

## MPhil in Statistical Science, Projects

The MPhil project is an extended piece of individual work, carried out under the supervision of your MPhil supervisor. Most of the work for your project will be done between late October and the end of April, concurrently with lectures during term time, and in the Christmas and Easter vacations. You are required to produce a written report of around 50 pages on your project. The first draft of your report should be given to your supervisor for their comments by 30th April 2005. **The deadline for handing in the final version of your report will be a date in late June/early July**, and this date will be announced during the Michaelmas Term. The report will be formally assessed by the examiners. You are also required to produce a poster about your project and hand this in on the same date in June/July. There is a prize for the best poster. In addition, you will be asked to give a short talk on your project in June.

A list of MPhil projects for 2005–6 is given in this booklet. Each project has a title and a short description, together with the name(s) of the project supervisor(s).

Usually, at least one of the named project supervisors is a full-time staff member of the Statistical Laboratory. In these cases, this person will oversee your academic progress in the MPhil as well as supervising your project. He or she will be your first point of contact with the Laboratory, and will see you regularly throughout your MPhil course. For those projects where the named supervisors are not full-time Laboratory members, your MPhil supervisor will be assigned separately.

Please look at the list of projects, and complete the form at the back of the booklet to give us some idea of your interests. **Please hand in the completed form to Julia Blackwell in D1.17 by 4pm on Thursday 6th October 2005.**

MPHil in Statistical Science, Project Titles for 2005-06

1. Statistical problems in bio-informatics  
..... Dr P.M.E.Altham & Dr J.Huppert
2. The effects of CCTV on Crime: Analysis of an English national quasi-experimental multi-site evaluation ..... Prof D.P.Farrington & Dr P.M.E.Altham
3. Modifiers of Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers  
..... Prof. S.P.Brooks & Dr Antonis Antoniou
4. Hierarchical Bayesian Modelling for Predicting the Functional Consequences of Amino Acid Polymorphisms ..... Prof. S.P.Brooks
5. Inverting a neutrally buoyant gas dispersion process from the Earth's surface into the lower atmospheric boundary layer: what distribution of sources best explains our concentration data? ..... Prof. S.P. Brooks & Bill Hirst
6. Higher order numerical schemes in Finance ..... Dr P.Friz
7. Volatility Derivatives ..... Dr P.Friz
8. Disordered Physical Systems ..... Prof. G.R.Grimmett
9. Simulation of directed last-passage percolation..... Dr S.Großkinsky
10. Statistical Studies of Traffic Flow in Road Networks Automatic Bottleneck Detection Algorithms ..... Prof. F.P.Kelly & Dr R.J.Gibbens
11. Modelling Interest-Rate Data ..... Dr D.P.Kennedy
12. Calculating derivatives of parameters in population balances or "What are detergents sensitive to?" ..... Dr J.R.Norris & Dr M.Kraft
13. Parameter estimation for random processes ..... Dr J.R.Norris
14. Analysing Asset Class Data ..... Dr S.M.Pitts & Dr M.Lakshmanan
15. Numerical methods in ruin theory ..... Dr S.M.Pitts
16. Identifying outliers in routine hospital outcomes... Dr S.M.Pitts & Dr D.Ohlssen
17. Macro-Economic Factors and the Cox-Ingersoll-Ross Yield Curve Model  
..... Prof. L.C.G.Rogers & Jagjit Chadha
18. Range-Based Correlation Estimation for Market Price Data  
..... Prof. L.C.G.Rogers & Nicholas Brown
19. Optimal bandwidth choice in classification problems  
..... Dr R.J.Samworth

- 20. A comparison of different nonparametric regression techniques  
..... Dr R.J.Samworth
- 21. Complexity and queues in modern industrial systems  
..... Prof. Y.M.Suhov & G.D.M.Frizelle
- 22. Re-insurance and large deviations ..... Prof. Y.M.Suhov & Dr M.Kelbert
- 23. Additivity problems for quantum information channels ..... Prof. Y.M.Suhov
- 24. Models of Recovery Rates ..... Dr M.Tehranchi & Dr P.Cotton
- 25. Pricing Options by Monte Carlo Simulations ..... Dr M.Tehranchi
- 26. Metrics for Portfolio Evaluation ..... Dr M.Tehranchi & Dr D.Wischik
- 27. Numerical Solution of the Broadcasting Problem when Peers have Differing Upload Rates ..... Prof. R.W.Weber & J. Mundunger
- 28. A Study of Heavy-tailed Inter-event Distributions  
..... Prof. R.R.Weber & Dr A.Chaintreau

**1. Statistical problems in bio-informatics**

..... **Dr P.M.E. Altham & Dr Julian Huppert**

There is an international project called ENCODE (<http://genome.ucsc.edu/ENCODE>) which aims to provide a detailed analysis of 1genome, understanding how those genes are controlled, and the role of the 'junk' DNA. Dr Huppert is particularly interested in DNA which is not in the form of a double helix: how do these other structures relate to the info in ENCODE. For example, do they turn genes on and off?

ENCODE provides a very large library of publicly available data, The first task of the student, after absorbing the necessary biological background, will be to provide (using R) relevant summaries of selected datasets, and how they associate with 4-stranded (ie G-Quadruplex) DNA.

This project will be jointly supervised with Dr Julian Huppert at the Wellcome Trust Sanger Centre.

**2. The effects of CCTV on Crime: Analysis of an English national quasi-experimental multi-site evaluation** ..... **Prof D.P.Farrington & Dr P.M.E.Altham**

We have data from 12 CCTV schemes, from residential areas, town and city centres, and a city hospital. Police data on monthly crime numbers were collected before and after the implementation of CCTV in target (i.e., with CCTV), control (comparable but without CCTV) and buffer (surrounding, but no CCTV) areas, and in the corresponding police Divisions. Is CCTV effective in reducing crime? To answer this question, we need to adopt suitable models, which take account of factors such as

- (i) the heterogeneity of the different areas in the study
- (ii) the possible seasonal effects on crime numbers
- (iii) the possible overdispersion of the data relative to the Poisson distribution.

**3. Modifiers of Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers**

..... **Prof. S.P.Brooks & Dr A.Antoniou**

BRCA1 and BRCA2 are the most important breast cancer susceptibility genes identified to date. Mutations in these genes confer increased risks of breast and ovarian cancer. A number of studies have estimated the breast and ovarian cancer risks associated with BRCA1 and BRCA2 mutations, but there is substantial variability in these risks.

This project seeks to analyse data arising from the EMBRACE study, an ongoing epidemiological study of BRCA1 and BRCA2 mutation families in the UK and Ireland (<http://www.srl.cam.ac.uk/genepi/embraceindex.htm>). The aim is to look at factors that modify the breast cancer risks conferred by these mutations.

A number of risk factors are known to be associated with breast cancer in the general population,

but their effect on breast cancer risk in BRCA1 and BRCA2 mutation carriers is still conflicting and not well documented (among the results published so far). The aim of this project is to investigate the effect of some of those risk factors on the breast cancer risk among mutation carriers. Such modifiers could include: lifestyle factors (e.g., smoking, alcohol consumption, Body Mass index); reproductive factors (e.g., age at menarche i.e. the age when periods start, number of pregnancies, breast feeding, menopausal status); hormonal factors (e.g., oral contraceptive use, hormone replacement therapy etc); and prophylactic surgery (e.g. removal of ovaries).

The data consist of both affected and unaffected carriers who were sampled under non-standard designs. The complexity here is that mutations in BRCA1 and BRCA2 are rare and population based studies would identify very small numbers of carriers. Therefore, EMBRACE collects data on carriers identified through ongoing genetic testing programs. This allows large numbers of carriers to be identified rapidly. However, genetic testing is targeted at individuals with strong family history of breast cancer, which in turn means the sampling of carriers is not random with respect to disease status. Potential analyses need to address the issues of: (i) Over-sampling of affected individuals, and (ii) non-independence of subjects as carriers may come from the same family.

In this project we will use a range of statistical techniques to determine the extent to which the risk factors listed above modify the breast cancer risks among carriers, whilst accounting for the issues of over-sampling and dependence between subjects. There will be the potential to implement a range of techniques taught within the MPhil course as well as new techniques that will require a degree of independent learning from relevant books and papers.

The project will be jointly supervised with Dr Antonis Antoniou at Strangeways Research Laboratory in Cambridge.

#### **4. Hierarchical Bayesian Modelling for Predicting the Functional Consequences of Amino Acid Polymorphisms .....Prof. S.P.Brooks**

Genetic polymorphisms in deoxyribonucleic acid (DNA) coding regions may have a phenotypic effect on the carrier, e.g. by influencing susceptibility to disease. Detection of deleterious mutations via association studies is hampered by the large number of candidate sites; therefore methods are needed to narrow down the search to the most promising sites. For this, a possible approach is to use structural and sequence-based information of the encoded protein to predict whether a mutation at a particular site is likely to disrupt the functionality of the protein itself.

In this project we will look at the use of hierarchical Bayesian multivariate adaptive regression spline (BMARS) model for supervised learning in this context and assess their predictive performance by using data from mutagenesis experiments on lac repressor and lysozyme proteins.

The project will be strongly based around the work of Verzilli et al. (2005) who provide an analysis of these data together with a model-fitting code written for the statistical package R. We will be looking at alternative implementations of the methodology (alternative link functions, prior structures etc.) as well as assessing sensitivity and performance. There will be ample opportunity within the project for the investigation and implementation of cutting-edge statistical techniques as well as opportunities to learn more about this exciting application area.

#### **Reference**

Verzilli, C.J., Whittaker, J.C., Stallard, N. and Chasman, D. (2005) A hierarchical Bayesian model

for predicting the functional consequences of amino-acid polymorphisms. *Applied Statistics*, **54**, p191-206. (Available at <http://www.statslab.cam.ac.uk/~steve/Proj1.pdf>)

**5. Inverting a neutrally buoyant gas dispersion process from the Earth's surface into the lower atmospheric boundary layer: what distribution of sources best explains our concentration data? ..... Prof. S.P. Brooks & Bill Hirst**

LightTouch uses ultra-sensitive laser spectroscopic techniques to measure ethane gas concentrations in the atmospheric boundary layer to an accuracy 50PPT. We combine these measurements with wind velocity data and a gas dispersion model to invert the dispersion process and map surface ethane flux: thereby locating ethane gas seepages, which are associated with hydrocarbon accumulations at depth. We are very interested to identify different/better approaches to solve the inverse problem of locating the sources responsible for the concentration measurements obtained at different locations, times, wind conditions. So far we have had encouraging results from: a Bayesian Markov Chain Monte-Carlo method (Massive Inference/ Maxent), a Monte-Carlo least squares minimisation approach (incorporating different regularisations schemes and simulated annealing), and a Probabilistic Inversion method. Currently, our problem is ill-conditioned, very sparse and under constrained. We have to make full use of physical information relevant to the process: source positivity, time invariance, maximum expected source strengths and source strength distributions. We do have data sets for which we know the correct answer; and others for which we have limited independent corroboration of the answers obtained by our existing methods. Upcoming experimental field trials will be yielding further calibration data sets of much greater coverage and data quality.

For this project a detailed description of a recently developed Probabilistic Inversion approach will be provided. The project task is to look at how the implementation of this promising approach could be enhanced or extended. Some suggestions already exist: these will need to be evaluated along with the student's own ideas. The work will likely involve: exploring genetic algorithms, applying importance sampling, measures of convergence ... and things we haven't thought of! This is a real problem that will be used and has potential for publication.

The description of the Probabilistic Inversion approach can be downloaded from:

<http://ssor.twi.tudelft.nl/~andrey/Thesis.doc>

Other background material on the overall application (includes a short video) can be viewed at: <http://www.physics.gla.ac.uk/Optics/projects/oilProspection/>

This project will be jointly supervised with Bill Hirst from Shell Global Solutions.

**6. Higher order numerical schemes in Finance ..... Dr P.Friz**

The Milstein scheme is a 2nd order scheme for stochastic differential equations. In presence of 2 or more factors, this scheme requires simulation of the 2nd iterated Wiener integral, known as Lvy area. See [Numerical Solution of Stochastic Differential Equations, Peter E. Kloeden and Eckhard Platen; 1992].

- (1) In an elliptic setting, a new method proposes to recover Milstein accuracy without the costly simulation of the Lvy-area. See [Stochastic Calculus of Variations in Mathematical Finance, P. Malliavin and A. Thalmaier; 2005] and the references therein.

- (2) A completely different approach is based on cubature on Wiener space. See [*Cubature on Wiener space, T. Lyons and N. Victoir Proc. R. Soc. Lond. A (2004) 460, 169-198*] and [*Weak Approximation of Stochastic Differential Equations and Application to Derivative Pricing, with S. Ninomiya and N. Victoir; 2005*].

The project should contain (a) a survey of these new techniques (b) a construction of suited examples and (c) an implementation to demonstrate the numerical benefits.

**7. Volatility Derivatives ..... Dr P.Friz**

Volatility derivatives are products that provide exposure to the realised volatilities or variances of asset returns, while avoiding direct exposure to the underlying assets themselves. These products are attractive to investors who either wish to hedge volatility risk or who wish to take a view on future realised volatilities. For traders, the most popular such product is the Variance Swap for which a stable hedge (and therefore price) is available. Volatility swaps, on the other hand, are popular among clients but prices appears to be very model dependent.

The project should contain (a) a survey of the literature (b) discussion which models are (un)suited for volatility pricing (c) a comparison of volatility swaps under local and stochastic volatility, including numerical implementation.

**Reference**

J. Gatheral: CASE STUDIES IN FINANCIAL MODELING, available at

[http://www.math.nyu.edu/fellows\\_fin\\_math/gatheral/case\\_studies.html](http://www.math.nyu.edu/fellows_fin_math/gatheral/case_studies.html)

**8. Disordered Physical Systems ..... Prof. G.R.Grimmett**

Professor Grimmett is willing to supervise MPhil projects within his own area of research interest which he describes as the theory of 'disordered physical systems'. This is the study of stochastic processes in finite-dimensional Euclidean spaces, and is motivated by practical problems arising in more applied sciences including physics and biology. Examples include the percolation model and the contact model, being models respectively for a disordered medium and the spread of epidemics. These systems give insight into the behaviour of models for ferromagnetism, such as the Ising and Potts models. They also provide a host of really captivating and well motivated problems in probability.

A further interest is the application of modern probability to the structure of the integers, thereby obtaining contemporary insight into probabilistic number theory.

Further details can be obtained from Professor Grimmett's Web page.

<http://www.statslab.cam.ac.uk/~grg/>

**9. Simulation of directed last-passage percolation ..... Dr S.Großkinsky**

Directed last-passage percolation concerns directed paths of maximal weight between given end-points on the lattice  $\mathbb{Z}^d$ . It is one of many variants of first-passage percolation, a traditional and widely studied model of statistical mechanics. Recently, there has been particular interest in the case  $d = 2$ , since there are various links to other models, such as systems of queues in tandem, the asymmetric simple exclusion process (ASEP), random matrices and certain growth models.

A summary recommended for introductory reading can be found in [1], complementary references on the above connections in [2, 3]. In the following the model and possible projects are shortly explained.

### Last-passage percolation

To all sites  $z$  of the lattice  $\mathbb{Z}_+^2$  associate weights  $X(z) \geq 0$ , which are i.i.d. random variables with distribution  $P$ . Consider  $\Pi(z)$  to be the set of all directed paths  $\pi$  from the origin to the point  $z$ . Then

$$T(z) = \max_{\pi \in \Pi(z)} \sum_{v \in \pi} X(v)$$

is the *last-passage time* from the origin to  $z$  and the maximizer  $\pi^* \in \Pi(z)$  (which may be not unique) is called the *optimal path*.

### Typical questions of interest are:

- limiting distributions for appropriately rescaled passage-times  $T(z)/|z|$  for  $|z| \rightarrow \infty$ , depending on the direction of the limit
- scaling laws for the shape of optimal paths, such as transversal fluctuations

There have been recent rigorous results on such questions in the case when  $P$  is an exponential or a geometric distribution. For example, for all  $x, y > 0$  the distribution of

$$\frac{T([xn], [yn]) - n g(x, y)}{n^{1/3}}$$

converges to the *Tracy-Widom distribution*, which arises also as the limiting distribution of the largest eigenvalue of a GUE random matrix. While the asymptotic shape function  $g(x, y)$  depends on the distribution  $P$  (known exactly in the above cases), the scaling exponent  $1/3$  of fluctuations and the Tracy-Widom distribution are believed to be universal, i.e., valid for a general class of distributions  $P$ .

### The project

Due to the recursive structure ( $z = (z_1, z_2)$ )

$$T(z_1, z_2) = \max \{T(z_1 - 1, z_2), T(z_1, z_2 - 1)\} + X(z_1, z_2) ,$$

the problem is very suitable to be simulated on a computer. In a first step, numerical results could be compared to one of the exactly solved cases, to get an idea about the accuracy of the approximation and finite size effects. Secondly, one of the various open questions could be attacked, such as the universality of the Tracy-Widom distribution for different weight distributions. Possible problems in this context are

1. Compare the limiting distribution of  $T(z)$  with Tracy-Widom if  $P$  is for example a power law (or any other distribution).
2. If  $P$  is Bernoulli with probability  $p$ , the limiting distribution of  $T(z)$  is expected to be Tracy-Widom for  $p \rightarrow 0$  and Gaussian for large  $p$ . The crossover could be checked numerically, looking at the expected value  $\mathbb{E}T(z)/|z|$  for  $|z| \rightarrow \infty$  as a function of  $p$ .

These questions appear to be suitable, since  $T(z)$  is a quantity easy to access numerically. But in case of interest, one could also study the shape of the optimal paths as mentioned above.

Technically, the project involves writing a computer program which produces large two dimensional arrays of the last-passage times  $T(z)$ . This is not very complex, but involves (reliable) generation of random numbers from general distributions.

In a second step, the data have to be analysed, i.e., the limiting distribution or the mean value of  $T(z)/|z|$  for  $|z| \rightarrow \infty$  have to be estimated from a sequence of finite samples. This involves finite size scaling and some elementary statistical tools.

## References

- [1] James B. Martin, Last passage percolation with general weight distribution, [www.liafa.jussieu.fr/~martin/papers.html](http://www.liafa.jussieu.fr/~martin/papers.html) (2005)
- [2] David Aldous and Persi Diaconis, Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem, *Bull. Amer. Math. Soc. (N.S.)* **36**, 413-432 (1999)
- [3] Jinho Baik, Limiting distribution of last passage percolation models, [math.PR/0310347](http://math.PR/0310347)

## 10. Statistical Studies of Traffic Flow in Road Networks Automatic Bottleneck Detection Algorithms .....Prof. F.P.Kelly & Dr R.J.Gibbens

Chen et al (2004) use a variety of statistical techniques to investigate traffic flow in the Los Angeles freeway network using data gathered from loop detectors placed under the roadway. In particular, they study the spatial-temporal patterns of traffic flow and propose automatic bottleneck detection algorithms.

Recently, loop detector data has become available for portions of the UK motorway network. In particular, several years of per minute traffic data is available for the highly congested south west quadrant of the M25 motorway. In a separate paper, Chen et al (2003) describe related techniques to handle missing values in such data.

There is a pressing need to better understand the conditions under which our motorways operate and to derive methodologies using raw data gathered at the roadside to create valuable information for all road users and providers.

This project would follow on from the project of MPhil student Wiebke Werft in 2004/5 on journey time prediction algorithms from loop detector data. An informal presentation of her project is due to appear in a forthcoming issue of the RSS's Significance magazine (2005, vol 2, part 3).

The principal aim of this project are to build on the methodologies presented by Chen et al (2004) and to apply them to the UK data.

This project will be jointly supervised with Dr Richard Gibbens at the Computer Laboratory.

## References

- C. Chen, A. Skabardonis and P. Varaiya (2004) "Systematic Identification of freeway bottlenecks". Transportation Reserach Board, 83rd Annual Meeting, Washington, DC.

[[http://pems.eecs.berkeley.edu/Resources/Papers/chen\\_bottlenecks.pdf](http://pems.eecs.berkeley.edu/Resources/Papers/chen_bottlenecks.pdf)]

C. Chen, J. Kwon, J. Rice, A. Skabardonis and P. Varaiya (2003) "Detecting errors and imputing missing data for single loop surveillance systems". Transportation Reserach Board, 82nd Annual Meeting, Washington, DC.

[[http://pems.eecs.berkeley.edu/Resources/Papers/loops\\_asv2.pdf](http://pems.eecs.berkeley.edu/Resources/Papers/loops_asv2.pdf)]

Richard Gibbens and Wiebke Werft (2005) "Data gold mining: MIDAS and journey time predictors". Significance, vol 2 part 3. Forthcoming.

## **11. Modelling Interest-Rate Data ..... Dr D.P.Kennedy**

This project will consider aspects of fitting various interest rate models to historic bond price data and the pricing of derivatives on the associated markets. The exact slant of the investigation will be decided in conjunction with the student to match to his/her interests. To obtain an idea of the flavour of the work that might be undertaken you may consult previous M.Phil dissertations by L. Yee Tan (2003), Y. Yamamoto (2004) and A. Fisher (2005). Background knowledge in mathematical finance at the level of the course Advanced Financial Models will be required.

## **12. Calculating derivatives of parameters in population balances or "What are detergents sensitive to?" ..... Dr J.R. Norris & Dr M.Kraft**

This is a challenging and open-ended project which addresses an important practical problem. Some understanding of Markov chains and Monte Carlo methods is required. The material on jump processes and differential equation limits from the course Stochastic Calculus and Applications is directly relevant – and should be discussed with Dr Norris at an early stage.

In this project we want to investigate and develop Monte Carlo methods for the calculation of derivatives, with respect to model parameters, for population balance equations. Population balance equations appear in many areas of science and for this project we shall study an application from Chemical Engineering. The final aim of the project is to apply a number of numerical methods to agglomeration problems which occur during the production of detergents: the calculation of derivatives will be a key step in optimizing model parameters. The detergent particles can be described by the amount of solid, liquid, and pore volume as well as the degree of chemical reaction between liquid and solid within the particles. The particle properties influence the quality of the end product and are dependent on operational parameters, for instance the speed of the mixer in which binder and solids mix.

The agglomeration of detergent particles can be described by a generalized coagulation equation for which stochastic particle methods have been used to approximate its solution. For this system a number of heuristic particle methods exist to calculate the derivatives of the solution of the population balance equation with respect to the parameters occurring in the coagulation rates.

In the course of the project the following steps need to be completed. First, we need to find a formal mathematical framework in which to describe these algorithms using the language of probability theory. In a second step we need to implement these algorithms and study the numerical performance for simple test cases. Finally the 'best' algorithm will be used to study the agglomeration of detergents. We shall perform a sensitivity analysis and use experimental data provided by Procter and Gamble to identify the numerical values of the model parameters.

The project will be jointly supervised with Dr Markus Kraft from University of Cambridge, Department of Chemical Engineering

**13. Parameter estimation for random processes .....Dr J.R.Norris**

The project will investigate and implement techniques of parameter estimation for a range of random processes, which will include Markov chains and Brownian motion, usually based on incomplete observation.

The student will first need to become acquainted with literature available on this problem and will need to develop a facility in simulation of some simple random processes.

The main content of the project is to devise, and to implement computationally, algorithms for parameter identification in at least two contexts of practical interest, to be agreed with the supervisor. There follow two possible examples.

A Markov chain model of chemical reactions: a large number of particles of various types are involved in reactions in which, typically, one or two particles of a given type change into one or more particles of another type. Suppose we observe the numbers of particles of each type at a sequence of times. How should we determine the reaction rates?

A Brownian motion model with variable volatility: consider a diffusion process without drift, whose diffusivity evolves over time according to a simple Markov processes, say a geometric Brownian motion or even just a two state Markov chain. We observe the diffusion process at a sequence of times. It is desired both to estimate the Markovian parameters of the diffusivity and to predict the current diffusivity.

**14. Analysing Asset Class Data ..... Dr S.M.Pitts & Dr M.Lakshmanan**

Asset Management is the branch of finance that involves making investment decision for clients who can either be institutional or retail in nature. The main areas of interest for a quantitative procedure in asset management are in performance analysis, risk management and portfolio construction. The focus of this project will involve statistical techniques used in each of these main areas.

Typically, the top-level investment process involves analysing and determining the exposure to the main asset classes, namely equity, fixed income and currency. The data set provided for this project will be market indices representing each of these asset classes.

The first part of the project will involve estimating the covariance matrix. Exponential weighted moving average is a commonly used method which involves giving more weight to the recent past. This is of particular interest to asset classes with short-memory like emerging markets. In addition, this method can be used as a forecasting tool for the covariance between assets. Another method that is used is Bootstrap estimation techniques. Estimation errors need to be analysed and assessed. The covariance thus estimated can be used for implicit factor analysis, risk management and portfolio construction.

The second leg of the project is to perform a time series analysis on the data set and use time series forecasting techniques such as GARCH for volatility forecasts. In addition, econometric analysis needs to be performed to understand the explicit factor exposures.

Finally, the project will investigate the non-linear dependence structure of the different asset classes

and uses techniques such as non-parametric correlation measures and eventually copula functions. These measures will then be used in the context of risk management to evaluate the Value-at-Risk of the portfolio.

This project will be supervised with Dr Meena Lakshmanan of Russell Investment Group, London.

### 15. Numerical methods in ruin theory ..... Dr S.M.Pitts

In the classical risk model in insurance mathematics, claims arrive at an insurance company in a Poisson process, and the surplus at time  $t$  is  $U(t) = u + ct - \sum_i^{N(t)} X_i$ , where  $u$  is the initial surplus,  $N(t)$  is the number of claims arrived by time  $t$ , and  $X_1, X_2, \dots$  are the successive claim sizes, assumed to be independent and identically distributed, and independent of the claim-arrivals process. The *ruin time* is defined to be  $\tau = \inf\{t > 0 : U(t) < 0\}$  (and is infinite if the set is empty). *Ruin theory* is concerned with quantities related to  $\tau$ , such as the distribution of  $\tau$ , its moments, the *probability of ruin* etc. This can also be studied in more general models such as the Sparre Andersen model, where claims arrive in a renewal process.

It turns out that there are not always easy explicit expressions for these ruin quantities, and so they have to be evaluated numerically. The aim of this project is to implement the fast Fourier transform algorithm for this, and to compare the results with other numerical methods and approximations, eg those in Dickson and Waters (2002) and the references there. The first step in the project will be to become proficient in using the fast Fourier transform for applied probability calculations, and then to work out the details of its applications in ruin theory. The comparisons with other numerical methods will involve learning what these are, and implementing them.

This project will need good computing and programming skills. At the end of it, you will have developed good numerical skills for applied probability, which will be useful in other models and situations, beyond the ruin theoretic applications studied in the project itself.

#### Reference

Dickson, D.C.M., and Waters, H.R. (2002) The distribution of the time to ruin in the classical risk model. *ASTIN Bulltin.* **32** 299–313.

### 16. Identifying outliers in routine hospital outcomes . . . Dr S.M.Pitts & Dr D.Ohlssen

The Health Care Commission use routine data to produce star ratings for Hospitals. The aim of this project will be to analyse some of the clinical outcome data used by the Health Care Commission with the aim of identifying odd results.

The main technical difficulty of the project will be robustly estimating a null distribution that can be used as an indication of oddness. This is not easy because we need some way of down weighting potential outliers. A further difficulty is that the data has a two level hierarchical structure with patient outcomes nested within hospitals. In the project we shall look at two approaches:

- (1) A simple winsorisation technique which involves replacing the most extreme outcomes with slightly less extreme outcomes. Clearly this approach is anti-conservative. This approach can easily be applied using R.
- (2) Using the t distribution to provide robust estimates of scale and location parameters for the null distribution. This method will require Markov chain Monte Carlo techniques to estimate the

parameters and will use the WinBUGS software.

**17. Macro-Economic Factors and the Cox-Ingersoll-Ross Yield Curve Model**

..... **Prof. L.C.G.Rogers & Jagjit Chadha**

We want to develop term structure pricing relationships that nest macroeconomic models of aggregate fluctuations within the framework of multi-factor models of the yield curve. We will adopt a no-arbitrage affine term structure model (see Dai and Singleton, 2000) where all factors can be interpreted in terms of the structural relations of a macro economic model. We will develop some inference on the appropriate macro model and the extent to which modelling macroeconomic factors can explain better yield curve dynamics.

This project will be jointly supervised with Jagjit Chadha at the University of St Andrews.

**Reference**

Dai, Q. and Singleton, K.J., *Specification analysis of affine term structure models* J.Finance, **55**, 1943–1978

**18. Range-Based Correlation Estimation for Market Price Data**

..... **Prof. L.C.G.Rogers & N.Brown**

Estimators of volatility are available that make use of the range of a price process over specified sub-intervals (e.g. the daily high and low of an assets trading price). Because they incorporate more information from the price process, they are generally more efficient than the sample standard deviation from daily observations. The objective is to derive an estimator for the correlation between two such processes that makes use of the ranges of both. It is relatively straightforward to derive an estimator that makes use of the range of the cross-product of the series, but in the case of historical financial data this is not obtainable, and so the estimator must use only the ranges for each individual process. The project would also require testing and comparison with other estimators, using data and facilities supplied by the sponsor.

This project will be jointly supervised with Nicholas Brown at BNP Paribas

**19. Optimal bandwidth choice in classification problems**

..... **Dr R.J.Samworth**

Suppose that we have samples of data  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  from two populations  $X$  and  $Y$ , and wish to classify a new observation  $z$  as coming from one or other of the populations. Such classification problems arise in a huge variety of applications (banks deciding whether or not to give a customer a loan, doctors making medical diagnoses, email filters deciding whether a message is real or spam, etc.).

One possible technique is to estimate the densities of the respective populations, and assign  $z$  to the population which has the higher density estimate at that point. When using a kernel density estimate, as discussed in the Statistical Theory course, this requires a choice of bandwidth, which determines how much the data in the sample is smoothed to produce the estimate. A great deal is known about the optimal choice of bandwidth for ordinary density estimation problems, but much less when the primary purpose of the estimation is for classification (though initial indications suggest the answers can be surprisingly different). After a little initial background reading, this

project would involve writing simulations to examine the optimal choice of bandwidth in a variety of classification contexts. It would be particularly interesting to see if the optimal bandwidth appears to be of the same order in the sample size as for ordinary density estimation.

**20. A comparison of different nonparametric regression techniques**

..... **Dr R.J.Samworth**

Regression problems are among the most common statistical problems encountered by scientists. The broad aim is to reconstruct a signal from noisy data (e.g. estimating electricity usage from measurements made at certain times), and this is achieved by some kind of smoothing technique. The great advantage of nonparametric methods is that they make very few assumptions about the shape of the underlying signal (we do not assume that the signal is a straight line or quadratic curve, for instance).

Recently, there have been great advances in the understanding of the mathematical properties of various nonparametric methods, of which some of the most popular include kernel smoothing, splines, Fourier analysis and wavelet methods. This project would compare the different techniques from a theoretical and/or practical perspective, highlighting the suitability of each for particular problems.

**21. Complexity and queues in modern industrial systems**

..... **Prof. Y.M.Suhov & G.D.M.Frizelle**

One of the goals of management of industrial systems (companies, plants, production shops, supply chains) is to achieve smooth running and to cut off unprofitable production lines. The diversity of the range and other specific aspects of modern production even within a relatively small unit make it difficult to create a universal approach to the problem of identifying a bottleneck in an industrial system. One possible way to analyse this problem is to consider ‘queues’ of various nature that often occur in systems, and to compare their entropies. These entropies (unconditional, conditional, mutual) are calculated from empirical distributions (histograms) generated from recorded results of observations. Such records have been performed on a number of industrial systems and accumulated by researchers at the Institute of Manufacturing, Department of Engineering. The project aims a scientific analysis of these data and drawing conclusions and recommendations for the management. The project work will not need visits to actual production sites, although such visits could be arranged. The emphasis of the work will be mainly on statistical modelling.

This project will be supervised jointly with Mr GDM Frizelle from the Department of Engineering.

**Recommended literature:**

G. Frizelle, Y. Suhov. An entropic measurement of queueing behaviour in a class of manufacturing operations. *Proc. Royal Soc. Lond., Ser A*, **457** (2001), 1579–1601.

G. Frizelle, Y. Suhov. Measurement of complex man-made systems. Manuscript available from the authors.

**22. Re-insurance and large deviations** ..... **Prof. Y.M.Suhov & Dr M.Kelbert**

The project aims to study various models of re-insurance in the regime of large deviations. In such models, a re-insurer (a large investor) committs himself to cover losses of an insurer (a small or

a medium size company) in the case of an unusually high level of claims sent to the latter (e.g., as a result of event of a rare, but possible, natural disaster). The rare event regime is studied by means of the theory of large deviations and is characterised by trajectories minimising a specific functional: the large deviation rate functional. This project will need a theoretical work based on large deviation theory, combined with a fair amount of creative numerical simulations.

This project will be jointly supervised with Dr M Kelbert of the Mathematics Department, University of Swansea.

**Recommended literature:**

J. Aquilina, M. Kelbert and Y. Suhov. On optimal strategies to deal with extreme regimes in insurance. *Quality Technology and Quantitative Management*, **1** (2004), 161 – 172.

M. Kelbert, I. Manolopoulou, I. Sazonov and Y. Suhov. Large deviations for a model of excess of loss in re-insurance. Manuscript available from the authors.

**23. Additivity problems for quantum information channels.....Prof. Y.M.Suhov**

Quantum information channels are a hot topic of research in the quantum information theory and quantum computing. A popular open problem here is to prove formulas for the so-called classical capacity of a memoryless channel. Mathematically, this is related to additivity properties of various characteristics of the channel, one of which is the minimum output entropy. While new rigorous results in this area are very difficult to achieve, a comprehensive numerical study of various classes of channels would provide the candidate with an exciting opportunity to explore completely new directions of research in a popular (and competitive) area bordering the information theory and quantum mechanics.

This project would be carried jointly by Prof Y Suhov (Statslab) and Mr GDM Frizelle, of Department of Engineering.

**Recommended literature:**

N. Datta, A.S. Holevo and Y. Suhov. Additivity and multiplicativity in transpose depolarizing channels. *Probl. Inform. Transmission* (2005), to appear. (Manuscript available from the authors.)

**24. Models of Recovery Rates.....Dr M.Tehranchi & Dr P.Cotton**

As the market for credit derivatives has grown in recent years, so too has the demand from the banking industry for good models of credit risk. The management of this risk depends on knowledge of the term structure of interest rates, the likelihood of default events, and the recovery rates of principal given a default. While models of the interest rate and default arrivals have been extensively studied in the mathematical finance literature, less is known about recovery rates. The goal of this project is to calibrate models of recovery rates to actual price data and to compare the stability and predictive power of different models.

This project will be supervised in collaboration with Dr. Peter Cotton of the Credit Derivatives Research Group at Morgan Stanley in New York.

**25. Pricing Options by Monte Carlo Simulations.....Dr M.Tehranchi**

A key result of financial mathematics is that the price of a financial asset can be expressed as the conditional expectation of an appropriate random variable. In rare cases, such as under the assumptions of the Black-Scholes model, the price is given by an explicit formula. In practice, however, the expected value usually is approximated by Monte Carlo simulations. Unfortunately, naive implementation of the Monte Carlo scheme can be computationally expensive. The goal of this project is to implement the recent proposals in the financial mathematics literature to use the methods of Malliavin calculus to increase the efficiency of the Monte Carlo computation of both conditional expectations and the sensitivity of these conditional expectations to model parameters.

**26. Metrics for Portfolio Evaluation .....Dr M.Tehranchi & Dr D.Wischik**

Current practice for portfolio evaluation is to get a large database of price movements for a variety of equities and then to construct a portfolio using subjective criteria based upon relevant knowledge and experience This portfolio is then tested in one of two ways: of two ways:

1. Predefine a fixed collection of time series, e.g. for market average, average of insurance companies, average of UK banks and then regress your portfolio against this average. The regression coefficients tell you how sensitive your portfolio is to each of those factors. [Problem: the coefficients depend very much on which factors you've chosen, and on their correlations.] You can also, calculate the standard deviation of the difference between your portfolio's price movements and that of a benchmark. You want this to be small, indicating that you're accurately tracking the benchmark. You can also attribute the variance of your portfolio's price movements to the various factors, and look for obvious outliers or anomalies.
2. Do a principal components analysis of the matrix of price movements for the entire market. Take say the 20 biggest components. Your portfolio can be expressed as a linear combination of these components. The components are hard to understand in themselves; so instead you use these components and weightings to answer more tangible questions like: if I increased my holdings in UK banks, how would this affect the standard deviation of tracking error? (tracking error = my price movements - benchmark price movements). This tells me how overexposed I am to UK banks.

This project aims to tackle the following questions:

- (i) What is the appropriate measure of risk (with respect to a benchmark)? Tracking error misses the point that we want overperformance. What measure of risk can we come up with to take into account a more sensible desired outcome? (e.g. queueing/dam/ruin models?)
- (ii) Within this measure of risk, how can we properly attribute meaning to different components of risk? Measures like "overweightness" and "sensitivity" go some way to this, but it is not clear what the practical implication of those measures is – in particular, because they only have meaning with respect to a pre-specified coordinate system.

The project's primary concern is with metrics. Current systems spend huge amounts of money on automated systems for "analysing" and "optimizing" portfolios – but there is no good way of understanding the output or determining (from that output) appropriate actions.

**27. Numerical Solution of the Broadcasting Problem when Peers have Differing Up-**

**load Rates ..... Prof. R.W.Weber & J.Mundunger**

This project is motivated by a desire to understand how files can be disseminated efficiently using peer-to-peer systems such as BitTorrent. It concerns the use of a mixed integer linear programming software to investigate the solution of a special makespan scheduling problem, known as the broadcasting problem. In this problem, a file is to be disseminated from a server to  $N$  end users (called peers) in the least possible time. The file is divided into  $M$  equal parts and once any peer has obtained a part he can upload that part to any of the other peers. As an example, suppose the upload rates of the server and peers are all the same: specifically, any file part can be uploaded from the server or a peer to another peer in 1 second. Then for  $M = N = 2$  the minimal dissemination time would be 3 seconds. To achieve this, in the first second the server uploads file part 1 to peer 1. In the next second the server uploads file part 2 to peer 1, and peer 1 uploads part 1 to peer 2. In the third second the server uploads part 2 to peer 2.

When the upload rates of the server and peers differ (say they are  $C_S, C_1, \dots, C_N$ ) the solution becomes very difficult. Now, peer  $i$  can upload a file part in time  $1/C_i$ . A recent PhD graduate, Jochen Mundinger, has described the full solution in the case  $N = 2$ , but only for the case of two differing parameters:  $C_S$  and  $C_1 = C_2$ . However, he shown that, in principle, the problem of minimizing the makespan can be formulated as a mixed integer linear program (MILP). The aim of the project is to implement this MILP within a commercial software package, and discover for how large values of  $M$  and  $N$  it is possible to calculate the optimal schedule, and learn, as far as possible, how the minimum makespan changes with  $M, N$  and the upload rates  $C_S, C_1, \dots, C_N$ . The student will have to experiment with choices for these parameters. By working on this project he or she will gain experience with using a mixed integer linear programming package.

**28. A Study of Heavy-tailed Inter-event Distributions  
..... Prof. R.R.Weber & Dr A.Chaintreau**

The project in the general area of queueing theory and simulation. It is motivated by the analysis of delays experienced in some mobile networking problem.

Specifically, there is empirical evidence that heavy-tailed distributions arise in contact processes between humans. Between a given pair of persons, the inter-contact times that are observed are well approximated by a power law (with infinite expectation) on a large range of values. It is therefore natural to truncate the distribution at some threshold value, corresponding to some long periodicity (e.g., one day or one week).

Motivated by the above, the aim of the project is to study how truncating the tail of such a heavy-tailed inter-event distribution affects the “sampled” inter-event distribution (i.e., the one we look at in the “waiting time paradox”), in continuous time. It will be interesting to discover which growth rate some functionals (e.g., expectation, median) may diverge for some distributions when the point of truncation is large.

The project will primarily involve computer simulation work and statistical analysis of the results and will be jointly supervised with Augustin Chaintreau from Intel Corp.

**MPhil in Statistical Science - Projects 2005-06**

Please complete the form below, giving up to four ranked project preferences, or alternatively giving your preferred area (such as “finance” or “medical”). Please hand in the completed form to Julia Blackwell in D1:17 by 4.00pm on Thursday 6th October 2005.

**EITHER;** list the titles and supervisors of your ranked project preferences (top preference first, maximum of four).

1. ....

.....

2. ....

.....

3. ....

.....

4. ....

.....

**OR;** give an indication of your preferred subject area for your project.

.....

.....

Name: .....

Email: .....