MPhil in Statistical Science 2007--8
Applied Projects (as summarised by their authors).

Cecilia Durieu	Comparison of investment strategies
Mary Gregory	Performance of multi-path routing in multi-hop wireless networks
Jonathan Joiner	Modelling volatility
Suhrid Joshi	Designing variable annuities with guaranteed minimum benefits
James Keough	Bayesian shrinkage regression in the estimation of covariances whose return histories have unequal length
Ying Ying Lai	Non-proportional hazards -- diagnosis and treatment
Sam Lees	Critical random transpositions
Andrew Lim	Dynamic stochastic programming models for pension funds
Fujia Liu	Nonparametric regression with applications to climatology data
Peter Rasmusen	Solving nonlinear PDE's by Monte Carlo methods
Constantinos Rossides	Parameter estimation for random processes
Sarju Shah	Option pricing with SciFinance
Marc Wilson	Bayesian methods for multi-object tracking

Cecilia Durieu **Comparison of investment strategies**

Investors hold positions in different financial assets and they periodically rebalance their portfolios to achieve a good performance. One approach to obtaining such a result is to find a model for the assets dynamics, to estimate the parameters of the model and then to use an optimal investment rule. In the finance industry, some approaches based on experience, such as graphical analysis, are also widely used.

First, this project looks at finding a model that fits well the assets returns using only historical data. According to the tests I run, ARMA(5,2) and ARMA(5,2)-GARCH(1,1) models fit particularly well the asset returns considered. Then, I studied the mean-variance problem and also shined a light on the role played by diversification in the reduction of risk. Finally, I simulated several investment strategies on real stocks from
FTSE100 and I compared the returns that would have been obtained with these portfolios. According to this example, it seems that allowing short-selling in the portfolio enables to obtain better performances but looks more risky. Moreover, the Merton portfolio seems to outperform the other strategies considered.

Mary Gregory **Performance of multi-path routing in multi-hop wireless networks**

The aim of this project is to find out more about the best way to transfer data in a multi-path multi-hop wireless network. It is known that in a wired network routing packets over multiple paths increases the robustness and performance of a network. However, in a wireless network data packets from each path may collide and the performance of the multi-path network may be worse than that of the single-path network.

Throughout the project an ALOHA model for the transfer of data through the network is assumed. In order to investigate when it is better to use a multi-path network we compare the maximum flow through two-path networks to the maximum flow through the equivalent single-path networks. Initially, the maximum flow through each network is found analytically. As the networks become more complex, it is simpler to find the maximum flows numerically and MATLAB is used to do this.

First, networks in which all nodes can communicate with all other nodes are considered. At this stage we look at specific networks with n nodes and compare the flow through the two-path network with the flow through the single-path equivalent. Then some of the interference between the two nodes is eliminated. In all cases, if we know the single-path network performs better than a two-path network, then we know that the single-path network performs better than any multi-path network with the same structure and patterns of interference.

Following this we go on to consider a multi-hop network in which each node can only communicate with the nodes nearest to it, a 1-hop network. Multi-hop networks are more realistic in this scenario as it is the only way data can get from the source to the destination and allows networks to cover longer distances. To begin with we continue to look at specific networks with n nodes and varying patterns of interference between the two paths. Following this we move on to networks with a large number of nodes. For this type of network an approximation for the flow through the single-path network is compared to a lower bound for the flow through the two-path

network. From this we can see when the two-path network has a greater maximum flow than the single-path network.

As an extension to the analysis there are a number of areas of possible further investigation. Some of these were discussed in the conclusion.

Jonathan Joiner **Modelling volatility**

This project was supervised by Dr. Tehranchi

Accurately modelling the stock market has been the goal of practitioners and academics for decades. Since Merton's 1973 paper, which introduced the Black-Scholes formula, modelling the volatility of the stock has been a focus of this work. In this project we look at different stock market models and how they treat volatility. We also compare how these models price a variety of exotic options in relation to each other.

In this project we focus on three different stock market models, namely Black-Scholes, Heston and S.A.B.R.. We then see how these models treat volatility and if that matches general observations, made by practitioners, on the stock market and what we have observed ourselves looking at the M.S.F.T. (Microsoft) stock. We also compare how each model that prices some set of the Vanilla options equivalently, prices exotic options and see if we can observe any patterns.

When looking at the M.S.F.T stock we noticed that the Black-Scholes implied volatility was certainly not constant, as specified by the model, but had an arc. This is known as the volatility smile. The Heston and S.A.B.R. models treat volatility as a stochastic process and when using appropriate parameters both models could display a similar volatility smile. The Heston model was the most successful in mimicking what was observed in the real life data. When comparing how the models price exotic options in relation to each other we observed definite patterns.

Suhrid Joshi **Designing variable annuities with guaranteed minimum benefits**

The last 20 years has seen an explosion in variable annuity products in the US and Japan. These have replaced conventional annuity products by combining the accumulation phase of traditional investment products with the decumulation phase of annuities. Customers are attracted to the possibility of this single cradle-to-grave investment product that acts as a source of income during their retirement years as well as protecting them against the uncertainties of life such as death, disability and financial risk. This, combined with the tax relief offered on them, means that variable annuities are an extremely marketable and profitable investment product.

Initial research suggests that variable annuities (VAs) have the potential to show similar expansion in the UK. If managed properly, they represent a significant new development for providers and customers alike. The research in this thesis is motivated by the analysis of the viability of VAs in the UK retirement product market. In order to design VAs, an efficient asset/liability management (ALM) model is developed and studied in detail. Asset return statistical models are calibrated to UK financial data and are then used to generate a large number of future market scenarios. These scenarios are used in a dynamic stochastic programming (DSP) model to develop a portfolio

strategy for the effective fund management of VA products. The focus of the study is on the ALM of two variable annuity products - a vanilla GMAB product and an exotic GMIB product, the first products launched in the US and Japan - and their suitability to the UK context.

James Keough **Bayesian shrinkage regression in the estimation of covariances whose return histories have unequal length**

Bayesian Shrinkage Regression in the estimation of the covariances between many assets whose return histories have unequal length

This project looks at the Bayesian posterior inference of covariances when there is a differing amount of historical data available. In particular, Gibbs Sampling (the theory of which is developed in the MPhil/Part III lecture course Monte Carlo Inference) is used to sample the posterior distribution for the parameters in a model similar to the standard linear model and then transformed to sample from distributions for individual covariances.

To start with, I consider the importance of parameter estimation for investment managers when analysing the risk of a portfolio of assets under the assumption that returns are multivariate normally distributed. The data available to the financial manager in the form of a time series may have a so-called monotone missingness pattern, so that the known asset returns stretch back to different times for the assets. The next section discusses methods (such as ridge regression and the lasso) which provide a penalty term for high values of the estimated regression coefficients in a linear model. These regressions are especially useful in producing low variance parameter estimates and particularly in cases when the number of parameters estimated are large (in this case, a large number of assets with covariances to be estimated). I then explain how these methods fit into a Bayesian framework and how they are maximisers of the posterior distributions given a particular prior setup. In analysing simulated data, I discover that these Bayesian methods (which invlove Gibbs Sampling from posterior distributions) have many of the desired properties required to estimate coefficients (comparable to other methods) and that they have advantages as well (particularly in giving credibility intervals for regression coefficients).

The final sections consider the actual data to which such methods might be applied in practice and the practical implications of their implementation (slow running time) as well as suggestions for possible improvements.

The computational side of the project relies extensively on R and some of the packages available for solving the maximum likelihood inference problem to estimate covariances (for example, the monomvn package written by Bobby Gramacy).

This project was supervised by Bobby Gramacy.

Ying Ying Lai **Non-proportional hazards -- diagnosis and treatment**

Most analyses of medical time-to-event data investigate the effect of potential predictor variables such as age, gender, treatment, extent of disease at baseline. People always have strong assumptions of the proportionality without much justification, are made about the form of the hazard function (the hazard function is the instantaneous risk of the event as a function of time since a starting time such as diagnosis). As an example: the hazard function for males is typically taken

to be a constant multiple of the hazard function for females. Assuming that hazard functions are proportional when they are not may lead to misleading conclusions being drawn from the statistical analysis.

In this project, two data sets were provided by Eastern Cancer Registration and Information Centre (ECRIC), and they are glioblastoma data (brain cancer data) and breast cancer data. From preliminary analyses for the data sets first, we reduced the sample size by deleting missing data and any trivial case for each prognostic variable. By using AIC method, we carried on the model selection step, which would give us several "interesting" variables, i.e those were more effective in the hazards model. Kaplan – Meier curves were applied widely next for presenting the overall survival status or some specific variable's of each cancer, also gave us rough visual sense about the proportionality held by each variable. In more theoretical way, we applied global test statistic to investigate the proportionalities for prognostic variables due to their p-values. Then we could generate Schoenfeld residual plots to see the behavior of proportionality or non-proportionality for each variable. After finding variables with non-proportionality, we introduced the Anderson - Gill formulation of the proportional hazards model as a counting process to investigate the non-proportionality more precisely.

All analyses were carried out using S-Plus and R.

Sam Lees **Critical random transpositions**

This project was supervised by Dr. N. Berestycki, and investigates the behaviour of the cycle structure of permutations generated by randomly selected transpositions.

It was conjectured that the cycle structure of a permutation undergoes a rapid phase change after a critical number of transpositions. If we consider n-permutations generated by cn transpositions, then this phase change occurs when $c \approx ½$.

The project is based on well known results from graph theory, a recent paper by N. Berestycki and R. Durrett, as well as a result by Oded Schramm in 2004, which shows that for $c \approx ½$, the scaled cycle sizes converge in distribution to a Poisson-Dirichlet random variable.

A result by Diaconis and Shahshahani is also demonstrated, which shows that after (n/2) log(n) transpositions, the system reaches equilibrium (so a permutation generated by (n/2) log(n) random transpositions is equivalent to one chosen at random from all n! possible).

This project verifies these results numerically as well as investigating the behaviour of the cycles when $c \approx ½$.

It is conjectured that when $c \approx ½$, the largest cycles have size of order $n^{⅔}$. This conjecture is supported by numerical simulations in the project and a result for the size of the critical region is developed.

R was used to generate the simulations used in the numerical aspects of the project.

Andrew Lim **Dynamic stochastic programming models for pension funds**

Dynamic stochastic programming is a technique for decision making under uncertainty. Pension funds exist to guarantee benefit payments to retired employees by investing in the financial markets, and they experience uncertainty from many sources including unknown future investment returns and actuarial factors. Like other institutional investors, a pension fund faces the problem of *asset liability management* (the process of managing and modelling its future financial position); this problem is *dynamic* as pension funds have to make many investment decisions over long periods of time, and *stochastic* since a particular sequence of decisions could produce a variety of different outcomes. Dynamic stochastic programming is therefore an appropriate framework for optimal investment -- that is, how to maximize returns while maintaining acceptably low levels of risk. The application of dynamic stochastic programming to financial planning is known as strategic dynamic financial analysis.

In this project we construct a strategic dynamic financial analysis system for pension fund investment. The data used are yearly simulations of asset classes which were generated by Watson Wyatt's Global Asset Model (GModel). We compare two methods of generating scenario trees from these simulations: modelling asset return behaviour as geometric Brownian motions and Ornstein-Uhlenbeck processes, and a ``rank-and-merge'' algorithm which takes averages of asset returns across several GModel simulations. We describe a simple investment model and implement it in optimization software. We then introduce transaction charges and investment constraints into the model and investigate the effects of using different utility functions. Finally, we look at how dynamic stochastic programming can be used in conjunction with real world data by performing a series of historical backtests, and analyse how effective this method is for generating investment portfolios for pension funds.

The optimization software used in this project was gspl by Cambridge Systems Associates (CSA). CSA's multigbm was used to simulate geometric Brownian motions and Ornstein-Uhlenbeck processes. The ``rank-and-merge'' algorithm and the backtesting scripts were implemented in Microsoft Visual Basic for Applications.

This project was supervised by Prof.~Michael Dempster and Yakoub Yakoubov.

Fujia Liu **Nonparametric regression with applications to climatology data**

This project was supervised by Dr Pat Altham, Dr Richard Samworth, and Dr Susan Pitts.

Inspired by the chapter on ‘Nonparametric Regression’ in Julian Faraway’s new book ‘Extending the Linear Model with R’, this project aims to find out the strengths and weaknesses of different nonparametric regression methods when applied to some real datasets. The main interest of this project is in solving regression problems. A simple and common approach is to estimate the function of interest from a parametric family. However, this can be inflexible and very much depends on the validity of the assumed function. Here, we explore a nonparametric approach, which involves choosing the function of interest to be smooth and continuous with no constraints on parameters. This approach provides more flexibility in a variety of situations, and this is what gives the fundamental motivation for the project.

To investigate the use of the nonparametric regression techniques in a more complete way, we choose two datasets from two different contexts, namely the climatology dataset ‘aatemp’ and the economic dataset ‘wages1833’. Therefore, this project is divided into two parts, and we carry out the analyses of the nonparametric regression approach with applications to the each dataset in turn.

Part One is on ‘Nonparametric Regression with Applications to Climatology Data’. After an introduction and description of the dataset ‘aatemp’, we start with the investigation of a failed simple linear model applied to this dataset. In the following chapters from 3 till 5, we will give the main theory of the nonparametric regression methodology before practising them with ‘aatemp’. The main methods used in this project are Kernel Estimators, Splines and Lowess, although there are other nonparametric regression schemes. For each technique, we discuss the effects of various parameters on determining the smoothness and goodness of fit for ‘aatemp’. We also find the disadvantage of using some popular criterion such as the cross-validation criterion of choosing a smoothing parameter in reality. On the whole, we apply the following estimators to the dataset: the Nadaraya-Watson kernel smoother, the sm.regression from the CV smoothing parameter, the smoothing splines, the linear regression splines, the cubic regression splines, and the lowess. In chapter 6, we draw conclusions about all the methods tried, and also make some careful discussions in response to missing data in a certain time period and the randomness of the original dataset.

Part Two is on ‘Nonparametric Regression with Applications to Economic Data’. To make this project consistent and easier to read, from chapter 7 to chapter 11, we follow exactly the same sequence of nonparametric regression study as in Part One with applications to the new dataset ‘wages1833’. We find the parametric linear regression is again unsatisfactory. The subsequent curves using kernel, splines and lowess methods fit much better. Unlike the previous dataset ‘aatemp’, this new dataset has a variable ‘weights’ which has to be taken into account in all the analyses. Therefore, the extra task in this case is to adapt all the commands to consider this variable, and hence generate the corresponding estimates accurately. Finally, chapter 12 draws conclusions about the applications to ‘wages1833’. We notice some similar problems as discovered in the previous dataset. We even try interpreting the statistical results from an economic view, making this project more relevant to the real life.

We use R to implement all the programmings. Since R command is a necessary part of research of this project, we will include it throughout this dissertation, and will only give edited R output.

Peter Rasmusen **Solving nonlinear PDE's by Monte Carlo methods**

The title of the project is "Solving Non-Linear Partial Differential Equations by Monte Carlo Methods".

The project is mainly based on the article "Second Order Backward Stochastic Differential Equations and Fully Non-Linear Parabolic PDEs" by P. Cheridito, H. Soner, N. Touzi and N. Victoir; a piece of work that was published in 2007.

Briefly the project is on numerically solving partial differential equations (PDEs) by solving a system of stochastic differential equations (SDEs) by forward and backward iterations. More precisely starting with a certain class of PDEs one can associate an ordinary SDE (an Itô }-process) and a so-called 2BSDE, which is a certain system of SDEs. It turns out that nice relations between solutions to the 2BSDE and the associated PDE can be shown and this provides a stochastic representation for solutions of fully non-linear parabolic PDEs. In the article an algorithm that solves the 2BSDE and hence the PDE is suggested but whether the method works still only stands as a conjecture. Shortly the idea is to approximate the strong solution of the SDE by the so-called Euler method by forward iteration. From here we can work backwards through the suggested algorithm thereby solving the 2BSDE and hence the PDE. The algorithm does involve estimation of conditional expectations, i.e one has to estimate stochastic variables. This is done by using that

conditional expectation is a projection onto the Hilbertspace $L^2(P)$. Since $L^2(P)$ is a Hilbertspace it has a countable orthonormal basis and the idea is to choose a finite sub-basis of an appropriate basis for $L^2(P)$ and approximate the conditional expectation by linear regression based on this sub-basis.

The main goal of the project was to implement this algorithm and empirically justifying that it works. For programming languages mainly R but also Matlab was used -- with access to a 64 bit computer memory capacity was not problem.

Constantinos Rossides **Parameter estimation for random processes**

The theory of random processes is a well-established area of mathematical statistics with a widely developed theoretical side. In this project we make an endevour to cover the basic concepts of some of the key topics of random processes such as Markov chains, Random walks, Renewal theory, Brownian motion and Diffusions.

Markov chains is perhaps the most interesting and fundamental class of random processes. In the first part of this project we develop a solid grounding on Markov chains by devising a facility in simulation for discrete and continuous-time Markov chains. We investigate properties of the limiting distribution and we show how renewal theory and central limit theorems can be applied to test the efficiency of our algorithms.

After getting some `hands-on experience' by implementing the Markov chain simulation algorithms, we restrict our discussion to applications concerning Brownian motion processes with constant and variable volatility. Stochastic volatility has always played a central role in asset pricing and financial markets in general. This is because it is allowing traders, investors and money-dealers to make more informed judgments on their trades, and analysts to set the prices of options and other derivative securities.

The importance of volatility drives the second part of this project, where simple code is used to unravel the dynamics that hide behind stochastic volatility estimation. By using an iterated cumulative sum of squares algorithm to identify variance changes in the sample path, we achieve to get accurate estimation results for a two-state piecewise constant volatility process. Beyond that, we make the problem harder and more realistic by modeling volatility by a random walk and geometric Brownian motion.

Undoubtedly, volatility estimation is a hot topic that prevails at every trading desk responsible for pricing derivatives. Consequently, it is vital for traders to be able to give precise and robust estimates in order for proper operations to govern the markets. By means of a set of experiments, we investigate various methods to optimise the estimation, and in what way these methods serve as a better solution to our problem.

This project was supervised by Prof. James Norris. All analysis was carried out using R.

Sarju Shah **Option pricing with SciFinance**

The purpose of this project is to harness the computational power provided by SciFinance's high level ASPEN programming environment to investigate some issues in the pricing of financial options. I will look firstly at pricing complex options under the simple Black-Scholes model, specifically vanilla American options and barrier options. Then I will investigate the pricing of vanilla options under more complicated stochastic volatility models, namely the Heston model and the SABR model.

In each case, I will first extend the theory developed in the Cambridge University MPhil Statistical Science course *Advanced Financial Models* to cover these more exotic pricing problems. Then I will illustrate these points with practical pricing run through SciFinance generated C code. I will look at both PDE and Monte Carlo pricers and aim to compare the performance of the two approaches in a variety of instructive contexts. Speed and accuracy are of critical importance in the financial industry and these will be the key measures of successful pricing.

Marc Wilson **Bayesian methods for multi-object tracking**

In engineering there is a class of problems called Multi-object tracking problems. A general method for solving this type of problem, called the Markov Chain Monte Carlo Data Association Algorithm (MCMCDA Algorithm), was introduced in 2004 and in my project I focus on implementing this algorithm using the engineering software Matlab. I also simulate data in order to present examples of how the MCMCDA Algorithm performs under different conditions.

I will present an example to illustrate the type of Multi-object tracking problem considered in my project, to illustrate how the problem can be modelled and how it can be solved using MCMCDA Algorithm. The details can be found in the report.

Using CCTV surveillance images from a particular camera which overlooks a large space, for example the concourse of a train station, we may wish to track the movement of people as they move through this space. Let's develop our story by supposing an explosion has sparked mass panic as well as damaging the CCTV camera. The CCTV footage contains distortion (periods with no picture) and therefore rather than capturing the continuous time movements of people within the train station the CCTV camera measures peoples position at discrete times. The footage that is not distorted is blurred and therefore there is an error attached to each positional measurement. Suppose further that the footage is so bad that people are indistinguishable and sometimes we even confuse inanimate objects with people.

How can we model such a complicated situation?

Let's introduce a stationary state space model which governs the movements of people from one time interval to the next where the time intervals are regularly spaced (every second say) and which also describes how the measurements are recorded. Under the assumption that there is mass panic it is reasonable to believe that the movements of people at the next time period will only depend on their current position and therefore this stationary model is reasonable.

Now suppose the number of people arriving into the train station at each second comes from a Poisson distribution with a known and fixed parameter and introduce a probability that a person leaves the train station at the end of each second which is also known.

Let's also introduce a probability that a person in the train station is detected and measured by the camera. That is, let's introduce a probability corresponding to the footage not being distorted at a particular time.

The measurements corresponding to the inanimate objects will be referred to as false alarms (which can be thought of as false measurements). The number of false alarms arriving at each second also has a Poisson distribution with known and fixed parameter. This creates an additional complication: the inclusion of false alarms means we do not know the actual number of people in the train station.

Can we use this model to gain insight into the movements of each individual?

We have a data set which contains measurements and the time the measurements were recorded but we do not know to whom the measurements belong. We can use the MCMCDA Algorithm to solve this data association problem.

The MCMCDA Algorithm is an iterative algorithm which resembles the Metroplis Hastings Algorithm. It constructs an Ergodic Markov Chain whose state space is the set of all possible partitions of the data into groups of measurements with each group describing the movements of an individual. A group of false alarms is also created.

Given the current partition of the data, the MCMCDA Algorithm proposes a new partition of the data by performing one of eight so called "moves" using the model described as well as a number of constraints to ensure the partition represents feasible movements. If the move is accepted the proposed partition becomes the current partition at the next iteration. This process is repeated until the Markov Chain converges to its stationary distribution.

This partition is the solution to the Multi-object tracking problem. It will provide us with our best estimate for the number of people in the train station and our best estimate for the movements of the people during the surveillance period.

Note that the solution is highly dependent on the probabilities and arrival rates which I described as being known. More likely they will be parameters in the model which we will have to estimate.

I show in the report that the MCMCDA Algorithm performs well when the number of false alarms is small and the probability of detection is high. The algorithm will be less accurate as the detection probability decreases and false alarm rate increases.