factorizes as $p(x,z) = p(x|z)\Pi_i p(z_i \mid z_{<i})$, where x are observations and z can be latent variables or model parameters. The previous formulation is sufficiently flexible to define models such as Bayesian neural networks, variational autoencoders, and stochastic processes such as hidden Markov models and variants thereof. Inference algorithms depend on whether latent variables are continuous (z_c) or discrete (z_d), so a Gibbs sampler alternating between $p(z_c|z_d,x)$,$p(z_d|z_c,x)$ used. Then, one resorts to custom inference algorithms for each conditional, e.g. Hamiltonian Monte Carlo or Stochastic Gradient Langevin Dynamics for z_c; Metropolis-Hastings or sum-product algorithms for z_d.

It is well-known that sampling efficiency varies highly depending on the parameterization (e.g. centered vs non-centered for continuous variables), and hyperparameters of the chosen sampler. To alleviate this problem, we propose a framework to embed samplers inside a variational approximate posterior $q(z|x)$, and then learn good reparameterizations and parameters while sampling from the posterior. We show that our automatic scheme offers greater performance compared to manual tuning in the previously mentioned models and illustrate it with relevant classes of stochastic process models.

Keywords: Probabilistic Programming, Markov Chain Monte Carlo, Variational Inference, Stochastic Processes

Analysis and test of asymmetric label switching

Aitor Gutiérrez-López, Javier Mediavilla-Relaño, Francisco J. González-Serrano, and Aníbal R. Figueiras-Vidal

Department of Signal Theory and Communications, Carlos III University of Madrid

anibalrfv@tsc.uc3m.es

Abstract: Breiman's label switching is an easy form of creating diversity to construct classification machine ensembles, whose present importance mainly comes from their effectiveness when working even with strong learners, such as Deep Neural Networks, to offer performance records in classification problems (see [1], for example). Considering problems with enough numbers of samples, a Bayesian analysis reveals that, applying Bregman surrogate costs for training, it is possible to obtain the "a posteriori" probabilities of the classes from the outputs of the label switching ensemble learners. This fact opens the door to designing new and more complex label switching mechanisms, that will offer performance advantage when dealing with some singular problems. In this contribution, we introduce an asymmetric switching technique for binary problems, carry out its Bayesian analysis, present the resulting aggregation schemes, and discuss the potential advantages this procedure offers. Some experiments with imbalanced real databases show the performance advantages of these algorithms and validate our analysis. We will close the presentation by suggesting how to reinforce the effectiveness of these designs and by mentioning other singular problems that can benefit from their application.

[1] Alvear-Sandoval, R.F., Figueiras-Vidal, A.R.: On building ensembles of stacked denoising auto-encoding classifiers and their further improvement. Information Fusion 39, 41-52 (2018)

Keywords: Bregman divergences, classification, machine ensembles

Stochastic Control of Dynamical Systems under uncertainty in the parameters

Pablo Angulo, Carlos García-Gutiérrez

Universidad Politécnica de Madrid

pablo.angulo@upm.es

Abstract: We consider some problems of optimal control where some parameters are fixed but unknown. We translate the problem into one of stochastic control in the belief space, which is tractable due to the existence of a sufficient statistic for those problems. As we observe the system evolution, our uncertainty over the unknown parameters is updated, which is reflected as different points in belief space. Even when the initial system is linear, the resulting stochastic control problem in belief space is not linear, so solutions must be approximated numerically.

This problem was motivated by some applications in econometrics where a model is used to make predictions and take decisions, but it is essential to consider trade offs between expected income and risk. In this work, the goal is to also take into account learning. We also show applications to other fields and, finally, we study the impact that good or bad prior information can have on the optimal paths.

Keywords: Dynamic programming, reinforcement learning, bayesian regression, dynamic linear models

Bayesian Estimation of a Bidimensional Markovian Arrival Process

Yoel G. Yera[1], Rosa E. Lillo[1;2] Pepa Ramírez-Cobo[3;4]

(1) Department of Statistics, Universidad Carlos III de Madrid, (2) UC3M-BS Institute of Financial Big Data Getafe, (3) Department of Statistics and Operations Research, Universidad de Cádiz, (4) IMUS, Institute of Mathematics, Univ. Sevilla

yyera@est-econ.uc3m.es

Abstract: Based on failure data in a two-dimensional context, this paper presents an extension to the two-dimensional case of the Markovian Arrival Process (MAP). The one-dimensional MAP has been proposed for the modeling of dependent and non-exponential inter-events times (inter-arrivals, inter-failures or inter-risk events). The new process maintains the marginal properties of a MAP but also allows dependence between the two inter-events sequences. The generalization is based on the use of the Marshall-Olkin exponential distribution. Using the Bayesian algorithm ABC, we show an estimation method that presents very good results for simulated data and for the real problem based on failures of a public transport company.

Keywords: Markovian Arrival Process (MAP), bivariate, ABC algorithm, real data

Representing Markov processes as dynamic pair-copula Bayesian networks

A. Kosgodagan, T. G. Yeungb, O. Morales-Napoles, B. Castanierd

Instituto de Ciencias Matematicas, Universidad Autonóma de Madrid, IMT-Atlantique, Nantes, France, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Netherlands, ISTIA, Université d'Angers, France

alex.kosgoda@gmail.com

Abstract: In multivariate statistics, recent attractive approaches refer to copula-based graphical models and specifically so-called pair-copula Bayesian networks. Their attractiveness is largely due to the flexibility that copula models provide, whereby the marginal distributions can be modelled arbitrarily, and any dependence captured by the copula. However, very little attention has been given for these models to _t within a full probabilistic framework and for which inference could desirably be used. In this paper, we first prove that any k-th order Markov process can be represented as a dynamic, pair-copula-based Bayesian network. Dependence is formulated as (conditional) pair time-copulas derived from the Markov process as well as the corresponding (conditional) rank correlation. Second, we explicitly show the requirements in order to perform analytical conditioning. We finally illustrate our findings through an example where Brownian motion is used to directly derive all of the pair-copula Bayesian network inputs.

Keywords: Pair-copula Bayesian networks, Copula modelling, Markov processes, Inference, Time-copula

Posters

Bayesian cumulative shrinkage for infinite factorizations

Sirio Legramanti [1], *Daniele Durante* [1] *and David B. Dunson* [2]

(1) Bocconi University, Milan, Italy, (2) Duke University

sirio.legramanti@phd.unibocconi.it

Abstract: There are a variety of Bayesian models relying on representations in which the dimension of the parameter space is, itself, unknown. For example, in factor analysis the number of latent variables is, in general, not known and has to be inferred from the data. Although classical shrinkage priors are useful in these situations, incorporating cumulative shrinkage can provide a more effective option which progressively penalizes more complex expansions. A successful proposal within this setting is the multiplicative gamma process. However, such a process is limited in scope, and has some drawbacks in terms of shrinkage properties and interpretability. We overcome these issues via a novel class of convex mixtures of spike and slab distributions assigning increasing mass to the spike through an adaptive function which grows with model complexity. This process has broader applicability, simple interpretation, parsimonious representation, and induces adaptive cumulative shrinkage of the terms associated with redundant, and potentially infinite, dimensions. Performance gains are illustrated in simulations.

Keywords*:* Factor analysis; Multiplicative gamma process; Poisson factorization; Spike and slab

Estimating marginal likelihoods by Gibbs sampling

F. Llorente [1], *L. Martino* [2], *D. Delgado* [1]

(1) Carlos III University of Madrid, (2) King Juan Carlos University

felloren@est-econ.uc3m.es

Abstract: Bayesian inference has become very popular in different fields such as machine learning and signal processing. Given a probabilistic model linking the data y with the variable of interest x (i.e., the conditional density p(y|x) a.k.a. the likelihood function) and a prior density p(x) decided by the user in advance, the Bayesian methods relies on the study of the posterior density p(x|y). However, in many applications, the analytical study of the posterior distributions is difficult or impossible. Moreover, in general, only the product p(x)p(y|x) (proportional to the posterior density p(x|y)) can be evaluated while the normalizing constant, i.e. $p(y) = \int p(x)p(y|x)dx$, is unknown. This normalizing constant p(y) is called marginal likelihood or Bayesian evidence, and it is useful for model selection purpose (i.e., to compare different models by means of Bayes' factors). To overcome these problems, one possibility is to employ conjugate models where prior and posterior distributions belong to the same family. However, conjugate models are not available or suitable in many realistic scenarios. An alternative is the use of Monte Carlo methods for approximating the posterior density by random samples. More specifically, in the last decades, Markov chain Monte Carlo sampling methods, such as the Metropolis-Hastings algorithm and the Gibbs sampler, have been successfully applied to perform approximate Bayesian inference helping popularize this paradigm. For instance, MCMC methods allow estimating moments of the posterior distribution by drawing correlated simples. The only requirement is being able to evaluate some function proportional to the posterior. However, the computation of the marginal likelihood p(y) is not straightforward using MCMC algorithms. This task becomes harder and harder as the dimension of the inference space grows. In this work, we describe a novel scheme for estimating the marginal likelihood p(y) by means of a Gibbs procedure. The novel technique produces samples distributed according to the posterior based on the Recycling Gibbs approach. Moreover, it yields a sequence of estimators of p(y) which

can be properly combined providing a final unique estimator. Keywords: Bayesian inference, Markov chain Monte Carlo (MCMC), Marginal likelihood, Model evidence, Model selection, Gibbs sampling.

Main References:

Martino, L., Elvira, V., & Camps-Valls, G. (2018). The recycling Gibbs sampler for efficient learning. Digital Signal Processing, 74, 1-13.

Robert, C., & Casella, G. (2013). Monte Carlo statistical methods. Springer Science & Business Media.

Heuristic Bayesian Optimization

Luis C. Jariego, Eduardo Garrido Merchán

Universidad Autónoma de Madrid

luis.jariego@estudiante.uam.es, eduardo.garrido@uam.es

Abstract: Bayesian Optimization (BO) methods optimize black-box functions. These black boxes do not have a known expression, are considered to be very expensive, and the observations may be corrupted by noise. In this context, BO arises as an exceptional approach to deal with this scenario, giving a good approximation to an optimum of the function in a small number of iterations. BO relies on a probabilistic model, typically a Gaussian Process (GP) that represents the uncertainty of the black box. Iteratively, BO builds an acquisition function (AF) over these GPs which is a criterion that tries to find a good trade-off between exploration and exploitation of the optimization problem. As there exist no best criterion for every possible problem, as states the no free lunch theorem, this work explores heuristics to try to adapt the BO AFs to every possible optimization problem. We are going to present heuristics in order to combine the Probability of Improvement (PI), Expected Improvement (EI) and GP Lower Confidence Bound (GP-LCB) AFs. These heuristics will try to combine these AFs by switching them according to the quality of the evaluations, mixing them by using adaptative or predefined weights in a single AF, choosing them at random or learning by a Machine learning model and a dataset of evaluated GPs which is the best AF for every possible condition of a GP. We expect to obtain improvements in the performance of the different standard criteria present in the SkOpt BO library with the presented heuristics in synthetic problems, benchmark optimization functions and real data scenarios.

Parallel Predictive Entropy Search for Multi-objective Bayesian Optimization with Constraints

Cesar Garrido Merchán, Daniel Hernández Lobato

Universidad Autónoma de Madrid

eduardo.garrido@uam.es , daniel.hernandez@uam.es

Abstract: This work presents PPESMOC, Parallel Predictive Entropy Search for Multiobjective Bayesian Optimization with Constraints. Both the objectives and the constraints are assumed to be expensive-to-evaluate black-box functions that lack an analytical expression. Furthermore, the evaluation process may be corrupted by noise. This is the typical setup assumed in Bayesian optimization, which prevents the use of standard optimization approaches. Unlike current state-of-the-art methods, PPESMOC suggests, at each iteration, not only one, but a set X of N points on which to evaluate the objectives and the constraints with the aim of solving the problem in a reduced number of iterations. That is, PPESMOC allows for batch Bayesian optimization, which is expected to be better when there are several computational resources available to evaluate the objectives and the constraints in parallel. Iteratively, PPESMOC fits a Gaussian process (GP) to the current observations of each objective and constraint. Using these models, PPESMOC builds an acquisition function a : RD×N → R (with D the dimensionality of the problem and N the batch size) whose maximum indicates the set of points X that are more promising to evaluate next. This acquisition function is then cheaply maximized since it only depends on the GPs and not on the actual black-boxes. After the evaluation is performed, the process repeats. When enough data has been collected like this, the GPs can be optimized to find a solution to the optimization problem. The acquisition function of PPESMOC, a(·), measures the expected reduction in the entropy of the Pareto set in the feasible space, after performing an evaluation of the black-boxes at the candidate set of points X . A smaller entropy means that the uncertainty of the solution of the optimization problem is smaller and hence we are closer to solving it. Unfortunately, this reduction in the entropy is intractable to compute. Nevertheless, the required computations can be efficiently approximated using the expectation propagation algorithm. Critically, given a candidate set of points

X, the gradient of the approximation of a (·) w.r.t. these points can be easily computed, which allows for an efficient optimization of the acquisition function. We plan to compare the performance of PPESMOC with a base-line strategy that chooses the set of candidate points X to evaluate next at random. We expect to obtain significantly better results in real and synthetic problems, which will illustrate the practical benefits of batch Bayesian optimization in the considered setting.

On-site surrogates for large-scale calibration

Jiangeng Huang, Robert B. Gramacy, Mickael Binois, and Mirko Libraschi

Abstract: Motivated by a challenging computer model calibration problem from the oil and gas industry, involving the design of a so-called honeycomb seal, we develop a new Bayesian calibration methodology to cope with limitations in the canonical apparatus stemming from several factors. We propose a new strategy of on-site experiment design and surrogate modeling to emulate a computer simulator acting on a high-dimensional input space that, although relatively speedy, is prone to numerical instabilities, missing data, and nonstationary dynamics.

Our aim is to strike a balance between data-faithful modeling and computational tractability within an overarching calibration framework---tuning the computer model to the outcome of a limited field experiment. Situating our on-site surrogates within the canonical calibration apparatus requires updates to that framework. In particular, we describe a novel yet intuitive Bayesian setup that carefully decomposes otherwise prohibitively large matrices by exploiting the sparse blockwise structure thus obtained. We illustrate empirically that this approach outperforms the canonical, stationary analog, and we summarize calibration results on a toy problem and on our motivating honeycomb example.

Using Correlated Topic Models in Automatic Job-Resume Matching

Vu-Hoang Le, Arthur White and Simon Wilson

School of Computer Science and Statistics – Trinity College Dublin

leho@tcd.ie, arwhite@tcd.ie. swilson@tcd.ie

Abstract: The related tasks of matching job descriptions and resumes efficiently are important for both organizations and individuals. Current solutions to these problems are predominantly based on manual handling through using keyword searches, which are inefficient due to the rich information and ambiguous structure of both documents and the large amount of data available online. Our research uses topic modelling to explore the structure of job descriptions and resumés, then performs an automatic job to resume matching using a similarity measure. We examine different topic modelling techniques, such as a variational approximation, in terms of scalability for implementing the Bayesian inference and for prediction quality. The experimental results show that automatic matching with the Correlated Topic Model, a hierarchical model that describes a document as a mixture of topics, with each topic defined as a distribution over words, can produce a match similar to manual matching while better in term of time and quantity. We show how our method is also a way to solve the cold-start problem in recommender systems, thus it can be used as a compliment to other job recommender systems.

Keywords: job-resume matching · bipartite graph · correlated topic model · Variational inference.

Mutually Exclusive Spike-and-Slab Priors for Structured Feature Selection in Multi-class Classification Problems

Alejandro Catalina-Feliú, Daniel Hernández Lobato

Universidad Autónoma de Madrid

alejandro.catalina@uam.es, daniel.hernandez@uam.es

Abstract: Many multi-class classification problems are characterized by a small number of observed instances N and a large number of explaining variables P. Under this setting, even a simple linear model is too complex and can lead to severe over-fitting. A strong regularization that has been proven to give better generalization results is to assume sparsity in the coefficients of the linear model. That is, many coefficient values are equal to zero. This has also the advantage of improving interpretability since only a reduced number of explaining attributes will contribute to the predictions made by the model. A very successful approach to introduce the sparsity assumption under a Bayesian setting is to assume a spike-and-slab prior for the model coefficients. Consider K different classes in the problem. Under this setting, there are K different hyperplanes to be estimated, i.e., one hyper-plane wk for each different class k = 1, . . . , K. Assume the labeling rule is yi = arg maxk x T i wk + k i , with k i some Gaussian noise. Let W ∈ RK×P be a matrix summarizing all the hyper-planes. Under a spike-and-slab prior, each entry of W, w k j , is assumed to be generated from either a Gaussian distribution with zero mean and fixed variance or from a point of probability mass at zero. Nevertheless, there are several different sparsity patterns that can be introduced in the matrix W. For example, one can consider that one can observe zeros arbitrarily at any entry. Another approach is to assume that row entries have to be either all equal to zero or different from zero. This is equivalent to assuming features relevant either for all K class labels or completely irrelevant at all. In this work, we explore a third sparsity pattern in W. Namely, that if one feature is relevant, it will only be relevant to discriminate a particular class label from all the others. This is equivalent to assuming that either all entries in a row of W are equal to zero, or only one entry in that row (i.e., the one corresponding to the class label) is different from zero. This has the advantage over the previous sparsity patterns of improving the interpretability of

the model. With this goal, we describe a novel spike-and-slab prior distribution for the model coefficients W. Exact Bayesian inference in a model under such a prior is intractable. Therefore, one has to use approximate methods in practice. For this, we describe an efficient algorithm based on expectation propagation. We plan to evaluate the performance of such a prior (and approximate inference method) in terms of prediction error and interpretability in several multi-class problems, involving real and synthetic data. A comparison with Markov Chain Monte Carlo methods will also be carried out to assess the bias in the approximation obtained by expectation propagation.

Life modelling from GPS traces

S. *Villanueva, V. Gallego, R. Naveiro, E. Nungesser, and A. Nuñez*

ICMAT-CSIC

ernesto.nungesser@icmat.es

Abstract: GPS data facilitates human behaviour understanding and tracking with potential uses for policy and marketing. In this paper, we first describe the preprocessing that we need to perform over GPS data to obtain meaningful data and then model such data with the aid of a hierarchical continuous time Markov chain model. We use the model to make relevant inferences and predictions from the sociological and marketing perspectives.

Keywords: GPS data, Continuous time Markov chains, Bayesian analysis

Inducing Point Methods for Gaussian Process Surrogates of Large-Scale Simulations

Austin Cole

Virginia Tech, USA

davidc89@vt.edu

Abstract: Gaussian processes (GPs) provide a flexible methodology for modeling complex surfaces. One challenge with GPs is the computational burden with an increasing sample size or number of dimensions. The machine learning community has turned to pseudo-inputs or inducing points to reduce the computational burden in such contexts. We seek to port this family of methods to build GP surrogates for noisy and even heteroskedastic stochastic processes, with extensions to sequential design and Bayesian optimization. We show that using inducing points extends the reach of GP surrogates in big simulation contexts and makes for efficient design and meta-modeling of large scale computer simulation experiments. Examples are provided for epidemiological, industrial, and financial applications.

Accelerating pseudo-marginal Metropolis-Hastings schemes for stochastic kinetic models

Tom Lowe and Andrew Golightly

Newcastle University

t.lowe@newcastle.ac.uk

Abstract: A stochastic kinetic model (SKM) typically refers to a reaction network, an associated rate law and a probabilistic description of the reaction dynamics. Reactions occur continuously in time with a reaction occurrence resulting in a discrete change to the system state. A Markov jump process (MJP) therefore provides a natural description of the time-course behaviour of the species involved in the reaction network. The resulting modelling framework is fairly flexible and consequently, has been used ubiquitously in areas such as epidemiology, population ecology and systems biology. Performing exact fully Bayesian inference is challenging due to the intractability of the observed data likelihood, necessitating the use of computationally intensive techniques such as pseudo-marginal Metropolis-Hastings (PMMH). We propose to increase the computational and statistical efficiency of this approach by leveraging the tractability of an inexpensive surrogate. The surrogate is used in three ways: in the design of a gradient based parameter proposal, to construct an appropriate bridge construct and in the first stage of a delayed acceptance step. We find that the resulting approach offers substantial gains in efficiency over a standard PMMH implementation.

Keywords: Stochastic kinetic model, Markov jump process, linear noise approximation, delayed acceptance

Large scale dynamic forecasting for distributed inventory management

Bruno Flores, Javier Martín Hernández

Aeroengy

bruno.flores@aeroengy.es

Abstract: We consider a large scale distributed inventory management problem in relation with a major retail company, with hundreds of shops each one with thousands of references. In our specific case, the shops have little stocking capacity beyond that available on shelves; moreover, inventory information is not fully reliable due to technical and human errors. The company has a hub based logistic system which facilitate product distribution in a moderate time (but not daily). Lack of products on the shelves induce economic and image losses. We describe the general inventory management problem used and focus on the entailed large scale dynamic forecasting problem. We focus on the influence of promotions on product demand.

Keywords: Retailing, Inventory management, Large scale forecasting, Bayesian analysis

Spectral Density-Based and Measure-Preserving ABC for Partially Observed Diffusion Processes

Irene Tubikanec, Evelyn Buckwar, Massimiliano Tamborrino

Institute for Stochastics, Johannes Kepler University Linz

irene.tubikanec@jku.at

Abstract: Let us consider a n-dimensional SDE whose solution process is only partially observed througha one-dimensional and parameter-dependent output process admitting an invariant distribution. We aim to infer the parameters from discrete time measurements of the invariant output process. Due to the increased model complexity, needed to understand and reproduce the real data, the underlying likelihood is often unknown or intractable. Among several likelihood-free inference methods, we focus on the Approximate Bayesian Computation (ABC) approach [1].

When applying ABC to stochastic processes, two difficulties arise. First, different realizations from the output process with the same choice of parameters may show a large variability, due to the stochasticity of the model. Second, exact simulation schemes are rarely available for general SDEs and, thus, a numerical method for the synthetic data generation within the ABC framework has to be derived. We tackle these issues as follows. To reduce the randomness coming from the underlying model, we propose to make use of the structural properties of the SDE, namely the existence of a unique invariant distribution. We map the synthetic data into their estimated invariant density and invariant spectral density, almost eliminating the variability in the data and, thus, making hidden information about the parameters accessible. Since our ABC algorithm is based on the structural property, it can only lead to successful inference when the invariant measure is preserved in the synthetic data simulation. To achieve this, we propose to use a structure-preserving numerical scheme that, differently from the commonly used Euler-Maruyama method, preserves the properties of the underlying SDE.

Here, we illustrate our proposed Spectral Density-Based and Measure-Preserving ABC Algorithm on the stochastic Jansen and Rit Neural Mass Model (JR-NMM) [2] and the stochastic FitzHugh-Nagumo model (FHN) [3]. Both models are ergodic, which results in the output process admitting an invariant measure. The use of numerical splitting schemes guarantees the preservation of the invariant distribution in the synthetic data generation step. With our new approach, we succeed in the simultaneous estimation of the three most crucial parameters of the JR-NMM and of all four parameters inherent in the FHN model. Finally, we apply our method to fit the JR-NMM to real EEG alpha-rhythmic recordings.

Keywords: Approximate Bayesian Computation. Stochastic differential equation. Invariant Measure.

References

[1] M.A. Beaumont, W. Zhang, D.J. Balding "Approximate Bayesian computation in population genetics". In: Genetics, 162(4), pp.2025-2035 (2002)

[2] M. Ableidinger, E. Buckwar, H. Hinterleitner. "A Stochastic Version of the Jansen and Rit Neural Mass Model: Analysis and Numerics." In: The Journal of Mathematical Neuroscience 7(8) (2017)

[3] J. R. Leon, A. Samson. "Hypoelliptic stochastic FitzHugh-Nagumo neuronal model: mixing, up-crossing and estimation of the spike rate." In: Annals of Applied Probability 28(4), pp.2243-2274 (2018)

[4] E. Buckwar, M. Tamborrino, I. Tubikanec. "Spectral Density-Based and Measure-Preserving ABC for Partially Observed Diffusion Processes. An Illustration on Hamiltonian SDEs." in preparation.

[5] E. Buckwar, A. Samson, M. Tamborrino, I. Tubikanec. "Spectral Density-Based and Measure-Preserving Approximate Bayesian Computation Applied to the Hypoelliptic Stochastic

FitzHugh-Nagumo Neuronal Model." *in preparation".*

Importance Sampling with the Integrated Nested Laplace Approximation

Virgilio Gómez-Rubio

Department of Mathematics, Universidad de Castilla-La Mancha

virgilio.gomez@uclm.es

Abstract: The integrated nested Laplace approximation (INLA, Rue et al., 2009) provides a framework for approximate Bayesian inference that is computationally faster that traditional Markov chain Monte Carlo (MCMC) methods. However, INLA is restricted to those models that can be expressed as a latent Gaussian Markov random field (GRMF).

Recently, Gómez-Rubio and Rue (2018) have shown that INLA can be combined with the Metropolis-Hastings algorithm to fit a wider class of models. In particular, those that can be expressed as a conditional (on some hyperparameters) latent GMRF. This algorithm, termed INLA within MCMC, works by splitting the set of hyperparameters so that MCMC is used to estimate the posterior distribution of a small subset of hyperparameters θ_1. At every step of the MCMC algorithm a model with INLA is fit conditional on the current value of θ_1, which also provides the (conditional on θ_1 posterior marginals of the remainder of hyperparameters in the model. Hence, the posterior marginals of the other set of hyperparameters are obtained by averaging their conditional marginals.

Although flexible, this method has two main problems: it is a sequential algorithm which does not scale well and it requires fitting a model with INLA at every step. This makes INLA within MCMC lose the computational speed of INLA.

In this work we propose replacing the Metropolis-Hastings algorithm by importance sampling (IS). This has the appealing of being a simple yet flexible method for Bayesian inference. In addition, IS can be easily implemented in parallel so that computational times can get close to the actual speed of the INLA method. IS is applied to obtain samples from the set θ_1 and importance weights are taken as the conditional marginal

likelihood (on θ_1) provided by INLA times the prior of θ_1. The posterior marginals of the other hyperparameters in the model are obtained by Bayesian model average of the conditional posterior marginals using the importance weights.

This news approach is illustrated using two examples in spatial econometrics and diseases mapping. In spatial econometrics the computation of the so called impacts (that measure the influence of covariates in neighbouring areas) requires low level multivariate inference, which is tackled effectively with IS and INLA. In disease mapping, we consider the class of multivariate models to study several diseases at the same time and use IS and INLA to estimate the joint posterior distribution of some dependence parameters among the diseases.

References:

Virgilio Gómez-Rubio and Håvard Rue (2018). Markov chain monte carlo with the integrated nested laplace approximation. Statistics and Computing 28 (5), 1033-1051.

Håvard Rue, Sara Martino, Nicolas Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71 (2), 319-392.

Adaptive Bayesian Gauss-Hermite quadrature

F. Llorente [1], L. Martino [2], Victor Elvira [3], D. Delgado [1]

(1) Carlos III University of Madrid, (2) King Juan Carlos University, (3) IMT Lille Douai - CRIStAL laboratory (CNRS 9189)

felloren@est-econ.uc3m.es

Abstract: Computing intractable integrals is a required task when performing Bayesian inference. For instance, the computation of expected values of some function (x) with respect to a posterior distribution (x), i.e.$\int (x)(x)dx$, is generally intractable. Deterministic quadrature methods represent an efficient alternative for this problem, when they can be applied. Monte Carlo methods, as Markov chain Monte Carlo (MCMC) methods are universal techniques (universal in the sense, they can be virtually always applied) used to approximate this kind of integrals by considering samples generated by a Markov chain with a stationary distribution pre-established by the user. In this work, we present an adaptive Bayesian quadrature scheme combining Gaussian Process (GP) regression methods with MCMC algorithms. In some sense, the proposed scheme can be considered as a multivariate extensions of the Sticky Metropolis techniques.

Keywords: Bayesian inference, Markov chain Monte Carlo (MCMC), Adaptive MCMC, GaussHermite quadrature, Gaussian processes, Bayesian quadrature

References:

1. D. Heestermans Svendsen, L. Martino, M. Campos-Taberner, J. Garcia-Haro, G. Camps-Valls, "Joint Gaussian Processes for Biophysical Parameter Retrieval", IEEE Transactions on Geoscience and Remote Sensing, Volume 56, Issue 3, Pages: 1718-1727, 2018.

2. Martino, L., Casarin, R., Leisen, F., & Luengo, D. (2018). Adaptive independent sticky MCMC algorithms. EURASIP Journal on Advances in Signal Processing, 2018(1), 5.

3. O'Hagan, A. (1991). Bayes–hermite quadrature. Journal of statistical planning and inference, 29(3), 245-260.

4. Liu, Q., & Pierce, D. A. (1994). A note on Gauss—Hermite quadrature. Biometrika, 81(3), 624-629.

Bayesian Sensitivity Analysis of Value-at-risk and expected shortfall of a GPD using distorted band classes

Eva L. Sanjuán, J. Pablo Arias-Nicolás, Jacinto Martín and Maribel Parra

Universidad de Extremadura, Departamento de Matemáticas

etlopez@unex.es

Abstract: Extreme value theory is a set of statistical tools employed for modeling and predicting the occurrence of rare events. It is well known that, when we consider the values of the sample space above a certain value, called threshold, the limit distribution function is a Generalized Pareto Distribution (GPD). Therefore, such distribution has been widely employed to model exceedances in several fields such as hydrology, finance, insurance and environmental science. Most useful and reliable risk measures are value-at-risk and expected shortfall. We set a distorted band class for the prior distribution based on concave and convex classical distortion functions. We compute the range of the parameters of the GPD, and the two risk measures under a Bayesian framework. Due to the fact that the distorted band depends on the election of the reference prior distribution (and also on the distorted functions considered), we illustrate with some numerical examples how this choice affects the calculation of the measures.

Keywords: Bayesian sensivity analysis, distortion functions, risk measures

Nonparametric Bayesian Functional Regression with application to shot put data

Alessandro Lanteri, Raffaele Argiento and Silvia Montagna

University of Torino

alessandro.lanteri@unito.it

Abstract: In sport analytics, there is often interest in predicting elite athlete's performance at a future sporting event given his/her competitive results tracked throughout the athlete's career and other (time-varying) covariates. Such predictions can be useful both for scouting purposes, and to build red flag indicators of unexpected increases in athlete performance for targeted anti-doping testing. We propose a predictive model for the longitudinal trajectory of athlete's performance where we characterize the curve with a sparse basis expansion allowing individual time-dependent covariates to impact the shape of the estimated trajectories. Moreover, we introduce random intercepts, distributed according to a nonparametric hierarchical process, in order to induce clustering while borrowing statistical information across curves. In particular, we assume a hierarchical normalized generalized gamma process to grants great flexibility in clustering and accuracy in prediction. We apply our model to a longitudinal study on shot put athletes, where their competitive results are tracked throughout their career.

Keywords: Bayesian Nonparametric, Functional Regression, Sport Data

Bayesian Dynamic Regression Trees

Michael Ferreira, Simon Wilson

SCSS, Trinity College Dublin

ferreima@tcd.ie, swilson@tcd.ie

Abstract: Bayesian Dynamic Regression Trees (BDRT) proposes the estimation of sequential data in a streaming setting by dynamically growing and mutating Bayesian regression trees. This ensemble of flexible regression models allows for the relationship between the covariates and m-dimensional response to evolve through a latent process. Given a set of known, constant parameters exact optimal inference can be performed on the latent variable via an implementation of the intermittent Kalman filter. Passing possible new predictors through learnt tree structures allows for sequential experimental design. Further, estimation at each partition can be modified to use the suite of non-linear Kalman filters, for instance the Extended and Unscented Kalman filters. Adaptive filtering can be implemented in a sub-optimal setting for estimation of, say, the state covariance matrix. Inference on the ensemble of trees to explore the space of partition structures is via MCMC sampling of reversible mutations of the tree structures. In the optimal case, an exact expression for the posterior weight of the trees is derived for use in model averaging. The algorithm is parallelised on a shared memory basis using OpenMP on a personal computer. A study of the model is performed on simulated streaming data sets. The model is then compared to other dynamic tree models in an online batch setting and a comparative assessment is provided.

Keywords: BART, sequential, streaming, non-stationary, non-linear, active learning