BISP7

Seventh Workshop on

BAYESIAN INFERENCE IN STOCHASTIC PROCESSES

Universidad Carlos III de Madrid, Getafe, Spain

September, 1-3, 2011

Poster Abstracts

Rupali Akerkar and Håvard Rue.

Department of Mathematical Sciences, NTNU, Norway.

Simultaneous Credible Bands for Non-Homogeneous Poisson Processes with application to survival analysis

In survival analysis, often there exists many covariates related to the event of interest but not all of them are equally important. The functional form of covariates are also not easy to judge, not even after estimating the functional form using, for example, INLA (Rue, Martino and Chopin, 2009). In this work, we investigate the use of simultaneous credible bands and posterior contour probabilities, computed within the INLA framework, to address these questions in the case of non-homogeneous Poisson processes. We illustrate our approach with examples.

Keywords: simultaneous credible bands; contour probabilities; non-homogeneous Poisson processes.

Louis JM Aslett and Simon Wilson

Trinity College, University of Dublin, Ireland

Inference for continuous-time Markov chains with absorbing states, and application to repairable redundant systems

The lifetime of a repairable redundant system finds natural expression through a Phase-type distribution. Motivated by simple examples in reliability theory, we examine the issues surrounding Bayesian inference for phase-type models. In particular, we build on the Markov chain Monte Carlo algorithm developed by Bladt et al. Since the physical makeup of a system is usually known, we examine imposing special structures on the underlying absorbing continuous-time Markov chain. Given the application area, censored data are common and their incorporation is addressed. Finally, with the restricted structure and type of data which is common computational bottlenecks arise in implementation, so approaches to address these are presented. This leads to a generally applicable modification of the original methodology which can be used in scientific applications where a Phase-type distribution provides a natural expression of the underlying stochastic process.

Keywords: Repairable redundant system; phase-type distribution; reliability theory.

Arnab Bhattacharya and Simon Wilson

Trinity College Dublin, Dublin, Ireland.

Inference for discrete dynamic state space models

This paper addresses the recursive Bayesian inference problem for a discrete time non-linear dynamic system commonly defined as a dynamic state-space model. This process evolves over time as a first order Markov process according to a transition density and the observations are assumed to be conditionally independent given the states and parameters. Bayes Law and the structure of the state-space model are used to sequentially update the posterior density of the model parameters as new observations arrive. In this talk various possible approximations are proposed and discussed that would allow fast functional approximation updates of the posterior distribution. These approximations rely on techniques such as the Kalman filter and its non-linear extensions, as well as the integrated nested Laplace approximation (Rue, Martino and Chopin, 2009). The approximate posterior is explored at a sufficient number of points on a grid which is computed at good evaluation points. The grid is re-assessed at each time point for addition/reduction of grid points. This new methodology of sequential updating makes the calculation of posterior both fast and accurate. It has been found to be extremely accurate for linear Gaussian models and it performance is also evaluated on several non-linear models.

Keywords: state space models; Kalman filter; integrated nested Laplace approximation; recursive updating.

Paul Blomstedt and Jukka Corander

Åbo Akademi University and University of Helsinki, Finland

Predictive Matching of Stochastic Processes

We consider the generic problem of comparing the degree of similarity between two stochastic processes. It is argued that the standard hypothesis testing framework does not provide suitable means for a statistical assessment of a match between the processes due to the asymmetric roles of null and alternative hypotheses. This issue is of particular concern for instance in forensic applications where a null hypothesis of similarity or equality is not acceptable. Here we introduce an approach for probabilistic matching based on predictive distributions derived from the observed processes. Predictive match probabilities are derived for stationary Markov chains of arbitrary finite order and for standard hidden Markov models. The work generalizes our earlier results on probabilistic matching of distributions for static random variables. We illustrate the derived approach by considering two forensic applications. The first one involves the matching of partial HIV sequences to determine whether two individuals have been infected by the same person. The second application concerns the matching of two intercepted noisy digital transmissions to assess the probability that they stem from the same source.

Keywords: hidden Markov models; higher order Markov chains; matching; predictive inference

Stefano Cabras[1], María Eugenia Castellanos[2] and Silvia Perra[1]

[1]Universitá de Cagliari, Italy and [2]Univerisidad Rey Juan Carlos, Madrid, Spain

A default Bayesian approach for model choice in Weibull regression

In this work we consider a Bayesian Weibull regression analysis for survival times and the variable selection problem for this model with right censored data. We approach model selection under an objective Bayesian point of view using the usual right-Haar measure for location-scale models. We compare a suitable set of models with respect to the same reference null model, i.e. the model with the intercept only. Model comparison is made via two versions of the intrinsic Bayes Factor (IBF) of Berger and Pericchi (JASA, 1996), namely the mean and median IBF. Calculus of IBF under censored data poses the problem of definition of the minimal training sample (MTS) because of the problems raised up in the paper of Berger and Pericchi (Ann.Stat, 2004). We address this problem using sequential MTS (SMTS) leading to a MTS with random size whose probability distribution is obtained analytically. We exploited parallel calculation to approximate the IBF because each MTS can be treated separately from the others and parallelization becomes straightforward. The behaviour of the IBF is analysed in a simulation study and applied to stage IV non-small cells lung cancer, which is the most common type of lung cancer.

Keywords: intrinsic Bayes factors; minimal training sample; non-small cells lung cancer; parallel computation; sequential sampling.

Ashley P. Ford and Gareth O. Roberts

Warwick University, UK

Indian Buffet Epidemics - A non-parametric Bayesian Approach to Modelling Heterogeneity

Analysis of data from epidemics shows that simple stochastic epidemic models often do not fit the observations and so predictions based on these models could be misleading.

The Indian Buffet Epidemic model has been developed to address the need for a model which is more suitable than assuming homogeneous mixing or an incorrect network model. The Indian Buffet Epidemic combines a bipartite network model with the Indian Buffet process, to provide a non-parametric model, which includes as special cases household models and other models for heterogeneity.

This poster describes the model and describes some of the difficulties in developing an effective MCMC algorithm. A grouped independence Metropolis-Hastings (GIMH) algorithm is described. The algorithm is demonstrated on a range of simulated data from both the true model and other epidemic models.

Keywords: Epidemic; MCMC; non-parametric; Bayesian

Roland Fried[1], Inoncent Agueusop[1], Bjőrn Bornkamp[2], Konstantinos Fokianos[3],
Jana Fruth[1] and Katja Ickstadt[1]

[1]TU Dortmund University, Germany, [2]Novartis, Basel, Switzerland, [3]University of Cyprus, Cyprus

Bayesian Outlier Detection in INGARCH Time Series

An INGARCH(1,1) model for time series of counts arising, e.g., in epidemiology assumes the observations to be Poisson distributed conditionally on the past, with the conditional mean being an affine-linear function of the previous observation and the previous conditional mean, with positive weights summing up to a value less than one.

We model additive outliers within such processes, assuming that we observe a contaminated process with additive Poisson distributed contamination, affecting each observation with a small probability. Additive outliers do not influence the dynamics of the process and can represent, e.g., measurement artifacts. They are difficult to handle within a non-Bayesian framework since the uncontaminated values entering the dynamics of the process at contaminated time points are unobserved. We have implemented a Bayesian version of this modified model in R, using gamma priors for the model parameters and a Dirichlet prior for the effects of the past. Values from the posterior distribution are simulated applying a componentwise Metropolis-Hastings algorithm.

We analyze real and simulated data sets and find Bayesian outlier detection based on posterior probabilities to work well if there is more than one additive outlier, i.e. if there is enough information for fitting the model with contamination.

Keywords: Generalized linear models; time series of counts; additive outliers; level shift.

Cristian L. Bayes[1], Jorge L. Bazán[1], **Catalina García**[2]

[1]Pontificia Universidad Católica de Lima, Peru, [2]Universidad de Granada, Spain

A Robust Regression Model for Proportions

A new regression model for proportions is presented by considering the Beta rectangular distribution proposed by Hahn (2008). This new model includes the Beta regression model, introduced by Ferrari and Cribari-Neto (2004) and the variable dispersion Beta regression model introduced by Smithson and Verkuilen (2006) as particular cases. Similar to Branscum, Johnson and Thurmond (2007) a Bayesian inference approach is adopted using a Markov Chain Monte Carlo

(MCMC) algorithms. A simulation study is carried out to show that the new model is less influenced on the estimation of regression parameters when outlying observations exist than the Beta regression model. In consequence, the new model is more general and
robust than usual.

Keywords: Proportions, Beta regression, Bayesian estimation, link function, MCMC

Vincent Garreta and John Haslett

Trinity College Dublin, Ireland

Inference for stochastic simulators: a review

We call `stochastic simulator' a chain of stochastic simulation mechanisms available under the form of a computer code. Such a simulator is effectively a *blackbox* and we assume that it defines a distribution of its outputs conditional on its inputs. This distribution and associated likelihoods are both *implicit* in the sense they cannot be evaluated pointwise, but only approximated, at the cost of many simulations.

In statistics and other fields of science, chains of conditional distributions define hierarchical models whose likelihood is --most of the time-- analytically intractable. Likelihood is thus *implicit* in the sense given before, and Monte Carlo-based inference methods in a Bayesian framework have proven to be efficient for the inference of such models. But for some hierarchical models, the latent space is so big and/or so constrained, that researchers recently proposed to consider the distribution itself as *implicit*, i.e. to `blackbox' the hierarchical model.

We define inference for models with *implicit distributions* and review the three methods available for inference: (a) approximate Bayesian computation, (b) indirect inference and, (c) emulation of stochastic simulators. We cross-interpret the way each method re-define the likelihood and we discuss their respective computational cost. We finally discuss recent proposals (Andrieu et al., 2010; Lindgren et al., 2011) opening the way to more accurate inference in larger spaces.

Key Words: Implicit distribution, implicit likelihood, approximate Bayesian computation, indirect inference, emulation

Claudio Macci

Universitá di Roma Tor Vergata, Italy

Extension of some recent large deviation results for posterior distributions

We present large deviation principles for sequences of posterior distributions. We consider a class of statistical models with positive parameter and some statistical models based on samples derived from stationary Gaussian processes, and from i.i.d. exponential power distributed samples. These large deviation principles are proved without any restriction on the prior distribution; in this way we extend some results in the literature.

Key words: large deviations; stationary Gaussian process; short-range and long-range dependence; exponential power distribution.

Tiep Mai and Simon Wilson

Trinity College Dublin, Ireland

Short-term Traffic Flow with Seasonal Vector Auto-Regressive Moving Average Model

In transportation, up until now, univariate time series are commonly used for short-term traffic flow forecasting in urban networks without considering the spatial effect of various junctions. In this poster, a k-dimension Seasonal Vector Auto-Regressive Moving Average (SVARMA) of additive form is used and compared with a univariate model. We adopt Bayesian approach and implement MCMC sampling to realize parameter estimation and prediction. In SVARMA, as the number of parameters is O(k^2) with the full matrix design and that makes the sampling infeasible, we utilize the first order dependency in which each junction flow depends only on its neighbours and reduce the dimension to O(k). The algorithm is tested with both simulation and real data set. By incorporating the spatial effect in SARMA, we want to improve the prediction power and use SVARMA as a stepping stone to a more general model which deals with spatial stop-and-go traffic events.

Keywords: Time series; seasonal VARMA; forecasting; Bayesian inference

Raquel Montes Diez and Alicia Quirós

Universidad Rey Juan Carlos, Madrid, Spain

Modelling the Hemodynamic Response in fMRI Using Gaussian Processes

During the last few decades our knowledge of the human brain has developed significantly as a result of new neuroimaging techniques, such as functional magnetic resonance imaging (fMRI). By observing the relation between a stimulus paradigm (in an experiment) and the changes in blood flow and blood oxygenation in the brain (known as hemodynamics), fMRI provides a measure of brain activation. The change in the ratio of oxygenated to deoxygenated blood is described by the so-called hemodynamic response function (HRF).

Modeling the HRF in fMRI experiments is therefore an important aspect of the analysis of data in functional neuroimaging. This has been done in the past using parametric response functions, typically including the Poisson, gamma or Gaussian densities. In this work, we consider the case in which the HRF is simply defined by a certain unknown function $z(\cdot)$. General Gaussian Processes theory presents an attractive way of expressing prior beliefs about the function $z(\cdot)$ and we show how, in this context, a combination of analytical methods may be used for making inference about the posterior predictive distribution of interest.

Results are shown on synthetic data and on real data from an event-related fMRI experiment.

Keywords: fMRI; Gaussian Processes.

Tomi Peltola, Pekka Marttinen and Aki Vehtari

Aalto University School of Science, Finland.

Observations from an Application of Bayesian Variable Selection to Searching for Additive and Dominant Effects in Genome-wide Data

A common strategy in searching for genomic variants associated to disease or other complex traits is to analyse the variants one at a time. However, many traits are thought to be affected by multiple variants, each contributing a small effect. The simultaneous analysis of all available variants, adhering more closely to the hypothesized genetic basis, has recently been made possible by developments in approximate computational methods and the increase in computational resources. We have studied Bayesian variable selection with a type of spike-and-slab prior in searching for additive and dominant genetic effects, utilizing an adaptive-during-burn-in Markov chain Monte Carlo scheme for the computation. We present

observations from simulations, characterizing and contrasting the behaviour against single-SNP analysis, and from an application to real data.

Keywords: Bayesian variable selection; model averaging; spike-and-slab prior; linear regression; genome-wide association analysis

C. Armero[1], S. Cabras[2], M.E. Castellanos[3], **S. Perra**[2], A. Quirós[3,1],
M.J. Oruezábal[4] and J. Sánchez-Rubio[4]

[1]Universitat de València, Spain, [2]Università di Cagliari, Italy, [3]Universidad Rey Juan Carlos, Madrid, Spain, [4]Hospital Infanta Cristina de Madrid, Spain

Bayesian multi-state models for assessing the progression of stage IV non small-cell lung cancer

Multi-state models are stochastic processes with a finite number of states which provide a general framework for modelling event history data. These models may bring out important insights which may be ignored when using an ordinary regression model. In the present work we analyse data corresponding to Stage IV non small-cell lung cancer that is the most advanced phase of this type of cancer. Medical treatments in this stage are oriented to stabilise the disease, improve survival and quality of life and avoid a fatal progression of the cancer. We model survival times through a non-homogeneous Markov chain with two transient states: cancer is stabilised (the initial one) and cancer has progressed, and an absorbing state which corresponds to death. In particular we use the disability model which is relevant for irreversible diseases when the disease increases the risk of death. Covariates may be incorporated through transition intensities to explain differences among individuals in the course of illness. In this work transition intensities between the states are modelled in terms of Weibull regression models from a Bayesian perspective. Data for the study come from the Hospital Infanta Cristina in Madrid (Spain).

Keywords: Disability model; hazard function; non-homogeneous Markov model; transition probabilities; Weibull regression.

Alicia Quirós Carretero

Universidad Rey Juan Carlos, Madrid, Spain

Dynamic linear models applied to the analysis of functional magnetic resonance imaging data

The signal-to-noise ratio (SNR), defined as the ratio of the signal variance to the variance of the system noise, is a broadly accepted measure for comparing the performance characteristics between different time series. In the functional Magnetic Resonance Imaging (fMRI) time series analysis, it is generally accepted that the SNR, defined as above, is larger for active than for non-active voxels, constituting, therefore, a key factor in the detection of brain activity. In this work, we use dynamic linear models to analyse a time series for a voxel of an fMRI in order to estimate its associated SNR. Further, we propose such estimated quantity as a tool for the detection of brain activity. It is of interest to note that by employing the approach presented, we are able to obtain the posterior distribution of the SNR and therefore to compute any probability of interest, with a minimum pre-processing of the images and making no assumptions about the stimulation paradigm. The analysis of a synthetic fMRI study shows the performance of the method proposed and further results are presented on the application of the model for the analysis of a real fMRI data set, in order to illustrate some practical issues.

Keywords: Brain activity detection; dynamic linear models; fMRI; signal-to-noise ratio.

Paul Fearnhead[1], Gareth O. Roberts[2], **Giorgos Sermaidis**[1] and Krysztoff Latuszynski[2]

[1]University of Lancaster, UK, [2]University of Warwick, UK.

Exact inference for discretely observed multivariate diffusion processes using the pseudo-marginal approach

Bayesian inference for di_usion processes modelled by multivariate stochastic differential equations (SDEs) is a very challenging task due to the intractability of the dynamics of the process and the unavailability of the transition density at times other than infinitesimally small. Most methods rely on high frequency imputation and discrete-time approximations of the continuous-time model, thus leading to biased inference. Utilising retrospective sampling techniques (Papaspiliopoulos and Roberts, 2008), we develop a novel continuous-time importance sampling framework for multivariate diffusions, which enables us to estimate unbiasedly the transition density of the process. The estimators are subsequently used to sample exact draws (i.e. free of any discretisation error) from the posterior distribution of the parameters using a pseudo-marginal (Andrieu and Roberts, 2009) MCMC algorithm.

Keywords: Stochastic differential equation; multivariate; diffusion; unbiased; transition density.

References

Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for e_cient Monte Carlo computations. *Ann. Statist.*, **37**, 697-725.

Papaspiliopoulos, O. and Roberts, G. O. (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169-186.

Jukka Sirén and Jukka Corander

University of Helsinki, Finland

Inferring Bacterial Population Phylogenies from MLST Data

In recent years, multilocus sequence typing (MLST) databases have been introduced for many bacterial species. While the loci chosen for them capture only small amount of genetic variation of the bacterial genomes, the quantity of data available facilitates accurate analysis of bacterial population structure.

We propose a method for inferring phylogenetic tree structures of bacterial populations from MLST data based on differences in allele frequencies between populations.The stochastic variation in allele frequencies along the branches of the tree is modelled using an approximation to the infinite alleles Wright-Fisher model. The approximation is obtained by separating the alleles in to two groups: those observed in several different populations and those private to a single population. This is motivated by the notion that only the frequencies of the alleles observed in multiple populations are informative for inferring the divergence times of populations.

We show how to obtain samples from the posterior distribution of the model parameters using an Adaptive Metropolis algorithm. We also discuss the possibilities to infer the topology of the phylogenetic tree. The model is applied to both simulated and real bacterial data sets.

Keywords: MLST; bacterial populations; population phylogenies; infinite alleles model

David Suda and Paul Fearnhead

University of Lancaster, UK

Asymptotic behaviour of Importance Samplers for Diffusions

One popular approach for estimating transition densities of diffusions, and for inference about the paths of a diffusion given its start and end points, is to use importance sampling. These methods propose paths from a proposal diffusion starting at the correct value, and conditioned to hit a specified value at a future time *T*. There have been a number of studies on this, including work by Durham and Gallant (2002), Deylon and Hu (2006) and Fearnhead (2008). Here we look at general methods for designing such proposal processes, and propose other importance samplers.

For these different importance sampling approaches, we then study theoretically their behaviour when *T* is large. We look at a property, which we call stability, of the distribution of importance sampling weights as *T* gets large; and give conditions on the target and proposal diffusion which ensure whether the importance sampler is stable or not. These results give insight into which types of proposal processes should be used for different types of targets. We sustain these theoretical find-outs with examples of specific diffusion/sampler combinations, and corresponding simulation results obtained.

Keywords: diffusions, importance sampling, Bayesian inference.

P. Volf

Technical University of Liberec, Czech Republic

On analysis of occurrence of records in a time series with a trend

We consider a random series and are interesting in the analysis and model of occurrence of extremal values. As we assume that the series has a trend, we first select a proper regression model. From it a Markov random walk of a sequence of records can be derived, describing the probability of record occurrence and the record increment. Our main concern is the model selection, model fitting, and also the problem of prediction. As the transition probabilities of the chain of records are not tractable analytically, we prefer the use of Bayes methodology connected with the MCMC procedures. Thus, from the sample representing the posteriordistribution of regression model parameters we can directly obtain a representation of predictive distributions in the chain of records. In the end, we study possibilities to adapt such models to real data, namely to the progression of records in athletic events.

Keywords: record; regression model; random walk; Bayes analysis; prediction.

Krzysztof Latuszyński, Gareth Roberts and **Katarzyna Wolny**

University of Warwick, UK

Geometric Ergodicity of Heterogeneously Scaled Metropolis-Adjusted Langevin Algorithms.

Metropolis-adjusted Langevin algorithms with Langevin diffusion with unitary diffusion coefficient (standard MALA) generally have good convergence properties and are faster than Random Walk Metropolis (RWM). However, they cannot be applied for some target distributions. Motivated by this imperfection and inspired by the recent work on Manifold Monte Carlo Methods of Girolami and Calderhead (2011), we investigate properties of MALA with the Langevin diffusion satisfying the following SDE:

$$dX_t = \left(\tfrac{1}{2}\, \sigma^2(X_t)\, (\log(\pi(X_t)))' + \sigma(X_t)\, \sigma(X_t)' \right) + \sigma(X_t) dW_t$$

We prove that manifold MALA is geometrically ergodic for a wider class of target distributions than standard MALA. Ergodic properties of RWM, standard MALA and manifold MALA are compared in the table below.

algorithm	**$0 < \beta < 1$**	**$\beta = 1$**	**$1 < \beta < 2$**	**$\beta = 2$**	**$2 < \beta$**
RWM	N	Y	Y	Y	Y
MALA	N	Y	Y	Y	N
MMALA	Y		Y	Y	Y

Table: Geometric ergodicity of RWM, standard MALA and manifold MALA for target $\pi(x) \propto \exp\{-|x|^\beta\}$. N = geometric ergodicity fails, Y = geometric ergodicity holds.

Keywords: geometric ergodicity; Metropolis-adjusted Langevin algorithm; Markov chain Monte Carlo; Langevin diffusion;

References:

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods: (with discussion). *J.R.S.S. B.*, **73**, 1-37.

Roberts, G.O. and Tweedie, R.L. (1996). Exponential Convergence of Langevin distributions and their discrete approximations (1996). *Bernoulli* **2**, 341-63.

Jukka Corander[1,3], Yaqiong Cui[1], Timo Koski[2] and **Jie Xiong**[1]

[1]University of Helsinki, Finland, [2]Royal Institute of Technology, Stockholm, Sweden, [3]Åbo Akademi University, Åbo, Finland

Predictive Sequential Classification

We develop inductive rules for sequential probabilistic supervised and semisupervised classification of data arising from multiple finite alphabets. The rules are based on combining a decision-theoretic formulation with predictive representation of data derived under random urn models and hidden Markov models, as well as generalized exchangeability. Our classification rules have attractive theoretical properties that conditional on the training data predict the

labels of sequentially observed items based on their features and the previously observed items whose true labels are unknown. Optimal classification rules are defined under different utility functions and we examine the behavior of the supervised rules as a function of the amount of sequential testing data to show the advantages of the simultaneous classifiers compared to a standard predictive classifier which assigns all test items independently. Asymptotic properties of the predictive sequential classifiers are investigated and we illustrate that the advantages of considering data from the previous items reaches a saturation when the amount of query data increases. Multiple possible stochastic learning algorithms are considered for implementing the sequential classifiers.

Keywords: Bayesian modeling; inductive learning; predictive inference; sequential probabilistic classification.