MPhil in Statistical Science 2009--10
Applied Projects (as summarised by their authors)

Name	Title
Alexis Bourel	Sequential Monte Carlo (SMC) methods (aka particle filters) for estimating static parameters
Mathieu Cambou	Algorithmic trading for high-frequency data
Yi King Chan (James)	Implied volatility asymptotics
Wing Lin Rita Cheung	From genes to function: statistical models for discovering molecular mechanisms underlying gene-disease associations
Yining Chen	A comparison of different nonparametric classification techniques
Johnny Kai Ming Chow	Statistical aspects of the epidemiology and treatment of cancer of the large bowel
Shuangzi Guo	Large deviation approach to extreme events in high-frequency trading
Sarah Hegerty	Predicting the probability of developing a successful therapy for metastatic pancreatic cancer
Helen Jordan	Sampling formulae for marine populations
Jean-Jacques Schraemli	Local and stochastic volatility models
Moxi Sun	Estimating the number of unseen species: dinosaurs, coins and Shakespeare's vocabulary
Soficlis Zambirinis	Models for predicting health outcomes
Chenye Zhang	Risk analysis of a multi-asset private wealth portfolio

Alexis Bourel
Sequential Monte Carlo (SMC) Methods (aka Particle Filters) for Estimating Static Parameters

This project aims to perform Bayesian inference on general nonlinear state-space models. To this end it introduces Particle Markov Chain Monte Carlo, a recent algorithm combining Sequential Monte Carlo (aka particle filters) and Markov Chain Monte Carlo. On the one hand, Sequential Monte Carlo can estimate quantities such as the likelihood for general, nonlinear models where more basic ideas in Monte Carlo methods such as Importance Sampling would fail on their own. The ability to compute the likelihood is needed in Markov Chain Monte Carlo to perform Bayesian inference, however it is normally intractable in nonlinear state-space models. Thus Sequential Monte Carlo can solve the limitations of Markov Chain Monte Carlo when combined therewith resulting in a Monte Carlo-within-Monte Carlo algorithm, offering a promising way to perform Bayesian parameter estimation on general state-space models.

The resulting Particle Markov Chain Monte Carlo algorithm is indeed a user-friendly way to perform Bayesian inference. Sequential Monte Carlo is combined within Markov Chain Monte Carlo, however it performs well with a minimal amount of work in choosing its proposal which can be taken to be the state-transition prior. Thus only little knowledge is required outside a few considerations such as the number of particles or the design of the Markov Chain Monte Carlo proposal. Particle Markov Chain Monte Carlo results are more reliable than those of the standard Metropolis-within-Gibbs sampler, one of the few alternatives for carrying out Bayesian inference on general state-space models.

However, some difficulties inherent to Markov Chain Monte Carlo methods do persist. While Sequential Monte Carlo does allow estimation of the likelihood at each iteration of Markov Chain Monte Carlo, any delay in doing so may result in excessive computational time overall. To address this, a few tricks are given in an implementation guide and in the appendix such as the use of parallel computing and adaptive proposals for Markov Chain Monte Carlo. Otherwise, a formal way to quantify uncertainty in the parameter estimation and a Bayesian model checking are explained in an attempt to show that these Monte Carlo methods can be carried out with a minimum of scientific rigor contrary to the claims of their opponents. Some remarks on good coding practice in MATLAB can help to reduce the computational time dramatically.

As an application, the Stochastic Volatility model is considered in order to fit some financial data. After a Bayesian model check validating this approach, the model parameters are estimated via Particle Markov Chain Monte Carlo assuming some vague prior distributions. For a certain range of data which is specified, the methodology developed in this project does prove advantageous for its use in Mean-Value optimization. It is noted that the same methodology would apply to higher order models which would likely prove more beneficial in practice.

The implementation is done in MATLAB with the Parallel Computing toolbox. Whenever possible, the use of the Statistics and Signal Processing toolboxes is avoided when functions of higher efficiency can be implemented. This project was supervised by Dr Singh.

Mathieu Cambou
Algorithmic Trading with High-Frequency Data

The ambition of this work is to present a comprehensive introduction to the use of trading algorithms with high-frequency data by giving examples and intuitions. There is no major mathematical tool employed for this purpose, we particularly focus on some market microstructure avoiding distributional assumptions. Note that R is the statistical software that we use in this project to treat the data, make plots, compute statistics, implement and execute the trading algorithms.

After presenting the data and the market under interest, we proceed to the cleaning and the preparation of our data set. This will allow us to make some observations on the distribution of quantities such as the bid-ask spreads, returns for different frequencies, ... We give a formal description of how trading algorithms can be constructed. Any trading strategy is composed of a set of tools that provides trading recommendations and assess about the model performance:
- Exposure calculator which gives recommendations on the direction of the trade and on the amount of the exposure;
- Return calculator computed for every deal according to an average price which is the price paid for achieving the current exposure;
- Stop-Loss detector which is triggered if the market moves substantially in an unexpected direction in order to prevent excessive loss of capital;
- Model performance measures such as

    - the total return;
    - the Profit/Loss ratio;

- the number of Stop-Loss hits.

Two main momentum based trading models are then developed. It is suggested from the results than the first algorithm is too much focused on following major trends and it fails to detect turning points that can be highly profitable in this high-frequency framework. The second algorithm is an evolution of the first one so that we now obtain a signal when a potential turning point is detected. In a last effort, we also adapt our reaction to Stop-Loss hits by never taking the same position twice in a row in such cases. That is, we stay neutral after a Stop-Loss hit as long as there is no new signal in the opposite direction. Although we have only been able to run the algorithms over six distinct days, we can say that the last version of the trading model shows characteristics of a successful trading model. It does not only yield positive results but it also seems to be well fitted to the market under study and it acts on it following the intuition we tried to translate into a programming language. It is important to state that heterogeneity of market has been an underlying motivation: there is no trading strategy that is absolutely better than other ones, which strategy to choose will depend on the trading and risk profile of the investor.

Chan Yi King, James
Implied Volatility Asymptotics

The financial market has undergone huge development over the decades. A diverse spectrum of innovative financial products has helped different investors to better manage their portfolios and monitor risks. One of the most widely used financial derivatives is vanilla call(put). They give the holder the right but not the obligation to buy(sell) an underlying asset at specific prices called strikes at specific maturity dates. Option pricing that arises from different models has induced an active area of research over the years.

Option prices are quoted in terms of the volatilities of the underlying assets by practitioners. The higher the volatility, the more expensive the option becomes since there is a higher chance for the option to be exercised to yield profit. To assess the volatility of a certain asset, one possible measure is to compute the historical volatility, however, a major drawback is that it measures how volatile the price was in the past but provide little information about its future. On the other hand, we can derive implied volatilities from traded market option prices, which in turn are determined by supply and demand. This makes implied volatility a forward-looking measure and better reflect how the market perceives the risk of investing in that particular asset in the future.

In this project, I would like to investigate some asymptotic properties of the option price and implied volatility under the Heston Model at extreme strikes or maturities. The bottom line is that we expect to see the implied volatility surface stabilizing and exhibiting a linear behaviour around those extreme parameter realms.

Yining Chen

A comparison between different nonparametric classification techniques

In this project, we consider the classification problem with two populations. Given samples of data $X_1, X_2, ..., X_m$ from the X population, and $Y_1, Y_2, ..., Y_n$ from the Y population, we wish to classify a new observation z as coming from one or other of the populations.

We broadly consider three types of classifiers, namely, the k-nearest-neighbour classifier (K-NN), the kernel density classifier (KDC) and the log-concave density classifier (LCDC). Their properties are demonstrated using simulated data, where we follow the Poisson interpretation of sample sizes (Hall *et al*., 2008). The risks of different classifiers are estimated in the simulation study via Monte Carlo method.

To choose the number of neighbours for the k-nearest-neighbour classifier, we use the bootstrap method proposed in Hall *et al*. (2008) that selects the empirically optimal k. We rerun their simulation study completely to improve the accuracy level of their results. We also note the possible bias of their method, and propose a simple modification to circumvent this problem.

To choose the bandwidth matrices for the kernel density classifier, we consider a two-stage plug-in rule and the bandwidth that minimizes the Mean Integrated Squared Error.

The log-concave density classifier is build based on the log-concave density maximum likelihood estimator (LCDMLE) studied in Cule *et al*. (2010). This classifier is fully automatic and requires no tuning parameters. Our simulation study suggests that this classifier doesn't perform well in small sample sizes. This is due to the fact that LCDMLE is only supported on the convex hull of the existing observations. We show that the probability of a new observation lying outside the convex hull is of $O(n^{-1})$ in one dimension, and is generally greater than $O(n^{-1})$ in higher dimensions. Moreover, we prove that this probability is the same for all densities in a location scale family. To improve the performance of LCDC, we consider two possible modifications: one uses 1-NN in case of a tie, the other smoothes LCDMLE via convolution.

Finally, we illustrate the performance of different classifiers on the Wisconsin breast cancer data set, with a focus on the decision boundaries generated by different classifiers.

All analyses were carried out using R, with packages: class, ks, logcondens and LogConcDEAD.

References:

[1] **CULE, M. L., SAMWORTH, R. J. AND STEWART, M. I.** (2010), Maximum likelihood estimation of a multidimensional log-concave density. J. Roy. Statist. Soc., Ser. B. (with discussion), to appear.

[2] **HALL, P., PARK, B. U. AND SAMWORTH, R. J.** (2008). Choice of neighbour order in nearest-neighbour classification. Ann. Statist. 36, 2135-2152.

Wing Lin Rita Cheung

From genes to function: statistical models for discovering molecular mechanisms underlying gene-disease association

In recent years, genome-wide association studies have successfully identified novel genes associated with certain complex diseases. However, not many functional explanations were given. The main aim of this project is to find out how many insights statistical analysis on genetics data can give to answer questions concerning the underlying molecular mechanisms of the protein products of the genes considered.

The idea of developing a model for extracting evidence of mechanistic interaction is motivated by the controversy of interpreting statistical interaction as biological interaction and also the inconsistency in the "usual" tests for presence of statistical interactions concerning an additive model and a multiplicative model under particular applications.

In this project, a decision-theoretic approach and the causal graphical models are proposed to induce a set of conditions for mechanistic factors between two factors and to justify all biological assumptions required.

The decision-theoretic approach incorporates two binary genetic variables and the framework was built under an interventional regime. Together with a rare disease assumption, conditions for co-action under the deterministic behaviour of the model are given in terms of parameters of a linear odds model, incorporating two binary factors.

However, quantities defined under the interventional regime may not be appropriate to be carried directly to the observational regime under certain situations. Causal graphical models are used to represent different scenarios in which by applying graphical rules, conditional independences among variables can be read off. Based on these conditional independences, we conclude conditions and biological assumptions that are required.

A retrospective study on Crohn's disease is included in which we illustrate the application of the methodologies proposed in this project. Genetics data of 1270 unrelated individuals with Crohn's disease (cases) and 1320 unrelated controls were analyzed, where the main focus was put on extracting evidence for gene-gene interactions that may be interpreted as joint involvements in a biological mechanism.

The statistical analysis was performed using R version 2.10.0.

This project was supervised by Prof. Berzuini and Dr. Susan Pitts.

Johnny Kai Ming Chow

Statistical aspects of the epidemiology and treatment of cancer of the large bowel

Cure modelling has been a major focus for the medical community in the past few decades as the curability of cancer is becoming more of a reality. Cure models are useful not only to predict the probability of cure for individual patients, but also allow us to draw conclusions about the specific factors affecting survival and cure separately. Several approaches to modelling cure have been proposed, with cure mixture models being the more popular choice due to its simplicity and usefulness. Cure mixture models assume the existence of two groups with different survival experiences. These mixture models allow cured and uncured patients to be modelled separately, while standard survival analysis techniques such as Cox fail to achieve this. In this paper, we shall focus on a logistic/Cox mixture model that is semi-parametric in nature. We will develop the theory and background for the model and illustrates its appropriateness by means of a simulation study. Then, we will apply the model on a colorectal cancer data set and subsequently evaluate the effectiveness and limitations of such cure mixture models. All of the programming is done with R.

Shuangzi Guo

Large deviation approach to extreme events in high-frequency trading

The project on **"Large deviation approach to extreme events in high-frequency trading"** supervised by Prof. Yurii Suhov concerns the application of large deviation theory on FTSE, gold and bunds high-frequency option prices. For each asset class around 2 million bid prices, accumulated from the limit order book of 6 to 8 random trading days, were analysed.

The large deviation theory has been a big research topic in the telecommunication business. Research papers from this area have been a good guideline for our analysis. For a queue defined by

$$Q_n = (Q_{n-1} + X_n)^+,$$

where (X_n , n ϵ Z) is a stationary ergodic sequence of random variables with $E(X_0) < 0$ the large deviation theory says that if the decay rate δ satisfies certain conditions then

$$\lim_{q\to\infty} \frac{1}{q} \log P(Q_0 \geq q) = -\delta\,.$$

We consider the limit order book as a queue formed by investors whose position in the queue is determined by their submitted bid and ask prices.

The large deviation theory is then tested on the deviances of the prices within specific quantiles from their mean. Before it can be applied, the option prices have to be calibrated first. We test several methods including subtracting the mean of intervals of different lengths from each tick to dividing this difference by the standard deviation of the price within that interval. Then we take an arbitrary cut-off point $c > 0$ such that we define the jump-up and down section as the intervals from c to the local maximum and the local maximum to c respectively. Similarly, the sections for the negative are defined by $-c$. These sections are divided into several quantiles and the mean and largest deviances can be determined for each quantile. The deviances are then separately collected for all i^{th} quantiles of the jump-up and

down-sections of the positive and negative extremes i.e. assuming sections are divided into 9 quantiles we get 9 x 2 x 2 different groups of deviations.
The large deviation theory then specifies the convergence/decay rate of these deviations which can be measured for our data sets. To assess whether the theory is valid for our data we plot the logarithm of the probability that the deviances within a group are bigger than some q against q. In case of validity the plot should show a straight line of slope δ as a direct consequence of the second equality mentioned above.

From our analysis conducted with the statistical language program R we found that the theory applies to FTSE and gold option prices to some extent where the decay rates δ varies with different procedures of calculating the deviances. More analysis is however needed for the bunds prices and for the ask as well as mid prices of all data sets.

Sarah Hegarty
Predicting the Probability of Developing a Successful Therapy for Metastatic Pancreatic Cancer

Drug testing is an extremely costly and lengthy process, with Phase III trials alone costing above $200,000,000 and lasting for 2 to 4 years, according to Amgen estimates. However, sheer amount of time and money dedicated to the testing of a drug does not guarantee the drug will prove successful in the trial. Thus, this large investment could be in vain. In fact, currently only 40% of the drugs being tested in Phase III trials make it to the market, according to Kola in [6]. This project aims to predict the probability of success in Phase III for a specific experimental drug developed by Amgen, a California-based biotechnology company with a statistical group located in Cambridge. This experimental drug is referred to as AMG and is intended to increase the overall survival for people suffering from metastatic pancreatic cancer. Taking the calculated probability of success for AMG and comparing it to 40%, the current industry-average of the probability of success in a Phase III study for a drug which has successfully completed a Phase II trial, Amgen can make a more informed decision about whether or not to continue to pursue AMG in a proposed Phase III study.

Firstly, Amgen performed a thorough perusal of the literature on pancreatic cancer drug studies from which various results were extracted, including the six month survival rate, log-hazard ratio and percentage of patients with metastatic pancreatic cancer.

In order to calculate this probability of success, Bayesian methods are employed. Thus, prior distributions for the treatment difference, that is the difference observed in the Phase II estimates of the six month survival rate for the control and active arms, are developed. Three distinct prior distributions are considered in the project in order to reflect the various prior beliefs about the efficacy of AMG. With these prior distributions established, a range of 9 treatment differences is combined with the prior distributions in order to develop a posterior distribution, and, subsequently, a posterior estimate, of the treatment difference in six month survival rates.

However, as aforementioned, AMG is intended to increase overall survival not just six month survival. Thus, the posterior estimates of the difference in six month survival rates must be

translated into estimates of some measure of the difference in overall survival rates. The measure of interest is the log-hazard ratio for overall survival which Amgen would use as its endpoint in the proposed Phase III study. In order to translate the estimates from differences in six month survival rates into log-hazard ratios for overall survival, a meta-regression is performed using the differences in six month survival rates and log-hazard ratios for overall survival extracted from the literature to establish a relationship between the two measures. Then, the estimates of the difference in six month survival rates are passed through this relationship to obtain the estimates of the log-hazard ratio for overall survival.

Finally, with these estimates of the log-hazard ratio for overall survival, the log-hazard ratio observed in the Phase III study can be simulated and an estimate of the true log-hazard ratio for overall survival based on the simulated values for the observed Phase III log-hazard ratio can be obtained. Counting the number of times this estimate shows a “success” significantly can then be used to predict the probability of success in the Phase III study for AMG.

In order to carry out this project, the statistical software packages WinBUGS and R are used. In addition, Excel is utilized to compile the data extracted from the literature.

This project was supervised by Dr. Tony Sabin of Amgen and Dr. Susan M. Pitts of the University of Cambridge.

Helen Jordan
Sampling formulae for marine populations

In this project, I considered genetic variation data from certain marine populations. I studied a specific gene from n individuals sampled from the population. Some of these individuals had exactly the same segment of DNA, while others had a variant of this gene (a different allele).

I wanted to model the amount of genetic diversity in the sample, in order to say something about the genealogy of the population. I studied two datasets, one from Pacific Oysters and one from Atlantic Cod. In the Oyster population, it has been suggested that sometimes one individual can have so many offspring that approximately 8% of the next generation will share the same parent. This means that the usual model of the genealogical tree (``Kingman's coalescent'') is not appropriate for these marine populations, and I had to look for a more suitable model.

If we suppose a certain population model in which distribution of the number of offspring of each individual has heavy tails (where ``how heavy'' the tails are depends on a parameter $\alpha$ then the genealogy converges to another coalescent (a ``Beta coalescent'' with parameter α so this is the model I used for the genealogical tree.

If we throw mutations at random onto a genealogical tree generated from Kingman's coalescent, the resulting distribution of the genetic variation data can be found in closed form. However, for a tree generated by a Beta coalescent, no such closed form formula is known. Instead, I made use of an asymptotic result, which holds as the sample size $n \to \infty$, to find a simple estimate of the parameter α from the genetic variation data. I wrote a program, using Python, to simulate Beta coalescents with various different values of the parameter α and with mutations added at different rates ρ. This gives us simulated genetic variation data for samples from a population with a genealogy generated by a Beta coalescent. I used these

simulated datasets to study (using R) how the estimator varied with α and ρ. I also wrote a program in Python to draw out realisations of the simulated Beta coalescents, which I ultimately used to give an idea of what the genealogical trees of the samples from the Oyster and Cod populations might actually look like.

I also estimated α and ρ using a method based on a recursion for the distribution of the genetic variation data. I wrote a program in Python to implement this algorithm. This method is very computationally expensive, and while it was feasible to use it in the case of a small sample (such as the Oyster data), large sample sizes (such as the Cod data) proved much more problematic.

I concluded that a Beta coalescent seems to be a good model of these marine populations, which have heavy tails in the offspring distribution.

Jean-Jacques Schraemli
Local and stochastic volatility models

Why do we need to go beyond Black-Scholes? This is the question that any student approaching local volatility (LV) and stochastic volatility (SV) models for the first time should have in mind.

In the opening section of this project we try to develop some intuitions. First, we seek a better understanding of how the market's implied probability distribution deviates from the Black-Scholes lognormal one. Then, we focus on how to use this information to evaluate derivatives which prices are not directly available in the market. Finally, we explain why it is crucial to develop a model for pricing exotic options which both:

1. is representative of the dynamics of the market;

2. is consistent with the volatility surface (i.e. vanilla options priced in this model exactly reproduce the volatility surface observed in the market).

A key aspect is that condition 2. implies that options providing payoffs at just one time (e.g. cash-or-nothing digital put options) are priced correctly (i.e. consistently with the market). However, the volatility surface does not directly provide enough information to uniquely determine the process by which the non-lognormal probability distributions arise. In other words, condition 2. does not necessary imply that the joint distribution of the asset price at two or more times is correct. Therefore, a number of different processes for the underlying asset can be postulated which would match the volatility surface and yet give different values for the same path-dependent option.

In Section 2 and 3 we clarify how LV and SV models go beyond Black-Scholes consistently with the above conditions. Fundamental results essential to the aim of the project are provided in both the LV and the SV setting. Particular attention is given to the Heston's (1993) assumption of the stochastic process that governs the volatility dynamics and his choice of equivalent martingale measure.

In Section 4 we discuss two basic numerical methods (and their implementation) for valuing derivatives when exact formulas are not available. Moreover, we analyze numerically (using

MATLAB 7.8.0) how some (path-dependent) exotic options prices in the Heston model differ from those in the LV model where the LV function is determined by the SV volatility surface (i.e. the volatility surface reproduced by the vanilla options priced in the Heston model). This is a simplified/idealized approach, since we do not deal with key matters as:

1. interpolation and extrapolation of the volatility surface (in the market implied volatility is usually observed for only a finite, potentially small, number of strike prices and maturity times);
2. estimation of the SV models free parameters (inverse, ill-posed problem).

However, for sensible choices of the SV free parameters we could for example assess the risk of using one model rather than the other.

Moxi Sun
Estimating the number of unseen species: dinosaurs, coins and Shakespeare's vocabulary

In my project I will discuss a particular type of statistical topic: estimating the number of undiscovered species based on the current existing data. This topic is very important in biology and ecology and therefore has drawn many discussions in various relevant literature. One famous example is dinosaur genera problem, which has been addressed by Wang and Dodson(2006). There are 527 currently discovered genera of dinosaurs in the entire Mesozoic era. Among the currently discovered fossils 309 genera of dinosaurs are known from only 1 individual, 55 genera are known from exactly 2 individuals, ... , 2 genera are known from 300 individuals. We wish to estimate the total genera of dinosaurs on the planet. There are many other similar examples and I will discuss them along with the dinosaur dataset in my project.

Thanks to its popularity, many approaches have been proposed to solve this problem. There are maximum likelihood based estimators that can be traced back to Yule and Greenwood (1920) and a nonparametric maximum likelihood estimator recently proposed by Norris and Pollock in 1998. Also there are coverage-based nonparametric estimators first discussed by Chao and Lee (1992). Finally the problem can be viewed from a different perspective of mark-recapture methodologies (See Burnham and Overton (1979), Boulinier et al. (1998) and Nichols et al. (1998a, b) for further details). In my project I will discuss three approaches: one is nonparametric and based on sample coverage, one by fitting a parametric model (In our case the Gamma-mixed Poisson model) to the data and one that has features of both of the above.

The aim of this project is to explain how the three methods mentioned above work. Then I will apply them to various datasets and give out their estimates. A limitation in Wang and Dodson's dinosaur paper is that they failed to attach any kind of precision to their estimations such as standard error or confidence interval. I will address this omission as well as doing similar analysis to other data.

The application of these approaches is carried out using a software called SPADE written in C. R is also used for statistical analysis.

Sofoclis P. Zambirinis
Models for predicting health outcomes

A cardiovascular risk-profiling system takes an individual's characteristics and uses a formula to derive a probability of, for example, experiencing a heart attack or stroke in the next 5 or 10 years. A number of such systems exist, generally constructed using a Cox regression analysis of a large dataset. In this project we will review a current system (Framingham) that uses Cox proportional hazards model to produce 10-year risks of cardiovascular events. Under the extra assumption of a piecewise constant hazard over one-year intervals, we can prove a formula that allows us to calculate 1-year risks and thus extend the current system.

We will review an already existing method for producing overall survival curves from the prospective of a whole population and use the theory of Competing Risks to give its theoretical description. We will then adapt the method to produce hazard curves from the prospective of a specific individual. This will give us the capability to construct a model that will allow an individual to see what the effect might be of changing some behaviors or taking medical treatment. We can then compare the effect of different treatments and the effect of starting treatment at different ages. The potential benefit can be expressed in terms of changes in: annual or cumulative risk of a cardiovascular event, annual or cumulative risk of death from any cause, and life expectancy. We can also explore the effect of using different thresholds to guide treatment decisions.

There may be a difference between the effect of changing a risk factor and naturally having a different level of that factor; for example people who stop smoking do not attain the same risk with someone who has never smoked before. We will explore the effect of an intervention, compare it with naturally occurring differences and pay particular attention to treat interventions appropriately during the whole implementation of the risk-profiling system.

The theoretical description of the system is accompanied by its implementation using programming in R.

Chenye Zhang
Risk analysis of a multi-asset private wealth portfolio

In selecting investment options, investors often use two approaches: top-down investing approach and bottom-up security selection. The former method allows investors to analyze the market in whole down to its individual stocks (Top-Down Investing); whereas the latter method enables investors to begin with individual stocks and move to expand to include the global economy (Bottom-Up Investing). This project will focus on the top-down asset allocation aspect.

The risk analysis of a portfolio begins by individual analysis, i.e. analyze each asset class separately. We first obtain the basic characteristics of the assets: mean, standard deviation (or volatility), skewness and kurtosis. We also calculate the VaR and CVaR for the assets. Next, we investigate the covariance and correlation between the asset classes in the portfolio, and finally obtain the VaR and CVaR for the portfolio.

This project is the risk analysis for a multi-asset private wealth portfolio provided by Vestra Wealth LLP. Each asset class in the portfolio is represented by an index. The project is organized as follows. In Chapter 1, we examine the indices in the portfolio individually. We first introduce the indices in Section 1.1. In Section 1.2, we describe the dataset in detail, and then produce histograms to study the distributions of the indices and perform the preliminary analysis, i.e. mean, standard deviation, skewness, and kurtosis calculations. Section 1.3 investigates the normality of each index using the Jarque-Bera test. Section 1.4 focuses on the correlation between returns in each return series. Section 1.5 calculates the VaR and CVaR of individual indices. Lastly, section 1.6 considers the variance ratio test. Starting from Chapter 2, we analyze the portfolio as a whole instead of doing individual index analysis. The Stambaugh's method, a method for investment analysis based on return histories that differ in length across assets, is used in Chapter 2 to calculate the covariance between the indices. Chapter 3 uses the Gaussian copula and t-copula to explore the different dependence structures between various asset classes, and performs VaR and CVaR calculations for the whole portfolio. Finally, Chapter 4 concludes the methods used in our risk analysis.

For this project, all analyses are performed in R.