THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT4799 Statistics Project (12 credits)
(Offered in 2021 - 2022 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. Statistical learning with fewer labeled data

Modern statistical learning models such as deep neural networks can achieve very good performances in a wide range of learning tasks such as speech recognition and image processing. However, these learning models typically rely on the availability of a large number of labeled training data, and in many learning tasks, labeled training data may be very difficult to obtain. For example, in image classification problems, the labels of the images in the training data set are usually given manually by human labelers, and thus can be expensive to obtain. To investigate this issue, this project aims at studying learning methods that work even when there is only a limited number of labeled data.

Possible directions include:
(1) Active learning. Active learning is a learning paradigm that actively queries the data label during training. It is known to be effective in reducing human labeling efforts by actively selecting the most informative examples to label.
(2) Semi-supervised learning. Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.
(3) Transfer learning/domain adaptation. Transfer learning applies the knowledge gained while solving one problem to a different but related problem. In this way, the learning of the related problem is possible even with limited data.

In this project, students will survey methods in the literature and run simulations/real-world experiments to compare the label complexity of different methods.

Requirement: Students are required to have basic knowledge in statistical learning and programming.

Supervisor: **Dr. Y. Cao**, saas@hku.hk, Dept of Statistics and Actuarial Science

3. Optimal Reinsurance Design

The objective of this project is to examine the optimal reinsurance strategies that best suits the risk profile and risk preference of insurance companies, subject to various business constraints and risk management considerations. The optimal design of reinsurance products, including both the indemnity structures and pricing, will also be investigated, especially under the framework of information asymmetry such as adverse selection.

Requirement: Knowledge in risk measures, a rudimentary knowledge in convex analysis and convex optimization.

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

4. Protein Structure Alignment and Clustering

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. SCOP2) classify protein structures manually by visual inspection, some databases (e.g. FSSP, CATH, 3Dee, Bio3D, etc) 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 and clustering 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 genetic association accounting for X-chromosome inactivation

X-chromosome inactivation (XCI) is an important epigenetic factor in the study of genetic disease or phenotype. XCI is the mechanism by which gene dosage compensation is achieved between male mammals with a single copy of the X-chromosome and female mammals with two copies of the X-chromosome by inactivation of one of the two copies of the X-chromosome during early embryonic development. In this project, we will propose statistical tests for association on the X-chromosome accounting for XCI. Based on the constructed models, we aim to establish a theoretical and empirical framework for X-chromosome association testing of common variants, allowing for different degrees of X-inactivation. The performance of the proposed tests will be investigated by simulation and the analysis of a data set. Students are expected to write computer programs using R. Knowledge in genetics is preferred but not necessary.

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

6. Machine learning methods for analysing single-cell genomic data

In recent years, the rapid development of single-cell sequencing technologies brings unprecedented opportunities to disentangle the heterogeneity in cell populations, including different immune cells, differentiation trajectory or cancer mutations. However, it remains highly challenging to decipher how a biological system functions and the underlying patterns of the data, not only because of high technical noise but also the high dimensions of gene or other molecular feature space.

Therefore, this project aims to develop machine learning methods, likely in a form of probabilistic models or deep learning methods, to analyse single-cell genomic data. Experimental data is available for both model validation and biological exploration. The student is expected to have interests in biomedical data, but no previous experience is required.

Requirement: Experience with Python or R programming.

Supervisor: **Dr. Y.H. Huang**, yuanhua@hku.hk, School of Biomedical Sciences & Dept of Statistics and Actuarial Science

7. Applications of unsupervised learning

Unsupervised learning aims at representing structure in the input data, often by means of features. The resulting features can be used as input for classification tasks or as initialization for further supervised learning. Traditional principal component analysis, factor analysis and independent component analysis are some examples. Deep learning methods including autoencoders, variational inference, variational autoencoders and GAN models are also developed.

The objective of the project is to explore and compare various unsupervised methods.

- Literature review of various unsupervised learning in the recent years.
- Apply to a real data set to identify any hidden features.
- Conduct simulation of various scenarios and evaluate the accuracy of models with various measures.

Requirement: Knowledge in Python,
Knowledge in multivariate statistics and machine learning,

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 R or Python.

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

9. Big Data Analytics in Securities Market

This project aims to discover new knowledge of market volatility of US securities and Hong Kong securities. Students will learn some skills of financial data analysis, and use big data to analyze the securities markets. Some statistical techniques, social media analysis, and/or artificial intelligence methods will be explored in this project. Students with some basic knowledge in statistics, data mining, text mining, and/or programming will take an advantage. Students are expected to be willing to learn new methods and skills.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

10. Building a Predictive Model to Determine Epidemic Disease Outbreak Risk

This project aims to develop a predictive model of epidemic disease outbreak risk. Students will research the environmental factors of epidemic disease outbreak, and explore to use a new AI method for building a predictive model. Students will learn the methods of contact analysis, network analysis, and artificial intelligence in this project. Students are expected to be interested in research, have some basic knowledge in programming, and use statistics and data mining software.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

11. Data Visualization in Global Market Analysis

This project aims to discover new business opportunities between countries by analyzing the import and export trading data. Some analysis, such as network analysis, demand and supply analysis, etc., will be done in this project. Students will develop a knowledge map to visualize the import and export trading patterns, and new discovery. Students will learn how to apply statistical methods and data analysis theories in business analysis. Students who have basic knowledge in using statistics and data mining software are preferred.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

12. 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

13. 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

14. Applications of Secure Blockchain Solution

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. Finally, we will construct practical blockchain solutions to current fintech problems.

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

15. Introduction to Quantum Computing Algorithms

First we begin with a basic understanding of quantum computing (QC). Then we move on to some popular QC algorithms, written in Javascript and Python. In addition to constructing these QC codes, we will also provide the meanings, purposes and theoretical bases of these QC codes.

The QC algorithms we will cover include: Deutsch-Jozsa Algorithm, Simon's Algorithm, Super Dense Coding, Period Finding, and Shor's Factoring Algorithm. The last one is particularly important in modern cryptography: given an integer which is a product of two distinct prime numbers, this algorithm finds one of its prime factors.

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

16. Statistical Inference for Tensor Data

Tensors have been used in many fields and have provided powerful applications in various practical domains. They generalize vectors and matrices and have been studied from different viewpoints. The study of tensor methods has a long history in statistics. In the era of big data, tensor data appear frequently in the forms of video data, spatio-temporal expression data, relationship data in recommending and mining, and latent variable models, from a vast range of statistical applications. However, the extension of methods for dealing with matrices to tensors is much more difficult than those from vectors to matrices. This project targets to several tensor-based statistical methods.

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

17. Modeling of Social Media Data

In this project, the students will implement latent Dirichlet allocation models to analyse social media data to discover hidden semantic structure in the social media. The students need to know python programming languages and data crawling skills.

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

18. 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

19. 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. (full book accessible online from HKUL by HKU students)
- de Weert, F. (2008). *Exotic Options Trading*. Wiley. (full book accessible online from HKUL by HKU students)
- Buchen, P. (2012). *An Introduction to Exotic Option Pricing*. CRC Press.
- Yen, J. and Lai, K. K. (2014). *Emerging Financial Derivatives: Understanding Exotic Options and Structured Products*. Routledge.
- Kyprianou, A., Schoutens, W., and Wilmott, P. (2005). *Exotic Option Pricing and Advanced Lévy Models*. Wiley.
- Hull, J. C. (2018). *Options, Futures, and Other Derivatives (10th Edition)*. Pearson.

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

20. On Randomized Algorithm to Graph Problems

Most graph problems are known to be NP complete or #P complete. While a deterministic linear time algorithm in solving those problem is not optimistic, there often exists linear time approximation algorithm by making use of randomization. The analysis of complexity and correctness often involves the use of probability theory. The student who takes this course is expected to study the basic theory in relation to randomized algorithm design and its application to some famous graph problems. All related literature will be provided.

Supervisor: **Dr. Jeff T.Y. Wong**, jefftywong@hku.hk, Dept of Statistics & Actuarial Science

21. Statistical Learning of Recurrent Events Data

Recurrent events data is an important type of survival data, which is frequently encountered in practice. There has been a vast literature on recurrent events data analysis. For example, well-known methods include modeling the intensity process of recurrent events and modeling the marginal hazard of each recurrent event or the gap time between recurrent events. In addition, recurrent events problems can also be fit into the paradigm of multi-state models, for which transition probability or transition intensity between states can be estimated to characterize event progression. Another popular approach for recurrent events data is to specify covariate effects on mean or rate functions of recurrent events. This type of approach is attractive because mean or rate functions are more intuitive to interpret than intensity or hazard functions. This project will assess and compare the above-mentioned approaches. In addition, we will also employ various machine learning/deep learning techniques and develop powerful predictive models for recurrent events data.

Requirement: R or Python.

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

22. Valuation of Equity-linked Insurance Products

Many insurance companies and financial institutions are involved in trading variable annuities and equity-indexed insurance contracts. The contractual structures of these products are more sophisticated than traditional insurance products. In particular, these products contain various exotic derivatives features. So, advanced quantitative tools and methods are required for valuation and risk management of these products. This project focuses on the valuation of this kind insurance products. I shall provide reading materials to the students, the students who take this project need to read related literature first, then summarize the main idea and methodologies. I shall also provide some research topics to the student.

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

23. Wishart Matrix 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

24. 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. Yin**, gyin@hku.hk, Dept of Statistics and Actuarial Science

25. Generalizable machine learning technology with application in medical image analysis

Medical imaging is a critical step in modern healthcare procedures and automatic medical image analysis is beneficial to computer-aided diagnosis, assessment, and therapy. While deep learning has achieved remarkable success in medical image analysis in recent years, most of the deep models lack the sense of reliability and generalization when facing new cases, which prohibits deep-learning-based methods from being translated into clinical practice. This project will explore and develop generalizable deep learning/machine learning techniques for medical image analysis. Real public benchmark data will be used for analysis and illustration.

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

26. Multi-task machine learning for joint diagnosis and prognosis of human cancers

With the tremendous development of artificial intelligence, many machine learning algorithms have been applied to the diagnosis of human cancers. Recently, rather than predicting categorical variables (e.g., stages and subtypes) as in cancer diagnosis, several prognosis prediction models basing on patients' survival information have been adopted to estimate the clinical outcome of cancer patients. However, most existing studies treat the diagnosis and prognosis tasks separately. This project will explore and develop multi-task multi-modal machine learning techniques for joint diagnosis and prognosis of cancers from multi-modality medical data, such as histopathological image and genomics data. Public real dataset will be used for analysis and illustration.

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

27. Optimality Studies with Dependent Risks

Due to the complexity of modern insurance and financial products, contemporary insurance risk models have taken many realistic features into consideration. In the actuarial literature, the incorporation of realistic features such as dividends, investment and reinsurance into the basic insurance risk process has generated a lot of interesting research on optimality in the past two decades. This project aims at studying optimal dividends, investment and/or reinsurance for an insurance risk models with dependent risks.

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

28. Non-stationary Multiple Output Gaussian Processes

This project focuses on developing a non-linear Gaussian process regression model with multiple outputs that can capture changes in the functional behavior over time (non-stationary behavior). Students will learn about Bayesian machine learning. The project will involve extensive amounts of programming and computation. Interested students are required to have experience in Python.

Supervisor: **Dr. Michael M.Y. Zhang**, mzhang18@hku.hk, Dept of Statistics and Actuarial Science

29. Forecasting Time Series: with Application to Stocks Trading

This project aims to forecast forward behavior of stock prices using neural networks. Simulated trading strategies based on the forecast results are also required.

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

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

30. Financial data analysis

This project aims to analyze the financial data by using the time series models, causal semantics, or machine learning techniques. Students are expected to use these methodologies to analyze real data sets, and develop useful trading algorithms.

Requirement: At least one programming language and knowledge about financial time series analysis

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

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