13/07/2018

THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT4799 Statistics Project (12 credits)

(Offered in 2018 – 2019 year long)

1. Mortality projection and longevity risk

Life insurance companies face different types of risks. Within the life annuity business we may find what is called 'longevity risk', which refers to the possibility that annuitants live longer than expected according to the life tables used for pricing, determination of benefits and technical reserves.

This situation represents a threat to life annuity business, and therefore we need to rely on projected life tables that account for the improvement in mortality, a fact that has been observed since the second half of the 20th Century in most developed countries.

The student who takes this project is expected to study the most widespread models in the literature for mortality projection in order to mitigate this type of risk.

Requirement: STAT3901 and STAT3909.

Supervisor: **Dr. A. Benchimol**, benchi@hku.hk, Dept of Statistics and Actuarial Science

2. Randomized algorithms

A randomized algorithm is an algorithm that employs a degree of randomness as part of its logic. This project studies selected randomized algorithms and related techniques, such as the Markov chain Monte Carlo method, Gibbs sampler, the Metropolis algorithm, the Propp–Wilson algorithm, coupling, sandwiching, and simulated annealing. Applications to approximate counting of combinatorial objects, simulations of Ising model, travelling salesman problem, etc, will be investigated. The student needs to have a good background in the theory of Markov chains. Programming skills are also essential.

Supervisor: **Dr. K.C. Cheung**, kccg@hku.hk, Dept of Statistics and Actuarial Science

3. Graphical Kernels and Gaussian Processes

Gaussian processes choose covariance function from the class of positive definite kernels. It is interesting to consider applications of Graphical Kernels on Gaussian processes. This study would like to have an in-depth review on Graphical Kernels and Gaussian Processes and their applications especially on Regression reinforcement learning.

Supervisor: **Dr. Simon K.C. Cheung,** simonkc@hku.hk, Dept of Statistics and Actuarial Science

4. Protein Structure Alignment

It has long been recognized that tertiary structure is more conserved than the amino acid sequence in protein evolution. The consequence of this is that proteins often adopt similar folds and the number of stable folds is limited. A classification of known three-dimensional (3D) structures of protein domains into fold families can assist molecular biologists and bioinformaticians in detecting the evolutionary relationship among proteins that are lacking of sequence similarity. With the rapid growth of the protein structure databases, numerous protein structure classification databases were constructed and made public. Whereas most of these databases (e.g. SCOP) classify protein structures manually by visual inspection, some databases (e.g. FSSP, 3Dee) organize structural clusters based on computational comparison algorithms. There is, however, no perfect agreement on these classifications of protein structures as different databases are constructed and maintained by different algorithms, under different standards and for different purposes. In this project, student will study the basic knowledge of structural alignment and explore different algorithms on aligning 3D structures of protein domains.

Requirement: Strong knowledge in programming language like C++ or R is a MUST. Some basic knowledge in biochemistry and computational algorithms such as dynamic programming and MCMC would be essential.

Supervisor: **Dr. Y.K. Chung**, yukchung@hku.hk, Dept of Statistics and Actuarial Science

5. Test for Parent-of-origin effects on the X-chromosome

Genomic imprinting is an important epigenetic factor in complex traits study, which has generally been examined by testing for parent-of-origin effects of alleles. In this project, we shall work on the detection of parent-of-origin effects on the X-chromosome, using the parental-asymmetry test based on case-parents trios and/or case-parent pairs data. Various other statistical tests would also be considered. The student is expected to write computer programs in R.

Supervisor: **Prof. Tony W.K. Fung**, wingfung@hku.hk, Dept of Statistics and Actuarial Science

6. Measurement Error Problem

The student will be asked to explore parametric, semi-parametric and non-parametric methods when the variables are measured with errors. The students will conduct numerical comparisons among various density estimation and regression methods. Hence, strong computational skill are required. High dimensional measurement error problem will be studied, and non-convex programming will be discussed.

Supervisor: **Dr. F. Jiang**, feijiang@hku.hk, Dept of Statistics and Actuarial Science

7. Latent Class Analysis

Latent class analysis is a method for analyzing the relationships among manifest data when some variables are unobserved. The unobserved variables are categorical, allowing the original dataset to be segmented into a number of exclusive and exhaustive subsets. This project is to explore some latent class analysis methods. It includes the evaluation of some existing algorithm and models as well as the determination of the number of latent classes and variable set.

Requirement: Knowledge of some computer programming languages is essential.

Supervisor: **Dr. C.W. Kwan**, cwkwan@hku.hk, Dept of Statistics and Actuarial Science

8. Analysis of correlated zero-inflated count data

In many medical and public health investigations, the count data encountered often exhibit an excess of zeros, and very frequently this type of data are collected on clusters of subjects or by repeated measurements on each subject. For example, in the analysis of medical expenditure, members in the same family may exhibit some correlation possibly due to housing locality, genetic predisposition, similar dietary and living habit. Ignoring such correlation may lead to misleading statistical inference. This project will survey the models and methods in the literature and apply them to a real data set.

Requirement: Knowledge in programming language like FORTRAN or C++.

Supervisor: **Dr. Eddy K.F. Lam**, hrntlkf@hku.hk, Dept of Statistics and Actuarial Science

9. Tail dependence and multivariate copula models

Copulas are multivariate distributions with standard uniform (i.e., Uniform(0,1)) margins, and are useful to describe the dependence characteristics among variables. When choosing an appropriate parametric copula family, one important characteristic to consider is the strength of tail dependence, or the likelihood of variables simultaneously taking large or small values. In this project, the student will explore some tail properties of various parametric copula models, and evaluate their suitability in modelling actual data sets based on various criteria.

Requirement: The student should be competent in computer programming. Knowledge in or willingness to learn the R programming language is essential.

Supervisor: **Dr. David Lee**, leedav@hku.hk, Dept of Statistics and Actuarial Science

10. Resampling methods for regression

Recent years have found increasing use of resampling methods in regression studies. Examples include the paired bootstrap, the residual bootstrap, the wild bootstrap, random perturbation, bagging, etc. In this project we explore their potential applications in contemporary regression settings where statistical inference remains prohibitively difficult.

Supervisor: **Prof. Stephen M.S. Lee**, smslee@hku.hk, Dept of Statistics and Actuarial Science

11. Security design in blockchain architecture

In this project we begin with a review of the basic architecture for blockchain in Python. This includes state transition rules, method for creating blocks, mechanisms for checking the validity of transactions, blocks, and the full chain. Next, we will create new blocks from data, validate the new blocks and add them to the existing blockchain.

Security is of the utmost importance in any blockchain architecture, in this project we will discuss 3 popular verification methods: public key cryptography, digital signature algorithm and trusted time-stamping. Despite the advanced level of technical sophistication, we will construct practical examples wherever possible.

Supervisor: **Dr. Eric A. L. Li**, ericli11@hku.hk, Dept of Statistics and Actuarial Science

12. Bootstrap approximation in time series modeling

The traditional time series modeling and further inference are based on the normality assumption or large enough sample size. In the real applications, the normality may be broken and the results may not be accurate for the moderate or small sample sizes. The bootstrap is a computer-intensive method, and the information in the real data is repeatedly used. Hence it may provide more accurate results. This project hopefully can train students for some bootstrap methods to dependent data, and some knowledge of computer languages such as FORTRAN or C is required since a little more computation will be involved.

Supervisor: **Dr. G.D. Li**, gdli@hku.hk, Dept of Statistics and Actuarial Science

13. DNA Sequencing Data

DNA sequencing data holds the promise of identifying causal rare variants associated with human traits and diseases with genetic background. In this project, we shall work on the detection of rare variants using next generation DNA sequencing data. The students are expected to have basic knowledge of kernel machine regression and statistical genetics, and should be strong in programming.

Supervisor: **Dr. Z. Liu**, zhhliu@hku.hk, Dept of Statistics and Actuarial Science

14. Deep Learning for NLP

Textual data could be found everywhere over the internet world, e.g. Facebook, twitter and web blogs. In the era of big data, more sophisticated techniques have been developed to handle this type of unstructured data or semi-structured data and extract hidden information for business applications. In this project, one of important area in textual analytics, Natural Language Processing (NLP) is implemented to analyze textual data which could be English language or any other languages. Due to huge data volume, traditional statistical techniques may not be capable of handling the task. Indeed, deep learning is one of the popular and effective machine learning techniques to analyze the sequential data. Among all those techniques, students are required to discover the insight of these algorithms. Certain extension could be found from these insights.

Requirement: Knowledge of R/python programming is required for this project.

Supervisor: **Dr. Gilbert C.S. Lui,** csglui@hku.hk, Dept of Statistics and Actuarial Science

15. Cointegration in financial analysis

The goal of this project is to test cointegration in financial time series. Students are required to have basic understanding of cointegration and some knowledge of computer programming.

Supervisor: **Dr. C. Wang**, stacw@hku.hk, Dept of Statistics and Actuarial Science

16. An Extensive Study on Exotic Options

The study of classical plain vanilla options has been a popular topic in financial engineering and risk management for many years. In contrast, the study of exotic options might not draw as much attention as their plain vanilla counterparts. The complicated payoff features and the complex product structures of exotic options keep them away from traditional elementary courses in financial risk modelling.

This project aims to explore the large class of exotic options with their classification, payoff strategies, pricing and modelling, as well as applicability to the real world.

Students taking this project are expected to have fundamental knowledge in option pricing and programming skills for extensive simulation study.

References:

- Bouzoubaa, M. and Osseiran, A. (2010). *Exotic Options and Hybrids: A Guide to Structuring, Pricing and Trading*. Wiley. http://onlinelibrary.wiley.com.eproxy2.lib.hku.hk/book/10.1002/9781119206965
- Hull, J. C. (2014). *Options, Futures, and Other Derivatives (9th Edition)*. Pearson.

Supervisor: **Dr. K.P. Wat**, watkp@hku.hk, Dept of Statistics and Actuarial Science

17. Investigation of Non-normality in a Simple Errors-in-variables Model

In a classical linear regression model, it is usually assumed that the predictive variable is not subject to any kind of random error. However, it is not always true in many applications. In addition, it is also a common practice to assume that the error in the regression model is normally distributed. Unfortunately, we may often find that most real data sets do not really exhibit such nice properties. In this project, student will investigate the non-normality situation where the errors in a regression model exist. Computer programming skill is required.

Requirement: Strong knowledge in computer programming and statistical simulation technique is a must.

Supervisor: **Dr. Raymond W.L. Wong**, rwong@hku.hk, Dept of Statistics & Actuarial Science

18. Deep Learning with Time-To-Event Data

Students will study the state of the art machine learning methods for time-to-event data and apply them to real financial and biomedical datasets. To appreciate the power and impact of deep learning in real applications in Finance and Medicine, it is of great interest to implement various machine learning methods in building up predictive models and assessing their practical performance. Students will survey an extensive literature and conduct extensive numerical studies in the investigation. The passion for programming and dedication to making impact on practice will be greatly appreciated.

Requirement: R (and Phython) programming.

Supervisor: **Dr. J.F. Xu**, xujf@hku.hk**,** Dept of Statistics & Actuarial Science

19. Stopped Stochastic Processes and Their Applications

In this project, the student will study various kinds stopped stochastic processes and their applications. There is no standard reference book on this topic. I shall provide the student some references. To take this course, the student needs to have some background on stochastic processes, such as, random walk, Brownian motion, Poisson process and basic stochastic calculus.

Supervisor: **Prof. H.L. Yang**, hlyang@hku.hk, Dept of Statistics & Actuarial Science

20. Wishart matrix, eigenvalue distribution and the Marchenko-Pastur law

Wishart matrix is a matrix model for sample covariance matrix from a multivariate normal distribution. It has a long history and many interesting results exist for its eigenvalues and eigenvectors. When the dimension increases to infinity, the empirical distribution of the eigenvalues converges to the celebrated Marchenko-Pastur law.

In this project, students will learn some basis theory on Wishart matrix and multivariate normal distributions. Some techniques from random matrix theory will be needed to derive the Marchenko-Pastur distribution. A good command of multivariate analysis and matrix algebra is required.

Supervisor: **Prof. Jeff J.F. Yao**, jeffyao@hku.hk, Dept of Statistics & Actuarial Science

21. Deep learning with application in artificial intelligence

This project will focus on extracting useful information from structured and unstructured data and formulating statistical models for inference and prediction. In particular, we will develop deep learning, including deep neural networks for imaging analysis and computer vision and natural language processing for text data analysis. Extensive computation will be needed and real data will be used for analysis and illustration.

Requirement: The student needs to have experience with Python and R programming.

Supervisor: **Prof. G.S. Yin**, gyin@hku.hk, Dept of Statistics and Actuarial Science

22. Electricity Load Forecasting using Deep Learning

Electricity load forecasting plays an essential role in the optimization system monitored by energy companies for power system scheduling. Accurate load forecasting can help companies to secure electricity supply and scheduling and reduce wastes since electricity is difficult to store. In this project, we will consider some deep learning models for short term load forecasting. Hourly load data in Hong Kong and elsewhere will be studied. Knowledge of R/Python is preferable.

Supervisor: **Dr. Philip L.H. Yu**, plhyu@hku.hk, Dept of Statistics and Actuarial Science

23. Pattern Recognition using Social Media

Identifying individuals at risk (such as suicide risk and preventable chronic illness) at an early stage is vital for intervention and prevention. As more and more people are using emergent online media, it is a great potential to provide an efficient way to reach out to hidden individuals at risk and to study their behavioral and linguistic characteristics through online social media. This project will extract profile and linguistic feature data from social media users and apply data mining methods to address the above problems.

Requirement: Knowledge of R/Python is preferable.

Supervisor: **Dr. Philip L.H. Yu**, plhyu@hku.hk, Dept of Statistics and Actuarial Science

24. Insurance Risk Models with Dependent Risks

In classical risk theory, the assumption of independence in the study of the surplus process of an insurance company plays an important role. Since this assumption is rather restrictive and unrealistic, insurance risk models with dependent risks have been studied extensively in the past few decades. In this project, a number of these models will be discussed. In particular, for each of these models, numerical and simulation studies will be carried out to assess the impact of the dependence structure on some actuarial quantities related to ruin.

Supervisor: **Prof. K.C. Yuen**, kcyuen@hku.hk, Dept of Statistics and Actuarial Science

25. Statistical Analysis of Large-scale Educational Assessment Datasets

In this project, we will explore international large-scale assessment datasets such as PISA, TIMSS and PIRLS with data manipulation and visualization techniques, then perform statistical analysis and predictive modeling. In particular, machine learning approaches will be investigated and compared to traditional modeling techniques such as item response theory and latent variable models. Proficient programming skills with R and/or Python are required.

Supervisor: **Dr. A.J. Zhang**, ajzhang@hku.hk, Dept of Statistics and Actuarial Science

26. Forecasting time series: with application to stocks trading

This project aims to forecast forward behavior of stock prices using either of the following techniques: threshold time series models, neural networks, random decision forests, support vector. Simulated trading strategies based on the forecast results are also required.

Requirement: Knowledge of course STAT3612 or STAT8017, and skills in statistical programming using either SAS, R, or C++.

Supervisor: **Dr. Z.Q. Zhang**, zhangz08@hku.hk, Dept of Statistics and Actuarial Science

27. Non-linear time series analysis

Non-linear time series models have achieved a great success in real applications. This project aims to give a study on the modelling and statistical inference of many non-linear time series models such as threshold AR, GARCH, and their variants. Students are expected to use these methodologies to analyze real data sets.

Requirement: Statistics and Matlab.

Supervisor: **Dr. K. Zhu**, mazhuke@hku.hk, Dept of Statistics and Actuarial Science

********** END **********