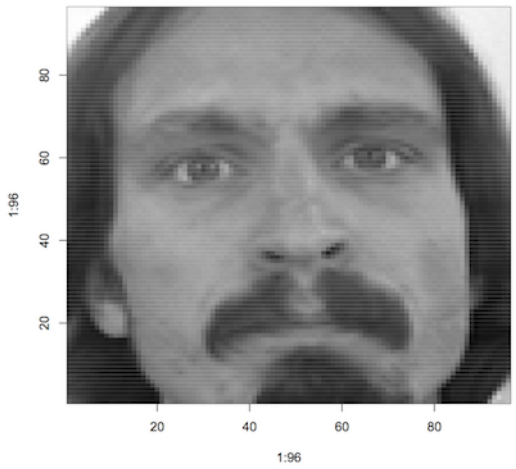
Facial Keypoints Detection

Supervisor: Ben Graham

Here is a picture of a face, stored as a 96x96 gray scale image:



That is not me by the way. There is a ongoing Kaggle competition [1], due to finish on 31 December 2014, where the challenge is to locate in the picture the facial keypoints: the eyes, eyebrows, nose, and so on.

Many modern cameras perform a similar task to help focus when taking pictures of people.

The aim of this project is to study the existing algorithms for facial keypoint detection and ~~win~~ compete in the Kaggle competition.

Experience with programming in R or Python required.

1. http://www.kaggle.com/c/facial-keypoints-detection

2. Paul Viola and Michael Jones, Robust Real-time Object Detection, International Journal of Computer Vision, 2001

Gradient descent algorithms

Supervisor: Ben Graham

A key component of many optimization algorithms is some form of gradient descent. For example, to train an artificial neural network (ANN) to preform a particular task, the network parameters must be fine tuned. For example, ANNs can be trained to do optical character recognition with accuracy exceeding 99% on some datasets, such as MNIST:

There are a large number of variations of the key principle: calculate a collection of derivatives and then move in that direction. The techniques of dropout and dropConnect for turning ANNs into ensemble learners require a drastic re-interpretation of how gradient descent works.

The goal of this project is to explore the family of gradient descent algorithms, and to see how they interact with dropout.

Experience with programming in R or Python required.

1. Improving neural networks by preventing co-adaptation of feature detectors Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R. Salakhutdinov http://arxiv.org/abs/1207.0580

2. Regularization of Neural Networks using DropConnect, Li Wan, Matthew Zeiler, Sixin Zhang, Yann LeCun, Rob Fergus http://cs.nyu.edu/~wanli/dropc/

3. No More Pesky Learning Rates Tom Schaul, Sixin Zhang, Yann LeCun http://arxiv.org/abs/1206.1106

The pylearn2 machine learning environment

Supervisor: Ben Graham

Pylearn2 is [1]:

> "... a machine learning research library. This does not just mean that it is a collection of machine learning algorithms that share a common API; it means that it has been designed for flexibility and extensibility in order to facilitate research projects that involve new or unusual use cases. In this paper we give a brief history of the library, an overview of its basic philosophy, a summary of the library's architecture, and a description of how the Pylearn2 community functions socially."

Pylearn2 is built on top of the Python programming language using the Theano array handling library.

The aim of this project is to explore how Pylearn2 can be applied to the problem of image recognition, for example how do you tell if a picture contains a dog or a cat. [4]

Familiarity with Python programming will be required to use Pylearn2.

1. Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. "Pylearn2: a machine learning research library". arXiv:1308.4214

2. http://deeplearning.net/software/pylearn2/

3. http://deeplearning.net/software/theano/

4. http://www.kaggle.com/c/dogs-vs-cats

Recognizing characters by their rough-path signatures

Supervisor: Ben Graham

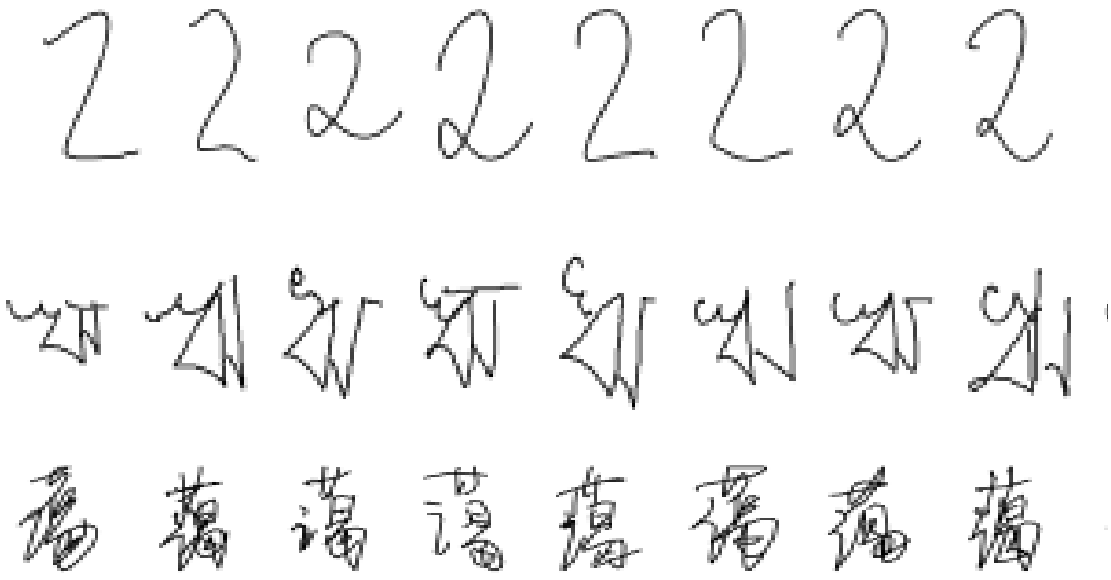
The "online character recognition problem" is the task of classifying a collection of paths representing a handwritten character. Rough paths theory provides a tool, the signature of iterated integrals,

$$X_{0,1}^n = \int_{0<u_1<\ldots<u_n<1} 1\, dX(u_1) \otimes \ldots \otimes dX(u_n) \in \mathbb{R}^{d^n}$$

for describing the shape of a path. If you think of the path as a driving signal for a differential equation

$$dY(t)f(Y(t)) = dX(t)$$

then the signature characterizes the $effect$ of the path. The major challenge is to identify pairs of signatures, that may be far apart in a Euclidean sense, but that look similar to a human reading them.



Experience with programming in Python required.

- Terry Lyons and Zhongmin Qian, System Control and Rough Paths, Clarendon Press, Oxford Mathematical Monographs
- Pen-Based Recognition of Handwritten Digits Data Set
  http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits
- Online Handwritten Assamese Characters Dataset Data Set
  http://archive.ics.uci.edu/ml/datasets/Online+Handwritten+Assamese+Characters+Dataset

Project: *Adaptive tempering for minimum variance estimation of model evidence.*

Supervisors:
Email c.oates@warwick.ac.uk, m.girolami@warwick.ac.uk.

Dr Chris Oates (Statistics)

Prof Mark Girolami (Statistics)

Background: Bayesian model selection relies on the "evidence" $p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta$, where y are the data and m is a candidate model. Yet for almost all models of interest, the evidence is unavailable in closed form and must be computed numerically. Recent advances in stochastic integration provide promising (low variance) estimators of evidence, but remain extremely computationally intensive and scale poorly to high-dimensional parameter spaces. This is essentially because the posterior $p(\theta|y, m)$ can be an extremely unpleasant object, even if the prior $p(\theta|m)$ is nice!

Objectives: This project seeks to develop an estimator for evidence that achieves the minimum possible variance among all estimators within its class. Specifically, we consider the class of tempered integration estimators of the form $\log p(y|m) = \int_0^1 E_{\theta|y,t} \log\{p(y|\theta, m)\}\, dt$ where the expectation is taken with respect to the "power posterior" $p_t(\theta|y) \propto p(y|\theta, m)^t p(\theta|m)$. The power posterior interpolates between the prior distribution ($t = 0$) and the posterior distribution ($t = 1$) over model parameters θ. Intuitively, this allows us to exploit some of the niceness of the prior when we have to deal with the unpleasant posterior, leading to more stable estimators [1]. Rewriting the estimator as $\log p(y|m) = \int_0^1 \left[\frac{E_{\theta|y,t} \log\{p(y|\theta,m)\}}{p(t)}\right] p(t)dt$ we see that we are free to choose the "temperature schedule" $p(t)$ that controls how quickly we move from the prior to the posterior. Empirical evidence suggests that choice of schedule has a significant influence on variance of the estimator $\hat{p}(y|m)$, yet in most applications the temperature schedule is chosen heuristically. In this project we will show that an optimal schedule exists that minimises variance of this estimator, and moreover this schedule can be computed "on-the-fly" using tools from adaptive ODE solvers.

"What the student will do": He/she will exploit recent advances in computational statistics [2,3] to develop state-of-the-art tools for Bayesian model selection. Estimation of evidence is an extremely important statistical challenge and, if completed to a high enough standard, there will be an opportunity for publication in a computational statistics journal. He/she can expect to gain a working knowledge of some advanced Markov chain Monte Carlo techniques and experience in applied Bayesian statistics. These skills are currently highly sought-after in the quantitative sciences and the student will be well prepared for their future research.

Prerequisites: A basic competence in a suitable programming language (MATLAB, R, C, etc.).

[1] Friel, N., & Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(3), 589-607.

[2] Calderhead, B., Girolami, M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data An. 53:4028-4045.

[3] Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science, 163-185.

Project: *Unbiased thermodynamic integration via Russian Roulette sampling*

Supervisors:
Email c.oates@warwick.ac.uk, m.girolami@warwick.ac.uk.

Dr Chris Oates (Statistics)

Prof Mark Girolami (Statistics)

Background: Bayesian model selection relies on the "evidence" $p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta$, where y are the data and m is a candidate model. Yet for almost all models of interest, the evidence is unavailable in closed form and must be computed numerically. Recent advances in stochastic integration provide promising (low variance) estimators of evidence. This project focuses on one such technique known as thermodynamic integration, that uses estimators of the form $\log p(y|m) = \int_0^1 E_{\theta|y,t} \log\{p(y|\theta, m)\}\, dt$ where the expectation is taken with respect to the "power posterior" $p_t(\theta|y) \propto p(y|\theta, m)^t p(\theta|m)$. (The power posterior interpolates between the prior distribution ($t = 0$) and the posterior distribution ($t = 1$) over model parameters θ. Intuitively, this leverages regularity of the prior to obtain more stable estimators of model evidence [1].) However, in practice the integral over t is computed numerically, which renders the estimator biased.

Objectives: This project seeks to develop an unbiased approach to thermodynamic integration. Such a methodology could be used throughout applied Bayesian statistics [2] and, if successful, would extend the class of models that are amenable to "exact" Bayesian analysis. To achieve this goal the student will exploit Russian Roulette sampling and related ideas from stochastic truncation of infinite series [3]:

Russian Roulette: Consider the infinite sum $\mathrm{S} = \sum \mathrm{s_j}$. Then the naïve "estimator" of S given by the finite truncation $\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{s_j}$ is biased in general. However, suppose $\mathrm{J} \in \{1,2, \dots\}$ is chosen randomly with p.m.f. $\mathrm{p(J = j) = p_j}$, and consider the estimator $\hat{\mathrm{S}}_{\mathrm{J}} = \frac{\mathrm{s_J}}{\mathrm{p_J}}$. Then we have $\mathrm{E}(\hat{\mathrm{S}}_{\mathrm{J}}) = \sum_{\mathrm{j}=1}^{\infty} \frac{\mathrm{s_j}}{\mathrm{p_j}} \mathrm{p_j} = \mathrm{S}$, i.e. an unbiased estimator that requires finite computation!

"What the student will do": He/she will exploit Russian Roulette sampling in the context of thermodynamic integration, expressing integrals over temperature t as infinite series expansions indexed by mesh refinement. He/she can expect to gain a working knowledge of some advanced Markov chain Monte Carlo techniques, exposure to elements from numerical analysis, and experience in applied Bayesian statistics. These skills are currently highly sought-after in the quantitative sciences and the student will be well prepared for their future research. Stochastic integration for Bayesian computation is an extremely active area of research here at Warwick and, if completed to a high enough standard, there will be an opportunity for publication in a computational statistics journal.

Prerequisites: A basic competence in a suitable programming language (MATLAB, R, C, etc.).

[1] Friel, N., & Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(3), 589-607.

[2] Calderhead, B., Girolami, M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data An. 53:4028-4045.

[3] Girolami, M., Lyne, A. M., Strathmann, H., Simpson, D., & Atchade, Y. (2013). Playing Russian Roulette with Intractable Likelihoods. arXiv preprint arXiv:1306.4032.

Project: *Zero-variance thermodynamic integration for Bayesian model selection*

Supervisors:
Email c.oates@warwick.ac.uk, m.girolami@warwick.ac.uk.

Dr Chris Oates (Statistics)

Prof Mark Girolami (Statistics)

Background: Bayesian model selection relies on the "evidence" $p(y|m) = \int p(y|\theta,m)p(\theta|m)d\theta$, where y are the data and m is a candidate model. Yet for almost all models of interest, the evidence is unavailable in closed form and must be computed numerically. Recent advances in stochastic integration provide promising (low variance) estimators of evidence [1], but remain extremely computationally intensive and scale poorly to high-dimensional parameter spaces.

Objectives: This project seeks to exploit the "zero-variance" technique [3] for reduced variance estimation of model evidence:

Zero-variance technique: Consider the problem of estimating the expectation $E[f(\theta)]$ by Monte Carlo $S_N^f = \sum_{j=1}^N f(\theta_j)$. Suppose $w = [w_1, \dots, w_m]$ are random variables with $E[w] = 0$. Then the function $g(\theta) = f(\theta) + a^T w$ satisfies $E[g(\theta)] = E[f(\theta)] + a^T E[w] = E[f(\theta)]$. Through careful choice of the "control variates" w we can design g such that $Var[S_N^g] < Var[S_N^f]$, i.e. the Monte Carlo estimator S_N^g of $E[g(\theta)]$ has lower variance than the original estimator S_N^f.

Zero-variance techniques have recently been exploited within advanced differential geometric Monte Carlo [2,3]. Here the student will extend this methodology to the important problem of estimating the evidence for Bayesian model selection, via thermodynamic integration:

Thermodynamic integration: Consider estimators of the form $\log p(y|m) = \int_0^1 E_{\theta|y,t} \log\{p(y|\theta,m)\}\, dt$ where the expectation is taken with respect to the "power posterior" $p_t(\theta|y) \propto p(y|\theta,m)^t p(\theta|m)$. The power posterior interpolates between the prior distribution ($t = 0$) and the posterior distribution ($t = 1$) over model parameters θ. Intuitively, this leverages regularity of the prior to obtain more stable estimators of model evidence [4].

"What the student will do": He/she will exploit the zero-variance technique in the context of thermodynamic integration, using control variates to obtain low-variance estimates for the expectations $E_{\theta|y,t} \log\{p(y|\theta,m)\}$ with respect to the power posterior. The methodology will be evaluated in an area of the student's interest, which could include generalised linear regression or parameter inference for dynamical systems, for example. He/she can expect to gain a working knowledge of advanced Markov chain Monte Carlo techniques and experience in applied Bayesian statistics. These skills are currently highly sought-after in the quantitative sciences and the student will be well prepared for their future research. If completed to a high enough standard, there will be an opportunity for publication in a computational statistics journal.

Prerequisites: A basic competence in a suitable programming language (MATLAB, R, C, etc.).

[1] Calderhead, B., Girolami, M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data An. 53:4028-4045.

[2] Papamarkou, T., Mira, A., Girolami, M. (2013). Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms.

[3] Girolami, M., Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Statist. Soc. B 73(2):1-37.

[4] Friel, N., & Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(3), 589-607.

Supervisor - Dr Krys Łatuszyński

Cortical Excitability and Sleep Disorders in Parkinson's Disease - Graphical Models approach

The aim of this project is to design a model for relating sleep disorders associated with the Parkinson disease to cortical excitability parameters measured during transcranial magnetic stimulation.

The project will aim at inferring structure of a graphical model (Bayesian Network and/or Markov Network) describing dependencies and independencies between cortical excitability parameters and sleep quality parameters. Implementing a range of algorithmic approaches that include classical graph algorithms as well as Markov chain Monte Carlo techniques, will be necessary.

The project will be co-supervised by a medical doctor, Dr Jakub Antczak (Sleep Center at the Kuchwald Hospital, Chemnitz, Germany and Department of Clinical Neurophysiology, Institute of Psychiatry and Neurology, Warsaw, Poland) and will require efficient communication in both medical and statistical language.

To work on this topic the student should have interest in:
Graphical modelling, Bayesian networks, MCMC, coding in R or other language, medical applications.
Prior familiarity with graphical models is not necessary but willingness to selfstudy them is essential.

Related literature includes: [DM12, Lau02] and selected chapters from one of the following books [KF09], [Whi09], [Smi10], [Lau96], [Edw00].

References

[DM12] Sophie Donnet and Jean-Michel Marin. An empirical bayes procedure for the selection of gaussian graphical models. *Statistics and Computing*, 22(5):1113–1123, 2012.

[Edw00] David Edwards. *Introduction to graphical modelling*. Springer, 2000.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[Lau96] Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.

[Lau02] Steffen Lilholt Lauritzen. *Lectures on contingency tables*. Institute of Mathematical Statistics, University of Copenhagen, 2002.

[Smi10] Jim Q Smith. *Bayesian decision analysis: principles and practice*. Cambridge University Press, 2010.

[Whi09] Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009.

Supervisor: Prof Christian Robert

Approximating marginal likelihoods by nested sampling

Marginal likelihoods are integral quantities customarily found in Bayesian testing of hypotheses and in Bayesian model comparison. While there are many simulation-based approches to the approximation of such integrals,

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta\,,$$

as shown in e.g. Chen et al. (2000) and Friel and Pettitt (2008), this project aims at evaluating and possibly improving the performances of the nested sampling method of Skilling (2006), which proceeds by slicing the likelihood into level sets, in settings where alternatives are available, as described in the survey of Cameron and Pettitt (2013). The recent modification of nested sampling made by Birge et al. (2012) will also be studied.

Prerequisites: Rudiments of Bayesian statistics, simulation methods, and programming skills

Specifics: Prof Robert will be available in Warwick on the weeks of June 30 and June 21, with email and Skype communication being possible most of August.

References

Birge, J. R., Chang, C. and Polson, N. G. (2012). Split Sampling: Expectations, Normalisation and Rare Events. *ArXiv e-prints*. `1212.0534`.

Cameron, E. and Pettitt, A. (2013). Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis. *ArXiv e-prints*. `1301.6450`.

Chen, M., Shao, Q. and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Friel, N. and Pettitt, A. (2008). Marginal likelihood estimation via power posteriors. *J. Royal Statist. Society Series B*, **70(3)** 589–607.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1(4)** 833–860.

Stochastic ranking processes

Supervisor: Dr David Croydon

The stochastic ranking process is a mathematical model that was introduced in order to describe the evolution over time of the rankings of books by a widely-used online bookshop (i.e. Amazon!). The initial aim of this project will be to review the mathematical results in this area (this will require understanding something about Poisson processes, and the convergence of probability measures). Following this, the student might investigate properties of the model, such as "how likely is it that the best/a particular book appears in the top 10 at any one time?", or explore whether the various modelling assumptions are reasonable.

Further background on the topic and references can be found in the short survey article 'Stochastic ranking process and web ranking numbers' by T. Hattori. This is available here:

web.econ.keio.ac.jp/staff/hattori/kyushu09.pdf

See also:

web.econ.keio.ac.jp/staff/hattori/amazone.htm

The student taking this project should have a strong background in stochastic processes, i.e. have taken ST406.

Adequacy of Bayesian massive multiple testing control

Supervisor: Dr David Rossell

Description: Currently many applications require testing thousands or even millions of hypothesis simultaneously. One important goal is to control the False Discovery Rate, i.e. the proportion of false positives amongst all tests that have been rejected. Within the Bayesian paradigm, the common view is that multiple testing is automatically taken care of as long as one uses an adequate prior structure. The goal is to study whether and when this is really the case, in particular focusing on the effect of the prior distribution on the parameters.

Main concepts: massive multiple testing, Bayesian hypothesis testing, False Discovery Rate control, prior distributions

Required knowledge: familiarity with classical hypothesis testing and Bayesian modelling. R programming.

References:

"An exploration of aspects of Bayesian multiple testing" (J Scott, JO Berger). Journal of Statistical Planning and Inference 136.7: 2144–62 (2006)

"Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem" (J Scott, JO Berger). The Annals of Statistics 38.5: 2587-2619 (2010).

On the use of non-local prior densities in Bayesian hypothesis tests. (VE Johnson, D Rossell). Journal of the Royal Statistical Society: Series B, 72,2:143-170 (2010)

The transition density function of a genetic diffusion process.

Supervisor : Dr Paul Jenkins

The time-evolution of the frequency of a gene in a large population is often modelled as a diffusion process. To understand it fully we'd like to write down the transition density function f(x,y,t), which gives the probability density that the frequency of the gene is now y given that it was at x a time t ago. In all but the simplest models this function is not known, but we can try to approximate it. One very interesting approach is to find a spectral eigenfunction representation of the function, i.e. to write down a PDE for the diffusion and then to seek a solution which is separable in y and t. The aim of this project is to write a survey of the use of transition density functions in genetic diffusion models. In particular, how does the use of duality arguments let us understand what our expressions actually mean? This has not been solved in every genetic model, so there is scope for the ambitious student to tackle these open problems. For example, how can we incorporate the biological process of recombination? A background in probability theory, and in stochastic processes in particular, would be advantageous.

Suggested reading:

Etheridge A. M. & Griffiths (2009). A coalescent dual process in a Moran model with genic selection. Theoretical Population Biology, 75, 320-330.

Griffiths R. C. & Spano D. (2010). Diffusion processes and coalescent trees. In Bingham, N. H. and Goldie, C. M., editors. Probability and Mathematical Genetics, Papers in Honour of Sir John Kingman, volume 378 of LMS Lecture Note Series, chapter 15, pages 358-375. Cambridge University Press.

Stopping time resampling

Supervisor: Dr Paul Jenkins

Stopping time resampling is an example of a simulation technique known as Sequential Monte Carlo (SMC). The aim is to sample from a sequence of increasingly intractable distributions, such as a posterior distribution as more and more data arrives. SMC is a general technique that propagates a collection of particles to provide an empirical approximation to the distributions of interest. It is surprisingly powerful and is used in many different applications, such as target tracking, protein folding, population genetics, and solving PDEs. Stopping time resampling takes the idea one step further by allowing particles to evolve at different speeds. However, little is known about the best way to achieve this. This project will address the question: How should we design our SMC algorithm so that the speeds of particles are optimal? It could be addressed theoretically but some simulation is probably inevitable, in which case some programming will be involved. ST407 Monte Carlo methods is useful background.

Further reading:

Doucet & Johansen (2011): A Tutorial on Particle Filtering and Smoothing: Fifteen years later. Chapter 8.2 in: The Oxford Handbook of Nonlinear Filtering, Oxford University Press.

Chen et al. (2005): Stopping-time resampling for sequential Monte Carlo methods, Journal of the Royal Statistical Society Series B 67:199-217.

Jenkins (2012): Stopping-time resampling and population genetic inference under coalescent models. Statistical applications in genetics and molecular biology, 11(1), Article 9.

Inside-Out: Characterisation of CT noise in projection and image space

Supervisor: Dr Julia Brettschneider & Dr Tom Nichols

Additive layer Manufacturing (ALM) is also known as "3D printing", a technique which creates objects directly from digital design data by the layer-wise addition of material. There is tremendous interest in this technology, especially because it can produce cost-effective small runs of objects with complex internal structure. A significant challenge at present, however, is the problem of determining the extent to which the digital design has been correctly realized: direct verification typically involves lengthy analysis of individual manufactured objects using Computed Tomography (CT) scans, slowing down the production process. This project is a part of a new collaboration between Statistics and the Warwick Manufacturing Group to create methods to speed up the quality assurance process.

The focus of this project is the exploration and identification of the noise in the CT raw and reconstructed data. It is hypothesized that use of raw CT "projection" data can be used to detect anomalies with out the time-consuming process of converting the 2D projection data into 3D images. But this requires characterisation of the noise in the projection data, which may be highly skewed due nature of CT (based on radioactive counts transmitted through the object). Time permitting, inferential procedures will be created that test for anomalies in the collected data.

A vital prerequisite for this project is data analysis experience as well as programming experience, ideally with R, but matlab, C or other languages would also be great. Experience with any sort of image data also helpful.

Residual analysis for spatial point processes

Supervisor: Dr Elke Thonnes

In this project the student will examine and evaluate methods of fitting spatial point processes to observed point patterns. The project requires an interest in computational work, working knowledge of R and a good background in probability theory and applied stochastic processes. The most relevant reference is the paper by A. Baddeley , R. Turner , J. Møller , M. Hazelton "Residual analysis for spatial point processes" which appeared in the *Journal of the Royal Statistical Society Series B,* 2005, vol. 67, issue 5, pages 617-666.

Modelling and Inference for time varying periodicity using Time Varying Autoregressive (TVAR) models

Supervisor: Dr Bärbel Finkenstädt

Many experiments investigating circadian clocks (periods of length circa 24 hours) in organisms are in conditions where (a) the exact period length is unknown and (b) the cyclic pattern might be affected by some treatment during the experiment. Hence, in order to be able to study the circadian clock in living organisms under varying experimental conditions, progression of a disease and/or treatments that change over time one has to extend the analysis to allow for a time varying frequency and phase. In this project we will investigate one suitable approach for time-varying spectral estimation namely the class of time-varying autoregressive (TVAR) models. TVAR models enable nonstationary spectral analysis by generating instantaneous estimates of the power spectrum thus providing a high time-frequency resolution of the spectrum while an analysis of the eigenstructure of the TVAR evolution matrix yields a decomposition of the signal into latent processes that are conceptually modeling sinusoids of time-varying amplitude, phase and frequency characteristics.

Dynamic linear model (DLM) theory can be used to achieve parameter estimation, smoothing and forecasting for this class of models.
TVAR state space models which have been successfully applied to EEG data (Prado and West 2010, Ting \etal 2011) and heart rate variability HRV (Tarvainen \etal 2006, Mendez \etal 2010). Experimental data on mouse activity during cancer development will be provided. Some familiarity with concepts of the DLM and state-space modeling is essential.

Visualization and inference for high-dimensional epigenetics data

Supervisor: Dr David Rossell & Professor Mark Girolami

Description: Large initiatives such as the ENCODE or modENCODE projects collect increasing amounts of data on the genome-wide locations of epigenetic factors in humans or other model organisms. Most approaches to the problem have been based on Hidden Markov Models. While these proved useful, their ability to combine data from different sources or to correct biases is limited. We recently developed chroGPS, a method to visualize and integrate large amounts of epigenetic data (Font-Burgada et al, 2013). The approach is based on dimensionality reduction and was formulated as a descriptive method that produces intuitive two or three-dimensional maps.

The goal of the MSc project is to provide methodology to compare maps between conditions, i.e. perform inference. We use a Bayesian probabilistic framework, which naturally enables uncertainty characterization and hypothesis testing. Specifically, we consider two families of probabilistic dimensionality-reduction models: the translation invariant Wishart-Dirichlet process (TIWD, Vogt et al 2010) and probabilistic self-organizing maps

Main concepts: dimensionality reduction, Bayesian model selection, self-organizing maps, Wishart-Dirichlet process

Required knowledge: familiarity with dimensionality reduction, Bayesian modelling and R programming.

References:

J. Font-Burgada, O. Reina, D. Rossell, F. Azorin. *ChroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome.* Nucleic Acids Research, 2013, doi: 10.103/nar/gkt1186, 1-12.

J.E. Vogt, S. Prabhakaran, T.J. Fuchs, V. Roth. *The translation-invariant Wishart-Dirichlet process for clustering distance data.* Proceedings of the 27th International Conference on Machine Learning, 2010.

Did first-millennium English builders use a standard unit of measurement?

Supervisor: Prof. W.S.Kendall

Professor John Blair of Oxford University asked me whether it was possible to answer this question statistically, using measurements taken from ground plans of Anglo-Saxon churches and other buildings. Historians and archaeologists who have looked at these ground plans have become convinced that the builders used a unit of measurement which was (somehow) fixed nationally; the Saxon perch (for example, Huggins, 1991). Adapting a simulation-based approach pioneered by my father (D.G.Kendall, 1974) in a different archaeological context, I have developed a statistical approach to this question. The project is about understanding the statistical background to the question, verifying the results for yourself and considering extended datasets, and considering links to other statistical techniques.

While this project is firmly locked in to the historical period 800-1100 AD, some of the fundamental statistical questions are related to statistical issues underlying, for example, the hunt for the Higgs boson.

In this project you will:

(a) confirm my results using statistical simulation with the help of the statistical package R (or S);

(b) compare and contrast a Bayesian approach due to Freeman (1976);

(c) review other recent archaeological applications of these approaches (example: Pakkanen, 2004);

(d) examine links to multiple hypothesis testing, using references which I will supply.

To undertake this project successfully, you should be confident in the use of R (or S) and have taken an undergraduate module in complex analysis. To get a taste of the project, look through the D.G.Kendall (1974) paper, which is very readable and written for a non-statistical audience.

Indicative Bibliography

P. R. Freeman, â€œA Bayesian analysis of the megalithic yard,â€ Journal of the Royal Statistical Society. Series A, vol. 139, no. 1, pp. 20-55, 1976.

P. J. Huggins, â€œAnglo-Saxon Timber Building Measurements: Recent Results,â€ Medieval Archaeology, vol. 35, pp. 6-28, 1991.

D. G. Kendall, â€œHunting Quanta,â€ Philosophical Transactions of the Royal Society, London, Series A, vol. 276, no. 1257, pp. 231-266, May 1974.

J. Pakkanen, â€œThe Toumba Building at Lefkandi: A Statistical Method for Detecting a Design-Unit,â€ The Annual of the British School at Athens, vol. 99, pp. 257-271, 2004.

Tempering Strategies for Sequential Monte Carlo

Supervisor: Dr Adam Johansen

When performing Bayesian inference using so-called sequential Monte Carlo methods in which we obtained weighted samples from a sequence of intermediate distributions as a tool to help us sample from the actual posterior distributions of interest, there are two main strategies for specifying this sequence of distributions. One is to gradually introduce observations so that each distribution in the sequence is an "intermediate posterior" distribution depending upon only a subset of the data. The other is to raise the likelihood to a power which increases gradually from 0 to 1 along the sequence of distributions so that the intermediate distributions are influenced by all of the observations but to a lesser degree than they are under the posterior distribution.

This project will investigate the relative strengths of the two strategies and if time allows consider intelligent combinations of the two approaches.

Recommended Reading:

N. Chopin. A sequential particle filter method for static models. Biometrika, 9(3):539â€“551, 2002.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. Journal of the Royal Statistical Society B, 633):411â€“436, 2006.

Pre-requisites: ST407 Monte Carlo methods will be /very/ useful.

The project /will/ involve implementing Monte Carlo algorithms and familiarity with a programming language (R would be sufficient) is essential.

Errors-in-variables regression and network inference

Supervisor: Dr Simon Spencer

Errors-in-variables regression is where the predictors and the response both suffer from measurement error. This project will begin by reviewing errors-in-variables regression approaches and comparing the performance of variable selection approaches for these models using simulated data. Secondly, an appropriate method will be selected for an application involving network inference with regression models. This project will require a good knowledge of linear models and use the statistical programming language R.

Section 9.5 of Greene, WH (1993). Econometric Analysis. Macmillan

---

Statistical inference for partially observed epidemic models

Supervisor: Dr Simon Spencer

Epidemic models can be used to investigate the mode of transmission of an infection and to test the effectiveness of intervention strategies before they are implemented in the real world. However, for a lot of diseases it can be very difficult to know exactly who has and has not been infected. For example about one half of influenza infections are thought to be asymptomatic, so that even the person infected doesn't know that it has happened. This project will investigate when it is possible to learn about these hidden infections using a simple stochastic model. The techniques will then be tested against a real influenza dataset. The model will be fitted using Markov chain Monte Carlo methods, and so knowledge of Markov chains and Bayesian statistics will be useful. This project will involve some computing using the statistical programming language R.

Andersson H and Britton T (2000). Stochastic epidemic models and their statistical analysis. Lecture notes in Statistics, Springer.

The Cramer-Lundberg model

Supervisor: Dr Larbi Alili

The aim of the model is to review the Cramer-Lundberg model in risk theory. The project should review the properties of compound Poisson processes. It should also include the differential equations method for the ruin probability, the adjustment coefficient method, the severity of ruin and the different methods for approximating the characteristic exponent. The probability of time to ruin and seal's formula should also be studied.

References:

Cramer, H. (1955) Collective Risk Theory. Skandia Jubilee Volume, Stockholm.

Schmidli, H. (2013) Risk Theory, Lecture notes. Available at http://www.math.ku.dk/~schmidli/rt.pdf

---

The Ammeter Risk model

Supervisor: Dr Larbi Alili

The aim of the project is to review the Ammeter model. This is similar to the Cramer-Lundberg model but with the Poisson process replaced by a mixed Poisson risk process. The project should include a short review of mixed Poisson processes. The student should review the adjustment coefficient method, the sub-exponential case and the finite time Lundberg inequality.

References:

Grandell, J. (1997) Mixed Poisson processes, Chapman & Hall, London.

Schmidli, H. (2013) Risk Theory, Lecture notes. Available at http://www.math.ku.dk/~schmidli/rt.pdf

---

The renewal risk model

Supervisor: Dr Larbi Alili

This is similar the Cramer-Lundberg model but the number of claims is a renewal process instead of a Poisson process. The project should include a review on renewal processes and martingales. It should also present the Sparre Anderson model and include the adjustment coefficient method and Lundberg's inequality. The Cramer-Lundberg approximation method and the sub exponential size distributions should also be presented

References:

Anderson, E. S. (1957) On the collective theory of risk in the case of contagion between the claims. Transactions XVth International Congress of Actuaries, New York, II.

Embrechts, P. and Veraverbeke, N. (1982) Estimates for the probability of ruin with special emphasis on the possibility of large claims. Insuarance Math. Econom. 1.

Schmidli, H. (2013) Risk Theory, Lecture notes. Available at http://www.math.ku.dk/~schmidli/rt.pdf

Nikos Zygouras (Email: N.Zygouras@warwick.ac.uk, Office C0.11, Dept. Statistics)

Project I: Queueing Theory.

Queueing theory is a central object in probability, operations research and engineering. It models arrivals and services in a network of systems (communication packages, internet etc.).

The student will learn the fundamentals of queueing theory and of the analysis of M/M/1 queues and will specialise on an appropriate application. Emphasis will be given into the algebraic structure, related to $(\max, +)$ algebras, underlying the solvability of some queuing models [BCOQ].

Project II: Central Limit Theorems and Wiener expansions.

The most classical result in probability theory is the Central Limit Theorem, originally stated as the fact that the sum of independent random variables is asymptotically Gaussian. However, there are many situations, e.g. when one is concerned with various estimators, where the quantity under study is not a linear functions of IID variables and still one obtains an asymptotic Gaussian behaviour. What is the structure that allows for asymptotic normality ? An important result in this direction was the $4th$ Moment Theorem [NP], which states that a random variable (in a, so called, fixed *chaos order*) is asymptotic Gaussian if *just* its $4th$ moments converges to 3.

The project will be concerned with understanding this theorem and seek various applications.

References

[BCOQ] F. Baccelli, G. Cohen, G.J. Olsder and J.-P. Quadrat. *Synchronization and Linearity: An Algebra for Discrete Event Systems.* Wiley, 1992.

[NP] D. Nualart, G. Peccati *Central limit theorems for sequences of multiple stochastic integrals* Ann. Probab.Volume 33, Number 1 (2005

[PT] G. Peccati, M. Taqqu *Wiener Chaos: Moments, Cumulants and Diagrams. (A survey with computer implementation)* Springer Verlag (2010) ISBN: 978-88-470-1678-1

Cluster/Feature models in statistical machine learning. Distributional properties and inferential power.

Supervisor: Dr Dario Spano

Cluster/feature models arise in statistical machine learning as a tool to infer the way in which a population can be clustered according to some (possibly multi-dimensional) criterion. If the criterion is one-dimensional, one only talks about clustering model, i.e. there is only one rule to judge if two items are considered either as equivalent or as distinct. In cluster/feature models, any two individual may be equivalent in some feature but distinct in some other. In Statistical machine learning, a powerful sampling scheme has been introduced to account for models where the number of criteria (features) is unknown a priori and potentially infinite. This sampling scheme is known as the Indian Buffet Process. Its structure can be interpreted in terms of the categories of Bayesian Nonparametric inference, and in this context it can be thought of as a multi-feature extension of the Blackwell-Mac Queen sampling scheme, which corresponds to the most powerful class of Bayesian Nonparametric prior measure, the Dirichlet Process prior.

This project aims at investigating the general distributional properties of the Indian Buffet process, and at identifying general ways the IBP can be extended to more flexible feature models. The investigation will be carried out at the level of the large sample limit (random measure), or at the finite-sample level (combinatorial properties of sampling distribution), or at both levels, starting from recent works of T. Broderick, J Pitman and M. Jordan. The inferential power of such model will also be studied and applications considered.

Requisites for this dissertation: familiarity with Bayesian modelling and with probability theory at an advanced level (desirable: confidence with material taught at M-level probability courses).

References:

1. Broderick, T, Pitman, J, and Jordan, MI.
   Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 2013.

1. Broderick, T, Boyd, N, Wibisono, A, Wilson, AC, and Jordan, MI.
   Streaming variational Bayes. *Neural Information Processing Systems*, 2013.