ISBIS 2016

Book of Abstracts

ISBIS 2016 Meeting
on
Statistics in Business and Industry

June 8th-10th, 2016

Barcelona, Spain

Book of Abstracts - ISBIS 2016

Meeting on Statistics in Business and Industry

Teresa A. Oliveira, Universidade Aberta, Portugal

Amílcar Oliveira, Universidade Aberta, Portugal

Rahim Mahmoudvand, Bu Ali Sina University, Iran

Nalini Ravishankar, University of Connecticut, USA

David Banks, Duke University, USA

(Authors)

Editor: Universidade Aberta de Lisboa

e-Book: Published by Universidade Aberta de Lisboa

ISBN: 978-972-674-795-6

e-Book Title:
[Book of abstracts - ISBIS 2016 Meeting on Statistics in Business and Industry]

Scientific Program Committee

David Banks, Duke University

Amílcar Oliveira, Universidade Aberta and CEAUL

Teresa A. Oliveira, Universidade Aberta and CEAUL

Nalini Ravishankar, University of Connecticut

Xavier Tort Martorell, Universitat Politécnica de Catalunya, Barcelona TECH

Martina Vandebroek, KU Leuven

Vincenzo Esposito Vinzi, ESSEC Business School

Preface

Dear Participants, Colleagues and Friends,

WELCOME to the 2016 International Symposium on Business and Industrial Statistics. We are delighted to celebrate the continued success of our society, and its work to draw together the international community of statisticians, both academics and industry professionals, who share our goal of making statistics the foundation for decision making in business and related applications.

Barcelona is a great city, and the host institution has been gracious and generous. We hope you will have a wonderful time and enjoy a productive conference.

Plenary Speakers

Marian Farah, The Climate Corporation
David Ríos Insua, Rey Juan Carlos University
Henry Wynn, London School of Economics and Political Science

Invited Speakers (Approximately)

Víctor Aguirre Torres - ITAM, Mexico
Gopalkrishnan Asha - Cochin University of Science and Technology, India
Louis Aslett - University of Oxford, UK
Olushina Olawale Awe - Obafemi Awolowo University, Nigeria
David Banks - Duke University, USA
Sanjib Basu - Northern Illinois University, USA
Souhaib Ben Taieb - Monash University, Australia
Sotirios Bersimis - University of Piraeus, Greece
Paulo Canas Rodrigues - Federal University of Bahia, Brazil
Mayerly Cano Arroyave - Universidad Nacional de Colombia, Colombia
Franco Caron - Politecnico di Milano, Italy
Hatice Oncel Cekim - Hacettepe University, Turkey
Cristina Corchero - Catalonia Institute for Energy Research,
Spain Marcel - Dettling Zurich University of Applied Sciences, Switzerland
Dipak Dey - University of Connecticut, USA
Tahir Ekin - Texas State University, USA
Alberto Ferrer-Riquelme - Universidad Politécnica de Valência, Spain
Luca Frigau - University of Cagliari, Italy
Jairo Fúquene - University of Warwick, UK
Jesus Garcia - UNICAMP, Brazil
Carolina García Martos - Universidad Politécnica de Madrid, Spain
Veronica Gonzalez-Lopez - UNICAMP, Brazil
Yannig Goude - EDF, France

Julieth Verónica Guarín Escudero - Universidad Nacional de Colombia, Colombia
Yves Laurent Grize - University of Applied Sciences, Switzerland
Jane L. Harvill - Baylor University, USA
Elisa Henning - Santa Catarina State University, Brazil
Javier Heredia Cervera - Universitat Politécnica de Catalunya. Barcelona TECH, Spain
Scott Holan - University of Missouri, USA
D.S. Hooda - GJ University of Science Technology, India
Tatsuya Ishikawa - IBM Research - Tokyo, Japan
Daniel Jeske - University of California Riverside, USA
Lao Kenao – SAKSS-Togo
Roselinde Kessels - University of Antwerp, Belgium
Ozan Kocadagli - Texas A M University, USA
Jitendra Kumar - Central University of Rajasthan, India
Debasis Kundu - IIT Kanpur, India
Karel Kupka - Trilobyte Statistical Software Ltd., Czech Republic
Paul Kvam - University of Richmond, USA
Wee-Yeap Lau - University of Malaya, Malaysia
Dennis Lin - Pennsylvania State University, USA
Hedibert Lopes - Insper, Brazil
Zagorka Lozanov-Crvenkovic Novi Sad University, Servia
Wendy Lou - University of Toronto, Canada
Francisco Louzada - University of São Paulo, Brazil
Rahim Mahmoudvand - Bu Ali Sina University, Iran
Suresh Malik - M. D. University Rohtak, India
William Meeker - Iowa State University, USA
Wanli Min - Alibaba, People's Republic of China
Fatemeh Nasiri - Mellat Insurance Company, Iran
Julie Novak - IBM Yorktown, USA
Amilcar Oliveira – Aberta University and CEAUL, Portugal
Teresa A. Oliveira – Aberta University and CEAUL, Portugal
Victor Leiva – Universidad Adolfo Ibáñez, Chile
Maria Ivette Gomes – Faculdade de Ciências, Universidade de Lisboa, Portugal
Olegbemi Olujimi – TNS RMS
Gamze Ozel - Hacettepe University, Turkey
Victor Pena Pizarro - Duke University, USA

Vivek Raich - Government Holkar Science College, India
Balaji Raman - Cogitaas AVA, India
Nalini Ravishanker - University of Connecticut, USA
Fabrizio Ruggeri - CNR IMATI, Italy
Luigi Salmaso - University of Padova, Italy
Ismael Sánchez - University of Piura, Peru
Josep Sánchez - Universitat Politécnica de Catalunya. Barcelona TECH, Spain
Ken Sejling - Novo Nordisk, Denmark
Ananda Sen - University of Michigan, USA
Mathieu Sinn - IBM Research, Ireland
Nozer Singpurwalla - City University of Hong Kong, Hong Kong
Ehsan Soofi - University of Wisconsin-Milwaukee, USA
Refik Soyer - George Washington University, USA
Ansgar Steland - RWTH Aachen University, Germany
Dirk Surmann - Technical University of Dortmund, Germany
Carlos Trucíos - University of Campinas, Brazil
Willy Ugaz - Universidad Carlos III de Madrid, Spain
Martina Vandebroek - KU Leuven, Belgium
Alan Vazquez Alcocer - University of Antwerp, Belgium
Marina Vives Mestres - Universitat de Girona, Spain
Hongxia Yang - Yahoo!, USA
Emmanuel Yaschin - IBM, USA
Jian Zou - Worcester Polytechnic University, USA

Table of Content

KEYNOTE LECTURES

A Framework for Risk Analysis in Aviation Safety

David Rios Insua, Insitute of Mathematical Sciences, ICMAT-CSIC

Abstract: Aviation is a key industrial sector for global development. Safety is essential for its healthy growth and sustainability. However, its management is pervaded by simplistic methods mostly based on risk matrices. We provide a framework to support risk management decisions in aviation safety at state level and briefly describe RIMAS an architecture implementing the framework.

Digitizing Agriculture: A Tech Revolution to Feed the World

Marian Farah, The Climate Corporation

Abstract: Exploding population growth, rising incomes, and increasing dietary energy consumption are driving an unprecedented demand for food around the world. To keep up with this demand, crop production needs to increase dramatically. Precision agriculture is emerging as a viable solution that makes modern crop production more efficient. It derives optimized field-level management decisions from increasingly rich agricultural data. Given recent technological advancements, which have revolutionized the collection, access, and fast processing of data in agriculture, statisticians and machine learners have a unique opportunity to make a real impact on helping the world's farmers sustainably protect and improve their crop yields. Modeling, forecasting, and uncertainty quantification of key processes that impact crop yields are becoming significant areas of research. This talk includes examples of how The Climate Corporation combines hyper-local weather monitoring, high-resolution weather simulations, agronomic modeling, and remotely sensed measurements to deliver digital tools that help farmers improve productivity by making better informed operating decisions. The company's unique technologies help the global $3 trillion agriculture industry to stabilize and improve yields and, ultimately, help feed the world.

Smooth Supersaturated Models

Henry Wynn[1], Hugo Maruri-Aguilar[2], Peter Curtis[3]
[1]London School of Economics and Political Science
[2]Queen Mary University of London
[3]Ron Bates, Roll-Royce, plc Derby, UK)

Abstract: Super saturated (polynomial) models are those for which there are more basis terms than experimental design points. SSM is the subclass where the extra model degrees of freedom are used to optimally increase the smoothness of the models. As more basis terms are added SSM behave like splines, but, being polynomial, are more tractable analytically and can be defined for any region. The models are competitive with kriging for computer experiments and make sensitivity analysis straightforward. They go some way to solving the problem of finding optimal experimental designs for splines for arbitrary regions and for models constrained along sub-varieties.

ORGANIZED SESSIONS

SESSION A1: Flexible Modeling for Competing Risks

CHAIR: **Gopalkrishnan Asha**, Cochin Univ. of Science & Technology, nakannan@nsf.gov

Flexible Modeling of Competing Risks and Limited Failure Rate

Sanjib Basu, Northern Illinois University

Abstract: The cumulative incidence functions based approach to competing risks modeling has the advantage of providing direct inference on the failure probabilities from each risk. A unified competing risks limited failure model is proposed in this work where the cumulative incidence functions of the competing risks are directly modeled. The proposed model further accounts for the possibility of limited failure from one or more of the competing risks. Bayesian analyses of these models are explored, and conceptual, methodological and computational issues related to Bayesian model fitting and model selection are discussed. The performance of the proposed model is investigated in simulation studies and real data. (This is joint work with Qi Jiang, Northern Illinois University).

Analyzing Left Truncated, Right Censored Competing Risks Data

Debasis Kundu, IIT Kanpur, India

Abstract: In this talk I will be talking about the analysis of left truncated right censored competing risks data, based on Cox's latent failure time model assumption. It is assumed that the lifetime distributions of the latent causes of failure follow Weibull distribution with the same shape parameter but different scale parameters. Maximum likelihood estimators of the unknown parameters cannot be obtained in closed form, and we propose to use the expectation maximization algorithm to compute the maximum likelihood estimates. Bayesian inference of the unknown parameters are obtained based on the assumption that the shape parameter has a log-concave prior density function, and for the given shape parameter, the scale parameters have Dirichlet-Gamma priors. We propose to use importance sampling procedure to compute the Bayes estimates, and also to compute highest posterior density credible intervals. Monte Carlo simulations are performed to compute the performance of the different estimators, and one data analysis has been performed for illustrative purposes. This is joint work with Debanjan Mitra and Ayon Ganguly.

On Dependent Competing Risks

Ananda Sen, University of Michigan

Abstract: The topic of analyzing time-to-event data where individual units are subjected to multiple causes for the event occurrence has been well-studied for decades. In this framework a particular case that has received lion's share of attention is when the event is caused by the earliest onset of a cause, known as the case of competing risks. Earlier work in competing risks analysis utilized a series system (observing the minimum of several lifetimes) formulation in terms of latent event times. It is well known that such a formulation is fraught with the issue of identifiability, unless one can assume the different causes to act independently. In recent times, substantial efforts have been made to formulate a model that has direct links to the observables and avoids imposing a dependence structure on the causes. Such formulation, however, fails to understand the role dependence among the causes plays on differential association between cause specific life times and prognostic factors or prediction of future time to event. This talk will focus on some structural dependence on the competing risks and its consequence on the model inference. Spotlight will be on recurrent events, application of which is abundant in industrial and biomedical context.

SESSION B1: Big Data Analytics in the Technology Industry

CHAIR: **Jon Hosking**, Amazon USA

Estimating Rates of Rare Events through a Multidimensional Dynamic Hierarchical Bayesian Framework

Hongxia Yang, Yahoo!

Abstract: We consider the problem of estimating occurrence rates of rare events for extremely sparse data using pre-existing hierarchies and selected features to perform inference along multiple dimensions. In particular, we focus on the problem of estimating click rates for {Advertiser, Publisher, User} tuples where both the Advertisers and the Publishers are organized as hierarchies that capture broad contextual information at different levels of granularities. Typically, the click rates are low and the coverage of the hierarchies and dimensions is sparse. To overcome these difficulties, we decompose the joint prior of the three-dimensional Click- Through-Rate (CTR) using tensor decomposition and propose a Multidimensional Hierarchical Bayesian framework (abbreviated as MadHab). We set up a specific framework of each dimension to model dimension-specific

characteristics. More specifically, we consider the hierarchical beta process prior for the Advertiser dimension and for the Publisher dimension respectively and a feature-dependent mixture model for the User dimension. Besides the centralized implementation, we propose two distributed algorithms through MapReduce and Spark for inferences which make the model highly scalable and suited for large scale data mining applications. We demonstrate that on a real world ads campaign platform our framework can effectively discriminate extremely rare events in terms of their click propensity.

Quality and Process Monitoring for Mapping the World

Angela Schoergendorfer, Google Inc., Mountain View, CA, USA

Abstract: Google Maps is striving to map the entire world. This is a complex mission due to the scale of the problem and the challenges in merging data from a variety of sources, not all of them with high quality. As the real world is continuously changing, and as new sources of bad information (like spammers) evolve, the challenges that the Google Maps team faces are never ending.

Measuring and tracking the quality of the Maps data repository is a challenge in itself, as it requires benchmarking against the real world. This talk will outline how the Geo data analytics team is approaching this problem, by integrating small-scale studies on data quality and process reliability with large-scale monitoring and Bayesian modeling.

SESSION C1: Six Sigma

Org: **Birger Madsden**, Novozymes, A/S
Chair: **Marina Vives Mestres**, Universitat de Girona, marina.vives@udg.edu

Revisiting Six Sigma: Its Advantages and Disadvantages

Vladimir Zaiats, Universitat de Vic , Universitat Central de Catalunya, vladimir.zaiats@uvic.cat

Abstract: Six Sigma is well-known to be a methodology for improving quality by analysing data using statistical tools, aiming at detecting the causes of quality problems and implementing adequate controls. In vast bibliography on the subject some criticism on the lack of an effective guide to implementation of Six Sigma can be found. Our aim is to try to bring together the latest developments in the subject, in order to understand how Six Sigma can effectively be used.

Going Beyond Six Sigma – Global Data Insight & Analytics at the Ford Motor Company

Thomas Hochkirchen, Ford Motor Company, thochkir@ford.com

Abstract: In this talk, exciting developments within Ford Motor Company will be described – the creation of a new global skill team devoted to the systematic use of data and analytical methods to support better business decisions.We will describe how the ground for this development was paved within the corporate mindset – from the "Whiz Kids" in the 1940s via Deming and, of course, application of Six Sigma at large scale.

Recent Advances in Design of Computer Experiments

Dennis K. J. Lin, University Distinguished Professor of Statistics, The Pennsylvania State University, USA. *DKL5@psu.edu*

Abstract: Computer models have became a routine practice for understanding complicated physical phenomena. Specially-designed experiment is required to run these computer experiments much more efficiently. Space-filling designs, such as Uniform Design or Latin hypercube (LHC) designs have recently found wide applications in running computer experiments. However, the original construction of LHCs by mating factors randomly is susceptible to having potential correlations among input factors. It is thus desirable to have an orthogonal Latin hypercube design. A series of orthogonal LHC have been constructed to be suitably applied to various types of computer models. This includes *regular (first-order and second-order) orthogonal LHC, nested orthogonal LHC, sliced orthogonal LHC, uniform sliced LHC,* as well as *orthogonal LHC for computer models with both qualitative and quantitative variables.* Recent developments on these newly constructed designs will be reviewed and discussed, from both theoretical and application perspectives.

This talk is based upon some initial results of my long time collaboration efforts with a computer experiment research team at Nankai University (Tianjing, China), led by Professors Minqian Liu and Jianfeng Yang. Their efforts must be acknowledged.

SESSION A2: Modeling Electricity Demand

Org/Chair: **Bei Chen**, IBM Ireland

Forecasting Algorithms for Hierarchical Electricity Demand Data

Souhaib Ben Taieb, Monash University

Abstract: Forecasting electricity demand is critical for electric utilities in order to undertake appropriate planning of generation and distribution. Recently, the large-scale deployment of smart electricity meters has made available a large amount of time series data representing household electricity demand at intervals from 1 minute to one hour. Electricity demand forecasts at the household level can be particularly useful for evaluating demand response programs as well as for improving forecasts at aggregated levels. Time series of electricity demand can often be represented in a hierarchical or grouped structure. For example, the electricity demand for a whole country can be disaggregated by states, cities, regions and homes. In order to allow consistent decisions over different levels of the hierarchy, the forecasts for the disaggregated series are usually required to add up exactly to the forecasts of the aggregated series, a constraint known as aggregate consistency. The aggregate consistent forecasts can be computed by first forecasting all series at all levels of the hierarchy. Then a regression procedure can be used to reconcile these forecasts, where the estimated coefficients represent the aggregate consistent forecasts. However, since electricity demand is positive, the reconciliation procedure must guarantee the non-negativity of the estimated coefficients. I will present and compare the performance of different point forecasting algorithms using an electricity smart meter dataset, and discuss some challenges in reconciling probabilistic forecasts.

Ensemble Methods for Energy Forecasting

Yannig Goude, Chef de projet Statistiques pour le management d'énergie-Expert Prévision, Électricité de France (EDF), France

Abstract: Ensemble methods are very popular machine learning methods, well known to achieve good results on very different data sets at a relatively low cost in terms of modelization effort for the statistician. We propose here to compare different recent ensemble methods on different energy data sets. We consider both off-line and on-line prediction framework.

Modelling Electricity Demand in Smart Grids: Data, Trends and Use Cases

Mathieu Sinn, Research staff member and Manager Exploratory Predictive Analytics, IBM Research, Ireland

Abstract

SESSION B2: Reliability I

Org/Chair: **Nalini Ravishanker**, University of Connecticut

Wear of Cylinder Liners in Ships: One Dataset, Many Models

Fabrizio Ruggeri, IMATI

Abstract: The talk will present data about wear of cylinder liners in ships and a selection of the models which have been used so far to model such process.

Bayesian Reliability Analysis in Dynamic Environments

Refik Soyer, George Washington University

Abstract: In this talk we consider systems operating under a dynamic environment which causes changes in the failure characteristics of the system. We discuss different modeling strategies to describe the evolution of the dynamic environment and develop Bayesian analysis of the models using Markov chain Monte Carlo methods and data augmentation techniques. We present illustrations from repairable systems using data from software testing and rail road track maintenance.

Estimating a Parametric Component Lifetime Distribution from a Collection of Superimposed Renewal Processes

William Meeker, Iowa State University
Co-Authors: Wei Zhang, Ye Tian, and Luis Escobar

Abstract: Maintenance data can be used to make inferences about the lifetime distribution of system components. Typically a fleet contains multiple systems. Within each system there is a set of nominally identical replaceable components of particular interest (e.g., two automobile head- lights, eight DIMM modules in a computing server, sixteen cylinders in a locomotive engine). For each component replacement event, there is system-level information that a component was replaced, but not information on which particular component was replaced. Thus the observed data is a collection of superpositions of renewal processes (SRP), one for each system in the fleet. This paper proposes a procedure for estimating the component lifetime distribution using the aggregated event data from a fleet of systems. We show how to compute the likelihood function for the collection of SRPs and provide suggestions for efficient computations. We compare performance of this incomplete-data ML estimator with the complete-data ML estimator and study the performance of confidence interval methods for estimating quantiles of the lifetime distribution of the component.

SESSION C2: Statistical Methods in Medicine

Org: **Daniel Jeske**, University of California–Riverside
Chair: **Luca Frigau**, University of Cagliari

Applications of Compound Patterns for Adaptive Monitoring Schemes

Wendy Lou, University of Toronto

Abstract: Motivated by practical applications involving sequential monitoring for data quality requiring timely decision making, statistical approaches based on runs and patterns will be the focus of the presentation. The challenges associated with practical issues will be discussed first, followed by possible statistical solutions incorporating flexible strategies. Real examples, including for biological samples and infectious diseases, will be given to illustrate the methodology.

Performance Monitoring in Health Services

Sotirios Bersimis, University of Piraeus

Abstract: Health and health service monitoring is among the most promising research area today and the world work towards efficient and cost effective health care. This paper deals with monitoring health service performance using more than one performance outcome variable (multi-attribute processes), which is common in most health services. Although monitoring whether a health service changes or improves over time is important this is well covered in the current literature. Therefore this paper focuses on comparing similar health services in terms of their performance. The proposed procedure is based on an appropriate control chart. The paper deals with firstly the case when no risk adjustment is required because the health services being compared treat the same patient case-mix which does not vary over time. Secondly it deals with comparing health services where risk adjustment is required because the patient case-mix they service do differ because they service either very different geographical locations or service very different demographics of the same population.

Maximizing the Usefulness of Statistical Classifiers for Two Populations with Some Illustrative Applications

Daniel Jeske, University of California–Riverside

Abstract: The usefulness of two-class statistical classifiers is limited when one or both of the conditional misclassification rates is unacceptably high. Incorporating a neutral zone region into the classifier provides a mechanism to refer ambiguous cases to follow-up where additional information might be obtained to clarify the classification decision. Through the use of the neutral zone region, the conditional misclassification rates can be controlled and the classifier becomes useful. Three real-life examples, including applications to prostate cancer and kidney dysfunction following heart surgery, are used to illustrate how neutral zone regions can salvage some utility from disappointing classifiers that would otherwise be completely abandoned. [This is joint work with Dr. Steven Smith, at City of Hope National Medical Center, Duarte, California.]

SESSION A3: Project Management

Org/Chair: **Fabrizio Ruggeri**, IMATI

Project Management: A Bayesian Approach

Franco Caron, Politecnico di Milano

Abstract: A reliable "Estimate at Completion" from the early stage of project execution is essential in order to enable efficient and proactive project management. The non-repetitive and uncertain ature of projects and the involvement of multiple stakeholders require the use and integration of multiple informative sources in order to provide accurate forecasts. The paper deals with multiple objectives: introducing the need for the identification and utilization of all the available knowledge in order to improve the forecasting process; developing a Bayesian approach in order to integrate the diverse knowledge sources; exploring the integration of data records and experts' judgment related to the ongoing project; exploring the integration of data records related to projects completed in the past and to the ongoing project and finally developing a Bayesian model capable of using three different knowledge sources: data records and experts' judgments related to the ongoing project and data records related to similar projects completed in the past. The model has been tested in a set of large

and complex projects in the Oil&Gas industry. The results show a higher forecasting accuracy of the Bayesian model compared to the traditional Earned Value Management (EVM) methodology.

Project Risk Management under Dynamic Environments

Janne Kettunen, George Washington University
Co-authors: Fabrizio Ruggeri and Refik Soyer

Abstract: 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.

Organization and Coordination of Project Actors in order to Cope with Complexity-Related Phenomena

Hadi Jaber, University Paris-Salary, CentraleSupelec, Laboratoire Génie Industriel
Franck MARLE, University Paris-Salary, CentraleSupelec, Laboratoire Génie Industriel
Ludovic-Alexandre VIDAL, University Paris-Salary, CentraleSupelec, Laboratoire Génie Industriel

Abstract: We introduced a clustering methodology to propose groups of actors in new product development projects, especially for the actors involved in many deliverable-related interdependencies in different phases of the project life cycle. This permits an increase in the coordination between interdependent actors who are not always formally connected via the hierarchical structure of the project organization. This allows the project organization to be actually closer to what a networked structure should be. Since the clustering approach encourages people to

meet together and communicate/coordinate better, we consider that the overall communication/coordination performance improvement is proportional to the performance of our algorithms. Indeed, the amount of interactions within the clusters (which is maximal) is a factual parameter. It determines a maximum potential for communication and coordination within clusters and a minimum risk of non-communication and/or lack of coordination at the interfaces between clusters.

Our contribution is a three-stage process for clustering a network of project elements. The first stage is information gathering, data input and parameter definition. The second stage consists of running each algorithm many times with several problem configurations. Afterwards, a number of clustered solutions is obtained, with quality indicators for each solution and for each cluster in the solution. In addition, a frequency analysis is done to indicate the number of times that each couple of elements (i.e., actors in our case study) were put together in a clustered solution. The idea is that the more often pairs of actors are proposed together in the different configurations, then the more robust the decision of putting them together in the final solution is. The third stage is the post processing of the obtained results; this is done by combining extractions of particular clusters or pieces of clusters from different solutions. This combination is based on the quality indicators and the frequency analysis on the results (i.e., the number of times the couple of actors were put together). A hybrid solution, which meets, at best, the needs of the decision maker, is built using a mix of best clusters from all configurations. This approach has been illustrated through actual data in a new product development project in the automotive industry, by grouping people according to interdependencies, changing more or less the way that actors were initially organized.

SESSION B3: Time Series in Industry

Org: **Emmanuel Yashchin**, IBM
Chair**: Jane L. Harvill**, Baylor University

Monotoring High-Frequency Time Series for Quality Control

Ansgar Steland and Annabel Prause

Institute of Statistics, RWTH Aachen University

Abstract: We study a nonparametric monitoring procedure, namely a control chart based on a nonparametric estimator for the underlying signal. The estimator is derived from the cardinal series associated to a bandlimited signal, which is behind the classical theory that such signals can be exactly reconstructed from their sampled function values. In the presence of random noise one has to post-filter the series to obtain a statistically sound method. It is assumed that the signal is sampled equidistantly and we allow for a high-frequency sampling scheme (infill asymptotics) which fills the nonnegative time axis asymptotically. Serial correlations are taken into account by estimating the long run variance of the data stream by a VARMA approach.
The procedure is applied to real high-frequency data streams of logging data from a photovoltaic system, representing the current from eight strings of connected modules.
As such photovoltaic data has a complex structure and is difficult to model, we propose to analyze the differences with respect to a reference string (gold standard).

Bayesian Time Series Forecasting for Hierarchically Structured Organizations

Julie Novak, IBM Yorktown

Abstract: An important task for any large-scale business is to prepare forecasts of business metrics, such as revenue, cost, and event occurrences, at different time horizons (e.g. weekly or quarterly intervals). Often these business organizations are structured in a hierarchical manner by line of business, division, geography, product line or a combination thereof. In many situations projections for these business metrics may have been obtained independently and for each level of the hierarchy. The problem with forecasts produced in this way is that there is no guarantee that forecasts are aggregate consistent according to the hierarchical structure of the business, while remaining as accurate as possible. In addition, it is often important for the organization to achieve accurate forecasts at certain levels of the hierarchy according to the needs of users. We propose a Bayesian hierarchical method that will treat the "base" forecasts (those which were initially provided) as observed data which are then updated and obey the hierarchical organizational structure. In addition, by leveraging the prior covariance matrix, we are able to set up a heterogeneous loss function to obtain higher accuracy at the levels prescribed by the user. 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.

SESSION C3: Reliability II

Chair: **Roman Viveros-Aguilera**, McMaster University

Cryptographically Secure Multiparty Evaluation of System Reliability

Louis J. M. Aslett, University of Oxford

Abstract: A company may consider the design (structure) of their engineered system to be a trade secret and so be unwilling to release it to component manufacturers, while at the same time componente manufacturers are frequently unwilling to release anything more than mean-time-to-failure figures for components. These two opposing goals lead to a situation in which it would seem unrealistic to achieve a full probabilistic reliability assessment and to honour the privacy requirements of all parties. However, this talk will present recente developments in cryptography which, when combined with recent advances in reliability theory, allows almost total privacy to be maintained. Thus, the system designer does not have to reveal their trade secret design and the manufacturer can retain component test data in-house, yet a full Bayesian posterior predictive system survival curve can still be constructed.

Modeling Load Sharing Systems with Frailty

Gopalkrishnan Asha, Cochin University of Science & Technology

Abstract: The concept of load sharing with frailty and observed covariates is very much of interest to engineers, biologists, statisticians working in the area of reliability and survival analysis. The main idea of the present paper is to introduce a general class of bivariate distributions using proportional hazards models to model a two component load sharing system using the same idea of Freund (1961,Journal of the American Statistical Association, 56(296):971-977) incorporating frailty and covariates. As a particular case we study a model using positive stable frailty with covariates and cumulative baseline hazard as bivariate Weibull distribution (Lu, 1989 IEEE Transactions on Reliability, 38(5):615-619). Various reliability properties and characterizations of the proposed model are presented. Estimation procedures are developed. Simulation studies are carried out to show the efficiency of our estimation procedures. Finally we analyze a load sharing industry data using this model and state our conclusions.

What effects of public expenditure in the framework of comprehensive Africa agriculture development programme (CAADP)? Case of Togo from 2010 to 2013

Lao KENAO, Lome, TOGO - kenaolao@yahoo.fr, Ministry of agriculture, Coordinator of National Strategic Analysis and Knowledge Support System (SAKSS-Togo)

Abstract: In this article, I analyze the effect of the government's efforts in the framework of CAADP for implementing its National Agriculture and Food Security Investment Program (NAFSIP) using COFOG methodology, calculation of Gini index and elasticies relative to expenditure. Results point out that an evolution of Maputo ratio between 2010 et 2013 is noticed though there are a flexing en 2011. Public support at crop production has more important effect on agriculture growth comparatively to others sub-sectors but dwells very vulnerable of climate change. The analysis of elasticities show that when agriculture public expenditure increases of 1%, the crop's production increases of 2.27% and induces agriculture growth of 0.59%. The effect in the attempt of the respect of Maputo's declaration is remarkable as much as in average 6.8% public allocated resources of State for the agriculture permitted to reache an agriculture growth of 4.1%. However, the share of agriculture public expenditure for the research dwells greatly feeble and therefore to ensure agriculture development, food security and poverty reduction require a national système of agriculture research right developed and a bearing of the levels of investment and adequate humans capacities.

Keywords: Effect, ratio of Maputo, CAADP, Agriculture growth

SESSION A4: ASMBI Special Session

Org/Chair: **Fabrizio Ruggeri**, IMATI

Discussant: Refik Soyer, The George Washington University
Discussant: Hedibert Lopes, Insper

Customer Level Modeling of Multivariate Count Time Series

Nalini Ravishanker, University of Connecticut, Storrs

Abstract: Discrete-valued time series modeling is emerging as an important area for many applications, as discussed in the recent CRC *Handbook of Discrete-valued Time Series*. Specifically, there is increasing interest in modeling univariate and multivariate time series of counts responses on several subjects as a function of subject-specific and/or time-dependent covariates. This talk presents a Bayesian framework for estimation and prediction by assuming a multivariate Poisson sampling distribution for the count responses and by fitting a hierarchical dynamic model which incorporates the temporal dependence as well as dependence between the components of the response vector. We illustrate this on ecology data to model count responses on different gastropod species. We also discuss a level correlated model (LCM) which enables us to account for association among the components of the response vector, possible overdispersion, and allows us combine different marginal count distributions and to build a hierarchical model for the vector time series of counts. We discuss the use of R-INLA for fast implementation of this flexible framework. We illustrate this on a marketing data set, by modeling the monthly prescription counts by physicians of a focal drug from a multinational pharmaceutical firm along with monthly counts of other competing drugs with sizable market share for the same therapeutic category. This is joint work with Volodymyr Serhiyenko and Rajkumar Venkatesan.

SESSION B4: Algorithms

Chair: **Olushina Olawale Awe**, Obafemi Awolowo University

Continuous Learning Algorithm on Skewed Data for Optimal Media Purchase

Balaji Raman, Cogitaas AVA

Abstract: A consumer appliance company has been buying media spots for direct selling. Spots are purchased every month. The observed data is not only fractional but also predicted to be biased. The MER (media efficiency ratio) is low, showing that spots bought in the past may have a bias towards weaker sales. Moreover, the MER values were highly volatile along multiple axes of day, time and channel. The business question was to increase spend efficiency by focusing on spots with high likelihood of sales. Problems of assigning data structures and then making inferences on all factors that drive MER, required solutions at two levels. For data structuring, the key component of our algorithm is homogeneous MER clusters. Clustering based on scale mixture of skew normal yielded significantly lower misclassification error rate of MER when compared to applying K-means or mixture of normal distributions. MER clusters were modelled as a function of channels, time and day along with other derived features. An ensemble of models comprising naïve Bayes, adaptive boosting, neural networks and random forests was developed. This methodology allowed for a search process of fundamental relations between variables for making inferences. A wide range of methods was also used since we did not have full data on required key variables to determine a single set of inference. Inferences were drawn on all possible combinations of input variables – and recommendations were provided on which

channels to focus? Which time spots to air infomercial? What is the role of advertising stock over time? Combination of new channels and time spots previously not considered.

Our algorithm has been successful in identifying spots skewed to stronger sales. It is being progressively applied with striking business results for a very large US corporation's Japanese arm. MER, which is ratio of sales to spend has risen by 40% in three months and % achieved the metric value required % for profitability.

iSports: A Web-Oriented Statistical Expert System for Talent Identification in Soccer

Francisco Louzada, University of Sao Paulo

Abstract: Nowadays soccer is the most practiced sport in the world and moves a multimillionaire business. Therefore, a club that is able to recruit and develop talented players to theirs fullest potential has a lot of advantages and economic benefits. However, in most clubs the players are selected through scouts and coaches recommendation, with predictive success based mostly on intuition than other objective criteria. In addition, it is known that talent development and identification is a multifactorial process involving many characteristics. To this end, this paper proposes the creation of performance indicators based on multivariate statistical analysis, which includes principal components and factor analysis, performed to construct physical, technical and general score, and copula modelling, proposed to create a consistency index, which generalizes the Z score method. With these indicators, a web-oriented statistical expert system for analyzing sport data in real time via R software is proposed as a powerful tool for talent identification in soccer. This system, the so called iSports, allows the monitoring and continuous comparison of athletes in a simple and efficient way, taking into account essentials aspects, as well as identifying candidate talented that have above the average performance, that is, who stand out from the studied population of soccer players. In order to promote and popularize the access of information and the statistical science applied in the sports context, the iSports system can be used in any training center, impacting the increase of knowledge of the athletes in training phase at any school, city or region. Real data on a Brazilian sport data is used to illustrate the iSport features.

SESSION C4: Design of Experiments

Org: **Birger Madsden**, Novozymes, A/S
Chair: **Martina Vandebroek**, KU Leuven

Using DOE in the Development of Pharma Products

Ken Sejling, Novo Nordisk

Abstract: This talk gives a presentation of how statistical specialists have used different ways of spreading and encouraging the use of statistical analyses and design of experiment methodology in the CMC development organization of a pharmaceutical company. To motivate and obtain an increased use of statistics it has been necessary to play on several strings ranging from teaching applied statistics along with practical exercises in the use of statistical software (JMP®), on-site statistical consultancy, writing statistical guidelines for recurring tasks, building network of statistically experienced chemists and documenting statistical results in reports reviewed by colleagues.

Design of Discrete Choice Experiments in the Presence of Many Attributes and a No-Choice Option

Roselinde Kessels, University of Antwerp

Abstract: We show how to optimally design a discrete choice experiment (DCE) with a no-choice option for estimating a nested logit model when a large number of attributes are under study. As motivating example, we present a DCE to identify and quantify the determinants that influence the competitive position of the coach bus as transport mode for medium-distance travel by Belgians. We measured the attractiveness of different bus services for different destinations (Lille, Amsterdam, Cologne, Paris and Frankfurt) by having participants choose their preferred bus trip out of two bus trips, while still allowing them to also choose not to take the bus but any other transport mode comprised by the no-choice option. Each bus trip is a combination of levels of seven attributes: price, duration, and comfort attributes including wifi, leg space, catering, entertainment and individual power outlet. Varying the levels of all seven attributes of the bus trips in the choice sets for respondent evaluation would be cognitively too demanding. To reduce the complexity of the choice tasks, we created an optimal design with partial profiles in which the levels of only four of the seven attributes vary within every choice set. The levels of the other three attributes are kept constant, but are still shown to the respondents to reflect real-world bus trip alternatives.

Extending Definitive Screening Designs by Concatenation

Alan Vazquez Alcocer, Peter Goos, Eric Schoen

Abstract: Definitive screening designs permit the screening of m quantitative factors in 2m+1 runs. The main effects are orthogonal to each other and to quadratic effects and two-factor interactions, while second-order effects are never fully aliased. Unfortunately, the performance of a standard definitive screening design can deteriorate when more than just a few effects are active. To alleviate this problem, we concatenate two definitive screening designs. The concatenated design improves the good statistical features of its parent designs. The concatenation employs an algorithm that minimizes the aliasing among pairs of second-order effects using foldover techniques and column permutations for one of the parent designs. We study the statistical properties of the new definitive screening designs and compare them to the best alternatives in the literature. The resulting designs bridge the gap between the ordinary definitive screening designs and traditional response surface designs.

Key Words: Three-level designs, Local search, Conference matrix

SESSION A5: Time Series II

Chair: **Jian Zou**, Worcester Polytechnic University

Spline Backfitted Kernel Forecasting for Functional-Coefficient Autoregressive Models

Jane L. Harvill (Baylor University), Joshua D. Patrick (Baylor University) and Justin Sims (Francis Marion)

Abstract: We propose three methods for forecasting a time series modeled using a functional coefficient autoregressive model (FCAR) fit via spline-backfitted local linear (SBLL) smoothing. The three methods are a ``naive" plug-in method, a bootstrap method, and a multistage method. We present asymptotic results of the SBLL estimation method for FCAR models and show the estimators are oracally efficient. The three forecasting methods are compared to local linear forecasts through simulation. We find that the bootstrap method outperforms the other two methods under the assumption that the coefficient functions are second-order and Lipschitz continuous and the series length and the order of the model are small. In all other circumstances, we find the naive method is preferable due to its performance in prediction error and computational speed. We apply the naive and multistage methods to solar irradiance data and compare forecasts based on our method to those of a linear AR model, the model most commonly applied in the solar energy literature. The ability to

accurately forecast irradiance greatly improves utility-scale plants to manage the sources of supply, and helps reduce energy costs.

On Volatility Ranking of Time-Varying Parameters in Dynamic Linear Models

O. Olawale Awe[1] and A. Adedayo Adepoju[2]
[1]Department of Mathematics (Statistics Unit), Obafemi Awolowo University, Ile-Ife, Nigeria.
[2]Department of Statistics, University of Ibadan, Ibadan, Nigeria.

Abstract: In the estimation of discount-weighted dynamic linear models, the choice of the evolution variance often play important roles in forecasting as it enables the computation of the onestep-ahead mean squared prediction error vectors. In order to reduce the complexity often involved in estimating the evolution variance, which plague the existing literature, it is usually discounted. In this paper, we present a recursive Bayesian algorithm for estimating and choosing optimal discount of values for the evolution variance simultaneously. Our algorithm makes optimal choice by cross-validating the discount value with the Mean Squared Prediction Error (MSPE) via a grid search. In our empirical analyses, we found that, for a range of simulated time series, the proposed algorithm estimated time-varying parameters with discount values (lambda) of the evolution variance in the range (0.50 <= lamda <= 0.75 <= lambda <=0.99) which can be approximated to 1. These optimal discount values were used to rank the volatility of the parameters associated with oil and gold prices over time.

A New Parsimonious Vector Forecasting Model in Singular Spectrum Analysis

Rahim Mahmoudvand
Department of Statistics, Bu-Ali Sina University, Hamedan, Iran
E-mail: r.mahmodvand@gmail.com

Abstract: In this paper, we introduce a new algorithm for vector forecasting in the singular spectrum analysis. This algorithm enables SSA users to use two different values for the window length

parameter: one for reconstruction and another for forecasting. Our results on both real and simulated data support the idea of using this new algorithm.

SESSION B5: CRA-ISI Special Session – Quality Control and Risk

Org/Chair: **Teresa A. Oliveira**, Universidade Aberta

Inflated Pareto Processes in Statistical Quality Control

M. Ivette Gomes[1,3] , Fernanda Otília Figueiredo[2,3] and Adelaide Maria Figueiredo,[2,4,3]
[1]Universidade de Lisboa, Faculdade de Ciências, Portugal
[2]Universidade do Porto, Faculdade de Economia, Portugal
[3]CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[4]LIAAD–INESC TEC Porto, Universidade do Porto, Portugal

E-mail addresses: ivette.gomes@fc.ul.pt ; otilia@fep.up.pt ; adelaide@fep.up.pt

Abstract: In food and other manufacturing industries, it is important to control the presence of some chemical substances in the raw material and, in some cases, along the different phases of its production. Indeed, such substances can strongly affect the quality of the final products. To measure the concentration of such substances, chromatography analyses are usually performed on samples of items taken from large batches. Next, and on the basis of the obtained measurements, we have to conclude about the absence or presence of such substances, and then to decide for the acceptance or rejection of the corresponding lots. However most of the chromatographs in use do not have sufficient precision to detect very low or high concentrations of such substances, and as a result, the data set of the measurements suggest an underlying inflated continuous distribution. In this work we highlight the adequacy of a heavy right-tailed parent, the inflated Pareto distribution, to model such type of data, and we define and evaluate acceptance sampling plans under this distributional setup. In a previous study, Figueiredo et al. [1] used the bootstrap methodology combined with Monte-Carlo simulations to evaluate the performance of complex acceptance sampling plans in the detection of chemical substances in lots of raw-material. Now, such evaluation is performed in terms of the probability of acceptance of the lots and its average outgoing quality level.

Keywords: Acceptance sampling; inflated Pareto distribution; statistical quality control; variables sampling plans.

Acknowledgements: Research partially funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the projects UID/MAT/00006/2013 and FCOMP-01-0124-FEDER-037281.

References

[1] Figueiredo, F., Figueiredo, A. and Gomes, M.I. (2014) . Comparison of Sampling Plans by Variables using the Bootstrap and Monte Carlo Simulations. AIP Conference Proceedings 1618, 535–538.

Control charts under measurement errors and non-normality with applications to pharmaceutical companies

Víctor Leiva[1], Helton Saulo[2], Juan Vega[3]

[1]Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile
[2] Institute of Mathematics and Statistics, Universidad Federal de Goiás, Brasil
[3] School of Industrial Engineering, Computing and Systems, Universidad de Tarapacá, Chile

Abstract: In statistics for quality, control charts are widely used to monitor the performance of production processes in companies. Observations of a same specimen often differ due to errors inherent in the measurement process, which occurs with frequency in the chemical industry and, particularly, in pharmaceutical companies. In addition, in practice, normality of the quality characteristic is rather a few common situation instead of being "the rule". However, practitioners of these charts in the pharmaceutical companies continue to ignore the lack of normality and measurement errors, which obviously leads to wrong decisions. In this paper, we present a control chart with measurement errors and apply it to real-world data of the pharmaceutical industry. Some issues about non-normality in control charts also are discussed.

Keywords: measurement error models, pharmaceutical processes, quality control, variance components

p-Charts for Attribute Control

Teresa A. Oliveira[1,2], Víctor Leiva[3]
[1]Universidade Aberta, Portugal
[2]CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[3]Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile

Abstract: In this talk we present charts for attribute control by means of the proportion (*p*) of defective items, named *p*-charts. Such charts are often used in statistical quality control to monitor the behavior of production processes by means of *p* over time. When the sample size (*n*) is constant, charts for atribute control related to the number (*np*) of defective items, named *np*-charts, are a good alternative to the *p*-charts. An update for *p*-charts and some recent ideas on the topic are provided considering life distributions. An implementation in the R software is discussed using examples.

Keywords: binomial distribution, life distributions, *p*-charts, R software, statistical attribute control

Acknowledgements: Research partially funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] Birnbaum, Z.W. & Saunders, S.C. (1969). A new family of life distributions, Journal of Applied Probability, 6, 319-327.
[2] Davis, D.J. (1952). An analysis of some failure data, Journal of American Statistical Association, 47, 113-150.
[3] Efron, B. & Tibshirani, R.J. (1993). An Introduction to the Bootstrap, Chapman and Hall, New York.
[4] Leiva, V., Soto, G., Cabrera, E. & Cabrera, G. (2011). New control charts based on the Birnbaum-Saunders distribution and their implementation, Colombian Journal of Statistics, 34, 147-176.
[5] Marshall, A.W. & Olkin, I. (2007). Life Distributions, Springer, New York.
[6] Montgomery, D.C. (2005). Introduction to Statistical Quality Control, Wiley, New York.
[7] Team (2014). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, available at URL www.R-project.org.
[8] Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control, R Journal, 4, 11-17.

np-Charts for Attribute Control

Amílcar Oliveira[1,2], Víctor Leiva[3]
[1]Universidade Aberta, Portugal
[2]CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[3]Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile

Abstract: In this talk we introduce charts for attribute control considering the number of defective items, named *np*-charts. These charts are preferable to the charts for proportion (*p*) of defective items, named *p*-charts, when the sample size (*n*) remains constant for all of subgroups. The benefits of using *np*-charts over *p*-charts are an easier interpretation and the fact that no calculation is required for each sample

result. We provide an update for *np*-charts and some recent ideas on the topic based on life distributions, as well as an implementation in the R software using examples with data on attributes and lifetimes.

Keywords: binomial distribution, life distributions, *np*-charts, R software, statistical attribute control

Acknowledgements: Research partially funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] Birnbaum, Z.W. & Saunders, S.C. (1969). A new family of life distributions, Journal of Applied Probability, 6, 319-327.
[2] Davis, D.J. (1952). An analysis of some failure data, Journal of American Statistical Association, 47, 113-150.
[3] Efron, B. & Tibshirani, R.J. (1993). An Introduction to the Bootstrap, Chapman and Hall, New York.
[4] Leiva, V., Soto, G., Cabrera, E. & Cabrera, G. (2011). New control charts based on the Birnbaum-Saunders distribution and their implementation, Colombian Journal of Statistics, 34, 147-176.
[5] Marshall, A.W. & Olkin, I. (2007). Life Distributions, Springer, New York.
[6] Montgomery, D.C. (2005). Introduction to Statistical Quality Control, Wiley, New York.
[7] Team (2014). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, available at URL www.R-project.org.
[8] Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control, R Journal, 4, 11-17.

Qr Codes structures: Algorithms, connections and applications

Carla Francisco[1] and Teresa A. Oliveira[1,2]

[1]Departamento de Ciências e Tecnologia, Universidade Aberta, Portugal
[2]CEAUL- Centro de Estatística e Aplicações da Universidade de Lisboa

Email adresses: carla.francisco.1@ulaval.ca ; teresa.oliveira@uab.pt

Abstract: In this paper, we review the Qr Codes structures and we approach an algorithm for the authentication Codes. Combinatorial designs play an important role in cryptology. Balanced Incomplete Block Designs (BIBDs) are very well known as a tool to solve rising problems in this area. We will describe what a QR code is made of and the two stages involved in it. We also illustrate it with some features of a QR code and their emerging applications on the security held so that they can be

used to prevent malicious hackers activities. Applications of R software for BIBDs development will be discussed.

Keywords: Algorithms, BIBDs, Qr Codes structures.

Acknowledgements: This work was partially sponsored by Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] B. Levin, F. Stewart, and L. Chao, Resource-limited growth, competition, and predation: a model, and experimental studies with bacteria and bacteriophage, Amer. Naturalist 111, 3-24, 1977.
[2] B. Levin and R. Lenski, Constraints on the evolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities, Amer. Naturalist 125(4), 585-602, 1985.
[3] Cain J.W., Mathematical Models in the Sciences, In: Molecular Life Sciences: An Encyclopedic Reference, Springer, 2014.
[4] Francisco, C., Delineamento Experimental em Blocos Incompletos: Estudo de Casos Particulares; Master Thesis; Open University; Lisbon, Portugal, 2014.
Available in: http://repositorioaberto.uab.pt/handle/10400.2/4413.
[5] Hillis, W. Daniel. Why physicists like models and why biologists should, Current Biology, 1993.
[6] Levins, Richard., The Strategy of Model Building in Population Biology, American Scientist, Vol. 54, No. 4, December, 1966.
[7] Neimark, J. I. Mathematical Models in Natural Science and Engineering, Berlin: Springer, 2003.
[8] Rowbotton, D. P., Models in Biology and Physics: What's the Diference? ; Foundations of Science, 2009, Volume 14, Issue 4, pp 281-294.
[9] Sontag, E.D. Molecular Systems Biology and Control, European Journal of
Control 2005; 11:396-435.

Applied scientific computing over the Web: robust methods in Acceptance Sampling for weibull variables

Elisabete Carolino[1], Miguel Casquilho[2], M. Rosário Ramos[3], Isabel Barão[4]
[1]Escola Superior de Tecnologias da Saúde de Lisboa (ESTeSL-IPL), Portugal, lizcarolino@gmail.com
[2]Universidade de Lisboa and Centro de Recursos Naturais e Ambiente (CERENA), Portugal, mcasquilho@tecnico.ulisboa.pt
[3]Universidade Aberta and Centro de Matemática, Aplicações Fundamentais e Investigação Operacional (CMAF-CIO), Portugal, mariar.ramos@uab.pt
[4]Faculdade de Ciências and Centro de Matemática, Aplicações Fundamentais e Investigação Operacional (CMAF-CIO), Universidade de Lisboa (ULisboa), Portugal, mibarao@ciencias.ulisboa.pt

Abstract: Acceptance sampling (AS) is used to inspect the process input or the process output. A sampling plan is designed to determine a procedure that, if applied to a series of lots of a given quality, and based on sampling information, leads to a specified risk of accepting or rejecting the lots. Classic AS by variables assumes Gaussian distribution, as treated in industry standards, which is sometimes an abusive assumption leading to wrong decisions. AS for variables with asymmetric and/or heavy tailed distributions is then a relevant topic. The case of Weibull distribution is treated here and specific AS plans are derived. As an alternative, traditional plans are used with robust estimators. The estimators are total median and the sample median for location and a modified version of the sample standard deviation and Total Range for scale estimates. The problem of determining AS plans by variables is addressed for Weibull distribution with unknown parameters. The aim is to apply scientific computing over the Web with remote servers in order to conduct a simulation study to compare the performance of these methods. Classical plans, specific plans and plans using the robust estimates for location, scale and degrees of asymmetry are compared.

Key-Words: statistical quality control; acceptance sampling by variables; robust methods, simulation, scientific computing on the web.

Application of Factorial Designs with computer simulation in the process of motors calibration

Bruna de Freitas Zappelino, **Elisa Henning**, Teresa A. Oliveira, Olga Maria Carvalho Formigoni Walter

Abstract: Industrial experiments are performed by companies in order to improve the characteristics of product quality and manufacturing processes. Using the techniques of planning and analysis of the experiments, also known as DOE (Design of Experiments), the number of experiments to be performed can be reduced in order to determine which factors significantly affect the response variable and the amplitude of the effects. In this sense, this work aims to apply the experimental design techniques to improve industrial quality and reduced process variability. As part of this goal, we developed a technique of applying Factor 2^k and Response Surface Methodology in the electric motor housings calibration process. This experimental study identifies the most important parameters to minimize the eccentricity, evaluating geometric characteristics of the engine casings and the calibration process in which the part is subjected. At the same time, optimal levels were identified for the adjustment of the parameters evaluated for the process, in order to operate according to the established by the company. The experimental procedure was followed by computer simulations using the *Arena® software* to validate the study of this process. With the simulation results, predictions can be justified and the improvements are still carried out before deployment, thus to save time and money is possible. The results achieved through this work show how technical planning allied to experiments administration of manufacturing can facilitate the improvement of the production process, through quality and productivity gains.

Keywords: Design of Experiments, Factorial Design 2^k, Calibration Process, Computer Simulation.

Acknowledgements: Research partially funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

SESSION C5: Business Statistics

Chair: **Daniel Jeske**, University of California–Riverside

Prediction of the Number of Additional Failures Using a Bayesian Approach

Víctor Aguirre Torres, ITAM

Abstract: The problem arises when trying to predict the number of stainless steel tubes that will fail in a nuclear power plant heat exchanger. The observed datum is the number of tubes that failed after 3 years of use. It is required to predict the additional number of failures after 10 years of use. Originally, the two frequentists solutions proposed made the assumption that the time to failure of the tube had a Weibull distribution where the shape parameter was known. The new Bayesian approach of prediction allows the possibility of uncertainty on both the shape and scale parameters of the Weibull distribution. The approach makes use of a Markov Chain Monte Carlo sequence to obtain the relevant

posterior distribution to produce the predictions. The method does not require large samples or that the probabilities of failure be small as in the case of the frequentists methods.

Characterizing Business Resilience Using SVM-Based Predictive Modeling

Karel Kupka, Trilobyte Statistical Software, Ltd.
Co-authors: Rudrajeet Pal, A.P. Aneja and J. Militky

Abstract: Business resilience has gained prominence, in academia and practice, vis-à-vis the heightened challenges recently faced by organizations, e.g. financial crisis. Developing resilience by thriving or bouncing back from crises yields sound business health in the future.

However extant scholarly discussion on predictive modelling of economic resilience is rather limited, while business health studies are mainly limited to bankruptcy failure predictions. These studies mostly utilize financial snapshots (based on only few years data) to construct the predictive models hence are static in nature (Balcaen and Ooghe 2006). Several assumptions underpin these static models, e.g. considering failure as a steady process devoid of organizational history (Appiah et al. 2015, du Jardin and Séverin 2011). Even though, few recent studies (cf. du Jardin and Séverin (2011), Chen et al. (2013) etc.) have designed a "trajectory of corporate collapse" to forecast the changes in firms' financial health, using various 'expert systems' like self-organizing maps (SOM) based upon unsupervised neural network approach, these studies still interpret the findings largely for predicting bankruptcy (a 'state') rather than drawing inference on the economic growth or recovery patterns (a 'trajectory') of organizations – a key to generate resilience. Neither these studies utilize longitudinal financial data (spanning over many years) to capture the dynamics of corporate history required to build resilience of organizations in reality.

In this context, our paper proposes developing a predictive econometric model of business resilience by using 'expert' SVM method. The expanded predictor based on financial ratios highlighted by Altman (1968)'s Z-score also takes into consideration the corporate dynamics (first and second derivatives). Historical financial data is gathered from 198 firms representing 26 Dow Jones industrial sectors, and starting from 1960s.

Our prediction model achieved comparatively high predictive accuracy of ---- (for a forecasting horizon of ----- years) and is comparable to similar studies. However, the main contribution of the paper is in proposing four archetypical patterns in business health trajectories, derived from the historical hind-sight, defined by tendency-dynamics combinations and is essential to characterize business resilience as follows:

Business Health (at T = t+1) = Business Health (T = 0 to t) + Resilience function

These four typical situations range from the most pessimistic case (tendency = Down, dynamics = Down) to the most promising (Up-Up). The four archetypes can be used to explain four resilience functions, viz. (i): up-up as sustainable resilience, (ii) up-down as short-term resilience, till t = T, (iii) down-up as resilience in near-future, at t = T, and (iv) down-down as lack of resilience.

Does the Effect of Portfolio Diversification Exist Among Style Indices? Evidence from MSCI Growth Style in Asian Markets

Wee-Yap Lau, University of Malaya

Abstract: This study investigates the dynamic linkages of equity style índices in selected Asian equity markets on whether there exists portfolio diversification effect if fund managers decide to invest in equity markets along the growth style indices in Asian markets. This study looks into the long-run and short-run linkages among the equity style indices from nine selected Asian countries. In order to examine the impact of the 2008 global financial crisis on these markets, the sample is split into three sub-periods: pre-crisis (2002-07), crisis (2007-09) and post-crisis (2009-15). Notably, we found that first, empirical results indicate that the dynamic linkages among Asian markets have become more intensified after the crisis. Second, Singapore is the most influential market within the Asian regional markets in the short run. Other established markets such as Hong Kong and Japan are not as influential as previously thought. Third, the restriction tests reveal that investors still derive benefits of portfolio diversification in the presence of cointegration, and after the financial crisis, there are greater diversification opportunities as more countries are not cointegrated. The ramification from these findings is significant. This study fills the gap of current literature of equity style where portfolio diversification effect has not been investigated. Overall, this study has provided invaluable insights for investors and fund managers on the existence of the economic content of equity style where the opportunity of portfolio diversification exists.

SESSION A6: Reliability III

Org/Chair: **Yves Grize**, Zurich University of Applied Sciences

Energy Optimization for Vessel Operations by Planned Maintenance

Marcel Dettling, Zurich University of Applied Sciences

Abstract: The operation of large ocean going vessels requires an enormous amount of energy in the form of heavy oil. On one hand, this has ecological implications such as the emission of greenhouse gases; on the other hand bunker cost is by far the biggest monetary factor when operating vessels. Our ideas target the optimal maintenance of the vessel hull and propeller in order to reduce fuel costs and pollutant emissions. Every time a vessel is under anchor, organisms will attach to its body and propeller. Studies have shown that poor hull and propeller condition will lead up to 30% more fuel consumption compared to a clean state. With our research, we first aim for estimating the additional fuel consumption due to the unobservable hull and propeller fouling from longitudinal operational data that are collected from various sources (fuel consumption on board, GPS position data, weather and oceanographic data, et cetera). At the same time, the effect of maintenance actions such as propeller polishing, hull cleaning or dry docking the vessel are pinned down with statistical methods. Finally, optimization methods are used for finding an optimal maintenance plan for any given vessel, including future fuel, maintenance and opportunity costs.

Synthetic Bayesian Experts

David Banks, Duke University
Disc: Henry Wynn, London School of Economics & Political Science

Abstract: Public policy relies strongly upon expert opinion, especially in risk assessment for rare events. But expert opinion is often inconsistent, both within and
between experts. We therefore develop a statistical model for the elicited opinions, and use that to borrow strength across the responses through an exchangeable prior. Several versions of that prior are considered; the most advanced uses covariate information on the experts to characterize their areas of agreement and disagreement, which ultimately allows the estimation of the opinion of a synthetic expert whose covariates are selected by
the analyst. This result depends upon a novel technique that incorporates the background information of the expert using hierarchical Dirichlet regression and a latent space model.

SESSION B6: Markov Chains

Org: **Emmanuel Yashchin**, IBM
Chair: **Paul Kvam**, University of Richmond

Hidden Markov Models for Life Pattern Recognition

Tatsuya Ishikawa, IBM Research–Tokyo

Abstract: Sensors are becoming ubiquitous in our daily lives. Time series data generated by these sensors often reflect our life patterns. Our objective is modeling such patterns, in particular typical behaviors on a 24-hour basis, from time series sensor data. In this talk, hidden Markov models for life pattern recognition that treat timestamps of data as an observation in addition to ordinary observations is considered. By regarding time as a torus-valued random variable, a homogeneous model with a cyclical structure can be learned. This is an alternative method to the commonly used hidden Markov models with cyclical effects in which, for example, the transition probabilities are periodic functions of time. The applicability of the new formulation is evaluated by comparison with existing models.

Full Interaction Partition Estimation in Stochastic Processes

Jesus Garcia, UNICAMP

Abstract: Consider Xt as being a multivariate Markov process on a finite alphabet A. The marginal processes of Xt interact depending on the past states of Xt. We introduce in this paper a consistent strategy to find the groups of independent marginal processes, conditioned to parts of the state space, in which the strings in the same part, of the state space, share the same transition probability to the next symbol on the alphabet A. The groups of conditionally independente marginal processes will be the interaction structure of Xt. The theoretical results introduced in this paper ensure through the Bayesian Information Criterion, that for a sample size large enough the estimation strategy allow to recover the true conditional interaction structure of Xt. Moreover, by construction, the strategy is also capable to catch mutual independence between the marginal processes of Xt. We use this methodology to identify independente groups of series from a total of 4 series with a high financial impact in the Brazilian stock market.

Adaptive Procedures for the EWMA Control Chart

Willy Ugaz Sánchez, Carlos III de Madrid University, wugaz@est-econ.uc3m.es

Co-authors: Ismael Sánchez, Universidad Carlos III de Madrid and Universidad de Piura, ismael@est-econ.uc3m.es
Andrés Alonso, Universidad Carlos III de Madrid, amalonso@est-econ.uc3m.es

Abstract: Time-weighted charts like EWMA or CUSUM are designed to be optimal to detect a specific shift. If they are designed to detect, for instance, a very small shift, they can be inefficient to detect moderate or large shifts. In the literature, several alternatives have been proposed to circumvent this limitation, like the use of control charts with variable parameters or adaptive control charts. This paper has as main goal to propose some adaptive EWMA control charts (AEWMA) based on the assessment of a potential misadjustment in the mean process or in the dispersion process, which is translated into a time-varying smoothing parameter. The resulting control charts can be seen as a smooth combination between Shewhart and EWMA control charts that can be efficient for a wide range of shifts. Markov chain procedures are established to analyze and design the proposed charts. Comparisons with other adaptive and traditional control charts show the advantages of the proposals.

SESSION C6: y-BIS Special Session

Org/Chair: **Kristina Lurz**, prognostica GmbH

Classification of EEG Signals for Detection of Epileptic Seizures using Hybrid Artificial Intelligent Techniques

Ozan Kocadagli, Department of Engineering Technology and Industrial Distribution, Texas A&M University, College Station, TX, USA, ozan.kocadagli@tamu.edu

Abstract: This study presents an efficient approach that provides an accurate classification of Electroencephalogram (EEG) signals for detection of epileptic seizures. Essentially, this approach is based on an automated multi resolution signal processing technique and artificial neural networks (ANNs). In this approach, while discrete wavelet transforms (DWT) allows the feature extraction from EEG signals, ANNs deal with classifying EEG signals with respect to the predetermined classes of epileptic and non-epileptic signals. In analysis, a benchmark EEG data is used and ANNs are trained by various gradient based algorithms considering early stopping, cross-validation and information criteria. As a result, this approach not only allows the analysts to make deeply analysis of EEG signals for

epilepsy detection, but also provides the best model configuration for ANNs in terms of reliability and complexity.

Keywords: Classification of EEG Signals, Epileptic Seizures, Discrete Wavelet Transform, Artificial Neural Networks, Gradient Based Learning Algorithms.

Model the System, Not the Data - Leverage System Knowledge in Statistical Analysis

Dirk Surmann, Technical University Dortmund

Abstract

Empirical Bayes Model Selection: Some Known Results, a New Prior, and Open Problems

Victor Pena Pizarro, Duke University

Abstract

SESSION A7: Special Talks

Efficiency of Various Designs for Assessing Preference Heterogeneity

Martina Vandebroek, KU Leuven

Abstract: Practitioners have frequently used the multinomial logit (MNL) model in the context of discrete choice experiments. As these models assume that all persons use the same partworths to assess the values of different product attributes, it is inappropriate to describe reality. Nowadays, the heterogeneity in consumers' preferences is mainly analyzed through the mixed logit model which estimates the distribution of the preference parameters in the population. Although efficient choice designs have been developed for the MNL model and several other closely related models, not much progress has been made in the construction of optimal designs for the more complicated models. We investigated how efficient the classical orthogonal designs and the designs optimized for various choice models are for estimating the mixed logit model. In this talk we will presente the results of this large simulation study.

Bonus-Malus Systems in Automobile Insurance: Past, Present and Future

Rahim Mahmoudvand[1] and **Fatemeh Nasiri[2]**,
[1]Department of Statistics, Bu-Ali Sina University, Hamedan, Iran
[2]Mellat Insurance Company

Abstract: A Bonus-Malus System (BMS) is one of the ratemaking methods generally used in automobile insurance. BMS penalizes policyholders at faults for one or more claims by a premium surcharge, known as malus, and rewards the policyholders with a claim-free year by granting discount in the premium say bonus. An optimal BMS should be possibly fair for policyholders and financially balanced for the insurer.

While much progress has been made on the development of BMS but for practical applications, many questions remain, including the reliability and efficiency of the systems. To deal with system efficiency 'rigorously' an open portfolio assumption must be considered. The approximation techniques proposed for this purpose are largely at a conceptual stage.

Keywords: Automobile insurance, Bonus-Malus System, Portfolio

Point of Scale Differences of Rating Scales for the Liking Score as the Determinant of Efficiency in Penalty Analysis of Liking Score and Other Sensory Attributes of Potatoes Chips on Just About Right Scale

Olegbemi Olujimi, TNS RMS

Abstract: One of the important factors in product innovation or development is product attributes optimisation. Here, consumer's acceptance or rejection of the product is determined for the product formulation or development. Penalty Analysis on the other hand is the statistical technique commonly used by food scientists and consumer product development experts to evaluate the consumer behaviour in either penalising the product or not, due to the responses of the tested consumers on the product tested for the overall liking and the Just About Right (JAR) attributes of the product. Researchers normally used 5 point Just About Right scale but oftentimes different point of scale for rating scale on liking score were used. This study evaluates the effect of different points of scale for liking score and the sensory attribute with just about right scale on the Penalty Analysis. For the efficient evaluation, three different rating scales of Overall liking (dependent variable) for the potato chips were selected: 5 points scale, 7 points scale and 9 points scale against the sensory attributes that are independent variables of this Potato chips: Saltiness, Sweetness, Acidity, Crunchiness which are on JAR scale. These independent variables were evaluated on the each of overall liking scores separately. After running the Penalty Analysis for the each of the liking scores on different point scales, it was found out that at $P < 0.0001$, the result of the Penalty Analysis run on both 9 points scale and 7 points scale of overall liking score were the same but these two were different from the one run of 5 Point scale of the liking score.

SESSION B7: Time Series Modeling and Prediction

Org/Chair: **Paulo Canas Rodrigues**, Federal University of Bahia; University of Tampere

Bias Correction for Dynamic Factor Models

Carolina García Martos, Universidad Politécnica de Madrid
Co-authors: A.M. Alonso, G. Bastos

Abstract: In this paper, we consider forecasts of multivariate time series that follow a dynamic factor model. We obtain interval forecasts for the time series by using bootstrap. In particular, we consider the setting of factors which are dominated by highly persistent AR processes, and samples that are

rather small. Therefore, the factors' AR coefficients are estimated using small sample bias correction techniques.

A Monte Carlo study reveals that bias-correcting the AR coefficients of the factors allows to obtain better results in terms of interval coverage.

As expected, the simulation reveals that bias-correction is more successful for smaller samples.

Results are gathered assuming the AR order and number of factors are known as well as unknown. As an application, we employ data on industrial production (486 monthly observations of the Industrial Production Index, IPI, from January, 1975, to June, 2015) for 13 European countries.

Recent Advances in Singular Spectrum Analysis

Paulo Canas Rodrigues, Federal University of Bahia, Brazil and
University of Tampere, Finland

Abstract: Singular spectrum analysis (SSA) is a nonparametric approach to analyze time series data. SSA is an extension of principal component analysis that allows the decomposition of the original (univariate) time series into a sum of independent components, which can be interpreted as trend, oscillatory and noise components. In this talk we present an overview of singular spectrum analysis and show its usefulness in several fields of research such as climatology, econometrics and industrial production.

Fuzzy Graph with Application to Solve Task Scheduling Problem

Vivek Raich, Government Holkar Science College, Indore 2.
Co-authors: Shweta Rai and D.S. Hooda

Abstract: The concept of obtaining fuzzy sum of fuzzy colorings problem has a novel application in scheduling theory. The problem of scheduling N jobs on a single machine and obtaining the minimum value of the job completion times is equivalent to finding the fuzzy chromatic sum of the fuzzy graph

modeled for this problem. In the present paper the task scheduling problem is solved by using fuzzy graph.

Key words and phrases: Fuzzy Graph, k-fuzzy coloring, Chromatic number, chromatic fuzzy sum and Γ -chromatic sum

References

[1] Eslahchi. C and B. N. Onagh, "Vertex Strength of Fuzzy Graphs", International Journal of Mathematics and Mathematics Sciences. Volume2006.
[2] Kabuki. E, "The Chromatic sum of a Graph", PhD dissertation Western Michigan University, Michigan, 1989.
[3]Munoz. S, T. Ortuno, J. Ramirez and J. Yanez, "Coloring Fuzzy Graphs", Omega 32 Pp.211-221, 2005.
[4] Pardlos PM, Mavridou T, Xue J. "The Graph Coloring Problem: a bibliographic survey". In Du Dz, Pardalos PM, Editoers, handbook of combinational optimization, Boston Kluwer Academic Publishers, vol.2.1998.
[5] Senthilraj. S, "On the matrix of chromatic joins" International Journal of Applied Theoratical and Information Technology Asian Research Publication Agency Network, Vol. 4 No. 3, Pp 106-110, 2008.
[6] Senthilraj. S, "Edge Critical Graph with double domination" Conference Proceeding of NCCT-08, Pp.51, 2008.
[7] Senthilraj. S, "Total Domination Number of Planer Graph" Conference Proceeding of NCCMMGP-08, Pp.01-17, 2008.
[8] Senthilraj. S, "The N Tuple Dominating of Graph with Algorithmic Approach" Conference Proceeding of ICMCS-09, Pp. 69-71, vol.1.1, 2009.
[9] Senthilraj. S, "Dominating Related Parameters on Bounds", International Journal on Computer Engineering & Information Technology, SERC, State of California (C.A.) USA Publication, vol. 4No.6, Pp:34-42, 2009.
[10] Senthilraj. S, "Fuzzy Graph Application of Job Allocation", International Journal of Engineering & Innovative Technology, vol.1, 2012.
[11] V.Nivethan, Dr.(Mrs) A.Parvathi," Fuzzy Total Coloring and Chromatic Number of a Complete Fuzzy Graph", International Journal of Emerging Trends in Engineering and Development Issue 3, Vol.6 (November 2013).

SESSION A8: ENBIS Spring Meeting I

Org/Chair: **Xavier Tort-Martorell**, Universitat Politécnica de Catalunya. Barcelona TECH

Understanding Compositional Eggs and Bananas (workshop)

Marina Vives Mestres, Universitat de Girona

Abstract: Compositional Data (CoDa) consists of multivariate data with strictly positive components representing parts of a whole and usually adding to a constant sum (e.g. 1 or 100). Natural examples consist of chemical formulations, food composition, impurity profile and mixtures. Specific statistical methods are necessary to deal with CoDa because of the constant sum restriction and the particular geometry of its restricted sample space.

In this seminar, we first introduce the concept of Compositional Data and its sample space. Then we show the need to use multiplicative tools or, equivalently, to do an analysis based on log ratios of components. To better understand the effect of the log ratio representation we use the software CoDaPack and very simple examples. We outline the advantages of the log ratio approach and its coherence and compare classical statistical analysis and log ratio analysis in the cases of: principal component analysis, linear discriminant analysis, linear regression models and cluster analysis. We also discuss the problem of zeros and how to deal with them.

We finish this introductory session with an example of application in the field of customer survey analysis. Specifically, we analyse the annual customer satisfaction survey of the ABC Company presented and analysed in detail in the book edited by Kenett and Salini (2011). The questionnaire consists of an assessment of overall satisfaction evaluated on a five-point anchored scale, so that it can be analysed from a CoDa perspective, and almost 50 statements with two types of scores: an evaluation score and a measure of item importance. Other questions such as repurchasing intentions and descriptive variables for each customer are used in analysing the ABC dataset.

References: Kenett, RS., Salini, S. (2011). Modern Analysis of Customer Satisfaction Surveys: with applications using R. Chichester: UK. John Wiley and Sons.

SESSION B8: Bayesian Methods

Chair: **Hedibert Lopes**, INSPER

Bayesian Inference for Ordinal-Response State Space Mixed Models with Stochastic Volatility

Dipak Dey, University of Connecticut

Abstract: We propose a state space mixed model with stochastic volatility for ordinal-valued time series, where the inverse link function is assumed to be a normal cumulative distribution function. We also consider two semiparametric extensions of this new class of parametric models; flexible modeling of the inverse link function and wider class of the distribution of the error variance of the state space distribution, using a Dirichlet Process prior. For parameter estimation, we design efficient Markov chain Monte Carlo algorithms while we conduct model comparison. We illustrate our methods with a simulation study and na empirical application to stock market contagion.

BIC-Based Estimation in N-grams Dynamic Hierarchical Bayesian Framework

Veronica Gonzalez-Lopez, UNICAMP

Abstract: The Bayesian Information Criterion (BIC) was identified as a consistent method of estimation of Variable Length Markov Chains (VLMC), by Csiszar and Talata, 2006. Since that, this criterion was explored by Garcia and Gonzalez-Lopez under a more general approach to obtain the optimal estimation of N-grams. The state space of a N-gram is composed by strings of size N-1 and the optimal estimation of a N-gram is given by the BIC-based estimation of a Partition Markov Model (PMM), which is a generalization of a VLMC. A PMM states the state space is divided in parts conforming a partition, the elements in each part of the partition, share the same transition probability to the next symbol in the process. The BIC allows to obtain the optimal partition of a N-gram. We show in this paper several tools BIC-based and under the scope of a PMM. The BiC-based approach allows (i) to define a distance between parts of a partition of the state space of a N-gram, (ii) to formulate a criterion of proximity between N-grams, (iii) to detect change point in a N-gram, (iv) to reveal the interaction structure in multivariate N-grams. In this way, in this paper we give an overview of the potential of the BIC-based procedures under a PMM.

SESSION C8: Applicable Statistics

Chair: **David Banks**, Duke University

Evaluating the Effectiveness of an Image Segmentation method

Luca Frigau, University of Cagliari
Co-authors: Conversano C., Mola F., University of Cagliari
Contacts: {frigau, conversa, mola}@unica.it

Abstract: Many image segmentation algorithms have been proposed to partition an image into foreground regions of interest and background regions to be ignored. We focus on examples related to images of botanical seeds presented to evaluate, from a statistical perspective, the effectiveness of the results provided by several image segmentation methods. More precisely, we assume that the separation of background pixels from foreground ones operated by a segmentation method needs to be further validated since, particularly for complex, it is very difficult to distinguish between the two categories even by a human eye or by powerful zooming. In this respect, the idea is to use a classification method, or classifier, in order to assess the degree of reliability of the separation between background and foreground pixels obtained from a standard segmentation image method. To this end, the comparison is made by evaluating, through the use of different types of classifiers, the accuracy of an image segmentation process. The statistical analysis involves many different settings in which each specific pre-processing method is, in turn, considered as the reference pre-processing method in the image segmentation process and the output of the different approaches proposed for image segmentation is used as response variable. In practice, in each setting the response variable is binary and corresponds, for each individual pixel, to the background/foreground assignment deriving from a specific segmentation method. The classification task is to ask a classifier to predict in the most accurate way the pixel category on the basis of the RGB intensities deriving form a specific pre-processing method. If a classifier is able to correctly predict all the available pixels, the relative segmentation method is 100% reliable. Thus, the more accurate is a classifier the more reliable is the pre-processing method at hand.

Statistical Monitoring of the Growth of Shrimp in an Aquaculture System

Ismael Sánchez, Universidad de Piura and Universidad Carlos III de Madrid
Co-authors: Isabel González. Universidad de Piura, Susana Vegas. Universidad de Piura.

Abstract: This work presents a methodology for the statistical monitoring of the production of shrimps in an aquaculture system. The procedure monitors the growth of the shrimps, as well as other influencing variables such as oxygen levels, temperature or feeding factors. Shrimps are farmed in large ponds, like in a batch process. All the shrimps in each pond/batch have approximately the same age. Then, the average weekly weight of the shrimps is estimated in each pond. The statistical properties of this weight depend on the pond, the age, and the initial weight of the shrimps. Appropriate transformations are then proposed to make data from different ponds and age comparable, allowing to build a reference distribution to perform the statistical monitoring. The methodology is applied to a large shrimp farm located in the north of Peru.

Unsupervised Data Mining for Medical Fraud Assessment

Tahir Ekin, McCoy College of Business, Texas State University, t_e18@txstate.edu

Abstract: U.S. governmental agencies report that three to ten percent of the annual health care spending is lost to fraud, waste and abuse. These fraudulent transactions have direct cost implications to the tax-payers and diminish the quality of the medical services. This talk discusses the use of unsupervised data mining approaches as pre-screening tools to aid in medical fraud assessment. They can help identify the hidden patterns among providers and medical procedures via outlier detection and similarity assessment. We illustrate the utilization of the proposed methods using U.S. Medicare Part B data and discuss the potential insights.

SESSION A9: ENBIS Spring Meeting II

Chair: **Marina Vives Mestres**, Universitat de Girona

Statistical Methods in Emotional Product Design Following the Kansei Engineering Model (workshop)

Luís Marco Almagro, Universitat Politécnica de Catalunya. Barcelona TECH

Abstract: When customers are questioned on what they want, a list of needs normally referring to functionality is obtained. Designers and engineers can translate this voice of the customer into technical parameters, so that the product fulfills those needs. However, customers do not usually explain their emotional needs, probably because they are not aware of having them or are unable to tell which they are. Even when those emotional needs are discovered, it is not obvious which technical properties of the product will elicit those desired emotions.

Some markets are currently so crowded of similar products in terms of functionality that adding an "emotional touch" can make a difference. How do designers create "emotional products"? They usually rely on their intuition, creativity and experience. But they also use different qualitative and quantitative methods to collect information on how products are perceived and used. Several of these methods can be grouped under the umbrella term "emotional design" or "affective design". One of the methods is the so-called kansei engineering (KE).

Kansei engineering is a method for incorporating emotions in the product development phase. The main purpose is discovering which technical parameters of a product elicit the chosen emotions. The method was first proposed by Prof. Mitsuo Nagamachi in the 70's and 80's, but gained attention in this XXI century, in part due to work by Prof. Simon Schütte at Linköpings Universitet. KE studies are based on self-reporting emotional reactions with questionnaires (usually rating on Likert or semantic differential scales). A set of different prototypes is shown to participants in the study, and ratings are given on elicited emotions. Each emotion acts as a response in a design of experiments.

There is a large range of statistical tools commonly used in KE studies, mainly multivariate techniques and regression models. Data in KE studies have a great amount of variability, and as building prototypes is costly, there is always the attempt to discover a lot of things (having a lot of factors) but only a few runs in the experiment (probably too few!). All these issues pose interesting statistical challenges; in fact, kansei engineering is a discipline "in need of statistics".

This seminar will take the form of a workshop, where you will be asked to discuss and work with your colleagues. We will first cover the basic ideas behind kansei engineering studies, and present the model used to conduct them. After several examples, a real (simple) KE study will be prepared by participants in small groups. This small example will be used to discuss some statistical tools useful in KE. For instance, multivariate techniques for summarizing information, and for automatically detecting "crazy" participants, will be covered. As "customers" of KE studies are often designers, a great importance is

placed on visual presentation of results. Quantification theory type I (QT1), a special version of regression analysis commonly used in KE, which makes interpretation of results easier when all independent variables are categorical, will also be explained (the ideas behind QT1 are, in fact, useful far beyond KE studies).

When the seminar finishes, you will have a good understanding of what kansei engineering is, and how statistics can contribute to this field. Moreover, you will have learnt some "tricks", such as QT1, that make statistical output easier to interpret to a broad audience.

SESSION B9: Statistical Theory

Chair: **Vivek Raich**, Government Holkar Science College

Approximating Extreme Compound Distribution Quantiles Using a Multiplier Approach

Helgard Raubenheimer, North-West University

Abstract: A popular method in modelling the aggregate loss distribution in risk and insurance is the Loss Distribution Approach (LDA). For example many banks currently use the LDA for estimating economic and regulatory capital for operational risk under Basel's Advanced Measurement Approach. The aggregate loss distribution is a compound distribution resulting from a random sum of losses, where the losses are distributed according to some severity distribution and the number (of losses) distributed according to some frequency distribution. This paper studies the approximation of extreme quantiles of the aggregate loss distribution. A key application of this approximation is the estimation of the economic or regulatory capital in a particular operational risk category (ORC). We propose an approach to approximate the extreme quantile of the compound distribution using a combination of a multiplier and the less extreme quantile of the severity distribution. The proposed approximation is assessed via a simulation study.

Big Data and Multivariate Permutation Tests

Rosa Arboretti, **Luigi Salmaso**

Department of Civil and Environmental Engineering
Department of Management and Engineering
University of Padova
rosa.arboretti@unipd.it, luigi.salmaso@unipd.it

Abstract: In several application fields, e.g. genetics, image and functional analysis, several biomedical and social experimental and observational studies, etc. it may happen that the number of observed variables is much larger than that of subjects. It can be proved that, for a given and fixed number of subjects, when the number of variables increases and the noncentrality parameter of the underlying population distribution increases with respect to each added variable, then power of multivariate permutation tests based on Pesarin's combining functions [F. Pesarin, Multivariate Permutation Tests with Applications in Biostatistics, Wiley, Chichester, New York, 2001] is monotonically increasing. These results confirm and extend those presented by Blair et al. [A study of multivariate permutation tests which may replace Hotelling's T 2 test in prescribed circumstances, Multivariate Behav. Res. 29 (1994), pp. 141–163]. Moreover, they allow us to introduce the property of finite-sample consistency for those kinds of combination-based permutation tests.

Sufficient conditions are given in order that the rejection rate converges to one, for fixed sample sizes at any attainable α-values, when the number of variables diverges. A simulation study and a real case study are presented.

Information Theoretic Models for Dependence Analysis and Missing Data Evaluation

D.S. Hooda, GJ University of Science & Technology, ds_hooda@rediffmail.com

Abstract: In the present paper we derive a new information theoretic model for testing and measurement of dependence among attributes in a contingency table. A relationship between information theoretic measure and chi-square statistic is established and discussed with numerical problems. A new generalized information theoretic measure is defined and studied in details. Maximum entropy model for estimation of missing data in design of experiment is also explained.

SESSION C9: Process Chemometrics

Org: **Alberto Ferrer**, Politechnic University of Valencia
Chair: **Ehsan Soofi**, University of Wisconsin–Milwaukee

PCA-Based Monitoring of Time-Dependent, High-Dimensional Data

B. De Ketelaere[1,2], M. Hubert[2,3] and E. Schmitt[2,4]

[1] Department of Biosystems, Division MeBioS, KU Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium
[2] Leuven Statistics Research Centre, KU Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium
[3] Department of Mathematics, KU Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium
[4] Protix, Industriestraat 3, 5107 NC Dongen, The Netherlands.

Abstract: Modern industrial processes are typically highly automated and equipped with in-line sensor technologies that produce vast amounts of data in a short time. The result is the availability of large process streams that often display autocorrelation because of the fast sampling schemes. Additionally, in a substantial number of real-life processes, nonstationarity is introduced because of warmup/cooldown, machine wear and variability in input material. This scenario of multivariate, time-dependent data is one of the most challenging encountered in statistical process monitoring (SPM), but it is often overlooked, although the separate fields of multivariate SPM and SPM for autocorrelated data have received more attention during the last decade. Approaches which are based on latent variables, such as principal component analysis (PCA), are a valuable direction for handling the multivariate nature, but need to be extended to cope with the time-dependent behaviour. Dynamic principal-component analysis, recursive principal-component analysis, and moving-window principal-component analysis are such extensions to cope with time-dependent features. We present a short review of these methods and will provide real-data examples to help draw connections between the methods and the behaviour they display. As parameter selection for those methods is a challenging aspect for which literature is very limited, we will present possible routes for choosing them.

Keywords: Time-dependent data; Process monitoring; Principal Component Analysis

Latent Variable-based Multivariate Statistical Process Monitoring for Big Data Streams

Alberto Ferrer, Multivariate Statistical Engineering Group, Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, 46022 Valencia, Spain

Abstract: "Big data" is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. According to [1] it is a *blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications*. Big data is linked to a multi-Vs system: Volume, Velocity, Variety, Veracity and Value.

For many industrial companies, big data is the result of Industrial Internet of Things (IIoT) connecting intelligent physical entities (e.g. computers, sensors, devices…) to each other, Internet services and applications [2].

But big data is not synonym of success; the key issue is how to extract valuable information from the data. A lot of potential information coming from structured and unstructured data streams needs to be analysed to give organizations new insights about their products, customers and services. This can be particularly valuable when it is critical to maintain quality and uptime, such as in process monitoring applications, by quickly detecting and diagnosing abnormal activities.

Big data exhibit high volume and correlation, rank deficiency, low signal-to-noise ratio, complex and changing structure, and missing values. Classic univariate and multivariate statistical process control techniques are not feasible for Big Data streams. In this talk we illustrate the effectiveness of latent variable-based multivariate statistical process monitoring methods to analyze Big Data streams and visualize extracted information in a way that is easily interpreted and that is useful to real time process monitoring and fault detection.

References

[1] Wikipedia (search for "big data")

[2] D. White (2016) Big Data. What is it? *Chemical Engineering Progress*, 112: 32-35.

Multivariate Statistical Analysis and Monitoring of Petrochemical Manufacturing Processes

J.M. González-Martínez, Shell Global Solutions International B.V., , Shell Technology Centre Amsterdam, Grasweg 31, 1031 HW Amsterdam, The Netherlands

Abstract: One of the most critical challenges that the oil and gas industry must address in the next decades is ensuring acceptable product quality and productivity. The lack of in-depth real-time knowledge about the process state forces manufacturers to operate at too conservative, sub-optimal and not intensified regimes. In order to overcome this situation two main objectives need to be pursued: i) better process understanding to facilitate risk-based regulatory decisions and innovation, and ii) better handling of product quality assessment and accomplishment.

The application of multivariate statistical projection methods in the *Multivariate Statistical Process Control* (MSPC) framework that copes with the challenges of batch and continuous process plays a crucial role in the contemporary industry [1]. Multivariate statistical monitoring systems based on *Principal Component Analysis* (PCA) have shown an outstanding capability for anomaly detection and diagnosis in industrial batch processes since their pioneering proposal [2]. In order to develop monitoring systems with good performance in terms of fault detection and diagnosis, the PCA model should be estimated with stable parameters capturing the actual process dynamics [3, 4]. A suitable approach to model process dynamics is by augmenting the data collected from the process with variables lagged in time, the so-called *Lagged Measurement Vectors* (LMV). These augmented data are then used to fit the PCA model. However, the stability in PCA is reduced as the number of variables grows. Hence, the number of LMVs in the model should be chosen as a compromise between stability and dynamics modelling [3]. Otherwise, the capabilities of the monitoring system to detect complex abnormalities in an incipient and safe manner might be seriously affected [5].

In this contribution, the bilinear modelling of batch processes is reviewed, emphasizing challenges and contradictions found in the literature. Examples of bilinear modelling of processes will be shown by using realistic simulated and real data of petrochemical processes.

References

[1] A. Ferrer. Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly. *Process. Quality Engineering*, 19: 311-325, 2007.

[2] P. Nomikos, J.F. MacGregor. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*, 37: 41-59, 1995.

[3] J. Camacho, J. Picó, A. Ferrer. Bilinear Modelling of Batch Processes. Part I: Theoretical Discussions. *Journal of Chemometrics*, 22: 299-308, 2009.

[4] J.M. González-Martínez, J. Camacho, A. Ferrer. Bilinear Modelling of Batch Processes. Part III: Parameter Stability, *Journal of Chemometrics*, 28(1):10–27, 2014.

[5] J.M. González-Martínez, Advances on Bilinear Modeling of Biochemical Batch Processes. PhD thesis, 2015. DOI: 10.4995/Thesis/10251/55684.

SESSION A10: Bayesian Applications in Business and Industrial Statistics

Org/Chair: **Refik Soyer**, George Washington University

Semiparametric Inference for Means of Heavy-Tailed Distributions

Hedibert Lopes, Insper

Abstract: Heavy tailed distributions present a tough setting for inference. They are also common in industrial applications, particularly with Internet transaction datasets, and machine learners often analyze such data without considering the biases and risks associated with the misuse of standard tools. This paper outlines a procedure for inference about the mean of a (possibly conditional) heavy tailed distribution that combines nonparametric inference for the bulk of the support with parametric inference motivated from extreme value theory for the heavy tail. We derive analytic posterior conditional means and variances for the expected value of a heavy tailed distribution. We also introduce a simple class of independence Metropolis Hastings algorithms that sample from the distribution for tail parameters via adjustments to a parametric bootstrap. This connects our framework to frequentist inference through a semiparametric bootstrap, and we show consistency for the latter algorithm. We also describe the use of informative priors for shrinking tails across groups to an overall background tail. The work is illustrated and validated on 72 experiments involving data from hundreds of millions of users of eBay.com.

Ranking Forecasts by Stochastic Error Distance, Information Measures, and Reliability Notions

Ehsan Soofi, University of Wisconsin–Milwaukee

Abstract: This paper presents an interface between econometrics, information, and reliability theories. We build on the theory of stochastic error distance (SED) introduced by Diebold and Shin (2014) for ranking point forecasts based on divergence between the distributions of the forecast error and the error-free forecast. The basic SED is a representation of the mean absolute error (MAE). We identify sufficient conditions for equivalent ranking of forecasts by MAE, the error variance, and error entropy. Drawing from the reliability theory, we introduce the SED representation of the mean residual absolute error (MRAE) function. This measure is the risk of a loss function where forecast errors below a tolerance threshold are not penalized. The global risk of MRAE over all thresholds is the entropy functional of the survival function which was introduced in the information theory literature during the last decade. The SED and new measures are illustrated using broad families of models for error distributions. Consistent estimators of the MRAE and the entropy functional of the survival function are available in the reliability and information literatures. We investigate the agreement/disagreement between the empirical versions of the proposed measures, MAE, and the mean squared error through ranking the principal components of five return-forecasting factors for bond risk premia.

Sequential Bayesian Analysis of Multivariate Count Data

Refik Soyer, George Washington University

Abstract: We consider modeling of multivariate time-series of correlated counts which often arise in finance, operations and marketing applications. Dependence among series arises as a result of sharing a common environment. We consider a class of multivariate Poisson time series models by assuming a common environmental process modulating the rates of the individual series. This setup gives us a class of dynamic multivariate negative binomial time series. We develop Bayesian inference for these models using particle filtering and Markov chain Monte Carlo methods. A by-product of particle filtering in our set up is predictive likelihoods which we refer to as multivariate confluent hyper-geometric negative binomial distribution. We discuss issues of sequential filtering, smoothing and prediction and illustrate the proposed models using a simulated data set as well as actual data on weekly household shopping trips.

SESSION B10: TBD

Chair: **Veronica Gonzalez-Lopez**, UNICAMP

Superposed Log-Linear Processes for Modeling Repairable Artillery

Paul Kvam, University of Richmond

Abstract: We investigate complex repairable artillery systems that include several failure modes. We derive a superposed process based on a mixture of nonhomogeneous Poisson processes in a minimal repair model. This allows for a bathtub shaped failure intensity that models artillery data better than currently used methods. Method of maximum likelihood is used to estimate model parameters and construct confidence intervals for the cumulative intensity of the superposed process. We also propose an optimal maintenance policy for repairable systems with bathtub shaped intensity and apply it to the artillery failure data.

Measuring the Effect of Uncertainty in the Estimation of the Conditional Covariance Matrix in Portfolio Selection and Risk Measures

Carlos Trucíos, UNICAMP

Abstract: Many decisions in finance are based on the estimates of the conditional covariance matrix of time series returns; for instance, in portfolio choice and risk management. Although the decisions are based on estimated covariance matrix little is known about effects of estimation uncertainties in the decisions. In this paper we analyze these effects on portfolio selection and Value-at-Risk estimation. The uncertainty and the effect are assessed via bootstrap procedure. We also deal with the presence of outliers and suggest a method robust to outliers. The procedure is applied to the corrected dynamical conditional correlation model, but it can be applied to any other model.

Acknowledgements: The author acknowledge financial support from São Paulo Research Foundation (FAPESP) grant 2012/09596-0 and Laboratory EPIFISMA.

Classifying the Defectives in Pipe Industry Using Artificial Neural Networks and Logistic Regression Models

Nurbanu Bursa, Hacettepe University

Abstract: In recent years, steel pipe manufacturing industry is developing rapidly. One of the problems that encountered is defective manufacturing in this sector. In order to avoid this problem, types of pipes and their types of defectives were examined in a pipe factory in Ankara, Turkey. In this research, for the month of May 2015 defective data were used. To classify the defectives of pipes, artificial neural networks and logistic regression models were used. It was found that artificial neural networks have a higher correct classification rate than logistic regression models for pipe industry.

Keywords: Artificial neural networks, classifying, logistic regression models

SESSION C10: Stochastic Modeling

Org/Chair: **Nalini Ravishanker**, University of Connecticut

Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics

Scott Holan, Jonathan R. Bradley and Christopher K. Wikle

Abstract: Many data sources report related variables of interest that are also referenced over geographic regions and time; however, there are relatively few general statistical methods that one can readily use that incorporate these multivariate spatio-temporal dependencies. Additionally, many multivariate spatio-temporal areal data sets are extremely high dimensional, which leads to practical issues when formulating statistical models. For example, we analyze Quarterly Workforce Indicators (QWI) published by the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. QWIs are available by different variables, regions, and time points, resulting in millions of tabulations. Despite their already expansive coverage, by adopting a fully Bayesian framework, the scope of the QWIs can be extended to provide estimates of missing values along

with associated measures of uncertainty. Motivated by the LEHD, and other applications in federal statistics, we introduce the multivariate spatio-temporal mixed effects model (MSTM), which can be used to efficiently model high-dimensional multivariate spatio- temporal areal data sets. The proposed MSTM extends the notion of Moran's I basis functions to the multivariate spatio-temporal setting. This extension leads to several methodological contributions, including extremely effective dimension reduction, a dynamic linear model for multivariate spatio-temporal areal processes, and the reduction of a high-dimensional parameter space using a novel parameter model. (Joint work with: Jonathan R. Bradley and Christopher K. Wikle).

Conquering Big Data in Volatility Inference & Risk Management

Jian Zou, Worcester, Polytechnic Institute

Abstract: The field of high-frequency finance has experienced a rapid evolvement over the past few decades. One focus point is volatility modeling and analysis for big data setting. It plays a major role in finance and economics. In this talk, we focus on the statistical inference problem on large volatility matrix using high-frequency financial data, and propose a methodology to tackle this problem under various settings. We illustrate the methodology with the high-frequency price data on stocks traded in New York Stock Exchange in 2013. The theory and numerical results show that our approach perform well while pooling together the strengths of regularization and estimation from a high-frequency finance perspective.

CONTRIBUTED TALKS

Inference under Competing Risks for Step Stress Models

Nandini Kannan, National Science Foundation

Abstract: In reliability or survival analysis, researchers are often interested in the effects of different risk factors such as temperature, voltage, dose etc. on the lifetimes of experimental units. Accelerated testing allows the experimenter to increase the levels of these stress factors to obtain information on the parameters of the life distributions more quickly than would be possible under normal operating conditions. A special class of accelerated tests is the class of step-stress tests which allows the experimenter to increase the stress levels at fixed times during the experiment. In this talk, we introduce the Cumulative Risk Model, a new model for step-stress experiments, that generalizes the Cumulative Exposure Model discussed by Nelson, and further introduces a competing risks framework. Assuming different parametric forms for the hazard function, we derive the MLEs and Least Squares Estimators of the model parameters.

POSTERS

Multiple imputation with interval-censored data

Mario César Jaramillo[1], Mayerly Cano Arroyave[2], lmcanoa@unal.edu.co
[1]Associate professor, Universidad Nacional de Colombia, Sede Medellín.
[2]Student of M. Sc. in Statistics, Universidad Nacional de Colombia, Sede Medellín.

Abstract: Usually, the exact time at which an event occurs cannot be observed for several reasons; for instance, it is not possible to monitor constantly a characteristic of interest. This generates a phenomenon known as censoring that can be classified as left censored, right censored or interval censored. Conventionally the lower limit, the middle or the upper limit of the inspection interval has been used as failure time, this is known in the literature as a simple imputation and it has been much used because of its simplicity compared to other methods. However, these methods have problems of bias in the estimates of the survival function, especially when the intervals are large, or are of different lengths, this is why other estimation methods should be used to correct the bias of previous methods. This paper compares by simulation several multiple imputation methods for interval-censored data using auxiliary variables and simple imputation methods, which do not use these variables. These methods are intended to permit the use of standard estimation techniques for right censored data.

Key Words: Survival Analysis, Efficiency, Simple Imputation.

On choosing mixture components via non-local priors

PhD student: Jairo Fúquene
Co-authors: David Rossell and Mark Steel
Department of Statistics, University of Warwick, UK

Abstract: The traditional Bayesian model selection criteria to choose the number of components in mixture models can fail to enforce parsimony and result in poorly separated components of little practical use. On the other hand, non-local priors (NLPs) are a family of prior distributions that encourage choosing adequately simpler models by enforcing a separation between the probability models under consideration. We propose NLPs for mixture models leading to tractable expressions and define default prior parameters from subject-of-matter considerations. In the context of mixture models, NLPs encourage a separation between the components that leads to extra parsimony for the considered models and therefore to more interpretable clusters. Therefore we demonstrate a theoretical characterization of the sparsity induced by NLPs in mixture models for choosing the number of components. We also propose an importance sampling scheme to compute the integrated likelihood based on the Gibbs sampling output and EM algorithms useful for posterior inference and

cluster classification of the observations in mixture models under NLPs. We fully investigate the use of NLPS for choosing the number of components in multivariate Normal mixture models. Therefore, we propose a family of exchangeable moment priors for multivariate Normal mixture models and we compare their relative performance relative to their local priors (LPs) counterparts and the Bayesian Information Criterion (BIC). The proposal is illustrated using simulated and real data sets.

Keywords: Mixture models, Non-local priors, Integrated Likelihood, Bayes Factor.

Methods for Selecting an Appropriate Copula

Julieth Verónica Guarín Escudero[1], Mario César Jaramillo Elorza[2], Carlos Mario Lopera Gómez[3], jvguarine@unal.edu.co
[1]Student of M.Sc. in Statistics. Universidad Nacional de Colombia, Sede Medellín.
[2]Associate professor. Universidad Nacional de Colombia, Sede Medellín
[3]Associate professor. Universidad Nacional de Colombia, Sede Medellín

Abstract: Copulas have become a useful tool for modeling data when the dependence among random variables exists and the multivariate normality assumption is not fulfilled. The interest in modeling multivariate problems involving dependent variables arise in several areas, turning this methodology into a convenient way to model the dependence structure of random variables. However, in practice there is not a standard method for selecting a copula among several possible models, so that the choice of an appropriate copula is one of the greatest challenges confronting the researcher. In this work some graphical goodness of fit tests for copulas are discussed.

Keywords: Copula, Graphics, Dependence, Goodness of Fit.

Proposal to Monitor the Demand Forecast Error of a Product Applied to Refrigeration

Tatiana Cristina de Oliveira[1], **Elisa Henning**[2], Teresa A. Oliveira[3]
[1]Santa Catarina State University, Department of Production Engineering, Brazil, tatiana.co@hotmail.com
[2]Santa Catarina State University, Department of Mathematics, Brazil, elisa.henning@udesc.br

[3]Universidade Aberta and CEAUL, Teresa.Oliveira@uab.pt

Abstract: Demand forecasting has been a growing challenge for companies, and those that are able to do good forecasting avoid unnecessary costs, becoming more competitive in the market. There are several methods for demand forecasting, but they always have some error value. Methods for monitoring the forecast errors are able to identify when the change in demand has values beyond the expected range and indicate a need for revision of the model. This paper presents a study on the monitoring system of demand forecast errors of a company in the mechanical metalworking industry, specializing in equipment for refrigeration, located in southern Brazil. Demand forecasting in the company is done by looking at the composition of the sales force and historical analogy. Errors are evaluated by an internal indicator that is based on the mean absolute percentage error. In this paper, monthly demand data for a product in the period between 2013 and 2014 were analyzed. Two new models for monitoring errors for the product were proposed, using control charts and tracking signal charts. Another proposal was to revise the formula used by the organization to calculate the mean absolute percentage error. The proposed methods have reliable results when evaluating the demand forecast. It was found that the company can use the suggested changes at the same time, since they do not require much mathematical or computational effort, nor much time to analyze the results. Control of the demand forecasting processes will positively impact the control of the company's entire supply chain, ranging from inventory to human resources, thus justifying the importance of monitoring forecast errors.

Keywords: Forecast error. Control chart. Monitoring. Demand forecasting. Tracking Signal.

Acknowledgements: This work was partially sponsored by Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

Analysis of complexity of stock returns

Zagorka Lozanov-Crvenković[1], Emilija Nikolić-Đorić[2]
[1]Department of Mathematics and Informatics, Faculty of Sciences, Novi Sad University, Novi Sad, Serbia – zlc@dmi.uns.ac.rs
[2]Department of Agricultural Economics and Rural Sociology, Faculty of Agriculture, Novi Sad University, Novi Sad, Serbia – emily@polj.uns.ac.rs

Abstract: The traditional approach in analyzing stock returns is based on financial econometric models such as family of GARCH models, models of stochastic volatility, Black Scholes model.

The aim of this research is to explore alternative methodology used in Econophysics, in order to quantify the randomness degree in fluctuations of daily returns time series. As measures of uncertainty (complexity) or disorder in time series we use the Kolmogorov complexity based on the Lempel-Ziv algorithm, maximum Kolmogorov complexity, sample, permutation and Shannon entropy.

In empirical analysis developed market indices (Dow Jones, NASDAQ), market index of EU new state (CROBEX-Croatia) and market index of candidate state (BELEX-Serbia) were considered and compared. The sample period is 2007-2015, and three sub periods 2007-2009, 2010-2012 and 2013-2015, as well.

The relationship between measures of variability (relative standard deviation, robust coefficient of variation), measures of risk (Value of risk-VaR, Expected Shortfall) and measures of complexity are analyzed also.

Forecasting Industry Production Industry Production Index in Turkey

Hatice ONCEL CEKIM, Cem KADILAR
Department of Statistics, Hacettepe University, 06800, Ankara, Turkey

Abstract: The Industry Production Index (IPI) is used to measure the improvement of economy and the effects of economic policy making in short term. In monthly, IPI is computed by production units of the industrial sector such as Manufacturing industry (B), Mining and quarrying sector (C) and Electricity, gas, steam and air conditioning supply (D) of the General Industrial Classification of Economic Activities within the European Communities (NACE Rev. 2). Moreover, one of the most importantly short-term economic indicators is the IPI. It affects the fluctuations in the level of industrial output in economy. Therefore, it provides an advantage to the interpretation of the short time economic.

Time Series Analysis is one of the most important tools for modeling and forecasting indicators in the economy of a country. ARIMA models are the most popular and preferred technique in Time Series Analysis to predict the related series.
The aim of the article is to obtain short time forecasting model, using the ARIMA model, to forecast the IPI in Turkey. When the time series graph of the IPI is analyzed, the index reaches its lower values in January 2005, January 2006, and February 2009. It is clearly seen that the indicator steadily increases after January 2010, in other words, it has on a fast upward trend with oscillations between 2010 and 2016.

When we model the IPI series, we obtain the SARIMA$(1,1,2)(3,2,0)_{12}$ model using the ACF and PACF graphs. We get the estimates and forecasting values of the series with the help of this model. As seen forecasting values of the series, the IPI series will have also continued to rise with fluctuations by the end of 2016.

A New Statistical Distribution of Stock Exchange Price

Gamze Ozel, Selen Cakmakyapan
Department of Statistics, Hacettepe University, 06800, Ankara, Turkey

Abstract: Researchers continue to seek theoretical probability distribution models that fit the empirical distributions of changes in spot exchange rates. Better theoretical models of these empirical distributions should contribute to more accurate pricing models for exchange rates and improved test statistics for such models. For years researchers assumed that the empirical distribution of changes in exchange rates was best described by either a normal or lognormal probability distribution. Indeed, some current studies and most current tests automatically apply the logarithmic transformation to returns from spot, forward, and futures exchange rate changes; this transformation assumes, explicitly or implicitly, that the transformed data produces returns that are normally distributed. Recent empirical studies reject these assumed probability distributions; these studies do not conclusively agree on any single alternative model. These studies usually support either the mixed jump diffusion model, a discrete mixture of normal distributions (multinormal model), or some type of generalized autoregressive conditional heteroscedastic (GARCH) model.

The aim of the present study is to propose a new distribution and evaluate the suitability of a large number of pdfs, commonly used to model stock exchange price. Hence, dataset obtained from the Borsa Istanbul data set is modeled with probability distributions. Firstly, we proposed a new model. The new distribution has increasing and decreasing shapes for the hazard rate function. Then, we used the proposed and some other probability distributions to model Borsa Istanbul data set and we obtained the best fit with the proposed distribution in general.

Key Words: Generalized distribution, Stock Exchange Price, Maximum Likelihood Estimation

Exploring the rule Mathematical Modelling: Applications on Biology

Carla Francisco[1] and Teresa A. Oliveira[2]

[1]Departamento de Ciências e Tecnologia, Universidade Aberta, Portugal
[2]Departamento de Ciências e Tecnologia, Universidade Aberta e Centro de Estatística e Aplicações, Universidade de Lisboa, Portugal
carla.francisco.1@ulaval.ca ; teresa.oliveira@uab.pt

Abstract: The contribution of mathematics to the progress of many fields of biology is very well known. Physical concepts and mathematical methods such as determination of spaces, linear and nonlinear adjustments, models approaches, are very often the necessary tools to use on biological problems. However for biologists to follow these tools, it is not easy. It is necessary to have attention to the

complexity of the model, the number of free parameters and the number and type of assumptions; the ability to simulate the computer model; the representation of loyalty details of the modeled system; the range of situations for which the model can be applied; the universality of the model and the ability of the mathematical model to explain the natural phenomena; the availability of suitable mathematical methods and the power of computational tools available; the art of letting go the irrelevant detail for understand the problem and the art of keeping the necessary features to enable the understanding of the system. In this work some of these topics will be explored as well as the importance of establishing multidisciplinary research teams. We highlight, with examples, the role of mathematical modelling in biological sciences.

Keywords: Biological, Mathematical, Modelling Software.

Acknowledgements: This work was partially sponsored by Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] B. Levin, F. Stewart, and L. Chao, Resource-limited growth, competition, and predation: a model, and experimental studies with bacteria and bacteriophage, Amer. Naturalist 111, 3{24, 1977.

[2] B. Levin and R. Lenski, Constraints on the evolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities, Amer. Naturalist 125(4), 585{602, 1985.

[3] Cain J.W., Mathematical Models in the Sciences, In: Molecular Life Sciences: An Encyclopedic Reference, Springer, 2014.

[4] Francisco, C., Delineamento Experimental em Blocos Incompletos: Estudo de Casos Particulares; Master Thesis; Open University; Lisbon, Portugal, 2014.

Available in: http://repositorioaberto.uab.pt/handle/10400.2/4413.

[5] Hillis, W. Daniel. Why physicists like models and why biologists should, Current Biology, 1993.

[6] Levins, Richard., The Strategy of Model Building in Population Biology, American Scientist, Vol. 54, No. 4, December, 1966.

[7] Neimark, J. I. Mathematical Models in Natural Science and Engineering, Berlin: Springer, 2003.

[8] Rowbotton, D. P., Models in Biology and Physics: What's the Diference? Foundations of Science, 2009, Volume 14, Issue 4, pp 281-294.

[9] Sontag, E.D. Molecular Systems Biology and Control, European Journal of Control 2005; 11:396-435.

Elucidating the shelf-life kinetic of apple snack foodproduct by multivariate modeling: use of Orthogonal Partial Least Square (O-PLS)

Saavedra, J. and Cordova, A.

DATACHEM Agrofood: Data Analysis and Applied Chemometrics Group, Escuela de Ing. Alimentos. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.
* jorge.saavedra@pucv.cl

Abstract: A comparative study of shelf-life kinetic mechanism in a apple's snack product, was performed using two multivariate approaches. Samples were incubated at 18°C, 25 °C and 35 °C, for 18 months. Quality attributes were Aw, humidity, aroma, flavor, texture, color sensing and color DE. Data were arranged in matrix, which were submitted jointly and separately to a Principal Components Analysis (PCA) and Orthogonal Partial Least Square (O-PLS) later, which breaks down the joint variability of a phenomenon in sub-smaller space. All analysis was performed with SIMCA-P+12 statistical software.

The PCA explanatory models retained 2 PC which explained 83.1% of the total variability (PC1: 68% and PC2: 16.2%). PCA sorted the variability based on the time in the first component (t1). Inspecting the Contributions Plot the variability explained by the second component (t2) was related to different profiles of behavior for the 3 storage temperatures. Thus, for the treatment of 18°C, contribution is mainly explained by attribute Aw, while at 25°C the color and SO2 content acquired greater importance. At 35°C, greater contributions were associated with humidity and texture. This findings were constrasted with O-PLS methods, retained a main factor with 93.7% of the explained variability.

These findings suggest that for each storage temperatures, there were alternatives predominant mechanisms of deterioration. So the multivariate model reflected in terms of variability, biochemical phenomena associated with the deterioration of the product.

The kinetic parameters were computed in parallel with the PCA (first PC) and the main O-PLS factor (both related over time), so as to obtain km (reaction rate), Eam (activation energy, cal/mol), α' (acceleration factor) and cut-off criteria. At market conditions (18 °C), shelf life estimated was 18.1 months, at 25 °C, 18.2 months and 35 °C, 15.5 months. Due to the short difference in shelf-life between 18 °C and 35 °C it is possible to infer the high stability of the product. In addition, multivariate shelf-life values obtained by O-PLS was more suitable inasmuch first factor (associate with time) depict the maximum variability, as compared with univariate kinetics (where deterioration only contemplates the humidity as attribute) shown a estimated shelf- life mean of 17.4 months.

Thus, multivariate analysis (specially O-PLS) appears like a better tool to represente the complex changes in food products. So, the O-PLS method could estimates simultaneously the deterioration of all quality attributes, showing the interactions that occur between them.

Author Index