PROPOSED TOPICS FOR GRADUATES in MATHEMATICS and STATISTICAL SCIENCE

Abstract

This document is designed for graduate students at the Department of Mathematics, **Faculty of Science**, *Mahidol University*, Thailand.

The writing provides a few specific research proposals from Fall 2018 in the disciplines of Pure Mathematics, Data Analytics, Mathematical Statistics, Applied Statistics, Geostatistics, Statistical Quality & Process Control, Environmental Science and Industrial Mathematics.

Project's outcomes can be used in various domains and applications, such as software manufacturing, petroleum and reliability engineering, mathematical modeling in logistics, total quality management (TQM) and process control, statistical modeling in environmental science and transportation study.

CONTACT:

Lecturer	Man VM. Nguyen, Ph.D.
Work interests	Statistical Optimization, Computer Algebra, Data Analytics, Operations Research
Work unit	Department of Mathematics Room M 204/ 3, M building
Institution	Faculty of Science, Mahidol University
E-mail:	man.ngu@mahidol.edu

Contents

==========

Main theoretical themes include the following disciplines.

Pure Algebra Computer Algebra and Algebraic Statistics

Combinatorics for Quality control Group and Design theories for manufacturing

Discrete Optimization Integer programming, graph, algebraic and geometric views

Statistical Inference factorial, optimal and mixture designs in Applied Statistics

Statistical Modeling mostly multivariate data analysis

Geo-statistics modeling and prediction in environment and ecology

Probabilistic models Markov process, Poisson modeling and applications in OR

Bayesian data science Sequential analysis, quickest change point detection ...

Introduction- Guidelines

General requirements. After choosing suitable topics, graduates should prepare a progress report in max 4 weeks, then present during the weekly seminar (**Seminar for graduates**) at Department of Mathematics. A project report has to be submitted (preferably as a hard copy, or in PDF format electronically) to the advisor. Both master students and doctoral students have to follow the following guidance in the first phase of 9 months. Then the passed doctoral students must fulfill the department's regulations, like taking Qualifying Examination...

Each topic- case study (to be described in next sections) must structurally show 5 following parts:

1. **General information about you and your project**: who are you? what is your main points in the report
2. **Introduction- Motivation**: Brief your aim and/or subject (with basic concepts) to be investigated, plus a clear motivation. [Report 1, after 1 month]
3. **Content of your case study**: what is your (a bit more) detailed problem? How to describe a real-life problem by a mathematical/statistical model? [Report 2, after 3 months]
4. **Partially solving your problem** How to apply the ideas into your own case study with your own suggested solutions. [Report 3, after 6 months]
5. **Conducting/proving/computing and/or simulation**: Implement your chosen phenomenon/topic with software **R**, Maple, Matlab, Open Modelica, Singular or GAP. [Report 4 (final report) after 8 months]

Please include at least the following elements in your first three reports:
1) The title, authors and their affiliation, and abstract of your case study.
2) A clear description in words of the research problem.
3) Background info leading up to the research problem.
4) The potential mathematical model(s) to be exploited (if appropriate).
From the 4th report must express:
5) The mathematical approach/techniques used to solve the model(s).
6) The principal conclusions and observations resulted from the research.
7) Other elements should be included whenever appropriate.

To Ph.D. students, after 1 year they continues the work with developing their own solutions for the selected problem, possibly diversifying a little from the original.

See specific proposals in next pages.

1 Dixmier conjecture on Weyl algebra with two generators

Advisors: Dr. Man Nguyen

Target graduate: master students

- **Keywords**: pure mathematics, computer algebra, Lie & Weyl algebras
- **Research motivation- Aims**: Graduate will study Weyl algebras (the algebra of polynomial differential operators) with a finite generating set. Historically, in 1966 Jacques Dixmier conjectured that whether every algebra endomorphism of Weyl algebra with two generators over a characteristic zero field is an automorphism. It is well known that Dixmier conjecture is stably equivalent to the well known Jacobian conjecture, whereby the Jacobian conjecture itself is ranked number 16 in Stephen Smale's list of Mathematical Problems for the 21st Century.

  Graduate will extend the work [1] from the case of degrees of generators $p = 6$ and $q = 9$ to the case of $p = 8$ and $q = 12$.
- **Reference**: [1] *A survey on computational algebraic statistics and its applications* / Nguyen V. Minh Man, East-West J. of Mathematics: Vol. 19, No 2 (2017)

2 Farmland conservation by a simultaneous multiple-knapsack IP model

Advisors: Dr. Man Nguyen

Target graduate: master students

- **Keywords**: operations research (OR), benefit-cost ratio, farmland management, integer programming (IP), binary IP, simultaneous model
- **Research motivation- Aims**: We aim to study the concept of **benefit-cost ratio** targeting in its selection of which agricultural land to preserve from a pool of willing sellers. Benefit-cost ratio targeting has been a selection approach advocated by economists as a way of getting improved aggregate benefit results that can approach the optimal results described above with binary IP.

  Graduate will study popular methods, and combine them in a certain way to find optimal solutions. Key point of this study is also about exploiting realistic dataset of local (domestic) area/county/country.
- **Reference**: *Mathematical programming for agricultural, environmental, and resource economics* / Harry M. Kaiser, Kent D. Messer, chapter 7.

3 Probability perturbation method in Data integration

Advisors: Dr. Man Nguyen and Prof. Nabendu Pal (ULL, the US)

Target graduate: master students

- **Keywords**: Bayesian data science, spatial statistical modeling, inverse probabilistic modeling in petroleum engineering
- **Research motivation- Aims**: Graduate will study the method of probability perturbation, geostatistics methods, and then investigate applications in environmental & ecological modeling, or in petroleum engineering...
- Reference: Jef Caers, Stanford University

4 Quickest change point detection in Terrorist Group Detection

Advisors: Dr. Man Nguyen and Dr. Suntaree Unhapipat

Target graduate: master/ doctoral students

- **Keywords**: stochastic processes, Bayesian statistics, national security
- **Research motivation- Aims**: To develop models for the activity profile of a terrorist group, detecting sudden spurts and down-falls in this profile, and in general, tracking it over a period of time. Two known strategies for spurt detection and tracking would be reviewed and redeveloped here:

  a) a model-independent strategy that uses the exponential weighted moving-average (EWMA) filter to track the strength of the group as measured by the number of attacks perpetrated by it, and

  b) a state estimation strategy that exploits a type of Discrete Stochastic Process with the underlying Hidden Markov Model structure. Finite state cases as d=2, 3, 4 should be most concerned. A demo software must be implemented with realistic data obtained from local places.

- **Reference**:

  [1] *Hidden Markov models for the activity profile of terrorist groups/* Vasanthan Raghavan, Aram Galstyan, and Alexander G. Tartakovsky, the Annals of Applied Statistics, 2012

  [2] *Operations Research and Management Science Handbook/* Natarajan Gautam, CRC Press, 2008

5 Parameter estimation of Life distributions for insufficient data

Advisors: Dr. Man Nguyen and Dr. Suntaree Unhapipat

Target graduate: master students

- **Keywords**: likelihood, parameter estimation, in-complete data sets, reliability in insurance industry
- **Research motivation- Aims**: Graduate should conduct a case study in reliability engineering or insurance industry, by employing suitable methods of probabilistic modeling, efficient analysis of insufficient data-sets and parameter estimation
- **Reference**: Handbook of Statistics, Vol. 7/ C.R. Rao, Elsevier Science

6 Quickest change point detection in Risk Management

Advisors: Dr. Man Nguyen and Dr. Suntaree Unhapipat

Target graduate: master/ doctoral students

- **Keywords**: sequential analysis, risk prediction, change point detection
- **Research motivation- Aims**:

  Quickest detection is a fascinating area of sequential analysis that spans across various branches of science and engineering, such as control of industrial processes, military and environment science...

  Trainee would study the original QCD, different approaches to finding optimal answers in the multistage schemes. The Wald's multistage model and the Neyman-Pearson's multistage model will be the starting base for practical usages in environment or local industrial processes.

- **Reference**:

  [1] *Data-Efficient Quickest Change Detection with On–Off Observation Control/* Taposh Banerjee and Venugopal V. Veeravall, **Sequential Analysis**, 31: 40–77, 2012

  [2] *Quickest Detection Problems: Fifty Years Later/* Albert N. Shiryaev, **Sequential Analysis**, 29: 2010

7 Statistical Analysis and Designs for Pollution Assessment

Advisors: Dr. Man Nguyen, Prof. Nabendu Pal (ULL, the US) and Prof. Phu Le Vo (VNUHCM, Vietnam)

Target graduate: doctoral students

- **Keywords**: multivariate data, repeated measure designs, polluted water
- **Research motivation**: In biological (or environmental) science, the use of animal (or surveying site) in realistic investigations must be suitably controlled and reduced. As a result, statisticians have to work with a very small number n of subjects and must simultaneously cope with a large number of the dimension p of repeated measurements on each subject, i.e. $n < p$, this phenomenon has been named *high-dimensional data*.
- **Terms and data**: *High-dimensional repeated measures designs* (HD-RMD) is very important in scientific research, because often **a lot of factors** differently contribute to a response, or because we have more chances to **observe several times or conduct experiments** under the same conditions when collecting abundant data sets.

  Our specific environmental dataset consists of $p = 88$ groundwater samples from thirty-seven ($n = 37$) wells, surveyed in South Vietnam. The data table provides a valuable source for arsenic pollution assessment of shallow groundwater, and generally water quality assessment in Mekong Delta Region (MDR), Vietnam. This problem recently becomes urgent due to the shortage of water source from upper Mekong River Basin/ Region (flowing) to the lower Mekong River Basin, there local inhabitants have to switch to using ground water in their daily life as well as for industrial activities.

  Design the field data collection: We chose thirty-seven ($n = 37$) independent wells along one of the two major streams of Mekong river system, and our observations were partially collected at wells. In our theoretic setting, we would define:

  - The intervention factor A to be the seasons in Vietnam, then A has $a = 3$ levels, dry (January), dry changing-to rainy (May), and rainy (August or October).
  - The observation time point factor B currently to have only one point ($b = 1$, due to limited budget!).
  - For each well we expect to obtain $d = a \times b = 3 \times 1 = 3$ measurements, hence all 37 wells should give total 111 measurements.

  Which approach for MDR data?

But due to few local constraints we got only 88 groundwater samples (i.e. repeated measurements), missing data occurs, and worse, the data pattern **does not completely fit** to RMD terminology jargon.

Hence missing data occurred in our MDR data. However, we do not employ some techniques like imputation (already implemented in software **R**) to fill in empty values, since each measurement is a vector of arsenic, ammonia, potassium, manganese, and some other key heavy metal concentrations... which are matter for human beings and animal as well. These substances could harm local inhabitants when they use groundwater as a popular alternative for becoming-rare river water source.

- **Research questions**: Motivated from this MDR data, our few of concerns are:
  1/ Statistical view? Since seasonal effects could be significant, which intervention factor A should be defined?

  i/ What designs could be used for this data with missing data points?

  ii/ And what if $b > 1$ and the design is still unbalanced, i.e. some wells would be observed b times, while others would not be?

  2/ Environmental view? Should remediation for the management of sustainable groundwater use be shaped to mitigate **As** exposure in areas where **As**-contaminated groundwater is the main source for irrigation and domestic purposes?

- **Reference**:

[1] *Arsenic Methylation Dynamics in a Rice Paddy Soil Anaerobic Enrichment Culture*, Matthew C. Reid, Julien Maillard, Alexandre Bagnoud, Leia Falquet, and Phu Le Vo, Environmental Science and Technology. 2017, Vol. 51, 1054610554

[2] *Analysis of high-dimensional one group repeated measures designs*, Markus Pauly et al., Statistics, Taylor & Francis 2015, Vol. 49, No. 6, 1243–1261

8 Computational methods for mixed orthogonal array

Advisors: Dr. Man Nguyen and Prof. John Borkowski (MSU, Bozeman, Montana, USA)

Target graduate: doctoral students

- **Keywords**: computer algebra, experimental design, group-theoretic computation, industrial statistics
- **Research motivation- Aims**: Mixed orthogonal arrays have many useful properties in statistical quality control, industrial statistics and total quality management (TQM). Strength 4 OAs furthermore, allow us to theoretically separate all two-factor interactions during the analysis of data obtained from experimentation.

  In [1] a general method for computing OAs with strength $t \geq 2$ was found. We now want to find specific strength 4 designs, investigate it in practice (such as service, industrial manufacturing, actuarial science).
- **Reference**: [1] *A survey on computational algebraic statistics and its applications*, N. V. M. Man , East-West J. of Mathematics: Vol. 19, No 2 (2017) pp. 1-44

9 Air Pollution Modeling and Prediction using Co-kringing

Advisors: Dr. Man Nguyen, Prof. Nabendu Pal (ULL, the US) and Dr An Khuong Nguyen (VNUHCM, Vietnam)

Target graduate: doctoral students

- **Keywords**: co-kriging, multivariate data analysis, spatial prediction, variogram
- **Research motivation- Aims**: Environmental pollution currently is a critical concern from both social and scientific views around the world. Specifically, *air pollution* - caused mostly by transportation and industry - increasingly degrades environment quality, and leads to severe problems for inhabitant's health as well. The building of air quality monitoring stations is essential, but also difficult because of expensive installation costs, no good information of selected areas for installation in order to achieve precise results. Given a realistic dataset observed form a network of monitoring stations, we have a few concerned questions: firstly finding suitable statistical models, fitting those models with data sets observed with high precision; and secondly processing data missing at degraded stations.

  The major aim of this project is handling these questions, mathematically.
- **Reference**: *Air Pollution Prediction with Co-kringing: A Case study with PM_{10} data in Ho Chi Minh City*, Man Nguyen, Nhut C. Nguyen and Khuong A. Nguyen, submitted to Journal of Mathematics for Data Science

10 Probabilistic modeling of Bulk-arrival queues

Advisors: Dr. Man Nguyen and Dr. Suntaree Unhapipat

Target graduate: doctoral students

Keywords: probabilistic modeling, Poisson processes, maximum likelihood estimation (MLE), queueing systems, bulk-arrival queues, transportation science.

Introduction. We study statistical analysis of bulk-arrival queues, an important research in Queueing Theory; which is helpful for *Transportation Science*. When customers arrive at service facility in groups, we have **bulk-arrival queues**.

In those cases, the size of an arriving group may be considered as a random variable X governed by a probability distribution F_X (called batch size distribution), or a fixed number.

When the arrivals folow a Poisson process, and the size X of an arriving group is governed by certain probability distribution, let denote by $M^{[X]}/M/1$ our bulk-arrival queue.

Brief aims of the research. This research project aims to investigate techniques of probabilistic modeling, statistical estimation, and visualization for bulk-arrival queues. Computing the mean system size (the average number of customers in the system) L_S is useful in studying traffic jams, generally in *Transportation Science* .

Research questions include:

1. to compute or estimate the mean system size L_S and the mean queue size L_Q, for some typical probability distributions F_X; (single bulk queue)
2. to write a demo simulation soft visualizing a road segment consisting of one intersection, where arrivals at the intersection folow Poisson process; then investigate road segments consisting of at least two intersections, where arrivals follow Poisson processes with distinct rates;
3. to propose a model capturing cases where at least two bulk-arrival queues going on the same direction; (multi bulk-arrival queues)
4. to conduct a statistical analysis (e.g. confidence estimation) of L_S; and validate on real data. Simulate cases if time allows.

Reference:

[1] Asia Mathematical Conference (AMC) 2009 Proceeding

[2] *Probability and Statistics for Computer Scientists*, chapter 7; Michael Baron, 2nd Edition (2014), CRC Press, Taylor & Francis Group