International Symposium on Business and Industrial Statistics 2018

4-6 July 2018

University of Piraeus, Greece

by

International Society for Business and Industrial Statistics (ISBIS)

and

International Statistical Institute (ISI)

BOOK OF ABSTRACTS

July 2, 2018

Book of abstracts of the ISBIS 2018 Meeting on Statistics in Business and Industry

4-6 July 2018 – Piraeus, Greece
Edited by Dr Athanasios Sachlas, Member of the LOC and Ass. Prof. Georgios Tzavelas, Member of the LOC
http://conf.sta.unipi.gr/

Sponsors (in random order)

Scientific Program Committee (in alphabetical order)

- Sotiris Bersimis, University of Piraeus, Greece (Co-Chair)
- Dan Jeske, University of California, USA (Co-Chair)
- Irena Ograjensek, University of Ljubljana, Slovenia (Co-Chair)
- Ronald Does, University of Amsterdam, Netherlands
- Jonathan Hosking, Amazon USA
- Markos Koutras, University of Piraeus, Greece
- Ioanna Manolopoulou, University College of London, UK
- Marianthi Markatou, University at Buffalo, USA
- Angela Schoengendorfer, Google
- Rituparna Sen, Indian Statistical Institute, India
- Ross Sparks, CSIRO, Australia
- Kwok Tsui, City University of Hong Kong, Hong Kong
- William Woodall, Virginia Tech, USA

Organizing Committee (in alphabetical order)

- Sotiris Bersimis, University of Piraeus, Greece, (Co-Chair)
- Dan Jeske, University of California, USA (Co-Chair)
- Markos Koutras, University of Piraeus, Greece, (Co-Chair)
- Haralambos Evangelaras, University of Piraeus, Greece
- George Iliopoulos, University of Piraeus, Greece
- Aleka Kapatou, American University, USA
- Athanasios Sachlas, University of Piraeus, Greece
- Georgios Tzavelas, University of Piraeus, Greece

Preface

The 2018 International Symposium on Business and Industrial Statistics will be held at University of Piraeus, Greece, 4-6 July 2018.

The Symposium is organized by the International Society for Business and Industrial Statistics (ISBIS), which is an international society that is dedicated to the promotion of business and industrial statistics worldwide. This is interpreted broadly and includes areas such as financial and health services among others. ISBIS is part of a family of international associations that operate under the umbrella of the International Statistical Institute (ISI). ISBIS promotes applications, research, and best current practices in business and industrial statistics, facilitates technology transfer, and fosters communications among members and practitioners worldwide.

To keep up in today's competitive marketplace, enterprise business entities must be able to constantly transform and improve their business. In order to improve, enterprise business entities have started to integrate sophisticated business analytics and big data, internal and external, in their internal operational processes for sales, marketing, finance, management, procurement, etc. Statistical methodology can play a powerful role in developing such an effective business transformation. In this conference, we explore how businesses increase efficiency, support decision-making under uncertainty, improve business operations and ultimately transform their business by using statistics. The ISBIS Meeting focuses upon the use of statistics in business and industry.

The Scientific Program Committee July 2018

Programme

Wednesday July 4 2018

08.00 – 08.25 **Registration**
The registration desk will be opened during all day

08.30 – 09.00 **Welcoming Remarks (Room A)**
Sotiris Bersimis
Chair of the LOC and chair of the SPC
Nalini Ravishanker
President of the ISBIS
Fabrizio Ruggeri
ISI representative
Chair of the SPC
Markos V. Koutras
Vice-Rector of Finance, Planning and Development of the University of Piraeus
Co-chair of the LOC and member of the SPC
Athanasios Kyriazis
Head of the Department of Statistics & Insurance Science, University of Piraeus
Aggelos Kotios
Rector of the University of Piraeus
Panagiotis Kourouplis
Minister of Merchant Shipping

Keynote Address (Room A)

09.00 – 09.45 Evdokia Xekalaki
On Modeling Overdispersion

Statistics in Operations (Room A)

Chair: Refik Soyer

09.45 – 10.15 Vadim Sokolov
Deep Learning: A Bayesian Perspective

10.15 – 10.45 Ehsan Soofi
Maximum entropy demand models with newsvendor information

10.45 – 11.15 Christopher Glynn
Bayesian Modeling of Abandonments in Ticket Queues

Data Science in Industry (Room B)

Chair: Ioanna Manolopoulou

09.45 – 10.15 David Hoyle
Data analysis challenges in analysing users sessions from an e-Commerce platform

10.15 – 10.45 Gavin Whitaker
Using player abilities to predict football

10.45 – 11.15 Christoforos Anagnostopoulos
Weak Supervision for Real-World Statistical Machine Learning: what to do you when you have hardly any labeled data

Combinatorial Testing (Room C)

Chair: Raghu Kacker

09.45 – 10.15 Raghu Kacker
Combinatorial testing: an adaptation of design of experiments

10.15 – 10.45 Jeff Yu Lei
Combinatorial Testing-Based Fault Localization

10.45 – 11.15 Dimitris Simos
Combinatorial Testing Methods and Algorithms for Detecting Cryptographic Trojans

Bayesian Inference (Room D)

Chair: Aleka Kapatou

09.45 – 10.15 Didem Egemen
Bayesian Modeling of Virtual Age in Repairable Systems

10.15 – 10.45 Olawale Awe
Bayesian Estimation of Time-Varying Parameters in the Presence of Discounted Evolution Variance

10.45 – 11.15 Gary Sharp
Tolerance intervals as a method for the assessment of energy output of a photovoltaic system

11.15 – 11.45 **Coffee Break**

Recent Advances in Data Mining Applications (Room A)

Chair: Paulo Canas Rodrigues

11.45 – 12.15 Tahir Ekin
Topic Model based Prescription Fraud and Abuse Detection

12.15 – 12.45 Ozan Kocadagli
Classification of EEG Signals for Epileptic Seizures using Hybrid Artificial Neural Networks based Wavelet Transforms and Fuzzy Relations

12.45 – 13.15 Rahim Mahmoudvand
A Comparison of the Multivariate SSA Methods for Forecasting Mortality Rates

13.15 – 13.45 Paulo Canas Rodrigues
Singular spectrum analysis for long and contaminated time series

Computational Statistics (Room B)

Chair: Martina Vandebroek

11.45 – 12.15 Aviwe Gqwaka
Evaluation of selection tools for the inefficiency distribution in stochastic frontier models

12.15 – 12.45 Jani Deyzel
Viability Assessment of Photovoltaic Systems using Bootstrap-based Tolerance Intervals

12.45 – 13.15 Martina Vandebroek
Designing and conducting discrete choice experiments with the R-package idefix

13.15 – 13.45 —
—

Statistics in Business and Finance (Room C)

Chair: Ta-Hsi Li

11.45 – 12.15 Francesca Bassi
Patterns of adoption of circular economy in Europe

12.15 – 12.45 Renjie Lu
Buffered Vector Error-Correction Models

12.45 – 13.15 Ta-Hsin Li
Experiments with some interpretable neural network models for a customer classification problem in financial industry

ASMBI Session (Room D)

Chair: Fabrizio Ruggeri

11.45 – 11.50 Fabrizio Ruggeri
ASMBI: the ISBIS journal

11.50 – 12.40 Stergios Fotopoulos
Properties of the Multivariate Generalized Hyperbolic Laws

12.40 – 12.50 Rituparna Sen
Discussion

12.50 – 13.00 Vadim Sokolov
Discussion

13.00 – 13.10 Stergios Fotopoulos
Rejoinder

13.10 – 13.45 Floor discussion

13.45 – 15.00 **Lunch**

Special issues of Statistical Process Monitoring (Room A)

Chair: Stelios Psarakis

15.00 – 15.30 Athanasios Sachlas
Multivariate Risk-Adjusted Control Charts

15.30 – 16.00 Javier M. Moguerza
Unattended smoothing of nonlinear profiles using R

16.00 – 16.30 Theodoros Perdikis
Adaptive Schemes for the Multivariate Control Charts

Market Surveys-Business and Statistical Perspectives (Room B)

Chair: Nalini Ravishanker

15.00 – 15.30 Nicholas Fisher
The Good, the Bad, and the Horrible: Interpreting Net-Promoter Score and the Safety Attitudes Questionnaire in the light of good market research practice

15.30 – 16.00 Kamal Sen
Market surveys in emerging markets - perspectives from CPG industry

16.00 – 16.30 Kostas Kotopoulos
Challenges for the researchers to measure marketing effectiveness in the fmcg sector

Network Modelling (Room C)

Chair: James D. Wilson

15.00 – 15.30 James D. Wilson
Weighted Exponential Random Graph Models: Specification and Application

15.30 – 16.00 Can Le
Edge sampling using network local information

16.00 – 16.30 Siraj Sengupta
Anomaly detection in static networks

Recent developments in statistical process monitoring (Room D)

Chair: Ronald Does

15.00 – 15.30 Mandla D. Diko
Guaranteed In-Control Performance of the EWMA $\bar{X}$ Chart

15.30 – 16.00 Leo C.E. Huberts
The Performance of $\bar{X}$ Control Charts for Large, Non-Normally Distributed Datasets

16.00 – 16.30 Xin Zhou
Comparison of Control Charts for Short Run Monitoring of Fractional Nonconformance

16.30 – 17.00 **Coffee Break**

Advances in SPC (Room A)

Chair: Daniel Jeske

17.00 – 17.30 Sven Knoth
Calibrating EWMA control charts for dispersion in presence of parameter uncertainty

17.30 – 18.00 Daniel Jeske
Weighted EWMA Charts for Monitoring Type I Censored Weibull Lifetimes

18.00 – 18.30 Karel Kupka
Approaches in aggregating and scaling-up quality and capability metrics in a corporate structure

Count Time Series (Room B)

Chair: Robert Lund

17.00 – 17.30 Robert Lund
Some New Ways of Modeling Stationary Integer Count Time Series

17.30 – 18.00 James Livsey
Multivariate Count Time Series with Flexible Autocorrelations

18.00 – 18.30 Stefanos Kechagias
Latent Gaussian Count Time Series Modelling

Computationally Efficient Bayesian Modeling of Non-Gaussian Processes (Room C)

Chair: George Iliopoulos

17.00 – 17.30 Scott Holan
Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data

17.30 – 18.00 Refik Soyer
Bayesian Modeling of Non-Gaussian Multivariate Time Series

18.00 – 18.30 Kathy Ensor
Dynamic Continuous Time Models with Discrete Observation Paths of Different Frequencies with Application to Financial Models

18.30 – 19.30 **ISBIS Assembly (Room A)**

19.30 – 21.00 **Welcome Reception**

Thursday July 5 2018

08.30 – 09.00 **Registration**
The registration desk will be opened during all day

Keynote Address (Room A)

09.00 – 09.45 Ron Kenett
On information quality and customer surveys

Design of Experiments I (Room A)

Chair: Haralambos Evangelaras

09.45 – 10.15 Stelios Georgiou
Screening designs based on weighing matrices with added two-level categorical factors

10.15 – 10.45 Stella Stylianou
Construction and analysis of D-optimal edge designs

10.45 – 11.15 Kalliopi Mylona
Supersaturated split-plot experiments

Models for Monitoring Non-Industrial Processes and Applications in Health (Room B)

Chair: Sotiris Bersimis

09.45 – 10.15 Kostas Triantafyllopoulos
Bayesian inference of high dimensional spatio-temporal data

10.15 – 10.45 Polychronis Economou
Simultaneous Monitoring the Number and the Spatial Distribution of Disease Events

10.45 – 11.15 Sotiris Bersimis
Public Health Monitoring Using Scans and Control Charts

Financial Time Series (Room C)

Chair: Abraham Bovas

09.45 – 10.15 Nalini Ravishanker
Multiple Day Biclustering of High-frequency Financial Time Series

10.15 – 10.45 Abraham Bovas
George Box: The "Accidental Statistician" who revolutionized time series analysis

10.45 – 11.15 Aera Thavaneswaran
Generalized Financial Risk Forecasting via Estimating functions with Applications

11.15 – 11.45 **Coffee Break**

Design of Experiments II (Room A)

Chair: Kalliopi Mylona

11.45 – 12.15 Peter Goos
Row-Column Screening Designs

12.15 – 12.45 Antony Overstall
Bayesian design for intractable models

12.45 – 13.15 Vasiliki Koutra
An algorithmic approach for designing experiments on networks

13.15 – 13.45 Werner Mueller
Copula-based robust optimal block designs

Time series and Sampling (Room B)

Chair: Athanasios Sachlas

11.45 – 12.15 Barbara Kowalczyk
Improved Poisson and Negative Binomial Item Count Models for Eliciting Truthful Answers to Sensitive Questions

12.15 – 12.45 Olugbemi A. Olujimi
Social Media Platforms: Tools for Data Collection in Technology-Driven Marketing Research

12.45 – 13.15 Miltiadis S. Chalikias
Optimal Repeated Measurements Two Treatment Designs for Dependent Observations: The Case of AR(1)

13.15 – 13.45 —
—

New findings in the health state of populations and implications in the Health Systems (Room C)

Chair: Christos H. Skiadas

11.45 – 12.15 Christos H. Skiadas
Estimating the health state of populations: Implications in the Health Systems

12.15 – 12.45 Konstantinos Zafeiris
Comparing recent mortality and health experiences in Greece and Turkey

12.45 – 13.15 Fragiskos G. Bersimis
Indices in health related scientific fields

13.15 – 13.45 Yiannis Dimotikalis
Age Groups Demographics Based on Entropic Analytics

13.45 – 15.00 **Lunch and ASMBI Editorial Board Meeting (Room B)**

Topics in Industrial Statistics (Room A)

Chair: Marianthi Markatou

15.00 – 15.30 Elisavet M. Sofikitou
Some Bivariate Semiparametric Charts Based on Order Statistics

15.30 – 16.00 Lindsay Berry
Bayesian Forecasting for high dimensional time series counts

16.00 – 16.30 Nikolaos Demiris
Some challenges with large data and large models in Medical Statistics

Reliability (Room B)

Chair: Asha Gopalakrishnan

15.00 – 15.30 Maxim Finkelstein
Optimal mission duration for non-repairable and partially repairable systems

15.30 – 16.00 Jose Maria Sarabia
Aggregation of risks for lifetimes with mixture exponential distributions

16.00 – 16.30 Asha Gopalakrishnan
On the mean time to failure of an age replacement model

Advanced Methods in Statistical Process Control I (Room C)

Chair: Subha Chakraborti

15.00 – 15.30 Konstantinos Bourazas
A Bayesian self-starting Shiryaev statistic for Phase I data

15.30 – 16.00 Marien Graham
Nonparametric precedence control charts with improved runs-rules

16.00 – 16.30 Athanasios Rakitzis
Control Charts for the Simultaneous Monitoring of the Parameters of a Zero-Inflated Poisson Process Under Unknown Shifts

16.30 – 17.00 **Coffee Break**

Statistical Issues and Methods in Big Data (Room A)

Chair: Panagiotis Tsiamyrtzis

17.00 – 17.30 David Banks
Designing a Curriculum for Data Science

17.30 – 18.00 Foteini Panagou
Model Based Clustering through copulas for high dimensional data

18.00 – 18.30 Panagiotis Tsiamyrtzis
Statistical process control and monitoring in the big data era

Topics on applied statistical inference (Room B)

Chair: Markos V. Koutras

17.00 – 17.30 Marianthi Markatou
Clustering Mixed-Type Data

17.30 – 18.00 Ioannis S. Triantafyllou
Wilcoxon-type rank-sum statistics for selecting the best population: some advances

18.00 – 18.30 —
—

Advanced Methods in Statistical Process Control II (Room C)

Chair: Athanasios Rakitzis

17.00 – 17.30 Konstantinos Tasias
Integrated Production Process Optimization: A Bayesian Approach

17.30 – 18.00 Kim Phuc Tran
Deep Learning and Computer Vision for Quality Control: A Perspective

18.00 – 18.30 —
—

18.30 – 19.30 **ISBIS Council Meeting (Room A)**

20.30 – 24.00 **Conference Banquet**

Friday July 6 2018

08.30 – 09.00 **Registration**
The registration desk will be opened during all day

Keynote Address (Room A)

09.00 – 09.45 Subha Chakraborti
(Almost) Two Decades of Statistical Process Control Research: Some Reflections

Functional Time Series (Room A)

Chair: Rituparna Sen

09.45 – 10.15 *Variable selection for $C[0,1]$-valued AR processes using RKHS*
Title (invited talk)

10.15 – 10.45 Pramita Bagchi
Test for stationarity in functional time series

10.45 – 11.15 Rituparna Sen
Granger causality in yield curves of different markets

Innovative Applications of Statistical Methods in Industry (Room B)

Chair: Balaji Raman

09.45 – 10.15 Balaji Raman
Measuring Effectiveness of Trade Schemes in CPG Domain Using DLM

10.15 – 10.45 Pradeep Sridharan
Multi-Party computations for Privacy Aware collaborative Analytics

10.45 – 11.15 Wee-Yeap Lau
Impact of Unconventional Monetary Policy on Japanese financial markets 2010 - 2016: A Comparison between CME and QQE periods

Novel Bayesian Tools for Analyzing Complex Socio-Economic Data (Room C)

Chair: Xinyi Xu

09.45 – 10.15 Mario Peruggia
Bayesian models for response times in cognitive experiments

10.15 – 10.45 Xiaojing Wang
Detecting Earnings Management: A Novel Tobit Modeling Approach

10.45 – 11.15 Alessandra Mattei
Bayesian Inference for Sequential Treatments under Latent Sequential Ignorability

11.15 – 11.45 **Coffee Break**

Multivariate techniques (Room A)

Chair: Georgios Tzavelas

11.45 – 12.15 Luca Frigau
Comparison of Non-parametric Approaches in Classification of Seeds

12.15 – 12.45 Stephen France
Overlapping Clustering: A Framework, Software, and Empirical Analysis

12.45 – 13.15 Gonzalo Chebi
Penalized M-estimators for Logistic Regression

13.15 – 13.45 —
—

Statistical Applications in Health Care Services Management (Room B)

Chair: Vassilis Plagiannakos

11.45 – 12.15 Vassilis Plagiannakos
Machine learning techniques for the analysis of composite quality indicators

12.15 – 12.45 Theodoros Paschalis
Composite quality indicators for assessing healthcare provision

12.45 – 13.15 Christina Georgakopoulou
Assessing the quality of health services using composite indicators

13.15 – 13.45 Spiros Goulas
Applications of composite indicators for the longitudinal assessment of the quality of health services

Reliability and Statistical Process Control (Room C)

Chair: Min Xie

11.45 – 12.15 Min Xie
Some Approaches for the Monitoring Event Magnitude and Frequency

12.15 – 12.45 Baris Surucu
Hazard Rate Estimation for Location-Scale Families: Monotonic and Non-monotonic Structures

12.45 – 13.15 Min Gong
Comparison of Several Samplers in a Stochastic EM Algorithm for Modeling Imperfect Maintenance Actions

13.15 – 13.45 —
—

13.45 – 15.00 **Lunch**

ISBA-IS (Room A)

Chair: Refik Soyer

15.00 – 15.30 Sanjib Basu
Bayesian Variable Selection in Complex Lifetime Models

15.30 – 16.00 Joshua Landon
Bayesian Analysis of Markov Modulated Queues with Abandonment

16.00 – 16.30 Fabrizio Ruggeri
Project Risk Management under Dynamic Environments

Recent Advances in Time Series Analysis Methods and Applications (Room B)

Chair: Bo Lu

15.00 – 15.30 Ioannis Kamarianakis
Projection predictive variable selection for ARMA models

15.30 – 16.00 Beatriz (Stefa) Etchegaray Garcia
A Bayesian Model for Forecasting Hierarchically Structured Time Series

16.00 – 16.30 Bei Chen
Anomaly detection in time series using deep learning

Social Statistics (Room C)

Chair: Cleon Tsimbos

15.00 – 15.30 Per Eckefeldt
Assessing the macroeconomic and budgetary impact of an ageing population in the EU

15.30 – 16.00 Cleon Tsimbos
Estimating the effects of the recent financial crisis on the stillbirth rates employing distributed lag models and demographic decomposition techniques: the case of Greece

16.00 – 16.30 Georgia Verropoulou
The Effect of Wealth and Income on Depression across European Regions: an Analysis based on Instrumental Variable Probit Models

16.30 – 17.00 **Coffee Break**

Statistical modelling for inhomogeneous and structured data (Room A)

Chair: Ioanna Manolopoulou

17.00 – 17.30 Simon Lunagomez
A Class of Models for Multiple Networks Using Graph Distance

17.30 – 18.00 Ioanna Manolopoulou
Bayesian hierarchical modelling of sparse count processes with applications in retail analytics

18.00 – 18.30 Isadora Antoniano-Villalobos
Bayesian estimation of probabilistic sensitivity measures for computer experiments

Young Statisticians y-BIS Session (Room B)

Chair: Tahir Ekin

17.00 – 17.30 Ezgi Ozer
Detection of Epileptic Seizures using Deep Neural Networks Based on Discrete Wavelet Transforms

17.30 – 18.00 Ayfer Ezgi Yilmaz
Inter-rater Agreement and Adjusted Degree of Distinguishability for 2×2 *Tables*

18.00 – 18.30 Damla Ilter
Comparison of the Hybrid Artificial Intelligence Techniques for Credit Scoring

Evaluation and performance of decision making units (Room C)

Chair: Polychronis Economou

17.00 – 17.30 Eugenio Kahn Epprecht
Ensuring both the in-control and out-of-control Phase II performances of the S2 control chart with estimated variance

17.30 – 18.00 Sonia Malefaki
Studying the performance and the availability of multi-state deteriorating systems: The case study of a diesel engine system

18.00 – 18.30 —
—

18.30 – 19.30 **Closing Remarks (Room A)**

Contents

Abstracts

Weak Supervision for Real-World Statistical Machine Learning: what to do you when you have hardly any labeled data

Christoforos Anagnostopoulos

Mentat, UK

In many domains, the promise of disruptive innovation via the use of machine learning has not materialised due to the lack of the one resource that most machine learning models rely upon: labelled data. This is particularly true in domains where ground truth does not avail itself readily by way of a natural process or a business operation, but is rather the result of tedious manual labelling performed by human experts sifting through the data. One example is cybersecurity, where successful attacks are rare, and can go undetected for long periods of time, so that high-quality labels require a significant time investment from highly paid and exceptionally busy cyber analysts. Another example is natural language processing in niche areas that are awash with domain-specific jargon, such as police investigations. Such domains struggle to take advantage of the huge methodological and technological advances in supervised machine learning technology and are poorly served by most popular machine learning software packages. In this presentation, we take a step back to challenge the standard interface of “learning by example", and offer an alternative, more scalable way of incorporating expert opinion into a machine learning pipeline, known as weakly supervised learning. We discuss how this framework can be related to previous work, the new questions it poses, and the challenges and opportunities it presents us with from a technological perspective.

Bayesian estimation of probabilistic sensitivity measures for computer experiments

Isadora Antoniano-Villalobos, Emanuele Borgonovo and Xuefei Lu

Bocconi University, Italy

Simulation-based experiments have become increasingly important for risk evaluation and decision-making in a broad range of applications, in engineering, science and public policy. In the presence of uncertainty regarding the phenomenon under study and, in particular, of the simulation model inputs, a probabilistic approach to sensitivity analysis becomes crucial. A number of global sensitivity measures have been proposed in the literature, together with estimation methods designed to work at relatively low computational costs. First in line is the one-sample or given-data approach which relies on adequate partitions of the input space. We propose a Bayesian alternative for the estimation of several

sensitivity measures which shows a good performance on synthetic examples, specially for small sample sizes. Furthermore, we propose the use of a nonparametric approach for conditional density estimation which bypasses the need for pre-defined partitions, allowing the sharing of information across the entire input space through the underlying assumption of partial exchangeability. In both cases, the Bayesian paradigm ensures the quantification of the uncertainty in the estimation.

Bayesian Estimation of Time-Varying Parameters in the Presence of Discounted Evolution Variance

Olawale Awe

Department of Mathematical Sciences, Anchor University Lagos

The choice of the evolution variance in the estimation of dynamic linear models play important roles in forecasting as it enables the computation of the one-step-ahead mean squared prediction error vectors as the confidence bounds of the forecasts. In order to give a tractable structure to the dynamic model and reduce its complexity, the variance of the dynamic linear model is often discounted. In this paper, we develop an on-line, recursive Bayesian algorithm for estimation and optimal choice of discount factors for the evolution variance. It is empirically found that, for a range of simulated time series, the proposed algorithm estimated time-varying parameters with discount values (λ) of the evolution variance of the dynamic linear model in the range ($0.50 \leq \lambda \leq 0.75$) while discount values for the fixed parameter models falls mostly in the range ($0.90 \leq \lambda \leq 0.99$) which can be approximated to 1. This range of discount values agrees with existing studies and can be chosen for estimating parameters in dynamic linear models in order to reduce the complexity often associated with the unknown evolution variance.

Test for stationarity in functional time series

Pramita Bagchi

Department of Mathematics, Institute of Statistics, Ruhr-Universitat Bochum, Deutschland

We propose a new measure for stationarity of a functional time series, which is based on an explicit representation of the L^2-distance between the spectral density operator of a non-stationary process and its best (L^2-)approximation by a spectral density operator corresponding to a stationary process. This distance can easily be estimated by sums of Hilbert-Schmidt inner products of periodogram operators (evaluated at different frequencies), and asymptotic normality of an appropriately standardised version of the estimator can be established for the corresponding estimate under the null hypothesis and alternative. As a result we obtain confidence intervals for the discrepancy of the underlying process from a functional stationary process and a simple asymptotic frequency domain level α test (using the quantiles of the normal distribution) for the hypothesis of stationarity of functional time series. Moreover, the new methodology allows also to test precise hypotheses of the form “the functional time series is approximately stationarity", which

means that the new measure of stationarity is smaller than a given threshold. Thus in contrast to methods proposed in the literature our approach also allows to test for “relevant" deviations from stationarity.

We demonstrate in a small simulation study that the new method has very good finite sample properties and compare it with the currently available alternative procedures. Moreover, we apply our test to annual temperature curves.

Designing a Curriculum for Data Science

David Banks

Department of Statistical Science, Duke University, Durham, USA

Essentially all of our current PhD students will be asked to work with Big Data often in their careers. The types of analyses they do will vary widely according to the application, but they will all have one thing in common: Virtually none of their graduate training is directly pertinent to the work that will be needed. This talk describes the kinds of skill sets that the rising generation of statisticians will need, the kind of course work that they will not need, and the strategies for helping our students maintain relevance in the new age.

Patterns of adoption of circular economy in Europe

Francesca Bassi[1] and José G. Dias[2]

[1]*University of Padua, Italy*
[2]*ISCTE, Lisbon, Portugal*

The concept of circular economy was introduced at the end of the last century, the first scientific papers on the topic were published in the 80s and have been receiving an increased attention by scholars still today. This vast literature gives various definitions of circular economy. A popular definition is that which takes advantage of the easy-to-remember so-called 3R: reduction, reusing and recycling; this definition describes the practical approach to the concept. An important fact regarding circular economy is that it was formally adopted in 2002 by the Central Government of China as a new development strategy to protect the environment and limit the production of pollution. This results in a great number of scientific publications both on theoretical aspects and on practical implementations that focus on the Chinese area and/or are authored by Chinese researchers. However, the roots of the topic are in Europe and various areas of the developed world are more and more interested in it. Circular economy, for example, has become recently an important EU policy objective, in particular the guidelines advice on the fact that products should be redesigned so that they are easy to maintain, repair, remanufacture or recycle, which is another way to describe the 3R principles. Nevertheless, the implementation of circular economy is a challenging task given the fact that in industry and society the linear-mind set prevails. According to various researches, the benefits on the environment are easier to be perceived than the economic ones. Often, to implement circular economy

practices, industries have to afford extra investments that might not be considered as profitable. Again, the general perception is that policy initiatives to favour circular economy are missing around the globe. In Europe, current rules provide limited incentives for such market development.

This paper focuses on the implementation of circular economy practices in European enterprises, exploiting data collected within Eurobarometer surveys, which is a series of multi-topic surveys undertaken by the European Commission across EU member states. Specifically, this research studies the profile of industries in Europe prone to circular economy.

Both people's and industries' choices, behaviours life or production styles are recognised as vital in achieving sustainable development. Starting from this evidence, there exists an important number of papers that analyse the profiles of the so-called green consumers and on their behaviour regarding household waste reduction, reuse, recycling, green purchasing, in many areas of the world: UK, Sweden, Japan. For what concerns industries, published research refers either to specific economic sectors or to specific geographical areas. Our research has the advantage of disposing of data referring to all countries of the European Union and collected in industries operating in all economic sectors.

Our analyses take into account the hierarchical nature of the collected data: i.e., the fact that businesses are nested into European countries; in this way heterogeneity between different types of businesses and between different countries is considered.

Descriptive statistics show that circular economy practices are adopted by firms in all 28 European countries, however there exists difference both inside each country due to firms characteristics, as, for example, dimension, age, turnover, type of activity, and also between countries: not everywhere in Europe the same attention is given to environmental and energy saving practices. The estimation of multilevel ordinal regression models investigates the possible determinants of industries' attitudes towards resource efficiency and recycling estimating also the effect of differences among countries.

Bayesian Variable Selection in Complex Lifetime Models

Sanjib Basu

University of Illinois at Chicago

We consider the question of variable selection in complex models for lifetime data. This is often a difficult problem due to the inherent nonlinearity of time-to event models and the resulting non-conjugacy in their Bayesian analysis. Bayesian variable selection in lifetime data models often utilize cross-validated predictive model selection criteria which can be relatively easy to estimate for a given model. However, the performances of these criteria are not well-studied. An alternative criterion is based on the highest posterior model but its implementation can be difficult in non-conjugate lifetime models. In this presentation, we compare the performances of these different criteria in complex lifetime data models including models. We also propose an efficient variable selection method and illustrate its performance in simulation studies and real example.

Bayesian Forecasting for high dimensional time series counts

Lindsay Berry

Department of Statistical Science, Duke University, Durham, USA

Modeling and forecasting of high-dimensional time series of non-negative counts is a common interest for many retailers who rely on forecasts for inventory management, production planning, and marketing decisions. Motivated by the field of product demand forecasting, this paper presents the dynamic count mixture model (DCMM) to flexibly and efficiently model time series of counts through a mixture of Bernoulli and Poisson dynamic generalized linear models. The sequential learning and forecasting of the DCMM allows fast, parallel analysis in high-dimensional forecasting frameworks. We extend the DCMM to overdispersed counts through the addition of a random effect in the conditionally Poisson model. Given the time constraints in product demand forecasting, many forecasters run univariate models independently across products. However, forecasting accuracy may be improved by pooling information across groups of related series when estimating trends or seasonal components. We present an efficient multiscale framework, which incorporates cross series linkages while insulating the parallel estimation of the DCMM. We apply this framework to a case study of supermarket products aimed at forecasting the daily demand 1-14 days in the future.

Indices in health related scientific fields

Fragkiskos G. Bersimis

Harokopio University of Athens, Greece

Scoring and combining into one single diagnostic tool various clinical or biological human attributes is an essential procedure for the settlement of effective prevention strategies, in several health domains. These diagnostic tools, i.e. health related indices are used for measuring cardiovascular and cancer disease risk, various metabolic disorders and infant mortality risk, etc., based on demographic, dietary or biochemical characteristics. Health related scientists aim to use accurate health-related indices assisting to the early identification of a future patient and therefore to the initiation of therapeutical treatment that may prolong a patient's life and improve its quality. These screening tools could assist to the organization of appropriate public or private health programs for the reliable prevention of different diseases and meet the most pressing human needs for basic health services. Additionally, health-related indices may be useful for better distribution of public or private financial resources in the health sector by using fewer basic medical supplies and thereby to minimize the medical expenses as regards the unnecessary repeated clinical examinations due to misdiagnosis. More specifically, in health related industry, using indices with high sensitivity and high specificity is of great importance due to the fact that the probability of misclassifying the truly diseased people from the untruly ones is minimized, and therefore, the corresponding economical costs are minimized.

Public Health Monitoring Using Scans and Control Charts

Sotiris Bersimis[1], Athanasios Sachlas[1] and Polychronis Economou[2]

[1]*Department of Statistics & Insurance Science, University of Piraeus, Greece*
[2]*Department of Civil Engineering, University of Patras, Greece*

If we want to timely and efficiently detect disease outbreaks we should take into account both spatial and temporal dimensions. In this case, of interest are global changes in the number of new disease events on time and/or hotspots of disease events which may evolve into outbreaks. One of the key assumptions in monitoring public health is that under normal conditions events are uniformly distributed in the plane and there are not too many changes over time. In this work, we propose a new public health monitoring procedure which combines scans and control charts.

A Bayesian self-starting Shiryaev statistic for Phase I data

Konstantinos Bourazas[1] and Panagiotis Tsiamyrtzis[2]

In Statistical Process Control (SPC) our interest is, in detecting when a process deteriorates from its “in control" state, typically established during a phase I exercise. Thus the phase I data play a very crucial role, as it is assumed to be a random sample from the in control distribution and are used to calibrate a control chart that will evaluate the process in phase II. In this work, we focus our attention on detecting persistent shifts in the parameters of interest during phase I, where low volume data are available. We propose a Bayesian scheme, which is based on the cumulative posterior probability that a step change has already occurred. The proposed methodology is a generalization of Shiryaev’s process, as it allows both the parameters and shift magnitude to be unknown. Furthermore, the Shiryaev’s assumption that the prior probability on the location of the change point is constant will be relaxed. Posterior inference for the unknown parameters and the location of a (potential) change point will be provided. A real data set will illustrate the Bayesian self-starting Shiryaev’s scheme, while a simulation study will evaluate its performance against standard competitors in the case of Normal data.

George Box: The “Accidental Statistician" who revolutionized time series analysis

Abraham Bovas

University of Waterloo, Waterloo, Canada

George Edward Pelham Box was born on October 19, 1919 in Gravesend, Kent, England and died on March 28, 2013 in Madison, Wisconsin. George Box was one of the world's most leading statisticians and made path breaking contributions to many areas of statistics including design of experiments, robustness, Bayesian methods, time series analysis and forecasting, and quality improvement. This talk discusses his contributions to time series analysis and forecasting. His work in this area started in collaboration with Gwilym Jenkins in the early 1960's and continued over the next several decades. His contributions include the classic and extraordinarily influential book “Time Series Analysis: Forecasting and Control" written with Gwilym Jenkins and first published by Holden Day in 1970. His subsequent contributions to time series analysis include joint work with George Tiao, Daniel Pena and many former graduate students including this one. His work provided a unified framework for carrying out time series analysis in practice and laid the foundation for many new developments in various areas including financial econometrics.

Variable selection for $C[0,1]$-valued AR processes using RKHS

Beatriz Bueno

Autonomous University of Madrid, Madrid, Spain

We propose an extension of the variable selection methodology introduced in Berrendero et al. (2017). We focus on prediction of functional time series whose realizations belong to $C[0,1]$, the space of continuous functions. Our basic tool is the RKHS (reproducing kernel Hilbert space) associated with the covariance function of the underlying stochastic process. Using the so-called Loeve's isometry in this RKHS, we define a new dependence model with an autoregressive structure. Then, establishing a natural optimality criterion, we can find the most relevant points for the prediction of the curves. Using this variable selection approach we can circumvent some standard problems, like the non-invertibility of the covariance operator. A simulation study, including real data examples, is presented in order to evaluate the performance of the proposal.

(Almost) Two Decades of Statistical Process Control Research: Some Reflections

Subhabrata Chakraborti

Department of Information Systems, Statistics and Management Science University of Alabama, USA

Research in Statistical Process Control (SPC) has made huge strides in a number of traditional and non-traditional areas of applications over the last several years. From manufacturing to healthcare to social networking, to cybercrime and environmental monitoring, to name a few, a wide landscape has been covered and new vistas have been discovered with far reaching impacts. To this end, control charts have been among the most useful tools that have drawn a lot of interest. They have been studied in both parametric and non-parametric settings, for univariate and multivariate problems, by researchers from around the world. The result has been a tremendous and sustained growth of the research literature. In this presentation, we will take a look back at the last couple of decades of SPC research and look forward to the future. Within this context, some of the key contributions and milestones will be highlighted and discussed, with a focus on the univariate case. Finally, time permitting, some guidance will be provided on how to do effective research (how to start, read and formulate ideas, work out details and write) and how best to navigate the steps to publishing research.

Optimal Repeated Measurements Two Treatment Designs for Dependent Observations: The Case of AR(1)

Miltiadis S. Chalikias

University of West Attica, Greece

In this paper optimal Repeated Measurement Designs of 2 treatment designs are constructed for estimating direct effects, the case of Autoregressive(1) dependency is examined. The model is presented and the design that minimizes the variance of the estimated difference of the two treatments is examined. The optimal designs with dependent observations in a AR(1) model are presented.

Penalized M-estimators for Logistic Regression

A. Bianco, G. Boente and Gonzalo Chebi
Universidad de Buenos Aires - CONICET

Logistic regression is a widely studied problem in the literature. In order to overcome the influence of outliers, Bianco and Yohai (1996) proposed a robust version of the maximum likelihood estimator. However, several problems arises when the data dimension is large with respect to the sample size. For instance, the estimation typically over fit the data, the minimization algorithms become unstable and the estimator is not well defined when the sample size is smaller than the dimension.

Nowadays, dealing with high-dimensional data is a recurrent problem that cuts across modern statistics. One main feature of high dimension data is that the dimension p (number of covariates) is high, while the sample size n is relatively small. A popular way to treat this problem is to assume sparsity on the regression coefficient vector, i.e., only a small number of regression coefficients are different from zero. Sparse covariates are frequent in the classification problem and in this situation the task of variable selection may be also of interest. For this purpose, we introduce a family of penalized M-type estimators for the logistic regression parameter that are stable against atypical data. We explore different penalizations functions and we introduce the so-called sign penalization. This new penalty has the advantage that it does not shrink the estimated coefficients to 0 and that it depends on only one parameter.

Theoretical results regarding oracle properties as well as consistency and asymptotic distribution results are studied. Through a numerical study, we compare the finite sample performance of the proposal with different penalized estimators either robust and classical in different scenarios including clean and contaminated samples.

Anomaly detection in time series using deep learning

Bei Chen
IBM Research, Ireland

Anomaly detection has wide applications including fraud detection, energy consumption monitoring, automated trading, image processing, quality control, etc. Although a large number of algorithms exist in the literature, timely and accurate detection of anomalies remains to be a challenge. This talk will present some results of the ongoing effort on detecting anomalies (data and contextual) using fast deep neural network algorithms, which includes building an anomaly simulator using probabilistic models, tuning deep neural networks for the detection of synthetic anomalies, and applying the algorithms to IOT use cases.

Some challenges with large data and large models in Medical Statistics

Nikolaos Demiris

Department of Statistics, Athens University of Economics and Business, Greece

In this presentation we will discuss a number of issues arising in statistical work driven be medical applications. In particular, special emphasis will be placed on the synthesis of diverse sources of evidence, potentially of different quality. The scalability of some popular algorithms will be extensively dicussed. The last part of this presentation is concerned with applications driven by Greek data on dynamic movement networks and infectious diseases.

Viability Assessment of Photovoltaic Systems using Bootstrap-based Tolerance Intervals

Jani Deyzel, Chantelle Clohessy, Warren Brettenny, Ernest van Dyk

Department of Statistics, Nelson Mandela University, South Africa

Viability assessments are used by policy makers and investors to determine the profitability of renewable energy installations. The predicted energy output of these systems is crucial for these assessments. This study demonstrates the use of bootstrap-based tolerance intervals for the energy output of photovoltaic (PV) systems. In particular the β-expectation and (α, β) two-sided tolerance intervals are used to provide further insight into future energy outputs of the PV system. Decision-makers, policy-makers, as well as investors, will be able to make more informed decisions with the proposed assessment method.

Guaranteed In-Control Performance of the EWMA $\bar{X}$ Chart

Mandla D. Diko[1], Subha Chakraborti[1] and Ronald J.M.M. Does[1]

[1]*Department of Operations Management, University of Amsterdam, Amsterdam, The Netherlands*
[2]*Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, Alabama, USA*

Research on the performance evaluation and the design of the Phase II EWMA $\bar{X}$ chart have mainly focused on the marginal (unconditional) in-control average run-length (ARL_{IN}). Recent research in this area, when parameters are estimated, has emphasized the lack of in-control performance due to practitioner to practitioner variability and has advocated the study of the conditional in-control average run-length ($CARL_{IN}$) distribution. We study the performance and design of the EWMA $\bar{X}$ chart using the $CARL_{IN}$ distribution and the exceedance probability criterion (EPC). The CARLIN distribution is approximated by Markov Chain theory and Monte Carlo simulations. Our results indicate that, in-order to design charts that guarantee a specified EPC, more Phase I data are needed than previously recommended in the literature. A method for adjusting the Phase II EWMA $\bar{X}$

chart control limits, to achieve a specified EPC, is presented. This method does not involve bootstrapping, but produces results that are the same as some known analytical results. Tables of the new charting constants are provided. A thorough examination of the in-control and out-of-control performance of these constants is presented. Results show that, for moderate to large shifts, the performance of the new constants is quite satisfactory.

Age Groups Demographics Based on Entropic Analytics

Yiannis Dimotikalis[1] and Christos H. Skiadas[2]

[1]*Department of Business Administration, T.E.I. of Crete, Greece*
[2]*ManLab, Technical University of Crete, Greece*

In Business statistics populations usually divided in specific age groups to perform frequencies analysis. In this paper the commonly used 5 age groups: -17,18-30,31-45,46-65,65+, analyzed using the max entropy approach for several countries, using available data of country populations at Human Mortality Database (mortality.org). The used data includes population estimates starting in 19th century for several European countries. The max entropy principle used by defining a normalized entropy chart for the 5 age groups. The time evolution of age groups distribution seems to be very close to max entropy limit of Exponential (Geometric) distribution, tends to approach and pass the absolute max entropy point of Uniform distribution. The closeness to max entropy limit measured by Kullback-Leibler divergence from max entropy unconstrained Uniform (Theil inequality index) and constrained Geometric distributions. The observed differences of male and female population are investigated in detail. Also, some “crisis" time periods in the population data are pointed out and explained by political, social and health conditions. Concluding remarks and implications of those findings to human populations discussed, and some future tasks suggested.

Assessing the macroeconomic and budgetary impact of an ageing population in the EU

Giuseppe Carone and Per Eckefeldt

European Commission, Brussels, Belgium

The presentation is based on the Ageing report 2018 (to be released by the European Commission and the Economic policy Committee in May). The report present a comprehensive assessment of the macroeconomic (on growth, productivity, employment) and budgetary (on pension, health care, long-term care and education spending) impact of an ageing population over the period 2018-2070. The presentation will focus on the main methodological aspects of the long-run projections and the main outcome. The projected low birth rates, rising life expectancy and continuing inflow of migrants will result in an almost unchanged, but much older, total EU population by 2070. This means that the EU would move from having four persons of working-age (aged 15-64) for every person aged 65 or more in 2010 to a ratio of only two to one.

Potential economic growth is likely to be much lower than experienced in previous decades, and the need for public provision of age-related transfers and services will increase. On the one hand, we should take comfort by the fact that there has been considerable progress with structural reforms, notably in the field of pensions. The improvements are already visible, for instance employment rates have risen on account of pension reforms, especially among older workers.

On the other hand, the fiscal impact of ageing is still being projected to be substantial in almost all Member States, becoming apparent already over the course of the next decade. Overall, on the basis of current policies, age-related public expenditure is projected to increase considerably over the coming two decades, and also in a longer - term perspective in the EU and EA - especially through pension, health care and long-term care spending, but there are notable differences across Member States.

The demographic challenge affects not only pensions. Budgetary pressures on care-related expenditure - health care and long-term care - are also likely to increase. In this context, there will be a need to discuss how health care services should be organized in the future; an ageing population will exercise pressures for higher spending on health care. On top of this, technological progress is likely to push up costs further. It will therefore be important to ensure that these services are provided in an efficient manner. It will also prompt a discussion on the accessibility and financing of health care services. More broadly, in reforming of welfare systems, there is no one-size-fits-all solution for the EU Member States. Different countries need to find different solutions.

Simultaneous Monitoring the Number and the Spatial Distribution of Disease Events

Sotiris Bersimis[1], Athanasios Sachlas[1] and Polychronis Economou[2]

[1]*Department of Statistics & Insurance Science, University of Piraeus, Greece*
[2]*Department of Civil Engineering, University of Patras, Greece*

Early and efficient detection of any disease outbreaks requires the simultaneously monitoring of the Number and the Spatial Distribution of Disease Events. Any outbreak evolves changes in the number of new disease events and/or in their spatial distribution. The last may reflect the presence of one or more hotspots of disease events. A key assumption in biosurveilance is that under normal conditions events are uniformly distributed in the plane. In the present work, a new two-step monitoring procedure is proposed to monitor firstly the number of disease events through control charting and secondly the spatial distribution of disease events via convex hulls. Simulations results showed a remarkable performance of the new test under different outbreaks scenarios.

Bayesian Modeling of Virtual Age in Repairable Systems

Didem Egemen[1], Fabrizio Ruggeri[2] and Refik Soyer[1]

[1]*The George Washington University, Washington D.C., USA*
[2]*CNR IMATI, Milano, Italy*

In this study, repairable system models, which are subject to minimal, perfect or imperfect repairs upon each failure, are discussed and a unifying model covering all these type of models is presented. Moreover, some extensions of this general model are proposed. These models are generally marked point processes, $(T_1, Z_1), (T_2, Z_2), \ldots, (T_n, Z_n)$ where T_i's are failure times and Z_i's are repair choices. The marks of this marked point process, i.e. repair actions, are assumed to be unknown and unobservable so modeled as latent variables. According to the dependence structure of the latent variables various models are developed. For the statistical analysis of these models, Bayesian framework is presented and posterior distributions are obtained through Markov Chain Monte Carlo methods.

Topic Model based Prescription Fraud and Abuse Detection

Tahir Ekin[1] and Babak Zafari[2]

McCoy College of Business, Texas State University, San Marcos, Texas, USA
Babson College, Massachusetts, USA

Prescription fraud and abuse has been a pressing issue in the U.S. resulting in large financial losses and adverse effects on human health. The size and complexity of the healthcare systems as well as the cost of medical audits make use of statistical methods necessary to generate investigative leads in prescription audits. This paper proposes the use of topic models to analyze prescription data with an emphasis on opioid abuse. The proposed method can be used to group drugs with respect to the billing patterns, to display associations and to exhibit the potential outlier behavior of provider-drug pairs. In addition, overall prescription patterns of the providers are summarized using the distance based measure of Lorenz curve. The use of proposed method is illustrated by using real world Medicare Part D prescription data. It can enable medical auditors to identify leads for audits of providers prescribing excessive or medically unnecessary drugs.

Dynamic Continuous Time Models with Discrete Observation Paths of Different Frequencies with Application to Financial Models

Kathy Ensor

Department of Statistics, Rice University, Houston, USA

Does news arrival impact oil futures returns and volatility? Is the impact of negative and positive news on returns and volatility symmetric? This paper answers these questions. We show in a model free environment that CME oil futures returns are affected by news arrival. The impact of negative and positive news information has an asymmetric impact on returns and volatility. We find that the number of news items arriving per day varies over time and the arrival processes for negative and positive news are different. Persistence in the volatility of oil returns is affected by news arrival; furthermore, negative and positive news have different explanatory power on volatility clustering.

We incorporate these findings into a reduced form model to price oil futures that recognizes the asymmetric impact of negative and positive news. Empirical results show that the effects of negative and positive news are described by different processes. We find that a significant proportion of volatility can be explained by news arrival and that the impact of negative news is larger than that of positive news.

Ensuring both the in-control and out-of-control Phase II performances of the S2 control chart with estimated variance

Francisco Aparisi, Jaime Mosquera and Eugenio K. Epprecht

Pontifical Catholic University of Rio de Janeiro, Brazil

In the last three years there has been a surge of papers studying the effect of parameter estimation (in Phase I) on the Phase II performance of process control charts from the perspective of the conditional distribution of some measures of in-control (IC) performance (such as the false-alarm rate or the IC ARL). Although this perspective of conditional performance had been pioneered by a few authors around the early 2000's, only now there has been a resurgence of interest in this perspective. The majority of recent authors have used, as a criterion to determine the minimum number of Phase I samples that guarantee a desired in-control performance, the exceedance probability: a concept that (although somewhat more general than this) can be defined for example as the probability that the conditional IC ARL is not smaller than the minimum value tolerated by the chart's user. These authors calculated the minimum numbers of Phase I samples that guarantee a high (specified) value (say, 95%) for this exceedance probability.

The uncomfortable common conclusion of these works was that in many cases the required number of Phase I samples is prohibitively high and not available in typical applications. As a response to this issue, many authors proposed and calculated adjustments to the charts' control limits with the purpose of ensuring the desired exceedance probability with a practicable number of Phase I samples. The focus however has been only on the IC performance. Depending on the numbers of Phase I samples, these adjustments may result in an unacceptable deterioration of the chart's sensitivity to relevant shifts in the process parameters, provoking a substantial increase in the out-of-control (OOC) ARLs.

The user is then faced with a hard decision problem: how to balance the requirements of feasible number of Phase I samples (and duration of Phase I), and desired IC and OOC performances, all at the same time? Only some of the works have examined the effects of the adjustments on the OOC performance, and calculated the deterioration of the OOC ARLs in some cases, but this does not make easier the problem for the practitioner, who cannot easily assess the effect of varying the level of his/her requirement on the number of Phase I samples, and tolerances about the IC and about the OOC ARLs, in order to find a good compromise solution. With this in mind, we propose a different approach, where the user's requirements on the IC and OOC performances are used as active constraints for an optimization problem with control limit factors and the number of Phase I samples as decision variables. The solution gives the adjusted control limits and the minimum number of Phase I samples that guarantee simultaneously the IC and OOC performances. The solution found is the most efficient possible, and if the number of samples found is still not feasible, the user should relax some of the requirements on the IC and OOC performances, and re-optimize. We illustrate the approach with the Shewhart $S2$ chart, but a paper considering the X-bar chart has already been submitted and another one, on the joint X-bar and S charts, is in an advanced stage.

A Bayesian Model for Forecasting Hierarchically Structured Time Series

Beatriz (Stefa) Etchegaray Garcia

IBM Thomas J. Watson Research Center, USA

This talk will focus on the problem of statistical forecasting of hierarchically structured data within a Bayesian framework. We develop a novel approach to hierarchical forecasting that provides an organization with optimal forecasts that reflect their preferred levels of accuracy while maintaining the proper additive structure of the business. We start out by assuming that there is a method in place to obtain forecasts for each node of the hierarchy. We treat the forecasts initially provided as observed data. In order to obtain the desired relative accuracy prescribed by the user, we use the past accuracy of the forecasts at all levels to determine which ones to trust depending on their historical ability to predict. We use a context specific heterogeneous loss function to penalize errors differently depending on the context and which levels we care about most in terms of accuracy. To illustrate our method's ability to improve forecasts, we compare our method to competitor methods. We include two sets of studies: one with simulated data and one with real data.

Optimal mission duration for non-repairable and partially repairable systems

Maxim Finkelstein

Department of Mathematical Statistics, University of the Free State, South Africa

As a system failure during a mission can result in considerable penalties, at some instances it is more cost-effective to terminate operation of a system than to attempt to complete its mission. This paper analyzes the optimal mission duration for systems that operate in a random environment modeled by a Poisson shock process and can be minimally repaired during a mission. Two independent sources of failures are considered and for both cases, the failures are classified as minor or terminal in accordance with the Brown-Proschan model. Under certain assumptions, an optimal time of mission termination is obtained. It is shown that, if for some reason a termination is not technically possible at this optimal time, the mission should be terminated within a specific time interval and, if this is not possible, it should not be terminated beyond this interval. Illustrative examples are presented. The influence of mission and system parameters on the mission termination interval is demonstrated.

The Good, the Bad, and the Horrible: Interpreting Net-Promoter Score and the Safety Attitudes Questionnaire in the light of good market research practice

Nicholas I. Fisher

University of Sydney and ValueMetrics, Australia

Net-Promoter Score (NPS) is a ubiquitous, easily-collected market research metric, having displaced many complete market research processes. Unfortunately, this has been its sole success. It possesses few, if any, of the characteristics that might be regarded as highly desirable in a high-level market research metric; on the contrary, it's done considerable damage to companies, to their shareholders and to their customers. Given the current focus on the financial services sector and its systemic failures in delivering value to customers, it is high time to question reliance on NPS.

The Safety Attitudes Questionnaire is an instrument for assessing Safety Culture in the workplace, and is similarly wide-spread throughout industries where Safety is a critical issue. It has now been adapted to assess other forms of culture, such as Risk Culture. Unfortunately, it is also highly flawed, albeit for quite different reasons.

Examining these two methodologies through the lens of good market research practice brings their fundamental flaws into focus.

Properties of the Multivariate Generalized Hyperbolic Laws

Stergios Fotopoulos

Department of Finance and Management Science, Washington State University, USA

The purpose of this study is to characterize multivariate generalized hyperbolic (MGH) distributions and their conditionals by considering the MGH as a subclass of the mean-variance mixing of the multivariate normal law. The essential contribution here lies in expressing multivariate generalized hyperbolic densities by utilizing various integral representations of the Bessel function. Moreover, in a more convenient form these modified density representations are more advantageous for deriving limiting results. The forms are also convenient for studying the transient as well as tail behavior of multivariate generalized hyperbolic distributions. The results include the normal distribution as a limiting form for the MGH distribution. This means the MGH model can be considered for modeling not only high frequency data but also for modeling low frequency data. This is against the currently prevailing notion that the MGH model is relevant for modeling only high frequency data.

Overlapping Clustering: A Framework, Software, and Empirical Analysis

Stephen France

College of Business, Mississippi State University, USA

Overlapping clustering is a variant of clustering where each item may be a member of more than one cluster. It has found particular use in marketing segmentation, where products may be members of more than one usage segment. Overlapping clustering methods have been developed from different clustering traditions. Additive decomposition methods, such as ADCLUS and INDCLUS, are discrete variants of continuous mapping methods. Fuzzy clustering methods can generate overlapping clustering solutions by setting thresholds for cluster membership. Partitioning clustering methods, such as k-means clustering, can be extended to overlapping clustering by relaxing the cluster membership constraints.

The R software package and associated framework described in this talk implements overlapping clustering methods from all of these traditions and also implements the generalized omega metric for cluster validation. Attention is paid to the optimization of the models and to the issues of solution initialization and locally optimal solutions. Empirical work on both synthetic and real world datasets is described. Applications are given in customer and product segmentation.

Comparison of Non-parametric Approaches in Classification of Seeds

Luca Frigau

Department of Economic and Business Science, University of Cagliari, Italy

Nowadays, the use of digital image analysis for classification of seeds is essential in taxonomic studies. It allows to overcome the problems linked to manual classification, inasmuch it is labor-intensive, subjective, and suffers from inconsistencies and errors. A lot of information can be extracted from images such as size, texture and shape of seeds. Consequently, different statistical approaches can be applied depending on the variables involved in the analysis. In this paper, we compare different kind of non-parametric algorithms, including Sequential Automatic Search of Subsets of Classifiers (SASSC) and other approaches prevalently based on tree-based models. We search for the best combination of size, texture and shape of seeds to classify among 19 varieties of Sardinian and 4 varieties of international Prunus domestica.

Assessing the quality of health services using composite indicators

Christina Georgakopoulou[1], Theodoros Paschalis[1], Spyros Goulas[1], Sotiris Bersimis[1,2], Athanasios Sachlas[2] and Vassilis Plagiannakos[3]

[1]*National Organization for Provision of Health Services (NOPHS), Greece*
[2]*Department of Statistics and Insurance Science, University of Piraeus, Greece*
[3]*Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece*

The National Organization for the Provision of Healthcare Services (NOPHS) is the largest healthcare service provider in Greece. NOPHS participates as a partner in the European program CrowdHEALTH, which aims to create an integrated Information and Communication Technology (ICT) platform that incorporates advanced statistical analysis techniques for supporting decision-making through the exploitation of collective knowledge derived from multiple heterogeneous sources. In this presentation, we will present the mechanism that transforms information to audit in the framework of the program, in order to assess the health services provided by NOPHS.

Screening designs based on weighing matrices with added two-level categorical factors

Stelios Georgiou

RMIT University, Melbourne, Australia

Screening designs are used widely by industrial experimentation. Also, they are used to identify the most potential factors among a large number of factors. In a recent paper, Jones and Nachtsheim (2011) proposed a new class of design called definitive screening design (DSDs). Definitive screening designs offer a number of advantages over standard screening designs. They avoid the fully confounding of effects and can identify factors having a nonlinear effect on the response. These designs have very nice statistical properties and the easiest method to be constructed is by using weighing matrices. A limitation of these designs is that all factors must be quantitative. In this paper, we provide a new method that can transform those designs to accommodate some two level qualitative factors.

Bayesian Modeling of Abandonments in Ticket Queues

Christopher Glynn

Peter T. Paul College of Business and Economics, University of New Hampshire, USA

Text data is routinely collected by businesses, but it is often not fully utilized because extracting insight from documents collected over time is a challenging statistical learning problem. Dynamic topic models offer a probabilistic modeling framework to decompose a corpus of text documents into “topics", i.e., probability distributions over vocabulary terms, while simultaneously learning the temporal dynamics of the relative prevalence of these topics. We extend the dynamic topic model of Blei & Lafferty (2006) by fusing its multinomial factor model on topics with dynamic linear models that account for time trends and seasonality in topic prevalence. A Markov chain Monte Carlo (MCMC) algorithm that utilizes Polya-Gamma data augmentation is developed for posterior sampling, and we present an applied analysis of real estate listings from the housing website Zillow. Analysis of the Zillow corpus demonstrates that the method is able to learn seasonal patterns and locally linear trends in topic prevalence, providing new insight into dynamics of real estate markets.

Comparison of Several Samplers in a Stochastic EM Algorithm for Modeling Imperfect Maintenance Actions

Min Gong

Department of systems engineering and engineering management, City University of Hong Kong, China

Maintenance policy changes the lifetime behavior of a system by introducing one or more factors into the underlying model, named by improvement factors. Due to many uncontrollable reasons, these factors should be treated as random. As a result, these random factors lead to great difficulty in parameter estimation. EM algorithm is a standard approach to estimate parameter for model with hidden variables. We compute the expected value of the complete log-likelihood function with respect to the distribution of the hidden variable in the E step, and then maximize it in the M step. The maximum likelihood estimate of parameters can be found by iterating these two steps.

However, traditional EM algorithm has the pitfall that the rate of convergence can be painfully slow. To improve the behavior of the Monte Carlo EM algorithm, we here employ the idea of stochastic EM algorithm. Because the expectation in E step is an intractable integral, for certain high-dimensional problems, MCMC simulation is the only known technique capable of providing a solution within a reasonable time. We herein develop some sampling methods as implementation of MCMC simulation, such as Metropolis-Hastings sampler or approximate Bayesian computation method. Examples will be provided to check their feasibility and make comparison.

Row-Column Screening Designs

Peter Goos[1], Eric Schoen[1] and Nha Vo-Thanh[2]

[1]*KU Leuven*

[2]*University of Hohenheim*

Many experiments span multiple days, use material from several batches and/or involve more than one operator. In such scenarios, blocking the experiment is important. Many textbooks discuss how experiments should be blocked when there is a single blocking factor. It is, however, not uncommon to have more than one blocking factor in an experiment. In this talk, we discuss the problem of designing screening experiments involving two crossed blocking factors. The required experimental designs in the presence of two crossed blocking factors are generally named row-column designs.

We show how integer linear programing can be used to arrange any given two-level orthogonal screening design in rows and columns, so that the main effects can be estimated independently from the block effects, and so that as many two-factor interaction effects are estimable as possible.

The motivating examples for the talk are a 24-run and a 28-run screening experiment performed by a car tire manufacturer to study the impact of 12 two-level factors on the wear of tires. Since only a limited number of experimental runs can be performed per day, and since several drivers are used for the experiment, the experiments involve two crossed blocking factors.

On the mean time to failure of an age replacement model

Asha Gopalakrishnan

Department of Statistics, Cochin University of Science and Technology

The mean time to failure ($MTTF$)of a preventive maintenance model under age replacement at intervals T is denoted by $M(T)$ and defined as

$$M(T) = \frac{1}{F(t)} \int_0^T R(x)dx, \;\; T > 0$$

where $R(x) = 1 - F(x) = P(X > x)$ is the reliability function of a system with continuous lifetime. Properties of the $MTTF(M(T))$ in a preventive maintenance model under age replacement, including the fact that $M(T)$ characterizes the baseline survival function is studied extensively. Several implications of the behavior of $M(T)$ on the ageing aspects of the model are also investigated. A new class of distributions namely the Decreasing Mean Time to Failure class ($DMTTF$) is introduced and its members are characterized.A new stochastic order with respect to $DMTTF$ is also introduced and its implications with several stochastic orders present in literature are studied. The proposed model is modified for discrete time. Finally we apply a characterization developed to propose a non-parametric test for testing constant $MTTF$ against decreasing mean time to failure.

Applications of composite indicators for the longitudinal assessment of the quality of health services

Spyros Goulas[1], Theodoros Paschalis[1], Christina Georgakopoulou[1], Sotiris Bersimis[1,2], Athanasios Sachlas[2] and Vassilis Plagiannakos[3]

[1]*National Organization for Provision of Health Services (NOPHS), Greece*
[2]*Department of Statistics and Insurance Science, University of Piraeus,Greece*
[3]*Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece*

One of the goals of the National Organization for Provision of Health Services (NOPHS) is to evaluate the quality of health services that it provides over time, utilizing the collective knowledge derived from multiple heterogeneous sources. In this presentation, we will present the application of advanced statistical analysis techniques (e.g. Statistical Process Control techniques, control charts, etc.) in the database of NOPHS. The results enable NOPHS to make the best possible decisions.

Evaluation of selection tools for the inefficiency distribution in stochastic frontier models

Aviwe Gqwaka, Warren Brettenny and Gary Sharp

Department of Statistics, Nelson Mandela University, South Africa

Implementations of efficiency analyses span a variety of fields, where notable attention has been directed towards the business and manufacturing sector. Evaluations into business performances provide insights into the ability to frugally manage resources when providing services, while an economic objective (e.g. cost or profit) is targeted. In addition, peer-to-peer comparisons may be possible, allowing for the sharing of best practices. As performance metrics, efficiency measures are important to business evaluations. Stochastic frontier analysis (SFA) is often used to assess these efficiencies. This method envelops observed performances using some ideal frontier (or theoretical best practice) on which an estimable functional form is imposed. Direct comparison with business performances is thus possible where any deviation from the frontier implies that, to some degree, firms employ inefficient practices. Unique to SFA, these deviations are modelled by a composed error measure. This composed error consists of a random shock term and an inefficiency term, on which the parametric nature of the SFA method allows for the imposition of distributions. Owing to the consistent assumption of normality for the random shock term, the study solely focuses on the construction and estimation of the inefficiency term. In particular, the study investigates whether the choice of distribution to model inefficiency matters or not. The study argues that, since the true distribution of inefficiency is unknown, investigations into the selection of the most appropriate distribution to model inefficiency, after estimation, should be undertaken. This will further give credence to the suitability of the distribution(s) chosen. Investigations into the suitability of various distributions to model inefficiency are undertaken using two non-nested model selection tools as developed by Vuong (1989) and Clarke (2003). The ability of these tests to discriminate between distributions is investigated through Monte Carlo simulations.

As success indicators, efficiency measures play an important role in evaluating business performance. Assessing the appropriateness of the distribution to model these efficiencies is thus validated, especially when implications on efficiency estimates are considered.

Nonparametric precedence control charts with improved runs-rules

Marien A. Graham

Department of Science, Mathematics and Technology Education, University of Pretoria, South Africa

Nonparametric or distribution-free control charts are highly desirable since a minimal set of modeling assumptions are necessary for their implementation. Chakraborti, Van der Laan and Van de Wiel (2004) proposed a class of nonparametric Shewhart-type control charts, referred to as the precedence charts, using some order statistic of a Phase II sample as the charting statistic with the control limits constructed using reference sample from Phase I. Albers and Kallenberg (2008) proposed a comparable nonparametric Shewhart-type control chart where the charting statistic is the minimum of a Phase II sample. A

comparison between the precedence and minimum charts is done through simulation. Subsequently, their performance is improved by adding some runs-rules. A summary and some concluding remarks are given.

Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data

Scott Holan[1], Jonathan R. Bradley[1] and Christopher K. Wikle [2]

[1]*Florida State University*
[2]*University of Missouri*

We introduce a computationally efficient Bayesian model for predicting high-dimensional dependent count-valued data. In this setting, the Poisson data model with a latent Gaussian process model has become the de facto model. However, this model can be difficult to use in high dimensional settings, where the data may be tabulated over different variables, geographic regions, and times. These computational difficulties are further exacerbated by acknowledging that count-valued data are naturally non-Gaussian. Thus, many of the current approaches, in Bayesian inference, require one to carefully calibrate a Markov chain Monte Carlo (MCMC) technique. We avoid MCMC methods that require tuning by developing a new conjugate multivariate distribution. Specifically, we introduce a multivariate log-gamma distribution and provide substantial methodological development of independent interest including: results regarding conditional distributions, marginal distributions, an asymptotic relationship with the multivariate normal distribution, and full-conditional distributions for a Gibbs sampler. To incorporate dependence between variables, regions, and time points, a multivariate spatio-temporal mixed effects model (MSTM) is used. To demonstrate our methodology we use data obtained from the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. In particular, our approach is motivated by the LEHD's Quarterly Workforce Indicators (QWIs), which constitute current estimates of important US economic variables.

Data analysis challenges in analysing users sessions from an e-Commerce platform

David Hoyle

Autotrader, UK

Modern e-commerce platforms, such as that of AutoTrader UK, typically offer a variety of ways for users to access the platform, e.g. via desktop websites, mobile websites or native Apps. At AutoTrader UK we provide a platform for the buying and selling of vehicles. This generates a wealth of user data that enables us to tailor and improve the user-experience. Typical data analysis challenges that this opportunity raises include, identifying similar user-sessions for the purposes of segmenting or clustering users, predicting whether a user will perform a particular action, or predicting and displaying the content the user is most likely to find useful. The variety of user touch-points and the need to stay commercially competitive means our platform is constantly and rapidly evolving.

This can introduce additional data analysis challenges such as testing whether platform changes have genuinely improved the business performance, and also diagnosing causes when the platform performance is negatively impacted. In this talk we will explain some of the approaches we have adopted at AutoTrader UK to tackle these data analysis challenges. We will highlight the approaches we believe have been successful, as well as highlighting some of the problems that remain.

The Performance of $\bar{X}$ Control Charts for Large, Non-Normally Distributed Datasets

Leo C.E. Huberts, Marit Schoonhoven, Rob Goedhart, Mandla D. Diko and Ronald J.M.M. Does

Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Amsterdam, The Netherlands

Due to digitalization, many organizations possess large datasets. Furthermore, measurement data are often not normally distributed. However, when samples are sufficiently large, the Central Limit Theorem may be used for the sample means. In this article, we evaluate the use of the Central Limit Theorem for various distributions and sample sizes, as well as its effects on the performance of a Shewhart control chart for these large non-normally distributed datasets. To this end, we use the sample means as individual observations and a Shewhart control chart for individual observations in order to monitor processes. We study the unconditional performance, expressed as the expectation of the in-control Average Run Length (ARL), as well as the conditional performance, expressed as the probability that the control chart based on estimated parameters will have a lower in-control ARL than a specified desired in-control ARL. We use recently developed factors to correct the control limits in order to obtain a specified conditional or unconditional in-control performance. The results in this paper indicate that the $\bar{X}$ control chart should be applied with caution, even with large sample sizes.

Comparison of the Hybrid Artificial Intelligence Techniques for Credit Scoring

Damla Ilter[1] and Ozan Kocadagli[2]

[1]*Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey*
[2]*Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey*

The credit scoring is one of the major activities in the financial sector. Because of growing market and increasing the loan applications, the researchers in this field still continue its concern in terms of rating the applicants and assessing the credit amounts. To reduce the number of wrong decisions when financing the small and medium-sized enterprises (SMEs) as well as the individual customers, the decision makers focus on estimating more robust models. However, the traditional statistical methods are criticized due to various pre-requisites and linear approximations in the high dimensional and excessive nonlinear cases. In addition, the increase of data flow rate and dimension in the financial sector

prompts the researchers to improve the automatic decision support systems. For this reason, the hybrid artificial intelligence techniques play important roles to handle the credit scoring problems together with technological improvements. This study presents an efficient procedure based the hybrid artificial intelligence techniques in the context of the credit scoring for SMEs. In the analysis, the performance of various techniques is compared with each other over the credit dataset of SMEs. According to the analysis results, the proposed procedure not only allows making an efficient analysis of credit scoring for SMEs, but also provides the best model configurations in the context of reliability and complexity.

Weighted EWMA Charts for Monitoring Type I Censored Weibull Lifetimes

Shangjie Xu and Daniel R. Jeske

Department of Statistics, University of California, USA

In this paper, we first propose a control chart for monitoring the Weibull scale parameter in the context of Type I censored data, assuming the shape parameter is fixed. The proposed chart is a Shewhart-type chart based on a likelihood ratio test (LRT) that utilizes an exponentially weighted moving average of the log-likelihood. In previous literature, this type of chart is called a weighted exponentially weighted moving average (WEWMA) chart. The WEWMA chart is compared with a more standard EWMA chart and a CUSUM chart which were recently studied as alternative solutions to the monitoring problem. Numerical results show that the WEWMA chart often performs better than these two alternatives. Sensitivity studies show that the WEWMA chart is more robust to variations in batch sizes and censoring times. A rustresistant example is used to illustrate the proposed WEWMA chart. We extend the WEWMA and the CUSUM charts to the context where the shape parameter also needs to be monitored. Joint charts that simultaneously look for change in either parameter are proposed.

Combinatorial testing: an adaptation of design of experiments

Raghu N. Kacker[1], D. Richard Kuhn[1], Yu Lei[2] and Dimitris E. Simos[3]

[1]*National Institute of Standards and Technology, Gaithersburg, USA*

[1]*University of Texas, Arlington, USA*

[1]*SBA-Research, Vienna, Austria*

Combinatorial t-way testing (CT) is like a fractional-factorial design of experiment (DoE) method where a system is run for a suite of input test cases (DoE plan) and the output (response) data are used to make inference about the system. In both cases, combinatorial mathematics is used to define the suite of test cases. A DoE plan (suite of test cases) is developed for efficient estimation of the parameters of the statistical model used for inference. CT is a method to identify a combination of the test values of any t-variables (a small number) out of k variables (a large number) for which a software based system does not perform as expected. In CT, combinatorial mathematics is used to determine

a suite of test cases that covers all t-way combinations of the test values of k variables with a minimal number of test cases. For each test case, the expected (correct) observable performance of the system is pre-determined. The system passes a test case if the actual performance agrees with the expected. The pass/fail data for the suite of test cases is used to search for the t-way combinations that caused the system to fail. Generally, combinatorial test cases are subject to physical and logical constraints which render some combinations of test setting invalid. Such invalid combinations must be excluded from the suite of test cases. Investigations of actual software failures showed that pairwise (2-way) testing was useful but not always sufficient and combinatorial t-way testing for t greater than 2 is almost always needed. Combinatorial t-way testing has become practical because efficient and free downloadable tools for generating test suites with support of constraints have become available. So far, most, CT applications have focused on detecting general software bugs. However, interest in using CT to trigger security vulnerabilities has greatly increased.

Projection predictive variable selection for ARMA models

Ioannis Kamarianakis

School of Mathematics & Statistical Sciences, College of Liberal Arts and Sciences, Arizona State University, USA

Iterative subset selection in autoregressive moving-average (ARMA) modeling can be computationally intensive when the true ARMA orders are high. To alleviate this issue, this work identifies optimal ARMA models by using shrinkage estimates and latent innovation terms. Three shrinkage estimators are evaluated in detail in a series of Monte Carlo experiments and an application: a) The projection predictive variable selection method, combined with Horseshoe priors; b) adaptive LASSO and c) adaptive elastic net.

Latent Gaussian Count Time Series Modelling

Stefanos Kechagias, Yisu Jia, James Livsey, Robert Lund and Vladas Pipiras

SAS Institute, USA

This work proposes a novel modeling approach for stationary count time series data. A new model is constructed via a latent Gaussian process and a copula-type transformation, leading to a series with very flexible autocovariance structures, including long memory, that can have virtually any pre-specified marginal distribution, for example, the classical Poisson, generalized Poisson, negative binomial, and binomial count structures. Two estimation methods are considered: a least squares approach based on calculating the autocovariance function of a stationary count series using Hermite expansions, and a particle filtering approach approximating the full likelihood of the model, which extends some of the usual hidden Markov techniques to stationary processes. The performance of both approaches is demonstrated through simulations, and several real count time series are also analyzed, with the results compared to existing count time series modeling techniques found in the literature.

On information quality and customer surveys

Ron S. Kenett

KPA Ltd., Raanana, Israel
Samuel Neman Institute, Technion, Israel

Customer surveys rely on structured questions used to reflect on reality. Surveys, in general, rely on sample observations derived from a population frame that can be statistically analyzed. Eventually, a survey is judged by the quality of the information it provides. The talk will present how to apply the information quality (InfoQ) framework to discuss various aspects of customer survey design, deployment and analysis. InfoQ involves 8 dimensions: 1) Data resolution, 2) Data structure, 3) Data integration, 4) Temporal relevance, 5) Generalizability, 6) Chronology of data and goal, 7) Operationalization and 8) Communication. The goal is to provide with a customer survey, by properly addressing these dimensions, high InfoQ. Various models like CUB, Bayesian networks, Non-linear PCA and decision trees used in analyzing customer surveys will be presented from an InfoQ perspective.

Calibrating EWMA control charts for dispersion in presence of parameter uncertainty

Sven Knoth

Institute of Mathematics and Statistics, Department of Economics and Social Sciences, Helmut Schmidt University Hamburg, Germany

Most of the literature concerned with the design of control charts relies on perfect knowledge of the distribution for at least the good (so-called in-control) process. For instance, in order to monitor the variance of a normally distributed r.v., one usually assumes that the in-control variance σ_0^2 is known. Given the data is sampled in subgroups of size n, the EWMA control chart framework is given by:
$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2, \bar{X}_i = \sum_{j=1}^{n} X_{ij}$,
$Z_0 = z_0 = \sigma_0^2 = 1$, $Z_i = (1-\lambda)Z_{i-1} + \lambda S_i^2$ with $\lambda \in (0;1]$,
$L_{upper} = min\{i \geq 1 : Z_i > c_u\}$,
$L_{two} = min\{i \geq 1 : Z_i > c_u \text{ or } Z_i < c_l\}$.

The parameters $\lambda \in (0;1]$ and $c_u; c_l > 0$ are chosen to enable a certain useful detection performance (not too much false alarms and quick detection of changes). The most popular performance measure is the so-called Average Run Length (ARL), that is $E_{\sigma^2}(L)$ for the true variance σ^2. If the in-control variance, σ_0^2, has to be estimated by sampling data during a pre-run phase, then this uncertain estimate effects, of course, the behavior of the applied control chart. For example, the resulting in-control ARL becomes a random variable. Most of the papers about characterizing the uncertainty impact for dispersion control charts focus on the expected ARL. Some introduce a lower bound that is gone below with a small, pre-defined probability only. Here, we want to discuss two possible and presumably applicable methods of setting up exponentially weighted moving average (EWMA) S^2 charts that control the false alarm risk in a neat way.

Classification of EEG Signals for Epileptic Seizures using Hybrid Artificial Neural Networks based Wavelet Transforms and Fuzzy Relations

Ozan Kocadagli

Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey

This study presents an efficient procedure that provides an accurate classification of Electroencephalogram (EEG) signals for early detection of epileptic seizures. Essentially, this procedure hybridizes many tools such as artificial neural networks (ANNs), gradient based algorithms, genetic algorithms (GAs), discrete wavelet transforms (DWT) and fuzzy relations. In analysis, ANNs are trained by the gradient based algorithms and GAs over a benchmark data sets considering early stopping, cross-validation and information criteria. In order to ensure an accurate classification performance, the automated multiresolution signal processing technique splits EEG signals into the detailed partitions with different bandwidths, and then decomposes them into detail and approximation coefficients by means of DWT at the different decomposition levels. Thus, some specific latent features that characterize the nonlinear and dynamical structures of EEG signals are acquired from these coefficients. The fuzzy relations bring out the significant components by reducing the dimension of feature matrix. To detect the epileptic behaviors in EEG signals, these selected components are processed by ANNs based cross-entropy and information criteria. According to the analysis results, this approach not only allows making deeply analysis of EEG signals for detection of epilepsy, but also provides the best model configurations for ANNs in terms of reliability and complexity.

Challenges for the researchers to measure marketing effectiveness in the fmcg sector

Kostas Kotopoulos

IRI, Greece

Lately in the sector of fast moving consumer goods (fmcg) it is being observed a sharp increase on the volumes and diversity of data that are available to the researchers whose mission is to decipher the shopping habits, preferences and dispositions of consumers (and potential shoppers) towards brands. Similarly the data that is currently available to the marketing departments is undeniably richer than ever before facilitating their attempts to illuminate and measure the impact of the activities they employ in driving consumption and consumer engagement with the brands.

At the same time the technological advances has made it a necessity for both fmcg manufacturers and retailers to be well versed in taking real time decisions in marketing activation (recommendation engines, programmatic digital advertising etc). Paradoxically, despite the proliferation of the data sources the gap in information is not narrowing.

Therefore today's researchers need to keep up with new methodologies to tackle the continuously increasing data size, the integration and harmonization of disparate data sources, the speed in execution as well as the information gap.

An algorithmic approach for designing experiments on networks

Vasiliki Koutra

Department of Mathematics, King's College London, UK

Designing experiments on networks challenges the traditional design approaches and classical assumptions due to the interference among the interconnected experimental units as well as the design size. We suggest a novel algorithmic approach for obtaining efficient designs on networks within a practical time frame, by utilising the network topology and particularly its symmetries. We show that the decomposition of the graph based on its symmetries can substantially reduce the search time while maintaining the design efficiency at a sufficient level. This technique can be regarded as an essential step in the search for an optimal design on experimental units that are connected in a large network. We discuss several synthetic and real-world examples.

Improved Poisson and Negative Binomial Item Count Models for Eliciting Truthful Answers to Sensitive Questions

Barbara Kowalczyk[1] and Robert Wieczorkowski[2]

[1]*SGH Warsaw School of Economics, Poland*
[1]*Statistics Poland*

Reliable data on sensitive attributes, socially unaccepted features or illegal behaviors are very hard to obtain in direct questioning. Item count techniques (ICTs) pioneered by Miller (1984) are an example of indirect survey methods designed to deal with sensitive features. ICTs have many practical advantages and have been willingly used by applied researchers. Statistical theory for the classic item count technique was introduced by Imai (2011). Recently Tian et al. (2017) proposed new techniques called Poisson and negative binomial ICTs. The new methods give many opportunities for further theoretical and practical developments. But the methods are not very efficient, i.e. large sample sizes are required to obtain reliable precision. Efficiency is an important issue in indirect methods of questioning. Protection of respondents' privacy is usually achieved at the expense of the efficiency of the estimation. In the present paper we propose new improved Poisson and negative binomial ICTs in which two neutral questions are incorporated and each of the two subsamples serve both as a control and a treatment group. This procedure allows to largely increase efficiency of the estimation as compared to the classic Poisson and negative binomial ICTs and still protect respondents' privacy at the same level. Theoretical results presented in the paper are illustrated by comprehensive simulation studies.

Approaches in aggregating and scaling-up quality and capability metrics in a corporate structure

Karel Kupka

TriloByte Statistical Software

The contribution describes approaches in defining and aggregating quality metrics within structured corporate quality metric system and dealing standard as well as non-standard situation. The data source is physical production technology data from four different processes at different production lines in different plants and countries. To avoid overcomplication of the statistical models and possible interpretation, only the simplest concepts have been used included normal distribution model, a set of four consistent estimators of standard deviation and mean, usual C_p, C_{pk}, P_p, P_{pk} indices and a probability measure (PPM). The approaches are applied to standard data as well as data violating assumptions, like distributional model. It has been shown that redefining or extending quality criteria can help to use standard quality tools meaningfully even in the case of serious departure from standard method assumptions (as normality, homogeneity, independence). A new scalable measure of the process improvement potential has been suggested: Quality Improvement Potential Factor (QIPF). Among the addressed problems are: interpreting high capability values, split-multistream, parallel and serial aggregation, univariate and multivariate process capability scaling.

Bayesian Analysis of Markov Modulated Queues with Abandonment

Joshua Landon

George Washington University, USA

We consider a Markovian queueing model with abandonment where customer arrival, service and abandonment processes are all modulated by an external environmental process. The environmental process depicts all factors that affect the exponential arrival, service, and abandonment rates. Moreover, the environmental process is a hidden Markov process whose true state is not observable. Instead, our observations consist only of customer arrival, service and departure times during some period of time. The main objective is to conduct Bayesian analysis in order to infer the parameters of the stochastic system. This also includes the unknown dimension of the environmental process. We illustrate the implementation of our model and the Bayesian approach by using actual data on call centers.

Impact of Unconventional Monetary Policy on Japanese financial markets 2010 - 2016: A Comparison between CME and QQE periods

Wee-Yeap Lau

Faculty of Economics and Administration, University of Malaya

This study uses daily Japanese Government Bond (JGB) 2-Year Yield as a proxy of changes in Monetary Base to investigate the impact of Unconventional Monetary Policy on Japanese financial market across four different monetary easing regime, namely Comprehensive Monetary Easing (CME), first and second Quantitative and Qualitative Monetary Easing (QQE), and QQE with Negative Interest Rate (NIRP) from 2010 to 2016. Based on the Vector Autoregressive model (VAR), the results indicate: Firstly, there are short- run causality from JGB yield to Japan exchange rate in pre-QQE with NIRP period; Secondly, JGB yield granger cause stock price index across four different periods. This result reaffirms unconventional monetary policy in Japan influence the stock market through portfolio rebalancing effect; Thirdly, JGB yield granger cause banking sector index across four different periods. Lower JGB yield improves banks profit margin as lower interest is paid to bond holders; Fourthly, there are short run causality from banking sector index to stock price index in post CME period; Lastly, overnight call rate granger causes banking sector index in QQE during NIRP period. Our results reaffirm that the unconventional monetary policies is able to stimulate the financial market. QQE appears to be powerful than CME as the role of banking sector becomes more influential throughout the QQE period.

Edge sampling using network local information

Can Le

Department of Statistics, University of California, Davis, USA

Edge sampling is an important topic in network analysis. It provides a natural way to reduce network size while retaining desired features of the original network. Sampling methods that only use local information are common in practice as they do not require access to the entire network and can be parallelized easily. Despite promising empirical performance, most of these methods are derived from heuristic considerations and therefore still lack theoretical justification. To address this issue, we study in this paper a simple edge sampling scheme that uses network local information. We show that when local connectivity is sufficiently strong, the sampled network satisfies a strong spectral property. We quantify the strength of local connectivity by a global parameter and relate it to more common network statistics such as clustering coefficient and Ricci curvature. Based on this result, we also derive a condition under which a hypergraph can be sampled and reduced to a weighted network.

Combinatorial Testing-Based Fault Localization

Jeff Yu Lei

University of Texas, Arlington, USA

Combinatorial testing has been shown to be a very effective testing strategy. After a failure is detected by testing, the next task is fault localization, i.e., how to locate the fault that causes the failure. In this talk, we will discuss a fault localization approach we have developed that leverages the result of combinatorial testing. Our approach consists of two major steps. At the first step, we identify failure-inducing combinations in a combinatorial test set. A combination is failure-inducing if its existence causes a test to fail. Based on the execution result of a combinatorial test set, we produce a ranking of suspicious combinations in terms of their likelihood to be inducing. At the second step, we create a small group of tests from a given failure-inducing combination. In the group, one test is referred to as the core member, and it produces a failed execution. The other tests are referred to as the derived members which are similar to the core member but produce passed executions. The traces of these test executions are then analyzed to locate the faults. Experimental results show that our approach is very effective in that only a small number of additional tests are needed to locate the faults.

Experiments with some interpretable neural network models for a customer classification problem in financial industry

Ta-Hsin Li

IBM T. J. Watson Research Center, Yorktown Heights, USA

The success of neural network models in image and speech processing has generated great interests in expanding their applications to business problems, an example of which is classification of customers for marketing campaigns in banking industry. However, the black-box nature of traditional neural networks limits their applicability. In recent years, new designs of neural network models with interpretability in mind have been proposed. In this paper, we report some results of our experimentation with such a model using a dataset of bank marketing operation. In addition, we consider a new training strategy with the aim of achieving the desired interpretability. Finally, we propose alternative architectures and regularization techniques to improve the interpretability and classification performance.

Multivariate Count Time Series with Flexible Autocorrelations

James Livsey

Center for Statistical Research and Methodology, Census Bureau, USA

Count time series modeling is an active current area of statistical research. This talk examines a bivariate count time series with some curious statistical features: Saffir-Simpson Category 3 and stronger annual hurricane counts in the North Atlantic and Pacific Ocean Basins. To describe the severe hurricane counts in both basins simultaneously, a bivariate count time series model with Poisson marginal distributions is needed - one that permits possible negative cross correlations at lag zero between the series and non-zero correlations at decadal lags in each marginal series. We develop a stationary multivariate count time series model with Poisson marginal distributions and a flexible autocovariance structure. Moreover, autocorrelations can have long-range dependence. Our model is based on categorizing and super-positioning multivariate Gaussian time series. We derive the autocovariance function of the model and propose a method to estimate model parameters. In the end, we conclude that severe hurricane counts are indeed negatively correlated across the two ocean basins.

Buffered Vector Error-Correction Models

Renjie Lu and Philip L.H. Yu

Department of Statistics and Actuarial Science, The University of Hong Kong

This paper extends the buffered autoregressive model to the buffered vector error-correction model (VECM). Least squares estimation and a reducedrank estimation are discussed, and the consistency of the estimators on the delay parameter and threshold parameters is derived. We also propose a supWald test for the presence of buffer-type threshold effect. Under the null hypothesis of no threshold, the supWald test statistic converges to a function of Gaussian process. A bootstrap method is proposed to obtain the p-value for the supWald test. We investigate the effectiveness of our methods by simulation studies. We apply our model to study the monthly Federal bond rates of United States and identify evidences of buffering regimes in the bond rates.

A Class of Models for Multiple Networks Using Graph Distance

Simon Lunagomez, Sofia Olhede and Patrick Wolfe

Lancaster University, UK

In this work, we introduce a new class of models for multiple networks. The core idea of our approach is to parametrize a distribution in the space of labelled graphs in terms of a centroid (which is itself a network) and a parameter that controls the dispersion of the distribution with respect to that centroid. The dispersion of the distribution will be defined in terms of its entropy. This new class of models depends on a specific a metric. We provide a general approach for setting up Bayesian models for this class of models and general strategies for sampling from the posterior.

Some New Ways of Modeling Stationary Integer Count Time Series

Robert Lund

Department of Mathematical Sciences, Clemson University

This talk proposes some new but simple methods for modeling stationary time series of integer counts. Previous work has focused on thinning methods and classical autoregressive moving-average (ARMA) difference equations. In contrast, our methods bypass ARMA tactics by building the desired marginal distribution from independent stationary Bernoulli sequences. Time series with binomial, Poisson, negative binomial, and many other discrete marginal distributions are easily built. The models are naturally parsimonious, can have negative autocorrelations and/or long-memory features, and can be statistically fitted via classical one-step-ahead linear prediction techniques.

A Comparison of the Multivariate SSA Methods for Forecasting Mortality Rates

Rahim Mahmoudvand

Department of Statistics, Bu-Ali Sina University, Hamedan, Iran

Multivariate singular spectrum analysis (MSSA) is a non-parametric technique in the field of time series analysis. There are four variants of MSSA which differ in performance for forecasting. In this paper, we compare the performance of these variants for mortality forecasting. We consider a real data application with nine European countries: Belgium, Denmark, Finland, France, Italy, Netherlands, Norway, Sweden and Switzerland, over a period 1900-2009.

Studying the performance and the availability of multi-state deteriorating systems: The case study of a diesel engine system

Sonia Malefaki

Department of Mechanical Engineering and Aeronautics, University of Patras, Greece

An important characteristic of our modern times is the design of large scale, complexity and accuracy of mechanical and other systems which are extremely demanding in dependability. Thus, most of the contemporary technological systems are operating under multiple deterioration stages and have complicated structures. The performance and the availability of these systems are of great importance since their deterioration and/or failure may lead to important financial and/ or social losses. In order to improve the operation of such systems and increase their availability, maintenance actions can be adopted. However, although preventive maintenance improves the performance of a system, it incurs cost. Thus, a lot of research effort has been paid in finding an appropriate maintenance policy that manages to reduce the total operational cost, as a performance measure, and improve the availability of such a system.

A typical example of a system that functions in a certain amount of deterioration states is a diesel engine system and its subsystems that are subject to inspection and maintenance. Great emphasis has been given to the asymptotic behavior of this system, due to the fact that the majority of them are designed to operate for a long period of time. In this work the availability and the total operational cost of the diesel engine's subsystems will be studied in order to evaluate the aforementioned measures of a diesel engine's system. The system is regularly inspected and depending on its condition, either no action takes place or maintenance is carried out, either minimal or major. The proposed model takes also into account the scenario of imperfect and failed maintenance. Moreover, an optimal inspection and also maintenance policy which maximizes the availability of the system or minimizes the total operational cost is determined.

Bayesian hierarchical modelling of sparse count processes with applications in retail analytics

Ioanna Manolopoulou

Department of Statistical Science,University College of London, UK

In retail analytics, slow-moving-inventory (SMI) refers to goods which rarely sell, resulting in very sparse count processes. Forecasting the sales of such goods is challenging, because traditional predictive models rely on large enough sales volumes to be accurate. In this work, we develop modelling, inferential and predictive methods able to learn the dynamics of sparse count processes for SMI products with few to no sales. We flexibly introduce covariates into the self-exciting model for sparse processes of Porter et al., (2012). We extend the model to include a cross-excitation contribution that allows differing series to excite one another, capturing the process of intertwined contemporaneous excitation dynamics. We integrate individual products into a Bayesian hierarchical model that accommodates shrinkage and information passing across differing sparse count process, without requiring the data for each product to exist over the same time period. We illustrate our methods on a retail analytics dataset from a major supermarket chain in the UK.

Clustering Mixed-Type Data

Marianthi Markatou[1] and Alex Foss[2]

[1]*Department of Biostatistics, University at Buffalo, USA*
[2]*Sandia National Labs*

Despite the existence of a large number of clustering algorithms, clustering mixed interval (continuous) and categorical (nominal and/or ordinal) scale data remains a challenging problem. We show that current clustering methods for mixed-scale data suffer from at least one of two central challenges: 1) they are unable to equitably balance the contribution of continuous and categorical scale variables without strong parametric assumptions; 2) they are unable to properly handle data sets in which only a subset of variables are related to the underlying cluster structure of interest. We first develop KAMILA

(KAY-means for MIxed LArge data), a clustering method that addresses (1) and in many situations (2) without requiring strong assumptions. We next develop MEDEA (Multivariate Eigenvalue Decomposition Error Adjustment), a weighting scheme that addresses (2) even in the face of a large number of uninformative variables. We study theoretical aspects of our methods and demonstrate their performance using Monte Carlo simulations and real data sets.

Bayesian Inference for Sequential Treatments under Latent Sequential Ignorability

Alessandra Mattei[1], Federico Ricciardiy[2] and Fabrizia Mealli[1]

[1]*Department of Statistics, Computer Science, Applications, University of Florence, Italy*
[2]*Department of Statistical Science, University College London (UCL), UK*

We focus on causal inference for longitudinal treatments, where units are assigned to treatments at multiple time points, aiming to assess the effect of different treatment sequences on an outcome observed at a final point. A common assumption in similar studies is Sequential Ignorability (SI): treatment assignment at each time point is assumed independent of unobserved past and future potential outcomes given past observed outcomes and covariates. SI is questionable when treatment participation depends on individual choices, and treatment assignment may depend on unobservable quantities associated with future outcomes. We rely on Principal Stratification to formulate a relaxed version of SI: Latent Sequential Ignorability (LSI) assumes that treatment assignment is conditionally independent on future potential outcomes given past treatments, covariates and principal stratum membership, a latent variable defined by the joint value of observed and missing intermediate outcomes. We evaluate SI and LSI, using theoretical arguments and simulation studies to investigate the performance of the two assumptions when one holds and inference is conducted under both. Simulations show that when SI does not hold, inference performed under SI leads to misleading conclusions. Conversely, LSI generally leads to correct posterior distributions, irrespective of which assumption holds.We apply our framework to an illustrative example based on real data, in which we investigate the effects of interest free loans on firm employment policies.

Unattended smoothing of nonlinear profiles using R

Javier M. Moguerza[1], Emilio L. Cano[2] and Mariano Prieto Corcoba[3]

[1]*Rey Juan Carlos University*
[2]*University of Castilla-La Mancha*
[3]*ENUSA Industrias Avanzadas*

Nonlinear profiles appear in processes whose samples are suitable of being individually represented by a nonlinear function that characterises each occurrence of the process. Therefore, these samples cannot be adequately represented by a linear model, and nonlinear smoothing techniques are needed to represent free of noise prototypes of each occurrence. In this work, for the smooting procedure a Support Vector Machine (SVM) approach is followed, with the novelty that an unattended parameters setting option is incorporated. The methodology is included in a R package, which also allows to represent non-smoothed profiles in order to check the quality of the smoothing stage. In addition, an overall prototype and confidence bands, which are actually the counterparts of center line and control limits in classical control charts, can be estimated to monitor new profiles. The package also includes the option of plotting an easy-to-interpret nonlinear control chart.

Copula-based robust optimal block designs

Werner Mueller and D. Woods

Johannes Kepler University Linz, Austria

Blocking is often used to reduce known variability in designed experiments by collecting together homogeneous experimental units. A common modelling assumption for such experiments is that responses from units within a block are dependent. Accounting for such dependencies in both the design of the experiment and the modelling of the resulting data when the response is not normally distributed can be challenging, particularly in terms of the computation required to find an optimal design. The application of copulas and marginal modelling provides a computationally efficient approach for estimating population-average treatment effects. Motivated by an experiment from materials testing, we develop and demonstrate designs with blocks of size two using copula models. Such designs are also important in applications ranging from microarray experiments to experiments on human eyes or limbs with naturally occurring blocks of size two. We present methodology for design selection, make comparisons to existing approaches in the literature and assess the robustness of the designs to modelling assumptions.

Supersaturated split-plot experiments

Kalliopi Mylona [1,2], E. S. Matthews[3] and D. C. Woods[3]

[1]*Department of Statistics, Universidad Carlos III de Madrid, Spain*
[2]*Department of Mathematics, King's College London, UK*
[3]*Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK*

Supersaturated split-plot experiments combine two classes of designs which are important for industrial experimentation; screening designs and designs with restrictions on randomisation due to hard-to-change factors or two-stage processes. Although such designs are prevalent in industry, the literature on them is limited. We propose an optimal design approach and present Bayesian optimality criteria to find these designs. The analysis of supersaturated split-plot designs is complicated by the correlation of columns in the model matrix and the estimation of two variance components. We propose a novel analysis method for responses from these designs, which includes empirical Bayes and coordinate descent. Industrial examples from materials and pharmaceutical sciences are used to demonstrate new approaches to both the design and analysis of such supersaturated splitplot experiments.

Social Media Platforms: Tools for Data Collection in Technology-Driven Marketing Research

Olugbemi A. Olujimi[1] and Wajdi Ben Rejeb[2]

[1]*Marketing Science and Data Management Dept., FactInfo, Lagos Nigeria*
[2]*University of Roehampton London - Laureate Online Education National Institute of Applied Sciences & Technology, Tunis*

This study investigates the use of social media as alternate tools for data collection in Nigerian marketing research using four social media platforms (Facebook, Twitter, Skype and WhatsApp) and the paper-based questionnaire as the control. The ANOVA and Duncan's test carried out on the responses from the leading data collection platforms experimented showed them not to be significantly different with $P > 0.05$. The same result was also obtained from the various cross analysis performed on the survey data. This implied that within the limits of this experimentation, social media networking equally good for data collection as the traditional paper questionnaire. However the study also revealed a significant different with $P < 0.05$ on preference level for the future online social networking for marketing research purposes with the greater number of respondents preferring online networking for marketing research purposes in the future especially through WhatsApp and Facebook.

Bayesian design for intractable models

Antony Overstall

University of Southampton

Bayesian designs are found by maximising the expectation of a utility function where the utility function is chosen to represent the aim of the experiment. There are several hurdles to overcome when considering Bayesian design for intractable models. Firstly, common to nearly all Bayesian design problems, the expected utility function is not analytically tractable and requires approximation. Secondly, this approximate expected utility needs to be maximised over a potentially high-dimensional design space. To compound these problems, thirdly, the model is intractable, i.e. has no closed form. New approaches to maximise an approximation to the expected utility for intractable models are developed and applied to illustrative exemplar design problems with experimental aims of parameter estimation and model selection.

Detection of Epileptic Seizures using Deep Neural Networks Based on Discrete Wavelet Transforms

Ezgi Ozer[1] and Ozan Kocadagli[2]

Faculty of Engineering, Okan University, Istanbul, Turkey
Department of Statistics, Faculty of Science and Letters, Mimar Sinan Fine Arts University, Istanbul, Turkey

This study presents an efficient approach that ensures an accurate classification of Electroencephalogram (EEG) signals for detection of epileptic seizures. Essentially, this approach is based on the deep learning algorithms with discrete wavelet transforms (DWT's) and feature selection methods. In this analysis, the performance of deep neural networks (DNNs), support vector machines (SVM), regression tree (RT), Naïve Bayes Classifier (NBCs) and multivariate logistic regression (MLR) is compared with each other over a benchmark epilepsy data set. To ensure an efficient classification performance, the automated multi-resolution signal processing technique splits EEG signals into the detailed partitions with different bandwidths, and then decomposes them into detail and approximation coefficients using DWT. To control the complexity of model, the features obtained from DWT's are reduced by feature selection methods brining out significant components where they are used as the inputs in the estimation procedures. As a result, the proposed procedure not only ensures better performance than the other approaches in the literature in context of detection of epileptic seizures, but also provides estimating more reliable and robust models in term of reliability and complexity.

Model Based Clustering through copulas for high dimensional data

Dimitris Karlis[1], Fotini Panagou[1] and Ioannis Kosmidis[2]

[1]*Department of Statistics, Athens University of Economics and Business, Greece*
[1]*Department of Statistics, University of Warwick, Coventry, UK*

In a recent paper Kosmidis and Karlis (2016) proposed model based clustering based on multivariate distributions defined through copulas. This approach offers a number of advantages over existing methods mainly due to the flexibility to define appropriate models in certain different circumstances. In this talk we exploit the ideas of extending the approach for higher dimensions. The central idea is to use a Gaussian copula and implement the correlation matrix of the Gaussian copula through certain parsimonious representations giving rise to models of different complexity. We use two different approaches, the first makes use of factor analyzers based on the factor decomposition of the correlation matrix and the second is based on Choleski type decompositions. Application with real and simulated data will be also described.

Composite quality indicators for assessing healthcare provision

Theodoros Paschalis[1], Christina Georgakopoulou[1], Spyros Goulas[1], Athanasios Sachlas[2] and Sotiris Bersimis[1,2]

[1]*National Organization for Provision of Health Services (NOPHS), Greece*
[1]*Department of Statistics and Insurance Science, University of Piraeus, Greece*

Quality in providing healthcare services incorporates effectiveness, efficiency, safety, accessibility and development of person-centered services. By exploiting its large database, National Organization for the Provision of Healthcare Services (NOPHS), succeeds in continually developing composite quality indicators to evaluate the provided healthcare services. In this presentation, we will present the rationale of developing such indicators and how these indicators can be implemented by NOPHS to provide as more qualitative healthcare services for the benefit of patients.

Adaptive Schemes for the Multivariate Control Charts

Theodoros Perdikis and Stelios Psarakis

Department of Statistics, Athens University of Business and Economics, Greece

Control charts is the most effective and widely used tool of the Statistical Process Control. Over the past decades, in order to enhance the efficiency and performance of the control charts the use of the adaptive feature in the design parameters has been added. In this work, after a brief review of the adaptive univariate schemes, a detailed presentation, evaluation and comparison of the adaptive multivariate control schemes in the multivariate case is investigated.

Bayesian models for response times in cognitive experiments

Mario Peruggia, Peter Craigmile and Trisha Van Zandt
Department of Statistics, The Ohio State University, USA

How do we form memories of objects or scenes that we see? How do we recall those memories? How do we react to simple stimuli? How do environmental conditions interfere with the performance of these tasks? How accurate are we in performing these tasks? The hierarchical Bayesian framework provides a powerful and flexible set of tools for addressing these questions.

There are many reasons why the Bayesian approach is successful. A good Bayesian model can deal with subject heterogeneity in a coherent fashion that allows one to accommodate individual cognitive differences while zeroing in on the common features of inferential interest. The features of inferential interest can be made to relate to different aspects of the cognitive process which will typically depend on experimental and environmental conditions. The relations between model components and the cognitive process can be of a purely descriptive nature or can be representative of theoretical constructs reflective of the way in which the brain processes information. In the latter case, competing cognitive theories can be assessed and validated based on the quality of the model fit.

In this talk I present several challenging applications illustrating these and other facets of the Bayesian analysis of response time data, and I argue that understanding how people make simple decision is key to the successful modeling of more complex, real-word, decision problems.

Control Charts for the Simultaneous Monitoring of the Parameters of a Zero-Inflated Poisson Process Under Unknown Shifts

Athanasios Rakitzis
Department of Mathematics, University of Aegean, Karlovasi, Samos, Greece

The zero-inflated Poisson (ZIP) distribution is one of the most appropriate models for overdispersed data with an excessive number of zeros. Data of this type frequently arise in manufacturing processes with a low fraction of defective items. Zip model has two parameters, one is the probability of extra zeros and the other stands for the expected Poisson counts. In this work, we propose and study three new control charts for detecting changes in either of the two parameters of a zip process. The proposed schemes do not need any prior information related to shift size and can be used either for individual observations or the subgroup samples.Their performance is studied for various in-control and out-of-control scenarios via Monte-Carlo simulation. Comparisons with other competitive charts are also given. The results reveal that they are very effective in the detection of small and moderate shifts in process parameters. Finally the practical implementation of the proposed schemes is also illustrated through a real-data example.

Measuring Effectiveness of Trade Schemes in CPG Domain Using DLM

Balaji Raman

Cogitaas AVA, India

In emerging markets like India, unorganized retail contributes 75% - 80% of sales for a Consumer-Packaged Goods (CPG) firm. In addition to standard sales margin, retailers are incentivised by CPG firms to sell their products to consumers. These incentives are commonly known as trade promotions and they are one of the fastest growing marketing expenditure for a firm. Despite this, in-store promotions prove to be a black box for sales managers as there is no precise way to estimate returns on a trade investment in clusters of stores in different market segments and current methods can often be misleading and lead to non-optimal spending.

In this talk, we discuss the use of hierarchical dynamic linear models to determine the returns on trade investment for a brand across several types of retailers, channels and regions. Dynamic linear models are required for couple of purposes - a) estimating weekly or monthly consumer off-take from retailer purchase data (also referred to as secondary sales) and b) to account for changes over time in other marketing factors, like media spends, pricing and consumer promotions.

Multiple Day Biclustering of High-frequency Financial Time Series

Nalini Ravishanker

Department of Statistics, University of Connecticut, CT, USA

With recent technological advances, high-frequency transaction-by- transaction data are widely available to investors and researchers. To explore the microstructure of variability of stock prices on transaction level intraday data and to dynamically study patterns of co-movement over multiple trading days, we propose a *multiple day* time series biclustering algorithm (CC-MDTSB) which extends the time series biclustering algorithm (CC-TSB). For identifying biclusters within each trading day, our algorithm provides a faster alternative to the random replacement method in the CC-TSB algorithm. Moreover, our algorithm does not require prespecification of the number of biclusters for each trading day. Instead, we set a threshold on the number of stocks within the biclusters to yield an adaptive stopping criterion for multiple day analysis. An analysis of the biclusters determined over multiple trading days enables us to study the dynamic behavior of stocks over time. We effectively estimate the comovement probability of each m-tuple of stocks conditional on the other stocks within the dynamic biclusters, and propose a method to forecast comovement days using a nonparametric double exponential smoothing procedure. This is joint work with Jian Zou and Haitao Liu, Worcester Polytechnic Institute.

Singular spectrum analysis for long and contaminated time series

Paulo Canas Rodrigues

Center for Applied Statistics and Data Analytics (CAST), University of Tampere, Finland

Singular spectrum analysis (SSA) is a non-parametric method for time series analysis and forecasting that incorporates elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. Although this technique has shown to be advantageous over traditional model based methods, in particular, one of the steps of the SSA algorithm, which refers to the singular value decomposition (SVD) of the trajectory matrix, is highly sensitive to data contamination and also time consuming.

In this talk I will present (i) a randomized SSA algorithm which is an alternative to SSA for long time series that keeps the quality of the analysis; and (ii) a robust SSA algorithm, where a robust SVD procedure replaces the least-squares based SVD in the original SSA procedure, in order to reduce the effect of data contamination by outlying observations.

The SSA and the randomized SSA are compared in terms of quality of the model fit and forecasting accuracy, and computational time, via Monte Carlo simulations and real data about the daily prices of five of the major world commodities. The SSA and the robust SSA are compared in terms of the quality of the model fit via Monte Carlo simulations that contemplate both clean and noisy/contaminated time series, and considering a real data application where a data set from the energy sector is analyzed.

Project Risk Management under Dynamic Environments

Fabrizio Ruggeri

Italian National Research Council in Milano, Italy

We model activity durations in a project network over time when concurrent activities can be affected by common external factors, like financial or political crisis, social turmoil or environmental causes. Dependence of activity durations is therefore captured by a common random environment with a Markovian evolution. We discuss probabilistic implications of the dependence structure and how this can be used to assess activity durations and project completion time in a dynamic manner. We develop Bayesian inference for the model and illustrate its implementation by using data from a real life project network. The developed model can be beneficial for project managers in risk analysis and planning

---

Multivariate Risk-Adjusted Control Charts

Athanasios Sachlas[1,2], Sotiris Bersimis[2] and Stelios Psarakis[1]

[1]*Department of Statistics, Athens University of Business and Economics*
[2]*Department of Statistics and Insurance Science, University of Piraeus*

Risk-adjusted control charts take into account the varying health conditions of the patients. This type of control charts is a modification of standard control charts, which appeared in the bibliography mainly in the last two decades to improve the monitoring mainly of medical processes. Biswas and Kalbfleisch (2008) outlined a risk-adjusted CUSUM procedure based on the Cox model for a failure time outcome while Sego et al. (2009) proposed a risk-adjusted survival time CUSUM chart for monitoring a continuous, time-to-event variable that may be right-censored. Motivated by the above mentioned papers, in this work we present the multivariate risk adjusted control charts and some preliminary results on multivariate risk-adjusted EWMA control charts.

Aggregation of risks for lifetimes with mixture exponential distributions

Jose Maria Sarabia

Department of Economics, University of Cantabria

In reliability, there are situations where data can be modelled as conditionally independent rather than independent using mixtures. In this paper, aggregation models or risks for lifetime data that have conditionally exponential distributions are studied. We begin reviewing the properties of the multivariate mixture of exponentials including a characterization theorem in terms of the copula generator and the marginal distributions, dependence conditions (total positivity of order two in pairs and associated random variables), dependence measures, moments, copula (which is Archimedean) and other relevant features. We continue with the analytical formulation for the probability density function and the cumulative distribution function of the aggregated distribution.

Then, we study some specific multivariate models with lifetime of the type Pareto, Gamma, Weibull, inverse Gaussian mixture of exponentials (Whitmore and Lee, 1991) and other lifetime distributions. For these models, we obtain specific expressions for the aggregated distribution, and we study some of their main properties. Finally, some extensions of the baseline multivariate model are studied.

Market surveys in emerging markets - perspectives from CPG industry

Kamal Sen

Cogitaas AVA

Marketing surveys are an essential part of any consumer insights team in a CPG firm. Surveys track household purchase pattern, consumer offtake, media GRPs, brand equity measures. Marketing strategies and investment decisions are taken based on such surveys. Over the past couple of years, consumer tracking happens more frequently, thanks to advances in technology and neuromarketing has become a standard practice for insights team to study consumers' cognitive response to marketing stimuli.

It is of interest to know whether all these different forms of survey data lead to reliable statistical inference? What is sufficient data and robust data for business decisions from the statistical point of view? Is more data collection leading to better statistical inference in consumer industries? What steps are being taken to improve survey data for industrial use?

These are issues that will impact the present and future of analytics excellence in consumer industries.

Granger causality in yield curves of different markets

Rituparna Sen

ISI Chennai, India

We develop time series analysis of functional data observed discretely, treating the whole curve as a random realization from a distribution on functions that evolve over time. The method consists of principal components analysis of functional data and subsequently modeling the principal component scores as vector ARMA process. We justify the method by showing that an underlying ARMAH structure of the curves leads to a VARMA structure on the principal component scores. We derive asymptotic properties of the estimators, fits and forecast. For term structures of interest rates, this provides a unified framework for studying the time and maturity components of interest rates under one set-up with few parametric assumptions. We apply the method to the yield curves of USA and India. We compare our forecasts to the parametric model of Diebold and Li (2006). We then use the method of Granger causality on the VARMA of Principal components scores for the two markets to test if one market drives the other.

Anomaly detection in static networks

Srijan Sengupta
Virginia Tech, USA

Anomaly in networks refers to the situation where the networked system, or part of it, shows significant departure from regular or expected behavioral patterns. Anomalies in networks often imply illegal or disruptive activity by the actors in the network. There has been a lot of recent emphasis on developing network monitoring tools that can detect such anomalous activity. Networks can be static, where we have a single snapshot of the system, or dynamic, where we have network snapshots at several points in time. Anomalies can have different meanings in these two scenarios.

In static networks, anomaly typically means a local anomaly, in the form of a small anomalous subgraph which is significantly different from the rest of the network. Local anomalies are difficult to detect using simple network-level metrics since the anomalous subnetwork might be too small to cause significant changes to network-level metrics, e.g., network degree. Instead, such anomalies might be detectable if we monitor sub-network level metrics, e.g., degrees of all subgraphs. However, that option is computationally infeasible, as it involves computing total degrees for all $O(2^n)$ subgraphs of an n-node network.

We propose a novel anomaly detection method by using egonet p-values, where the egonet of a node is defined as the sub-network spanned by all neighbors of that node. Since there are exactly n egonets, the number of subgraphs being monitored is n, which is a relatively manageable number. We establish theoretical properties of the egonet method. We demonstrate its accuracy from simulation studies involving a broad range of statistical network models. We also illustrate the method on several well-studied network datasets.

Tolerance intervals as a method for the assessment of energy output of a photovoltaic system

Gary Sharp[1], Chantelle Clohessy[1], Johan Hugo[1] and Ernest van Dyk[2]
Department of Statistics, Nelson Mandela University, Port Elizabeth
Department of Physics, Nelson Mandela University, Port Elizabeth

Photovoltaic systems generate energy from the solar radiation of the sun. The radiation is converted to energy which is then used in the same manner as a traditional hydro electrical and fossil fuel energy generation system. The emergence of photovoltaic systems for energy generation provides opportunities to develop and propose assessment methodologies which can be used by investors to decide on the viability of the systems.

This study proposes the use of Bayesian tolerance intervals for the assessment of energy output of a photovoltaic energy system. Bayesian simulation methods are used to obtain variance components to estimate the tolerance intervals. This methodology is illustrated for an energy yield case study where data is simulated for a hypothetical scenario of a 1MW photovoltaic system. The simulation data is obtained from the software package, PVSyst (V6.39), a programme used extensively in industry for the explicit purpose of estimating energy yield information. The use of tolerance intervals provides a novel

approach to the assessment of energy generation, by including both the variability and uncertainty of the energy yield estimates.

Combinatorial Testing Methods and Algorithms for Detecting Cryptographic Trojans

Dimitris E. Simos[1], D. Richard Kuhn[2], Yu Lei[3] and Raghu N. Kacker[2]

SBA-Research, Vienna, Austria
National Institute of Standards and Technology, Gaithersburg, USA
University of Texas, Arlington, USA

Combinatorial methods have attracted attention as a means of providing strong assurance at reduced cost, but are these methods practical and cost-effective, in the case of malicious hardware logic detection. In this talk, we are concerned with the problem of detecting cryptographic Trojans that manifest as instances of malicious hardware on top of FPGA technologies. We will develop theoretical and algorithmic methods originating from the field of covering arrays, a special class of combinatorial designs, as a means to provide a fully automated testing framework capable of revealing hard to spot malicious instances of Trojans. We will further demonstrate that combinatorial testing provides the theoretical guarantees for exciting a Trojan of specific lengths by covering all input combinations. Our findings indicate that combinatorial testing constructs can improve the existing FPGA Trojan detection capabilities by reducing significantly the number of tests needed by several orders of magnitude.

This work is part of the *combinatorial security testing* framework developed by the authors, which asides providing mathematical guarantees for hardware Trojan detection also ensures quality assurance and effective re-verification for security testing of web applications, operating systems and communication protocols.

Estimating the health state of populations: Implications in the Health Systems

Christos H. Skiadas[1] and Charilaos Skiadas[2]

ManLab, Technical University of Crete, Greece
Department of Mathematics and Computer Science, Hanover College, USA

The establishment of a well-organized Country Health System is vital for the society development. The very important points are connected with the measurement of the functionality and the impact of the health system to the population longevity and well-being. The last three decades improvements in estimating the health state of a population provide efficient measures of several indicators as the Healthy Life Expectancy (HLE) and the Healthy Life Years Lost (HLYL) improving the classical Life Expectancy (LE) estimates. So far several methods are proposed and comparative tables for the countries are presented as for example the estimates of the World Health Organization or the Eurostat health statistics.

In this paper we present and analyze the existing methods and techniques for estimating the HLE and HLYL and present related applications in several countries and in

Greece. The related bibliography is summarized in our two recent books published as volumes 45 and 46 of “The Springer Series on Demographic Methods and Population Analysis".

Some Bivariate Semiparametric Charts Based on Order Statistics

Markos V. Koutras and Elisavet M. Sofikitou

Department of Statistics and Insurance Science, University of Piraeus, Greece

In the nonparametric process monitoring, the construction of control charts involves the use of two independent samples, the reference and the test sample. The former is exploited to determine appropriate control limits, while the latter is used to ascertain whether the out-of-control signaling rule is valid or not.

In the present work, two bivariate semiparametric (Shewhart-type) control charts are introduced in which both the location of a single pair of order statistics with respect to the X and Y test samples, as well as the number of observations of the test samples that lie between the control limits are taken into account. The proposed charts are quite effective for achieving a simultaneous monitoring of the process mean and variance.

The key advantage of these schemes is the fact that the FAR and the ARLin values are not affected by the choice of the marginal distributions. Indeed, these quantities are typically affected by the dependence structure of the monitored characteristics, as reflected on the associated copula. However, they are practically almost the same when different copulas are used and therefore the new charts can be used as fully nonparametric ones.

Expressions for the operating characteristic function, as well as the alarm rate are obtained. Moreover, exact formulae are provided for the FAR and a numerical study is carried out to assess the performance of the new charts. Finally, illustrative examples are provided for the implementation of the new charts in real-world data set related to biostatistics.

Deep Learning: A Bayesian Perspective

Vadim Sokolov

George Mason University, USA

Deep learning is a form of machine learning for nonlinear high dimensional pattern matching and prediction. By taking a Bayesian probabilistic perspective, we provide a number of insights into more efficient algorithms for optimisation and hyper-parameter tuning. Traditional high-dimensional data reduction techniques, such as principal component analysis (PCA), partial least squares (PLS), reduced rank regression (RRR), projection pursuit regression (PPR) are all shown to be shallow learners. Their deep learning counterparts exploit multiple deep layers of data reduction which provide predictive performance gains. Stochastic gradient descent (SGD) training optimisation and Dropout (DO) regularization provide estimation and variable selection. Bayesian regularization is central to finding weights and connections in networks to optimize the predictive bias-variance trade-off.

Maximum entropy demand models with newsvendor information

Amirsaman H. Bajgiran[1], Mahsa Mardikoraem[2] and Ehsan S. Soofi[2]

Department of Industrial Engineering, University of Wisconsin-Milwaukee
Lubar School of Business, University of Wisconsin-Milwaukee

The newsvendor model is characterized by an uncertain demand and fixed prices for a product and the optimal order quantity as a quantile of the demand distribution. To bypass assuming a complete probability distribution for the demand, researchers have resorted to deriving the optimal order quantity based on partial information in terms of some features of the demand distribution. The maximum entropy has been used for justifying the use some well-known model for the demand distribution, without connection to the newsvendor's optimal order quantity. This paper derives maximum entropy demand distributions using the optimal order quantity as partial information along with various types of moments. Inclusion of this information provides change point demand distribution at the optimal order quantity. The new ME model is an adjustment of the family of the maximum entropy models without the quantile information. A measure of information gain for the inclusion of the optimal order quantity is given. Some stochastic dominance results due to the inclusion of the quantile information for the demand and for the cost are presented.

Bayesian Modeling of Non-Gaussian Multivariate Time Series

Tevfik Aktekin[1], Nicholas G. Polson[2] and Refik Soyer[3]

University of New Hampshire
University of Chicago
George Washington University

Modeling of multivariate non Gaussian time series of correlated observations is considered. In so doing, we focus on time series from multivariate counts and durations. Dependence among series arises as a result of sharing a common dynamic environment. We discuss characteristics of the resulting multivariate time series models and develop Bayesian inference for them using particle filtering and Markov chain Monte Carlo methods. We illustrate application of the proposed approach using conditionally multivariate Poisson and gamma time series.

Multi-Party computations for Privacy Aware collaborative Analytics

Pradeep Sridharan Srinivas
University of Pune India, India

What is Multi-Party Computation (MPC)?

- MPC is a set of Techniques that help multiple Parties set of Techniques that help multiple Parties perform joint analytics **Characteristic of MPC:** Maintaining the confidentiality of each parties' data without exposing anything besides results.

Business Background

- Image classification plays an important role in autonomous driving to derive control inputs for the car.
- Trained models for image classification become more accurate with larger data sets. However it costs time and money to collect this data.
- MPC helps in saving time and cost to build large datasets for training models.
- Helps collecting critical data and features for required for automatic vehicles.

Typical Scenarios

- X1 and X2 want to jointly analyze a data set to identify an anomaly in a product. X2 owns the data related to this product, and X1 and X2 are in partial know how of the formula used to compute the anomaly
- But X2 does not want to give access to all the data and the formula related to anomaly computation
- X1 does not want to reveal to X2 the complete formula for anomaly computation
- But both X2 and X1 benefit from if the anomaly can be detected

Benefits

- X1 and X2 can perform such Joint Analytics while maintaining the confidentiality of their own data without exposing anything besides results using Multi-Party Computation
- The newest technology interventions that helps address the data gap in short time and cost effective way
- MPC is one of such technology intervention to play in a key role in this space to build critical data and features required for automatic vehicles

---

Construction and analysis of D-optimal edge designs

Stella Stylianou

School of Sciences, RMIT University, Melbourne, Australia

Edge designs are screening experimental designs that allow a model independent estimate of the set of relevant variables, thus providing more robustness than traditional designs. In this paper, new classes of D-optimal edge designs are constructed. This construction uses weighing matrices of order n and weight k together with permutation matrices of order n to obtain D-optimal edge designs. Linear and quadratic simulated screening scenarios are studied and compared using linear regression and edge designs analysis. An alternative method for constructing and analyzing expanded edge designs is introduced. This method provides a model-independent estimate of the set of active factors and also gives a linearity test for the underlying model.

Hazard Rate Estimation for Location-Scale Families: Monotonic and Non-monotonic Structures

Baris Surucu

Department of Statistics, Middle East Technical University, Ankara, Turkey

Hazard rate is known as the instantaneous probability of default at time t. It is of much use in numerous areas including finance, actuarial sciences, medicine and many more. The topic of estimating hazard rate function for different statistical distributions under various scenarios has received great attention in the literature. Depending on complete and censored sample structures, the estimation procedure changes. In this talk, we will show how the hazard rate function is estimated under these sampling schemes when the underlying distributions come from location-scale families with some unknown parameters. Moreover, the difference between the estimation of monotonic and non-monotonic hazard rate functions will be discussed. Some real life applications will also be presented during the talk.

Integrated Production Process Optimization: A Bayesian Approach

Konstantinos A. Tasias, George Nenes and Sofia Panagiotidou

University of Western Macedonia, Department of Mechanical Engineering, Kozani, Greece

This paper presents an integrated production, statistical process monitoring and condition-based maintenance model for imperfect processes subject to multiple assignable causes. The assignable causes are independent and may affect both the central tendency and the dispersion of the process. The operation of the proposed, fully adaptive, control scheme is modeled based on the Bayes theorem and the optimal inspection, maintenance and inventory policy are defined through economic and statistical criteria. The realistic assumption of imperfect preventive maintenance actions has also been considered.

Machine learning techniques for the analysis of composite quality indicators

Sotiris Tasoulis[3], Theodoros Paschalis[1], Christina Georgakopoulou[1], Spyros Goulas[1], Sotiris Bersimis[1,2], Athanasios Sachlas[2] and Vassilis Plagiannakos[3]

National Organization for Provision of Health Services (NOPHS), Greece
Department of Statistics and Insurance Science, University of Piraeus, Greece
Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece

Machine Learning is the science that gives computers the ability to learn without being explicitly programmed. In other words, Machine Learning explores the study and construction of algorithms that can learn from data and make predictions. In this study, we will present the results from the application of Machine Learning techniques to data available on the National Organization for Provision of Health Services (NOPHS). These results contribute to the optimal decision making and the promotion of the quality of health services provided by NOPHS.

Generalized Financial Risk Forecasting via Estimating functions with Applications

Aera Thavaneswaran

Department of Statistics, University of Manitoba

Recently, there has been a growing interest in VaR (value at risk) forecasts and model risk forecasts. In this paper, using estimating function approach, a new optimal volatility estimator is introduced and based on the recursive form of the estimator a data-driven generalized EWMA model for VaR forecast is proposed. An appropriate data-driven model for volatility is identified by the relationship between absolute deviation and standard deviation for symmetric distributions with finite variance. It is shown that the asymptotic variance of the proposed volatility estimator is smaller than that of conventional estimators and is more appropriate for financial data with larger kurtosis. For IBM, Microsoft, and Apple stocks the proposed method is used to identify the model, estimate the volatility and obtain the VaR forecasts. Optimality of the VaR forecast and the superiority of the approach are also discussed in some detail.

Deep Learning and Computer Vision for Quality Control: A Perspective

Kim Phuc Tran, Anne Cozol and Beatrice Vedsel

Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne-Sud, Vannes, France

Although traditional statistical process monitoring (SPM) has been widely used in manufacturing, these tools are not capable of handling the large streams of multivariate, videos and images data found in modern system. Additive manufacturing or micromanufacturing combined with fast multi-stream high-speed sensors are paving the way for a new generation of industrial big-data requiring novel approaches for real-time monitoring production processes. Although there are several methods for monitoring image data, most of these perform using multivariate image analysis, which affects detection accuracy and efficiency. In this research, we discuss several applications of machine learning, data mining and computer vision for quality control. We also highlight some application opportunities available in the use of deep learning, computer vision and control charts with videos and images data and provide some advice to practitioners. Our goal is to bring the research issues into better focus and encourage approaches development using deep learning and computer vision for quality control.

Bayesian inference of high dimensional spatio-temporal data

Kostas Triantafyllopoulos, Sofia Karadimitriou and Tim Heaton

University of Sheffield, UK

Over the last 20 years there has been an increased interest in data which vary over space and time. Examples of such data stem from environmental sciences, signal processing and recently astronomy, to name but a few. A typical target application is river flows, for which the flow is measured over a geographical space and over time. Such data can be high dimensional, partly due to the nature of the application and partly due to the advance of computer-aided data collection. The complexity introduced by the spatial structure as well as the high dimensionality of the data result in certain challenges for inference and forecasting.

In this talk we consider that observations are generated by a continuous process over the space (location) and a discrete one over time. This is a differential equation, which is approximated by using a wavelet decomposition into a deference equation. This is then put in state-space form by making use of the orthonormality of the wavelet decomposition. A hierarchical Bayesian dynamic linear model is proposed and inference follows by a suitable Markov chain Monte Carlo scheme. Considerations of this model include Gaussian and non-Gaussian models. For example, one possibility is to analyse count data that vary spatially and temporally. The proposed methodology is illustrated with simulated and real data.

Wilcoxon-type rank-sum statistics for selecting the best population: some advances

Markos V. Koutras [1] and Ioannis S. Triantafyllou [2]

[1] *Department of Statistics and Insurance Science, University of Piraeus, Greece*
[2] *Department of Computer Science & Biomedical Informatics, University of Thessaly, Greece*

In this article, we introduce three new nonparametric procedures based on modified Wilcoxon-type rank sum statistics, for testing the equality of two life-time distributions. The setup of the proposed testing processes is presented in detail and the exact null distribution of all Wilcoxon-type statistics utilized is studied. Closed formulae for the corresponding exact probability of correct selection of the best population are derived for the class of Lehmann alternatives. A detailed numerical study is carried out to elucidate the level of performance of the proposed testing schemes.

Statistical process control and monitoring in the big data era

Panagiotis Tsiamyrtzis

Department of Statistics, Athens University of Economics and Business, Greece

Statistical Process Control and Monitoring (SPC/M) is the area of statistics that has been traditionally used in the industry (and not only) to identify the presence of assignable causes of variation, or in more common terms, to infer when a process moves from the In Control (IC) to the Out Of Control (OOC) state. This typically translates to identification of transient or persistent parameter shifts under an assumed statistical model and various control charts have been developed over the years for this purpose. All the above assume the typical scheme of sequential sampling from a process with an interest to identify as soon as possible when we move to the OOC state, while we keep the false alarm rate at a low level. Nowadays, cheap sensors allow having even 100% inspection on a process, exploding the sample size. In this big data era several of the existing tools used in SPC/M are pushed to the limit and should be modified accordingly. In this work we will highlight the major issues rising for the SPC/M tools when we handle big data and we will suggest plots and methods that will help the visualization and evaluation of the process's quality in presence of big volumes of data.

Estimating the effects of the recent financial crisis on the stillbirth rates employing distributed lag models and demographic decomposition techniques: the case of Greece

Cleon Tsimbos[1] and Georgia Verropoulou[1,2]

University of Piraeus, Greece
Institute of Education, University of London

Scope In this paper we explore changes in stillbirth rates in Greece in the light of the recent economic recession employing linear regression and lag distributed regression models. We also propose a decomposition method to distinguish changes in 2006 and 2014 attributed to differentials in the levels of the stillbirth rates from variations due to the composition of births by certain demographic characteristics.

Data For the purpose of the analysis we use official vital registration data for the period 1995-2016 as well as microdata on livebirths and stillbirths recorded in 2006 and during 2010-2014; the microdata used in the analysis are unpublished and have been provided by the Hellenic Statistical Authority upon special request. In addition we also consider information on per capita GDP which is a well-known index reflecting the socioeconomic conditions of a country.

Methods To assess the impact of the recent financial crisis on the stillbirth rates we employ regression models with specific dummy variables as well as lag distributed models. We also propose a decomposition method relying on direct standardisation demographic techniques for discerning changes in stillbirth rates observed between two points in time.

Results The results show that in times of financial prosperity the relationship between economy and stillbirth rates is negative and statistically significant; in times of economic distress the favourable financial effects on stillbirth outcomes dissipate. The application of the proposed decomposition procedure reveals that the increase in the stillbirth rates between 2006 (3.34 per 1000) and 2014 (3.82 per 1000) is attributed mainly to changes in the composition of births by age of mother and by period of gestation.

Designing and conducting discrete choice experiments with the R-package idefix

Frits Traets and Martina Vandebroek

Faculty of Economics and Business, KU Leuven, Leuven, Belgium

Discrete choice experiments are widely used in a broad area of research fields to capture the preference structure of respondents. The design of such experiments will determine to a large extent the accuracy with which the preference parameters can be estimated. This presentation presents a new R-package, called idefix, which enables users to generate optimal designs for discrete choice experiments based on the multinomial logit model. In addition, the package provides the necessary tools to set up online surveys with the possibility of making use of the individual adaptive sequential Bayesian design approach for estimating the mixed logit model. After data collection the package can be used to transform the data into the necessary format in order to use existing estimation software in R.

The Effect of Wealth and Income on Depression across European Regions: an Analysis based on Instrumental Variable Probit Models

Georgia Verropoulou[1,2], Cleon Tsimbos[1] and Dimitrios Kourouklis[3]

[1]*University of Piraeus, Greece*
[2]*Institute of Education, University of London*
[3]*Frankfurt School of Finance and Management*

The paper examines the impact of wealth and income on the likelihood of depression across four European regions: Southern, Central/Eastern, Northern and Western Europe. These regions exhibit substantial differences in terms of standards of living as well as welfare regimes. To address possible effects, we use data from wave 6 of the Survey of Health, Ageing and Retirement in Europe (SHARE) which was carried out in 2015. The sample includes 60,864 persons aged 50 or higher, resident in 16 European countries. Depression is measured by a binary indicator, constructed using the EURO-D scale. The modelling strategy involves use of probit and instrumental variable (IV) probit models. IV probit estimations are employed to address issues of endogeneity in the analysis due to reverse causality and omitted variable bias. As instrument, the educational attainment of the partner of the respondent is used which relates both to income and wealth at household level while it is a plausibly exogenoussource of variation.

Detecting Earnings Management: A Novel Tobit Modeling Approach

Xiaojing Wang and Wuqing Wu

University of Connecticut, USA

In this paper, we focus on a special type of data, i.e., the data from Earnings Management, where in the middle of their distributions, there are measurements errors. We find that when the Earnings Management is used as a response variable in the analysis to estimate the degree of Earnings Management, it always yields the biased estimation. To solve this issue, we propose a new type of Tobit model named Type-III Tobit model, where both the upper and lower thresholds are unknown. We give the maximum likelihood estimation for the parameters in our model and derive the corresponding asymptotic normality for the estimators of the parameters. Numerical simulations are conducted to compare the behaviors of the ordinal least squares (OLS) method and the proposed Type-III Tobit model in fitting the Earnings Management data. These simulations illustrate that when there are measurement errors existing in the middle of the distributions of Earning Management, the proposed Type-III Tobit model has no bias in the fitting such data, while the results from OLS have systematic errors. We apply our model to an empirical analysis of earning management in listed companies from 12 countries. This analysis further shows that for estimating the parameters in the popular modified Jones model, OLS tends to underestimate the earning management compared with the proposed Tobit model. This discovery has significant impact on the research related to Earning Management.

Using player abilities to predict football

Gavin Whitaker

UCL and Stratagem, UK

We consider the task of determining a football player's ability for a given event type, for example, scoring a goal. We propose an interpretable Bayesian inference approach that centres on variational inference methods. We implement a Poisson model to capture occurrences of event types, from which we infer player abilities. Our approach also allows the visualisation of differences between players, for a specific ability, through the marginal posterior variational densities. We then use these inferred player abilities to capture a team's scoring rate (the rate at which they score goals) through a Bayesian hierarchical model. We demonstrate the resulting scheme on the English Premier League, capturing player abilities. Furthermore, Stratagem Technologies is using the output of the hierarchical model, in conjunction with proprietary data, to make informed trading decisions across both the over/under market (total goals scored), and the Asian handicap market (how many goals a team will win by). These methods can be combined with other tools at Stratagem's disposal to highlight the key areas on the pitch where players have most impact for these abilities, which we also demonstrate here.

Weighted Exponential Random Graph Models: Specification and Application

James D. Wilson

Department of Mathematics and Statistics, University of San Francisco, USA

The exponential random graph model (ERGM) is a popular and flexible tool for statistical inference with network data. However, a major limitation of the common formulation of the ERGM is that it can only be applied to networks with dichotomous edges. In this talk, we will present a flexible distribution on weighted networks that generalizes the ERGM to networks with continuous-valued weighted edges. Through the use of Markov Chain Monte Carlo methodology, we show how to efficiently simulate and estimate these generalized exponential random graph models (GERGMs). Furthermore, we will demonstrate how to utilize these types of models on correlation graphs which frequently arise in genetic and neurological applications. We will highlight the use of these models on the Default Mode Network and 9 other functional subnetworks of the brain measured from resting state fMRI data.

On Modeling Overdispersion

Evdokia Xekalaki

Department of Statistics, Athens University of Economics, Greece

Count data often manifest variability that exceeds what would be expected under a Poisson distribution. This phenomenon, commonly termed *overdispersion*, is particularly prevalent in applications, and is often remedied by switching from Poisson to negative binomial-type models. We will explore flexible alternatives to such models based on the generalized Waring distribution. Starting from simple univariate settings, we will gradually generalize to multivariate settings, adapting them to handle temporally evolving data, and, finally, to general spatial point processes. In doing so, we will highlight the advantages of using the generalised Waring relative to the negative binomial, which become increasingly salient as we move from simpler to more complex settings.

Some Approaches for the Monitoring Event Magnitude and Frequency

Min Xie, Ridwan A. Sanusi and Tahir Mahmood

City University of Hong Kong

Control charting techniques for monitoring the magnitude and frequency of an event is crucial in manufacturing and industrial areas. Traditionally, two isolated sequential stopping rules are employed for monitoring these variables separately. However, recent development in this area employs some unified approach to simultaneously monitor the magnitude and frequency of an event. In this talk, we will present some recent developments in the statistical process monitoring considering both the magnitude and frequency of events. In a recent study, we design an adaptive approach scheme for monitoring these variables simultaneously. The scheme combines a max and a distance-based statistics. It retains the advantages of both the Max-type and the Distance-type schemes for joint inference. The proposed scheme is very effective, efficient and competent in detecting a shift in the process distribution of magnitude or frequency of an event, or both. In addition, it is easier to implement. It has a distinct advantage over its traditional counterparts in detecting moderate to large shifts. Finally, we illustrate the implementation of the proposed scheme with a real dataset.

Inter-rater Agreement and Adjusted Degree of Distinguishability for 2×2 Tables

Ayfer Ezgi Yilmaz and Tulay Saracbasi

[1]*Department of Statistics, Faculty of Science, Hacettepe University, Beytepe-Ankara, Turkey*

In square contingency tables, analysis of agreement between the row and column classifications is mostly of interest. In such tables, kappa coefficient is used to summarize the degree of agreement between two raters. In addition to investigate the agreement among raters, the term of category distinguishability should be discussed. The aim of this study is to assess the agreement coefficients and degree of distinguishability together in 2×2 tables. The adjusted degree of distinguishability is suggested to overcome the problem of calculating the degree of distinguishability that falls outside the defined range. A simulation study is performed to compare the proposed adjusted degree of distinguishability and the classical degree of distinguishability. Furthermore, the interpretation levels for the degree of distinguishability are determined based on a simulation study. The results are discussed over numerical examples.

Comparing recent mortality and health experiences in Greece and Turkey

Konstantinos N. Zafeiris[1] and Christos Skiadas[2]

[1]*Laboratory of P. Anthropology, Department of History and Ethnology, Democritus University of Thrace, Greece*

[2]*ManLab, Technical University of Crete, Greece*

During the last years Greece and Turkey followed opposite course on economic grounds. In Greece, economic crisis afflicted severely the population and caused not only economical problems but also socio-cultural ones. In Turkey the steady economic development of the last years changes significantly the country, even if some signs of economic recession have been observed recently. The scope of this paper is to compare mortality and health experiences in the two countries. For that several parameters like life expectancy at birth and healthy years lost because of disabilities were used. Results indicate the existence of significant differences among the two countries which cannot be explained solely on levels of economic development.

Comparison of Control Charts for Short Run Monitoring of Fractional Nonconformance

Xin Zhou

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Measurement errors are considerable in the testing of food quality characteristics such as fat percentage because routine testing in the shopf-loor is based on high performance liquid chromatography (HPLC) technique. To adjust for the effect of measurement errors, Govindaraju and Jones (2015) proposed a fractional nonconformance (FNC) probability measure for known measurement error distributions. The FNC statistic was initially applied for acceptance sampling inspection and has been further implemented for short run process monitoring by Zhou et al. (2017).

Exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) chart accumulate information selectively from the past observations. Past research established that these charts are sensitive to detect of small shift levels (on the order of about 1.5sigma or less) compared to the ordinary Shewhart chart, but slower in detecting a very large abrupt shift in the process level.

Fractional nonconformance statistic was implemented on EWMA and CUSUM framework for short-run production monitoring under various process models. We found that EWMA and CUSUM chart are superior to Shewhart charting for detection of both small and big shifts (up to 3sigma) only when the false alarm rate is small ($\alpha < 1\%$). When false alarm rate is of the order 5%, the performances of EWMA, CUSUM and Shewhart charts are indistinguishable. For short production runs, a very small false alarm rate is not feasible, and hence the simpler Shewhart charting is adequate.

Author Index