ISBIS-2010: Book of Abstracts

Abstracts are ordered alphabetically by surname of first registered authors.
For papers with multiple authorship,
names of other registered authors are listed in the Index.

Robust Regression Methods for the Analysis of Unreplicated Factorials

Victor Aguirre
ITAM Mexico

Roman de la Vara
CIMAT

The existing methods for analyzing unreplicated factorial that do not contemplate the possibility of outliers in experimental data have a poor performance for detecting the active effects when that possibility becomes a reality. Some approaches to deal with this problem are found in the literature like: using a Bayesian approach to identify active effects and outlying observations; transforming the response by means of ranks; and detecting a gap in the estimates of the effects as an indication of an outlying observation. We propose an iterative procedure based on robust regression to estimate the effects in unreplicated experiments that allows the presence of contaminated data. We conducted a simulation to compare the power of our method with other procedures that permit the presence of outliers. The study showed that in general our method has a higher power to detect active effects compared to the other methods.

Improved Estimation Strategies in Partially Linear Models with Nonlinear Time Series Errors

Ejaz Ahmed*, Saber Fallahpour*
University of Windsor

Kjell Doksum
University of Wisconson, Madison

This paper is concerned with a partially linear regression model with serially correlated random errors which are unobservable and modeled by a random coefficient auto-regressive process. Our main objective is to provide natural adaptive estimators that significantly improve upon the classical procedures in the situation where some of the predictors are inactive that may not affect the association between the response and the main predictors. In the context of two competing regression models (full and sub-models), we consider shrinkage estimation strategy. The shrinkage estimators are shown to have higher efficiency than the classical estimators for a wide class of models. We develop

the properties of these estimators using the notion of asymptotic distributional risk. Further, we proposed absolute penalty type estimator (APE) for the regression parameters which is an extension of the LASSO method for linear models. The relative dominance picture of the estimators is established. Monte Carlo simulation experiments are conducted and the performance of each procedure is evaluated in terms of simulated mean squared error. The comparison reveals that the shrinkage strategy performs better than the APE strategy when, and only when, there are many inactive predictors in the model.

Enterprise Statistical Forecasting and Tracking Framework

Yasuo Amemiya
IBM T. J. Watson Research Center

Currently, considerable interest exists in creating and utilizing business analytics systems that have impact in managing complex enterprise businesses. Such systems are typically based on us various performance tracking and condition indicator metrics measured over regular time intervals, and on relevant statistical analysis procedures applied repeatedly. Two features of the problem which can present unique difficulty and can be of technical interest are multi-layered hierarchical enterprise structure and multiple-scaled repeated time dimension. A common framework for such systems involving statistical analysis methods appropriate for enterprise business management is presented and described. Statistical forecasting and tracking methods are illustrated as use-case examples.

The Pareto Frontier Approach for Selecting Designed Experiments

Christine Anderson-Cook*, Lu Lu*
Los Alamos National Laboratory

Timothy Robinson
University of Wyoming

The selection of a good design for an experiment should typically be based on balancing multiple competing objectives. Good estimation of model parameters, good prediction in the region of interest, protecting against bias from potential model misspecification and obtaining a good estimate of the natural variability of the underlying process can all be important considerations. In this talk, we present an approach using Pareto frontiers to identify all candidate designs which represent an optimal balancing of the selected criteria. This approach is related to, but more general than desirability

functions, and allows for different weightings of the criteria to be easily explored. A new algorithm for efficiently identifying the Pareto frontier is shown, and the method is illustrated with an example.

Contradictions of Industrial Development of Ukraine

Anna Anisimova
Donetsk National University

The development of Ukraine is possible only at the expense of realization of innovative development as extensive sources of economic development are reached now. Realization of such policy requires significant financing of innovative development of Ukraine and the contradiction decision between industrial and a sustainable development. Research objective is studying of contradictions of industrial development of Ukraine. As a result of research discrepancy between real economic circumstances and tasks in view of development of Ukraine is revealed. It is reasonable that Ukraine is at the stage of export-oriented growth and transition to a stage of the developed market is impossible without a stage of stimulation of the accelerated growth. To overcome dependence of economic development on production export is possible at the expense of domestic demand increase. However production export is carried out basically by metallurgy, heavy engineering and the chemical industry, i.e. branches which depend on extraction of not renewed resources and are extremely material-intensive and ecologically dangerous. Therefore "development" in this case means investment in a mineral industry so, the exhaustion of base of minerals that at modern level of specific consumption of materials of production will lead only to decrease in its efficiency. At the same time Ukraine has accepted the concept of a sustainable development which bases on position about preservation of resource base, having ratified 2/4/2004 the Kyoto protocol. Hence, without the problem decision material capacities of industry, Ukraine cannot solve a sustainable development and ecological equilibrium problem. In work introduction influence resource-saved and waste-free technologies on size of resource consuming with the regression analysis is investigated. As a result is revealed that introduction of each resource-saved process leads to resource consuming decrease on the average on 13 million hryvnas. That speaks about the efficiency and necessity of use of such technologies. Dependence of consumption of power resources on expenses for new technologies is researched for Ukraine. Model parameters prove that at increase of expenses at new technologies at 1 million hryvnas lead to increase of fuel and energy resources consumption on 237 thousand t. with slowing down of such growth in time. On the basis of model the extremum point of function is calculated. It allows to draw a conclusion: as soon as such expenses will exceed 395 million hryvnas, it will be sufficient for change of a dependence direction and will lead to decrease of consumed power resources. The problems of high specific consumption of materials, inconsistency of problems of development, insufficiency of financing of innovations in a kind resource-saved technologies

are revealed. Necessary level of financing of expenses for new technologies is evaluated (395 million hryvnas).

Critical Success Factors for the Successful Deployment of Lean Six Sigma in Healthcare

Jiju Antony
University of Strathclyde

The healthcare industry is constantly concerned about how to better streamline the services they provide to result in better patient care with less waste of resources. Decreasing the rate of medical errors while increasing the general efficiency of services provided are something that cannot be taken lightly when the lives of patients are at risk. Lean Six Sigma is a powerful methodology which integrates the best of both methodologies for improving the efficiency and effectives of healthcare processes. The purpose of Lean Thinking in healthcare is to create the environment for improving flow and eliminating waste whereas Six Sigma on the other hand helps to identify and quantify problems that are related to variation in processes. Lean always asks the question "why does this process exist at all?" What is the value in the value stream? Six Sigma starts with "how can we improve this process?" It does not ask "why does it exist at all?" Critical Success Factors (CSFs) are the essential ingredients that must be achieved by the company to produce the greatest competitive leverage. These factors represent the essential things without which the initiative stands a little chance of success. Each one must receive constant and careful attention from management as these are the areas that must 'go right' for the organisation to flourish. If results in these areas are not adequate then the efforts of the organisation will be less than desired. This paper presents the CSFs for the successful deployment of Lean Six Sigma (LSS) in UK NHS. A thorough literature review was carried out initially to understand the CSFs for the introduction and development of a LSS initiative in the context of public sector organisations across the UK. This included Police, City councils, Fire and Ambulance services and NHS. Both survey questionnaires and semi-structured interviews were designed for data collection. The analysis of the data has revealed that strong focus on the needs of patients, senior management commitment and support, strong and visionary leadership to continuous improvement programme, selection of a project champion to assist with project selection and execution, etc. were the most critical factors. It was also found that linking LSS to business strategy, LSS training, understanding of tools, techniques and methodology are the least important factors. The paper also illustrates the key challenges in the deployment of LSS in the healthcare context. Note: I am delivering this paper with Professor Ronald Does from the University of Amsterdam and Dr Shirley Coleman from the University of Newcastle.

What is Optimization of Control Charts?

Francisco Aparisi
Universidad Politecnica de Valencia, Spain

For many years an extensive research for developing new quality control charts has been made. However, many times the real use of these charts is not optimal, i.e., better performance could be achieved employing optimized schemes. This session shows how finding the best parameters of a control chart can be posed as an optimization problem. Some of these optimization problems are extremely difficult to solve, therefore, a powerful optimization tool like Genetics Algorithms is employed. During the talk some examples are shown, from the easiest optimization (best performance to detect a single shift) to multi-objective optimization, like finding the Pareto front for in-control and out-of-control regions problem.

Optimization of a Set of Multiple Poisson Statistical Quality Control Charts

Francisco Aparisi
Universidad Politecnica de Valencia, Spain

Eugenio Epprecht
Catholic University of Rio de Janeiro

Some industrial applications require controlling simultaneously several Poisson variables. Each of the variables is based on the sum of a common Poisson variable and an independent variable. Each chart has a probability of false alarm (in-control ARL) and a probability of detecting the shift in the variable (out-of-control ARL). An optimization is carried out using Genetic Algorithms and consists of obtaining the control limits for the multiple charts given an in-control ARL, minimizing the time to detect the shift in the process, minimizing the out-of-control ARL of the set of Poisson charts.

Statistics for Sets, Using the Distance Transform

Adrian Baddeley
CSIRO and University of Western Australia

The distance transform of a binary image is a discrete version of the distance function of a set. It can be computed very rapidly and is a very useful tool for image analysis, especially with large numbers of images. This talk presents a survey of techniques which use the distance transform as a statistical tool, defining the average of a random set, the variability of a random set, and the

mean square error in estimating a set. Applications include environmental assessment of aircraft flight paths and quantification of the prediction error in weather forecasts.

Use of Bayesian Stochastic Frontier Analysis Model to Create Efficiency Index for Brazilian Electricity Distribution Firms

Marcus Vinicius Pereira de Souza, Reinaldo Castro Souza, Madiagne Diallo,
Tara Keshar Nanda Baidya
PUC/RIO, BRASIL

Use of Bayesian Stochastic Frontier Analysis Model to create Efficiency Index for Brazilian Electricity Distribution firms The objective of our research is to create efficiency index for Brazilian electricity distributions firms using Bayesian Stochastic Frontier Analysis approach. Brazil generates about 60.000 MWmed of electricity, most of which are distributed by the 60 firms used in our sample. We have considered two different cost functions: Cobb-Douglas and translog, each of them involving one output and three inputs, which are operational expenditure (output), energy distributed by the firm, number of consumer units, network distribution length (3 inputs). The stochastic part is captured by two components: a technical inefficiency variable, that can take only non-negative values, and random error variable capturing all non-manageable risk shocks. The latter variable is assumed to have a normal distribution in the present case. Also it is assumed to be independent of the first variable. The technical inefficiency variable is assumed to follow a negative exponential distribution function. In another work we have assumed it to have a half-normal distribution, yet in another work we have assumed it to have a truncated normal distribution. There is no a priori justification for the selection of any particular distribution. These distributions are selected arbitrarily. We are using the Bayesian approach because of the uncertainty related to the cost function parameters and the variances. The parameters of the cost functions are assumed to have truncated normal distribution and the variance of the inefficiency variable to have negative exponential distribution and the variance of the random variable to have a Gamma distribution. Using the distribution of these parameters as prior distribution, we obtained the posterior distribution. We have used the WinBUGS software package to obtain the parameters of the frontier cost function. We found all the parameters of the cost function statistically significant. From the software package we also obtained the posterior mean which is used as Economic Efficiency Index to rank the Brazilian Electricity Distribution firms. The result, obtained by assuming translog cost functions, the negative exponential distribution for the technical inefficiency variable, the normal distribution for the random variable, truncated normal prior distribution for the cost function parameters, negative exponential prior distribution for the variance of the technical inefficiency variable, and Gamma prior distribution for the variance of the random error variable, will be shown in a table. We have done similar studies assuming other technical inefficiency variable distribution (e.g Half normal, truncated normal

distribution). The ranking does not change much. Similarly we have also compared with the results obtained by using the Data Envelopment Analysis method. Again the ranking does not change much.

Bayesian Borel Games

David Banks
Duke University

Emile Borel proposed the game _La Relance_, in which two players each ante a unit amount, and then draw independently and privately from a uniform distribuion on [0,1]. The first player may bet an amount b or fold; the second player may then fold or call. If both players bet, then the player with the larger draw wins the money. The game has been studied by Borel, von Neumann & Morgenstern, Bellman & Blackwell, Karlin & Restrepo, and Ferguson & Ferguson. This talk presents a Bayesian analysis of the problem, treating the initial problem, and variations with multiple players and continuous bets.

Implementing Classifier Fusion in Credit-scoring

Farid Beninel
ENSAI

The data in credit-scoring are often from a mixture of populations as borrowers customers of the bank and the borrowers not-customers. Also, individuals from the training sample may have been subject to various other external constraints (reimbursement process over different periods, ...). The practice shows that in these circumstances it is difficult to separate, credit worthy customers and the others. Using a single discrimination rule i.e, classical discrimination methods based on a single rule (logistics, LDA, SVM, ...) placed in competition, lead to similar results. We study the performance of methods combining several rules, in particular the fusion of classifiers. We implemente the idea of classifier fusion using different classifiers to predict the boorrowers behavior. We focus on the problem of selection of classifiers accuracy and diversity. We compare the different fusion methods. A first category of methods consist of the fusion by component i.e., scores of an individual are replaced by an aggregated score (max, min, mean, product). The second category, consist in methods based on the distance between profiles, i.e., between individual profile and profile of the class.

The Unbearable Transparency of Stein Estimation

Rudolf Beran
University of California, Davis

Charles Stein, whose 90th birthday was celebrated in early 2010, discovered in 1956 that the usual unbiased estimator for the mean vector of a multivariate normal distribution is inadmissible, under quadratic loss, if the dimension of the mean vector exceeds two. It has since been claimed that Stein\'s result is counter-intuitive, even paradoxical, and is not very useful. In response to assertions of paradox, Efron and Morris (1973) presented an alternative empirical Bayes approach to Stein estimation. Stigler (1990) gave another derivation based on a "Galtonian perspective". But surely Stein himself did not find his results paradoxical. Is it not more likely that assertions of "paradoxical" or "counter-intuitive" or "useless" have overlooked essential discussions in Stein's brilliantly written 1956 paper? This talk will sketch three such arguments: the asymptotic geometry of quadratic loss in high dimensions that makes Stein estimation transparent, the optimality results associated with Stein estimation, and the remarkable practical efficacy of multiple shrinkage. It will be seen how Stein's ideas underlie the success of certain modern estimators constructed through penalized least squares, or submodel selection, or local smoothing.

Using Exploration Trees to Analyze Data about Loan Applications

Petr Berka
University of Economics, Prague

Induction of decision trees belongs to the most popular algorithms used in machine learning and data mining. When building a decision tree, we recursively partition the attribute space in a top-down way at each branching node looking for best attribute to make a split. The quality of the attribute is evaluated on the basis of its ability to separate examples of different classes. This process will result in a single tree that can be use both for classification of new examples and for description the partitioning of the training set. But due to the used greedy search strategy, this tree need not to split the training data in the best way with respect to the classes. In the paper we propose an alternative approach that is related to the idea of finding all interesting relations (usually association rules, but in our case all interesting trees) in given data. When building the so called exploration trees, we consider not a single best attribute for branching but more "good" attributes for each split. This modification of the tree learning algorithm will result in more trees, each partitioning the data in a little different way giving thus alternative knowledge for segmentation of the data with respect to the classes. The exploration trees can of course be used for classification tasks as well; in this case we can use either single tree or work with a whole ensemble of created trees. The

proposed method will be compared with the "standard" C4.5 algorithm on several data sets from the loan application domain. One of the data sets comes from the ECML/PKDD Discovery Challenge workshop, the other sets are taken from the UCI Machine Learning Repository. The results show, that among the exploration trees created for different data sets, there was always a tree that better splits the training data (had higher classification accuracy on training data) than the tree created (by greedy search) using C4.5. We are aware of the fact, that trees in C4.5 are tuned to perform well on the testing data (to avoid over-fitting) but if the task is to find and describe segments of training data related to the class attribute, our method gives better results.

On-Line Spatiotemporal Methods of Air Quality Surveillance for Safeguarding Community Health

Sotiris Bersimis
University of Piraeus

Kostas Triantafyllopoulos
University of Sheffield

Air pollution consists mainly of the introduction of various chemicals into the atmosphere, causing severe damage to the environment. The atmosphere, an essential part of the environment, is a complex, dynamic natural gaseous system which vital role is to support life on earth. Recent research related to air pollution is focused on identifying the impact of air pollution to human health and finding possible ways for protecting public health. This is known as one of the most important research areas today, since, especially in cities, air is already unclean and the quality of the environment is constantly aggravated, so that community health is threatened in many ways. In this paper, we attempt to develop a unified framework for monitoring air quality variables in the time domain as well as in the spatial domain combining the use of time series forecasting, multivariate analysis and statistical process control methods, in order to establish an air quality surveillance system, for monitoring and diagnosing probable extreme air pollution levels in real time which may be incorporated in a public health surveillance strategy or used independently for safeguarding community health from threats due to instability of air quality.

Activity Time Allocation Models: A Comparison Between Maximum-Likelihood and Bayesian Estimators

***Massimiliano Bez**, Italo Meloni, Erika Spissu*
University of Cagliari

Discrete-continuous choice models have been extensively employed in transportation applications, in particular to analyze the activity time allocation behaviour of individuals which involves the decision of whether or not to participate in an activity (discrete choice) and the amount of time to allocate to this activity (continuous choice). Usually, multivariate extreme-value (MEV) and multivariate Tobit (MT) models are employed to represent the activity time allocation behaviors. Both structures in fact allow the analysts to address the individual utility maximization problem related to the discrete-continuous choice to participate and allocate time to J activities. Further the mixed version of these models captures the heterogeneity across individuals due to unobserved individual attributes that are correlated across alternatives. Technically, it translates on enabling the calculation of a multidimensional integral by quasi-Montecarlo methods that average simulated values of integrated functions across different independent variables. However, these methods present a number of shortcomings. First, the quasi-Montecarlo integration fails increasing the number of independent variables due to correlation presents in the quasi-random sequence needed by this method. Second, even if the likelihood is correctly calculated, there is still a problem of inference, because the most of commercial software uses two common numerical approximations of the parameter covariance matrix, but each method can give different estimates, and there is no way to distinguish between them using standard asymptotic theory. Third, maximum likelihood computation will typically be very slow because the log-likelihood function is not convex in the correlation parameters, and this requires manually restarting the optimization from different starting points to help find a global maximum. Last, the frequentist methods require a large sample to insure the adequacy of asymptotic approximations. In this paper, a MT model is estimated through a Bayesian approach based on Gibbs sampling. The proposed method replaces the maximum likelihood estimation of parameters with a sampling-based estimation via Markov chain Monte Carlo (MCMC) method using the idea of data augmentation. This approach presents some advantages. First, it avoids the direct evaluation of the nontrivial likelihood. Second, by posterior distribution it provides finite-sample inference of the parameters and hence is free from the use of asymptotic approximations, so it does not require large sample or the numerical approximations of covariance matrix. Third, this method avoids manually restarting the optimization from different starting points to help find a global maximum. This paper is aimed at checking the sampling-based MT model parameter estimations using a simulated time allocation data from a well-know model and finally at comparing Bayesian and maximum likelihood MT model estimations using a real time allocation data sample. The database used to estimate the model is drawn from the Multi-purpose "Time use" Survey conducted between 2002 and 2003 by Turin Town Council jointly with the Italian National Institute of Statistics (ISTAT), within its own territory and 14 neighbouring town councils.

High Content Cellular Screening for Biological Research and Drug Discovery

***Leanne Bischof**, Changming Sun, Ryan Lagerstrom, Dadong Wang*
CSIRO

In modern biology, important biological information is often captured in the form of images. Extracting the information from those images manually can be tedious and time consuming. There is increasing demand for software to perform this analysis automatically. Modern image analysis techniques are making it possible to automate even the most challenging applications. I will illustrate what is possible by referring to our work in High Content Analysis (HCA). HCA refers to the automated analysis of fluorescence microscope images of cells to extract visual features which quantify cellular morphology and function. High content analysis is used in the pharmaceutical industry to screen candidate drugs for efficacy and toxicity. It is increasingly being used in academia to expand the fundamental understanding of cellular biology. I will briefly mention some of the image segmentation and feature extraction techniques that we use and show a series of biological assays which require a range of analysis techniques. These HCA assays will include neurite outgrowth analysis in 2D and 3D, analysis of mixed cell populations (such as neuron-astrocyte co-cultures), and the analysis of protein translocation and sub-cellular localisation. Increasingly HCA is requiring the extraction of more and more complex features. This places demands on both the statistical analysis of those features and also on the speed of processing to extract them. I will also mention our use of GPU and multi-core based batch processing to achieve the throughput required by the pharmaceutical industry in screening campaigns involving hundreds of thousands of images.

Traditional Versus Fractal Analysis of the Foreign Exchange Rates

Maria Bohdalova, Michal Greguš
Comenius University

We will concern to the nonparametric method for modeling of the financial times series, in this paper. We use the concept of the fractal dimension to the measurement of the complexity of time series of observed financial data. The aim of this paper is distinguish between randomness and determinism of the financial information. We will compare fractal analysis and dynamic fractal analysis of the selected foreign exchange rates and Down Jones Industrial index. Fractal analysis was introduced into financial time series, by Mandelbrot and Peters. Due to financial crisis this theory was renewed. Fractal analysis indicate, that traditional econometric methods are inadequate for analyzing financial time series. Adequate analysis of the financial time series

allows us to predict precisely their future values and risks connected with portfolios that influenced.

Analyzing Policy Risk and Accounting for Strategy: Auctions in the National Airspace System

James Bono
American University

We examine the potential for simple auction mechanisms to efficiently allocate arrival and departure slots during Ground Delay Programs (GDPs). The analysis is conducted using a new approach to predicting strategic behavior called Predictive Game Theory (PGT) [see Wolpert (2010)]. The difference between PGT and the familiar Equilibrium Concept Approach (ECA) is that PGT models produce distribution-valued solution concepts rather than set-valued ones. The advantages of PGT over ECA in policy analysis and design are that PGT allows for decision-theoretic prediction and policy evaluation. Furthermore, PGT allows for a comprehensive account of risk, including two types of risk, systematic and modeling, that cannot be considered with the ECA. The results show that a second price GDP slot auction dominates a first price GDP slot auction in many decision-relevant categories, including higher expected efficiency, lower variance in efficiency, lower probability of significant efficiency loss and higher probability of significant efficiency gain. These findings are despite the fact that there is no *a priori* reason to expect the second price auction to be more efficient because none of the conventional reasons for preferring second price over first price auctions, i.e. dominant strategy implementability, apply to the GDP slot auction setting.

On the Reduction of a Spatial Monitoring Grid: Proposals and Applications to Semiconductor Processes

Riccardo Borgoni
Department of Statistics, University of Milano-Bicocca

Luigi Radaelli
SPC & Robustness Group, Numonyx, Italy

Valeria Tritto
IRER Istituto regionale di ricerca della Lombardia, Milano, Italy

To focus on the issue we are going to present, start considering the fabrication of integrated circuits (IC). The process consists of a sequence of several different physical and chemical steps performed on a thin silicon slice, called wafer. In some production processes it is necessary to grow a silicon oxide (SiO2) layer over the wafer surface. The thickness of this layer has, in general,

a target value. However the actual value one gets in practice is not constant over the wafer surface but it can randomly deviate from the target due to structural causes. In order to control the deposition process and ensure optimal IC performances, the SiO2 layer thickness across the wafer and across the boat must be carefully evaluated. For this reason, maps containing wafer coordinates, where thickness must be measured, are available. Assessing whether the deposited film thickness is uniform (or almost uniform) over the wafer surface is essential to further steps in chip manufacturing. Data collection procedures are time consuming and expensive: for this reason it is often worthy trying to reduce the number of points that are necessary to accurately reproduce the film thickness over the wafer surface. The issue previously addressed may be found in all those production processes where the quality of a surface/volume must be kept under control. Reducing a spatial monitoring grid requires to select a subset of the original measurement points in such a way that the "best possible" estimate of the variable of interest is returned. In collaboration with Numonyx, a worldwide semiconductor manufacturing company, we have applied the simulated annealing (SA) method and compared its performance with a new alternative method, we called ZBR (Borgoni R. et al, 2009). SA was firstly employed in spatial sampling by Sacks and Schiller (1988) and extended by Van Groenigen and Stein (1998). ZBR starts from assigning to the surface an optimal map, according to some criterion, and then it searches for that subset from the original map nearest too the optimal. In both cases the optimal solution and the maximum reduction factor of the original map go through the computation of a prediction error which has not a unique solution and mostly depends on what the experimenter wants to keep under control. SA has been used within a kriging predictor paradigm in order to exploit the spatial correlation that may be consistently estimated if data are available. ZBR has been thought for both/either those contexts where only the map is available and/or a parametric response surface over the wafer is a priori known/expected because of technological reasons. For the latter case the prediction error has been computed using a spatial parametric logvariance model. The performances of the two methods will be presented using data from real semiconductor processes.

Some Issues in the Design and Analysis of Industrial Split-Plot Experiments

John Brewster
University of Manitoba

In industrial experiments, some factors are often easier to vary than others. In many cases, this leads to restrictions on randomization and a split-plot structure. In this talk, we will examine some issues that arise is both the design and analysis of such experiments. At the design stage, these issues arise in designing screening and optimization experiments and in designing follow-up experiments. At the analysis stage, some of the issues revolve

around the use of inferential procedures that are not faithful to parameter constraints induced by the whole-plot and subplot variance components. The frequentist properties of alternative procedures that have both conditional and Bayesian interpretations will be presented.

Efficient Estimation of Learning Models

Laurent Calvet, Veronika Czellar
HEC Paris

In securities markets, investors do not directly observe the full state of the economy, but must typically infer it from available financial and macroeconomic news. The investors' sequential learning process can have profound implications for asset valuation and aggregate activity. For instance during the recent financial crisis, market participants had to impute the solvability of homeowners and financial institutions from the scarce data available, and the resulting uncertainty severely impacted liquidity and asset valuations. In asset pricing theory, a growing body of research has investigated the implications of investor learning (e.g. Brennan 1998; Brennan and Xia 2001; David 1997; Guidoli and Timmermann 2003; Lettau Ludvigson and Wachter 2008; Pastor and Veronesi 2009; Timmerman 1993, 1996; Veronesi 1999, 2000). For tractability reasons, the state of the economy, such as the growth rate of aggregate dividends, is assumed to switch between a small number of values. In order to obtain sizeable effects from learning, the models are calibrated at low (e.g. yearly) frequencies, a feature that considerably limits their applicability to real time forecasting. A solution to these difficulties is offered by the Markov-Switching Multifractal (MSM) introduced in Calvet and Fisher ("CF" 2001, 2004). MSM assumes that the economy is driven by a large number of components with heterogeneous frequencies. In volatility applications, MSM matches the extreme returns and intertwined volatility cycles of multiple durations exhibited by financial data. Furthermore, MSM paves the way for real-time applications of learning models. In this paper, we develop a toolkit of inference and forecasting methods for a large class of structural models with learning. Our approach builds on the observation that in many instances, the structural model is easy to simulate and its full-information version has a closed-form likelihood. We develop an accurate estimation method based on indirect inference (Smith 1993, Gouriéroux Monfort and Renault 1993, Czellar Karolyi and Ronchetti 2007, Czellar and Ronchetti 2009). An auxiliary estimator is defined by expanding the maximum likelihood estimator of the full-information economy with a set of statistics that investor learning is designed to capture. The indirect inference estimator is chosen so that the auxiliary estimator in simulated samples matches the auxiliary estimator computed from the data. We develop an indirect inference estimator for the multifrequency learning model of CF (2007), which we apply to the daily excess returns on a U.S. aggregate equity index between 1926 and 2008. Estimation accuracy is verified by Monte Carlo simulations, and filtering techniques are developed to impute investor beliefs

from market returns. We show that the learning model provides better forecasts of stock return volatility and value at risk than its full-information counterpart, and also outperforms some of the best reduced-form econometric models.

A Multifrequency Theory of the Interest Rate Term Structure

Laurent Calvet
HEC Paris

Adlai Fisher
University of British Columbia

Liuren Wu
Baruch College

The interest rate term structure responds to shocks of all frequencies. At high frequencies, large transactions of a particular fixed-income instrument can significantly move rates at the associated maturities, followed by quick dissipation along the yield curve through hedging practices. At intermediate horizons, central banks have historically implemented monetary policy through their influence on a short-term rate. Monetary surprises thereby directly impact the short-end of the yield curve, and spread across the whole term structure through their influence on market expectations of future short rate movements (e.g., Balduzzi, Bertola, and Foresi 1997, Piazzesi 2005, and Heidari and Wu 2009). In the long run, positive shocks to inflation raise the interest rate level across all maturities, whereas positive shocks to real output growth raise short-term rates more than long-term rates. In the dynamic term structure literature, no-arbitrage conditions impose strong restrictions on the cross-sectional relation between interest rates of different maturities. Despite the rapid progress in this literature over the past decade, the focus of empirical work remains on low-dimensional models, most typically with three factors. Substantially higher-dimensional affine models have not previously been considered practical because of the classic curse of dimensionality that plagues model identification. A generic three-factor model can have over 20 free parameters, many of which cannot be estimated with statistical significance, and the number of parameters grows approximately quadratically with the factor dimension. However, restricting attention to low-dimensional models may inhibit empirical performance in several areas. First, low-dimensional models face limitations in the cross-sectional fitting of observed interest rates across different maturities. Although the fitting errors can appear small relative to the average interest rate level, the errors can become economically significant when one forms interest rate portfolios to neutralize the exposure to low-frequency movements (Bali, Heidari, and Wu 2009) and similarly impact the pricing of interest rate options (Heidari and Wu 2008). Second, low-dimensional term structure models often imply high cross-correlations between interest rates changes of different maturities, but the actual cross-correlation estimates are often much lower (Dai and Singleton (2002)). Finally, low-

dimensional models generate poor forecasting performances, often worse than the performance of a simple random walk assumption (Duffee 2002). In this paper, we develop a class of affine term structure models that accommodates many interest-rate factors with a small number few parameters. The model builds on a short-rate cascade, a parsimonious recursive structure that naturally ranks the latent state variables by their rates of mean reversion, each revolving around the next lowest frequency factor equating to a level in the cascade. With appropriate assumptions on factor volatilities and risk premia, the model overcomes the curse of dimensionality associated with general affine models, permitting a finite parameter vector to describe term-structure dynamics for an arbitrary number of factors. Using a panel of 15 LIBOR and swap rates, we estimate models using from one to 15 latent factors and only five parameters. High-dimensional specifications substantially outperform lower-dimensional specifications both in- and out-of-sample. The in-sample fit is near exact, with absolute pricing errors averaging less than one basis point, permitting yield-curve stripping in an arbitrage-free, dynamically consistent environment. Out-of-sample interest rate forecasting shows significant improvements over traditional benchmarks, and cross-maturity correlations are more accurate than low-dimensional models.

Twin Picks: Disentangling the Determinants of Risk-Taking in Household Portfolios

Laurent Calvet
HEC Paris

Paolo Sodini
Stockholm School of Economics

This paper investigates the determinants of financial risk-taking in a dynamic panel containing the asset holdings of Swedish twins. We use this novel dataset to answer some key questions in financial economics. Does wealth drive the share of risky assets in the portfolios of individual investors? Is the financial wealth elasticity of the risky share homogenous in the population? How does the aggregate demand for risky assets respond to changes in the wealth distribution? The empirical household finance literature provides only partial answers to these questions. In cross-sections, richer and more educated investors are known to allocate a higher proportion of their financial wealth to risky assets than less sophisticated households (e.g. Calvet Campbell and Sodini "CCS", 2007, 2009a&b). In addition, the risky share has a negative cross-sectional relation to real estate holdings, leverage (Guiso Jappelli and Terlizzese 1996), and internal consumption habit (Lupton 2002). It is unclear, however, whether these variables directly impact portfolio choice, or simply proxy for latent traits such as ability, genes, risk aversion, or upbringing. Several recent papers suggest that panel data offer a possible solution to this identification problem (e.g. Brunnermeier and Nagel 2008, CCS 2009a, Chiappori and Paiella 2008). One difficulty with the dynamic panel approach is

that the researcher needs to control for household inertia by using instruments, and the results are sensitive to the validity of the instruments. In this paper, we consider an alternative estimation strategy based on the comparison of the financial portfolios held by twins. The analysis is made possible by a novel dataset containing the disaggregated portfolios and detailed characteristics of twins in Sweden. We observe the worldwide assets owned by each twin at the end of a tax year, including bank accounts, mutual funds and stocks but excluding retirement accounts. All holdings are reported at the asset level for the 1999–2002 period. We estimate panel regressions of the risky share on a broad set of household characteristics, and use yearly twin pair fixed effects to control for genes and shared background. We report a strong positive relation between risky asset market participation and financial wealth. Among participants, the average financial wealth elasticity of the risky share is significantly positive and estimated at 22%, which suggests that the average individual investor has decreasing relative risk aversion. Furthermore, the financial wealth elasticity of the risky share itself is heterogeneous across investors and varies strongly with characteristics. The elasticity decreases with financial wealth and human capital, and increases with habit, real estate wealth and household size. As a consequence, the elasticity of the aggregate demand for risky assets to exogenous wealth shocks is close to, but does not coincide with, the elasticity of a representative investor with constant relative risk aversion. We confirm the robustness of our results by running time-differenced instrumental variable regressions, and by controlling for zygosity, lifestyle, mental and physical health, the intensity of communication between twins, and measures of social interactions.

Box-Cox Transformation in Acceptance Sampling Plans

Elisabete Carolino
ESTeSL, IPL, Portugal

Isabel Barão
DEIO, FCUL, Portugal

In the quality control of a production process (of goods and services), from a statistical point of view, focus is either on the process itself with application of Statistical Process Control, or on its frontiers, with application of Acceptance Sampling (AS) – studied here – and Experimental Design. AS is used to inspect either the output process – final product – or the input – initial product. The purpose of AS is to determine a course of action, not to estimate lot quality. AS prescribes a procedure that, if applied to a series of lots, will give a specified risk of accepting lots of given quality. In other words, AS yields quality assurance. An AS plan merely accepts and rejects lots, considering sampling information. The AS by variables is based on the hypothesis that the observed quality characteristics follow a known distribution, namely the Gaussian distribution (classical case of the AS by variables – treated in classical standards). This is sometimes, however, an abusive assumption, that leads to wrong decisions. AS for non-Gaussian, mainly asymmetrical variables,

is thus relevant. When we have a non-Gaussian distribution we can build specific AS plans associated with that distribution. If the real distribution of data is very asymmetric and/or has heavy tails, but we are able to adequately model the data and estimate its parameters, which usually is not easy, we can use those specific AS plans. Alternatively, we can make the transformation of the original data into normal values through a transformation of the Box-Cox type, which requires no prior modeling process of the data and then use AS plans for the classical case – the Gaussian case. In this work we will address the problem of determining AS plans by variables for Extreme Value distributions, both methods being compared.

Applications of Chemometric Techniques for Geochemical Studies of Environmental Samples

Claudia Elena Casalino
Utrecht University

Over the last 20 years considerable studies have been carried out in the field of environmental research. Most often they have been devoted to the chemical characterization in terms of metals content with the aim of evaluating their impact and behaviour. To better understand and to interpret the large amount of results, chemometric methods have become an integral part of environmental studies. Specifically the methods of multivariate statistical analysis are essential tools that allow us to obtain a better visual representation of the results and to find out similarities and differences among samples and correlation between variables which would be more difficult to detect just observing the numbers in the tables. In this talk I will present examples of chemometric methods applied to studies related to environmental samples (e.g. marine sediment from Antarctica, estuarine sediment from UK). In particular applications of principal component analysis (PCA) and hierarchical cluster analysis (HCA) will be shown.

A Study on Estimating Methods for Aged 80 and Above

*Kee-Whan Kim, Hak-Min Lee****, Chun-Kyung Cha***
Korea University

The aging of population is an inevitable issue in developed countries. It also causes serious problem in South Korea. According to UN standards, we measure the degree of aging of society as the population aged over 65, but the population of people aged over 80 is a major concern in aging society. Since collecting and compiling statistics on very elderly people is more difficult than statistics on other ages, each country correct the data for right use or may use the estimated results. In South Korea, the population data like census data,

residential data and estimated population data from Korean Statistical Information Service (KOSIS) is inconsistent and is insufficient for studying people aged 80 and over. In this paper, we discuss some problems of high ages in 'Survey results of aged over 100' reported by Statistics Korea in 2006 and in the official population statistics in South Korea. To cope with this problem, we review estimating method for high ages. And we propose supplementary ways of estimating people aged 80 and over in South Korea with estimating results for them.

High Temperature Extreme Values In The Climate Change Context: EDF Experience

Christophe Chaussin, Sylvie Parey
EDF/R&D

Thi Thu Huong Hoang, Didier Dacunha-Castelle
Université Paris Sud

The dimensioning of buildings and industrial installations is based on rare, extreme events susceptible to occur during their lifetime and to which they must withstand. These rare levels are generally derived using the statistical extreme value theory. This theory however supposes that the studied series are stationary, which can no more be stated for water or air temperature, for example. Different ways of dealing with these problems have been proposed and tested at EDF/R&D: identification and extrapolation of trends in the parameters of the extreme value distributions, use of the link between the evolution of the mean and standard deviation of the central field and the evolution of extreme events. The main encountered difficulties and the proposed solutions will be illustrated with examples for water and/or air temperature series concerning the determination of high temperature return levels at different future time scales.

A Study of Multivariate Control Charts

Su-Tsu Chen
Department of Occupational Safety and Hygiene, Fooyin University

Jeh-Nan Pan
Department of Statistics, National Cheng-Kung University

For multivariate processes, various control charts including the MCUSUM and MEWMA charts were proposed to monitor small process mean shifts. However, the calculation of optimal parameters assumes that the distribution of a manufacturing process is known. This may not be true in practice. To obviate the difficulties of choosing optimal parameters when using control charts and

develop a simple yet user-friendly multivariate control chart, a multivariate extension of Shewhart-like CUSUM control chart will be proposed. The performance of this chart will be studied and compared with the MEWMA and MCUSUM charts in terms of in-control and out-of-control ARLs.

Confidence Intervals for Lognormal Means in Small Samples

Su-Tsu Chen, *Guan-Chyun Lin*
Department of Occupational Safety and Hygiene, Fooyin University

Trey-Shye Wang
Department of Biotechnology, Fooyin University

The lognormal distribution has frequently used to model data in environment, biomedicine, and even applied nuclear science and technology. The inference used to focus on arithmetic means and now focuses on confidence intervals. Several confidence interval estimation methods, including Cox's method, the bootstrap method, and other methods, are discussed by researchers. However, when only a small sample is obtained, these methods may not lead to a desired result such as that of a large sample. In this paper we discuss the coverage error of confidence intervals of Cox's method in small samples.

Stochastic Graph Models and Their Application in the UK 2001 Foot-And-Mouth Epidemic

Shojaeddin Chenouri, *Yasaman Hosseinkashi, Christopher Small*
University of Waterloo

A dynamic random graph is defined as a sequence of random graphs that changes over time. In this paper, we consider a statistical inference problem in dynamic random graphs when the edge information is not available. The problem is motivated by the foot and mouth disease (FMD) outbreak in UK 2001. A dynamic Euclidean graph model with a Markov property is introduced and applied in analyzing this epidemic. The model provides a probability distribution over unobserved infectious pathways (edges) based on the individual farm (vertex) attributes and their Euclidean distances. The cumulative resistance and threat associated with each farm is measured based on the indegree and outdegree of the dynamic graph. Also the basic reproductive number (R0) is estimated in a more general framework using the mean outdegree of the infectious network.

Measurement for Improvement in Practice

Shirley Coleman
Newcastle University

Six sigma adds strategic focus and management accountability to the application of statistical methods. Quality improvement projects benefit greatly when care is taken with this perspective. Six sigma includes many excellent statistical methods, for example, comparative tests, survey sampling, confidence intervals and designed experiments. Lean six sigma adds further excellent techniques, such as flow charting and value stream mapping. Lean methods have been widely embraced by the NHS in the UK. It appears that it is the lean tools, rather than the statistical tools which are most meaningful and accessible to healthcare staff. Statistical methods have been taught in many waves of quality improvement effort spanning many decades. Of all the statistical methods included in such training, statistical process control is the most widely appreciated. Charts are a familiar mode of data presentation in healthcare and SPC continues to be implemented in a wide range of healthcare situations. SPC incorporates many of the fundamentals of quality improvement programmes: strategic objectives, operational definitions, measurement systems analysis, team working, data collection and visualisation, variation and testing. SPC also leads to design of experiments as special causes are identified. ISRU are currently working alongside 14 trusts in UK in conjunction with the Patient Safety First project. Measurement for improvement was identified as the key initiative to be undertaken by a progressive NHS Trust in the NE of England. SPC is an important part of this initiative and the practical experiences of staff involved will be reviewed.

Analysis of Unreplicated Factorial Designs – Methods' Performance

Nuno Costa
Instituto Politécnico de Setúbal - Escola Superior de Tecnologia de Setúbal, Campus do IPS UNIDEMI – Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

José Palma
Instituto Politécnico de Setúbal - Escola Superior de Tecnologia de Setúbal, Campus do IPS

Zulema Pereira
UNIDEMI – Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

A technique widely used for analyzing unreplicated factorial designs with factors at two levels, which have been increasingly used in industrial settings due to economical and technical reasons, is the (half-)normal probability plot. Recently, Whitcomb and Oehlert (2008) presented an extension to the Daniel plot of effects that permits analyzing factors with more than two levels,

arguing that half-normal plot is less intimidating than numerical methods, is easy to understand and it provides a degree of protection against multiple testing. However, as the selection of active factors by visual inspection of the (half-)normal probability plots depends on the analyst's sensibility and knowledge of the process and product, these plots should preferably be supplemented with a more formal procedure (Olgín and Fearn, 1997). Such as shown by Adams (2002), fails occur in the separation of the significant factors from the insignificant ones merely by visual analysis of the normal probability plot, especially if this task is performed by novice users. Despite the improvements proposed by Zahn (1975) and Olgín and Fearn (1997), the formal procedure using the half-normal plot suggested by Daniel (1945) has been rarely used. In order to provide an alternative or complement to the graphical technique, various analytical methods have been put forward in literature. An extensive review and comparison on method's performance is provided by Chen and Kunert (2004) and Hamada and Balakrisnan (1998). While Hamada and Balakrisnan consider a different number of active effects of the same magnitude, Chen and Kunert (2004) consider more realistic situations that include a different number and magnitude of active effects. Both studies compare method's performance controlling them for a specific value of either IER (Individual Error Rate- the average proportion of inactive effects declared active) or EER (Experimental-Wise Error Rate) value. This is statistically sound, but comparing methods' performance on an exactly equal basis (without destroying their essence) is not possible because of the different forms of the methods (Hamada and Balakrisnan, 1998). To complement the published research works, we assessed methods' performance for different number and magnitude of active effects when no calibration is made in the method´s error rates. For that purpose we ran 5000 independent samples for each one of the different settings considered and used various criteria for assessing the method's performance. The results achieved are valid for two level factorial designs with 16- and 32-runs where active effects of the same and different magnitude are considered under the assumption of normality, independence and homogeneity of variance. Based on the results of this research we also propose a methodology for analyzing unreplicated factorial designs.

Data Quality Mining

Tamraparni Dasu
AT&T Labs Research

Data types and data structures are becoming increasingly complex, as they try to keep pace with evolving technologies and applications. Data streams, web logs, wikipedias, biomedical applications, video streams and social networking websites generate a mind boggling variety of data types. The result is an increase in the number and complexity of data quality problems. Traditional static notions of data quality no longer suffice. Data quality mining is the application of data mining techniques to manage, measure and improve data

quality in large data sets and data streams. This talk will provide a brief overview of the disciplines of exploratory data mining and data quality mining. It will introduce research, methods and tools that are derived from multiple disciplines such as statistics, computer science and databases, while focusing on applications and real world examples. The talk will conclude with a discussion of new directions for research, such as the detection and analysis of multi-dimensional and temporally correlated data glitches.

Nonstationary Extremes and the US Business Cycle

Miguel de Carvalho
Banco de Portugal / Universidade Nova de Lisboa

K. Feridun Turkman
Universidade de Lisboa

António Rua
Banco de Portugal - Department of Economic Studies

Considerable attention has been devoted to the statistical analysis of extreme events. Classical peaks over threshold methods are a popular modelling strategy for extreme value statistics of stationary data. For nonstationary series a variant of the peaks over threshold analysis is routinely applied using covariates as a means to overcome the lack of stationarity in the series of interest. In this paper we concern ourselves with extremes of possibly nonstationary processes. Given that our approach is, in some way, linked to the celebrated Box-Jenkins method, we refer to the procedure proposed and applied herein as Box-Jenkins-Pareto. Our procedure is particularly appropriate for settings where the parameter covariate model is non-trivial or when well qualified covariates are simply unavailable. We apply the Box-Jenkins-Pareto approach to the weekly number of unemployment insurance claims in the US and exploit the connection between threshold exceedances and the US business cycle.

Tracking the US Business Cycle with the Singular Spectrum Filter

***Miguel de Carvalho**, **Paulo C. Rodrigues**, António Rua*
Banco de Portugal / Universidade Nova de Lisboa

The monitoring of economic developments is of particular importance for policymakers, namely, central banks and fiscal authorities as well as for other economic agents such as financial intermediaries, firms and households. However, the assessment of the business cycle is not easy as the cyclical component is not an observable variable. This paper focuses on the business

cycle of one leading economies in the world – the US. We make use of singular spectrum methods in order to disentangle the most comprehensive measure of economic activity, Gross Domestic Product, in the several underlying components. The band-pass features of the singular spectrum filter are here exploited in order to discern the US business cycle. The cyclical component obtained is put at test through a comparison with the US economy contraction and expansion periods dated by the US Business Cycle Dating Committee of the National Bureau of Economic Research (NBER). Furthermore, a comparison with the most popular filters in the related literature, namely the Hodrick-Prescott and Baxter-King filters, is also enclosed here.

Bias Adjusting and Combining Realized Volatilities to Estimate Daily Volatility

Hennie Venter, ***Riaan de Jongh***
North-West University

Daily return volatility can be estimated by realized volatility type estimators if intra-day (high frequency) trade prices are available. Originally realized volatility was defined as the sum of the squared returns over short calendar time intervals during the trading day but currently many different types of realized measures are available. This raises the issue of how to handle the multitude of possibilities in applications. We propose methodology that considers the logarithm of daily volatility as a latent variable which relates to daily returns via a standard stochastic volatility model and to different log realized measures via linear models. In effect this enables us to combine different realized measures into an overall daily volatility estimator. Fitting such models is technically demanding: Kalman Filter and efficient importance sampling approaches are followed here. We evaluate and illustrate our proposals using LSE intra-day share price data.

Local Unbiased Estimation of the Tail Index and Its Application to Fire Insurance Data

Tertius de Wet
University of Stellenbosch

Yuri Goegebeur
University of Southern Denmark

Extreme value theory (EVT) plays an important role in many areas of finance, e.g. in value-at-risk, large losses, large insurance claims, etc. In many applications covariate information is available for improved estimation. In this talk we develop and study in the framework of Pareto-type distributions,

a class of nonparametric, unbiased kernel estimators for the conditional tail index. The estimators are obtained by local estimation of the conditional tail index using a moving window approach. Since these unbiased kernel estimators depend on a second order tail parameter, a locally consistent estimator for the latter is introduced and the asymptotic normality of the bias-corrected kernel estimators is derived using this local estimator. Some results of a finite sample simulation are given and an application of the results to a data set of fire insurance data is discussed.

Process Improvement in Healthcare

Ronald Does
University of Amsterdam

Perhaps the first connotation with the topic of healthcare improvement is innovation in medical science, including innovations in treatment protocols, medical equipment, and pharmaceuticals. This lecture, however, focuses on the improvement of healthcare by improving its delivery. Healthcare delivery concerns the operating routines in hospitals, including primary patient processes, medical support processes, and nonmedical support processes. Characteristics of these processes, such as their capacity, efficiency, and reliability, determine important performance dimensions of healthcare, such as throughput, patient safety, and waiting times. Ultimately, they have a substantial impact on patient satisfaction, cost, and the quality and timeliness of medical care. A process is a repeated series of activities that act on jobs, transforming input into output. Processes, thus, are operating routines in the vocabulary of evolutionary economics. The majority of activities in professional organizations are done as routines, and routinization (that is, turning into a process) of activities constitutes the most important form of storage of an organization's specific operational knowledge. In the process improvement paradigm, improvement originates in mapping, describing and measuring processes in an organization. Typical improvement points emerging in process improvement projects include. – Optimizing capacity and utilization of staff and equipment, ensuring a smooth workflow with acceptable waiting times, and reducing cost for personnel and equipment. – Reducing throughput times and waiting times by identifying bottlenecks and iterations in the processes. – Optimizing or introducing standardized routing through the process, such as introducing prioritization rules for queue management, introducing restrictions on the amount of work-in-process as in kanban and CONWIP, or replacing batch-wise work with a single-piece flow discipline. – Improving a process's reliability and safety by mitigating failure opportunities and making the process more robust. – Reducing cycle times per task by optimizing work methods and procedures. – Reducing length of stay of patients by better planning. In this lecture we give an overview of our experiences in many hospitals in the Netherlands and Belgium during the past ten years. It includes a description of generic improvement projects in healthcare with some real life examples.

Successive Events in Transportation of Dangerous Media

Gejza Dohnal
Czech Technical University in Prague

The notion of successive events involves such situations in which the strong dependence must be considered. On the contrary with a sequence of various traffic accidents in time, equipment failures or nonconforming product release, which occur independently in time, this kind of events comes collective, due to some common cause. As examples we can use some disasters in oil production, attacks of IT systems, wildfire spreading, epidemic progression and so on. The transportation of such dangerous media as oil, gas or chemicals, are in the focus of both transport companies and environmenalists. In the lecture, some type of events, known as disastrous, will be dealt with. Unlike a usual failure, disastrous events have the ability of dynamic spreading. However as a rule they cause nonreversible changes. Usually they are not repairable in a relatively short time. Their consequences are often deep-going. There are no doubts that an impact of disastrous events on human society can be severe. The model for the dynamic spreading of disastrous events as networked systems will be presented. We consider a disaster as a time sequence of single events, which spreads from an initiating event (parent, focus) to other nodes of network in a cascade-like manner. Considering the system of all objects which can be affected by a consequence of some initiative disastrous event as a whole along with all its macro-states, we can use markov chains to model events spreading. Then, we are able to compute some predictions such as the lifetime of the system, first affect time to selected object and others. This modeling can help us to make some preventive decision, optimize transport plans or to prepare disaster recovery plans. In the contribution, the model will be described and some computations will be outlined.

Review and Development of a Risk Rating System for Ocean Going Vessels

John Donnelly, Ross Sparks*, Gordon Sutton*
CSIRO Australia

CSIRO has been collaborating with RightShip Pty Ltd in Australia to review and develop the RightShip star rating system. This system allocates a star rating on a scale of one to five to certain commercial vessels as an estimate of their risk of encountering casualties and incidents which may hinder or prevent the successful transportation of cargo from one port to another. The current system rates vessels on selected criteria and then combines the resulting rating scores in an ad hoc but sensible way to obtain a final score which is converted to a star rating. Currently, the weightings, sensitivities and dependencies of the different variables in relation to the final star rating are determined by a combination of expert opinion and basic statistical techniques.

A vessel given a rating of five stars has a very low risk and hence would be a preferred choice for transporting cargo. The aim of the study was to: (1) Review the variables/parameters used in the current method and how the rating is determined. (2) Identify possible issues with the current methodology. (3) Focus on development and assessment of alternative models, giving details of analyses and selected algorithms. (4) Using a 90 day window, assess the performance of proposed algorithms against the current rating method and make recommendations as to which method or combination of methods add further value to the existing method. Problems encountered with the data, the results and outcomes of the study will be presented, including discussion of the various statistical models investigated.

A Generalized Brown-Proschan Model for Preventive and Corrective Maintenance

Laurent Doyen
Grenoble University

In the study of repairable systems reliability, the basic assumptions on repair efficiency are known as minimal repair or As Bad As Old (ABAO) and perfect repair or As Good As New (AGAN). Obviously, reality is between these two extreme cases. This is known as imperfect repair. One of the most famous imperfect repair model is the Brown-Proschan (1983)(BP) model in which the system is perfectly repaired with probability p and minimally repaired with probability $(1-p)$. This random effect of repair actions can be represented with random variables B(i): $B(i)=1$ if the i th repair is AGAN, and $B(i)=0$ if the i th repair is ABAO. Authors have usually supposed that the effect of each repair, B(i), is known. But, in practical case, repair effects are unknown. To our knowledge, only four papers deal with BP model with unknown repair effects. Lim (1998) has estimated the parameters of the first time to failure distribution and repair efficiency with the Expectation-Maximization algorithm. Lim and Lie (2000) have used a SEM algorithm in order to estimate the parameters of a generalization of the BP model that allows first-order dependency between two consecutive repair effects. But they have assumed that only some repair effects are unknown. Lim, Lu and Park (1998) have proposed another method based on Bayesian analysis: they have assumed a prior beta distribution for the parameter p. Langseth and Lindqvist (2004) have generalized the BP model in the case of imperfect preventive maintenance. They have proposed to estimate the parameters of the model (including the repair efficiency one) with the likelihood function. All these papers propose a single estimation of the parameter p that represent the average efficiency of every repair or maintenance actions. Then, the drawback of all these methods, and more generally of all maintenance efficiency estimation methods (Shin Lim Lie (1996), Jack (1998), Yun Choung (1999), Karminskiy Krivstov (2000), Doyen Gaudoin (2004)), is that the efficiency of each maintenance is never individually assessed. In this paper we propose a model, with Brown-Proschan PM effect and ABAO CM effect. Several methods based on likelihood

maximization are proposed to estimate the parameters of the first time to failure distribution and the PM efficiency parameter p. Thanks to this model and using a hidden reconstruction method, the efficiency of each preventive maintenance is individually assessed. Forecast reliability indicators are also derived. The framework proposed in this paper can take into account a left and right censorship of the failure process. Finally all the results are applied and validated on a real data set issued from the French electricity manufacturer EDF.

Modelling the Effects of Adverse Weather on the Power-Distribution Network

Carl Duckworth, *Neil Sandison, David Macleman*
Scottish and Southern Energy Plc

Carl Donovan, Monique Mackenzie
University of St Andrews

The overhead lines of a power distribution network are particularly susceptible to weather-related faults. Predicting the magnitude and locale of fault events is important for the pre-emptive deployment of resources, to minimise customer interruptions and the resulting loss of income and regulator penalties. We present the results of a multi-year study for collating extensive fault databases and building predictive models on the basis of observed and predicted weather events. The Scottish power network is considered here – a challenging environment to distribute power over, with both a sparse customer base and harsh weather. The models are based on the Generalized Additive Model (GAM) framework and provide improvements over the existing predictive tools. The problem is multi-faceted and practically difficult.

The Coming Revolution in Statistics

Lee Edlefsen
REvolution Computing

We are going to see a major revolution in statistical practice and theory over the next several years, spurred by disruptive technological change. For many years it has been assumed that increases in hardware speeds would allow us to continue using our existing algorithms and code as data set sizes increased. However, the explosion in the amount of digitized data combined with the inability of the speed of single cores and hard drives and of RAM to keep up have made it clear this is not the case. To deal with the amount of data becoming available, we must take advantage not only of multiple cores but also of multiple computers, and the vast majority of existing code for doing

statistical computations is incapable of doing either of these. I believe this fact will lead to revolutionary changes not only in statistical computation and statistical graphics but also in statistical theory, as we come to realize that huge samples are qualitatively different than small ones.

Monitoring the Intra-Stream Variance in Multiple Stream Processes

***Eugenio Epprecht**, Bruno Simões*
Catholic University of Rio de Janeiro

In a multiple stream process (MSP), a same quality variable is measured in several streams in parallel. The first tool proposed for monitoring MSPs was the Group Control Chart (GCC) by Boyd in 1950. Its efficiency is impaired by the presence of cross correlation between streams. A useful model for MSPs (Mortell and Runger, 1995) represents the value of the quality variable in each stream at any time t as the sum of a random variable (or stochastic process) b(t) that is common to all streams, plus the individual variation of each stream relative to b(t). Mortell and Runger proposed monitoring the individual streams components by the range Rt of the stream subgroup averages (or of the stream individual observations if this is the case) and also suggested, as an alternative (which they have not explored), a GCC of the residuals of the streams to the average of all streams at each sampling time. This alternative has been recently studied by Epprecht and Barbosa (ISBIS 2008). Runger, Alt and Montgomery (1996) proposed an alternative chart, S2, using as monitoring statistic the variance between streams at each sampling time. All these schemes were devoted to monitoring the mean of the individual components of the streams; to the best of our knowledge, no previous work considered the case of increases in the variance of a stream. We develop four different GCCs for monitoring the inner variability of the individual streams: a GCC of S2, the sample variances of each stream (which is not the same as Runger, Alt and Montgomery's S2); a GCC of EWMA(lnS2); a GCC of the Moving Ranges of the residuals of each stream to the estimate of b(t), and an EWMA version of it. The last two GCCs cater for the case where only individual observations are feasible, which is frequent with a large number of streams. We analyze the ARL performance of every one of these schemes, in a variety of situations. We also evaluate the ARL performance (in the case of increases in the variance of one stream) of the schemes designed for monitoring the means of individual streams. The fastest scheme is, as expected, the GCC of EWMA(lnS2). Somewhat surprisingly, its gains in performance relative to the GCC(S2) are not very large. These two schemes largely outperform all other ones. With individual observations, the GCCs based on moving ranges were in many cases analyzed outperformed by schemes designed for monitoring the mean (even though the latter, unlike the former, have not been optimized for changes in the variance).

A Comprehensive PLS Rationale for Multidimensional Blocks in Predictive Path Models

Vincenzo Esposito Vinzi
ESSEC Business School of Paris

Giorgio Russolillo
CEDRIC-CNAM Paris

When studying complex systems, the difficulty of analysis is mainly due to the theoretically hypothesized network of (eventually hidden) causal relationships between concepts indirectly measured by blocks of items. This leads to the problem of extracting information from uncertain models rather than modeling uncertainty. PLS Path Modeling (PLS-PM) is classically regarded as a component-based approach to causal networks and has been more recently revisited as a general framework for multi-block data analysis. Indeed we like to define the type of models dealt by PLS-PM as predictive path models thus focusing more on prediction than estimation accuracy. This approach boasts several applications in the domain of business and industry when multiple sets of variables are observed that underlie different concepts linked to each other by predictive relationships (e.g. in food product development, customer satisfaction, consumer behavior, sensory analysis just to mention a few). Two main modes exist to estimate the outer weights for PLS-PM components: Mode A (for outwards directed or reflective blocks) and Mode B (for inwards directed or formative blocks). In Mode A the generic outer weight used in the outer estimate of the latent variable (LV) is the regression coefficient of the simple linear regression of each manifest variable (MV) on the inner estimate of the corresponding LV. In Mode B, the outer weights are the regression coefficients of a multiple regression model of the inner estimate of each LV on its own MV. Multicollinearity problems (even between only a few MV) may lead to non-significant regression coefficients or to difficultly interpretable weights because of the difference in sign between the regression coefficient of a MV and its correlation with the LV. In order to overcome these drawbacks, we investigated a new way to calculate the outer weights. This approach integrates PLS Regression in order to exploit the shrinkage property of its estimators as well as the data analysis flavor of its results and interpretation tools. We propose two new modes for estimating outer weights in PLS-PM by PLS Regression: the PLScore Mode and the PLScow Mode. Both modes involve integrating a PLS Regression as an estimation technique within the outer estimation phase of PLS-PM. However, in PLScore Mode a PLS Regression is run under the classical PLS-PM constraints of unitary variance for the LV scores, while in PLScow Mode the outer weights are constrained to have a unitary norm thus importing the classical normalization constraints of PLS Regression. Applying PLScore Mode requires to choose a proper number of PLS components in the PLS regression for each block. This allows considering PLScore Mode as a fine tuning of the analysis between two extreme cases: classical Mode A in case of one PLS-component; classical Mode B in case of as many PLS-components as there are MVs in each block. We have empirically proved that PLScow Mode with one component yields the same solution as the recently proposed criterion-based approach by Tenenhaus & Tenenhaus

[PLS'09 conference in Beijing], called New Mode A. If more components are considered (while keeping the normalization constraint on the outer weights), PLScow Mode yields a new range of solutions between New Mode A (one PLS component) and a New Mode B (as many PLS components as there are MVs in a block). The criterion, if any, being optimized by the multi-component solutions of PLScow Mode still needs to be investigated. However, we have empirically shown that New Mode B performs very close to classical Mode B in terms of correlations between adjacent LVs and, in any case, better in terms of covariances. The number of components might be then chosen no longer by cross-validation but so as to maximize a specific fitting criterion, such as the GoF index.

Precision Testing of Iron Ore Samples - Problems with the ISO Protocol

Jim Everett
University of Western Australia

To estimate sampling, preparation and measurement precision, iron ore samples are split into halves at each of the three stages, yielding eight assays. Appropriate calculations of the differences between assay pairs are then made to estimate precision of each stage of sampling. The International Standard for checking precision in iron ore sampling (ISO3085) prescribes a method for identifying outliers that is ambiguous and possibly inappropriate. The procedure also makes an unnecessary implicit assumption, that the differences between assay pairs are normally distributed. If applied as specified, the procedure could lead to precision estimates well below the correct values, thus overestimating the sampling method's capability. This paper suggests an improved method for analysing the sampling, preparation and measurement precisions. A bootstrap procedure is used to estimate confidence limits for the calculated precision estimates. The Anderson-Darling statistic is used to test whether the assay pair differences have distributions significantly different from normal. The discussion is supported by extensive simulation modelling of realistic data.

Definition of Great Levels of Parameters of Eletrical Stunning Equipment Through Experimental Optimization

***Karla Faccio**, **Vera Lúcia Milani Martins**, Liane Werner*
UFRGS - Brazil

Nowadays companies are concerned with minimizing costs. During the production process the reduction of failures and identification of the factors affecting the quality of the product represent opportunities to obtain this goal.

The slaughter process of animals, in the food industry, influences the quality of the meat products. A special equipment is used in this process, the electrical stunning, whose function is to cause unconsciousness of the animal by means of electrical discharges in the heart and in the brain to further slaughter. However, if the parameters levels of the electrical stunning equipment are not well adjusted, the electrical discharge can result in reactions that cause hematomas on the animal with consequent loss in meat quality. This type of equipment creates significant losses by lack of determination the most appropriate levels of parameters such as voltage and frequency of discharge applied to the animals. For food industry, the slaughterhouse represents a little-explored portion with regard to improving the definition and control of critical parameters at slaughter moment, essential to the quality of the industrialized product. In this context, studies perform that identify the optimal configuration provides guaranty to the equipment developers on advantage customer requirements in less time and cost, acting in the project for improvement of the product. The techniques of experimental design allow planning a sequence of tests in organized manner, maximizing resources and time. The efficiency of these experimentations is superior in terms of information the any other unstructured sequence of trials (shape empirical) among the final product characteristics and their parameters. Based on this mathematical model identifies the best combination of values specified for each parameter in the product project, reflecting in attendance to the requirements demands by the market. Therefore, this paper aims to present the determination of optimal values of these parameters using design of experiments techniques. The central composite design is what that presents features relevant to this case once maximizes the information with reduced number of trials contemplating a larger number of levels per factor. The experiment is performed in live animals, so the design considers the smaller sample size while avoiding exhaustive testing realization. As a result, it can be observed the points of applicability and restrictions on the use of the above design for the slaughter of pigs and have been described what are optimal levels of the parameters studied for electrical stunning equipment that minimize losses in relation to pork quality.

The Modification of GCRMA Pre-Processing Method and Its Application

***Rohmatul Fajriyah**, Andrew P Harrison, Berthold Lausen*
Department of Mathematical Sciences, University of Essex

There are many methods of doing pre-processing the microarray data. It is ranging from the default method of affymetrix with MAS to the nonparametric method by Chen et al (2006). Despite of the drawback from each method, Irizarry et al (2004) has built the affycomp package in R to facilitate of the new method compares with the others. Irizarry et al (2004) used the Receiver Operating Curve (ROC) curve as a tool to assess which method is the best comparing with the existing one. Among the methods, the method based on

statistical model tends to be outperformed to the others. One of them is GC-Robust Multiarray Analysis (GCRMA) method. GCRMA has been used for many years in microarray data analysis to perform pre-processing. We realized that GCRMA produces the constant values more than other methods, in its final result. We have modified one step on the pre-processing algorithm of GCRMA and it successfully removed the constant values. Based on the affycomp's criteria, our modification perform slightly better than the original one. Furthermore, we used the modification method to detect the differentially expressed genes by implementing the nonparametrics method and Permax.

Multivariate Image Analysis: from Grey-Level to Hyperspectral (NIR) Images

Alberto Ferrer, *Jose Manuel Prats-Montalban*
Universidad Politécnica de Valencia

Image analysis is a wide denomination that encloses classical studies ranging from gray scale to hyperspectral images. Pioneering data treatments in image analysis were applied to simple images mainly for defect detection, segmentation and classification by computer scientists. From the late 80s, the chemometric community joined this field introducing powerful tools for image analysis, which were already in use for the study of classical spectroscopic data sets and were appropriately modified to fit the particular characteristics of image structures. These chemometric approaches adapt to images of all kinds, from the simplest to the hyperspectral images, and have provided new insights on the spatial and spectroscopic information of this kind of data sets. New fields opened by the introduction of chemometrics on image analysis are exploratory image analysis, multivariate statistical process control (monitoring), multivariate image regression or image resolution. This talks reviews the different techniques developed in image analysis and shows the evolution in the information provided by the different methodologies, which has been heavily pushed by the increasing complexity of the image measurements in the spatial and, particularly, in the spectral direction.

Control Charts Based on Quantiles

Fernanda Otilia Figueiredo
Faculdade de Economia da Universidade do Porto e CEAUL

Maria Ivette Gomes
CEAUL e DEIO, Universidade de Lisboa

Subha Chakraborti
University of Alabama

The Shewhart control charts based on summary statistics as well as the EWMA and the CUSUM charts, are usually implemented to detect changes in the mean value and/or the standard deviation of the data distribution and hardly detect changes occurred in the distributional shape. However, the possibility of having quite different distributions in terms of skewness and tail-weight, for instance, even when the mean value and the standard deviation conform to the specifications, reveals that it is important to detect possible changes in the distribution shape. We shall consider control charts based on the quantile function to detect changes in the distributional shape along the lines of the paper by Grimshaw and Alt (1997). To describe the data process we consider Fréchet and Generalized Pareto distributions, commonly used for modeling rare events in many areas of application, such as in Biostatistics, Insurance, Economics, Finance and Telecommunications, among others, and the Weibull distribution, commonly used to model asymmetric positive data in many life testing and reliability studies.

A Probabilistic Approach for the Classification of Variables

Adelaide Figueiredo
Faculdade de Economia da Universidade do Porto

Paulo Gomes
Comissão de Coordenação e Desenvolvimento Regional do Norte

We consider the multivariate data with n individuals described by p variables. In the classical approach the variables are fixed and the individuals are randomly selected from a population of individuals. We consider the dual approach where the individuals are fixed and the variables are randomly selected from a population of variables. We suppose that the variables are normalized and we associate to the sample of variables a probabilistic model. The aim of this presentation is to contribute to the problem of the dual inference in the multivariate statistical analysis and to give a contribution for the problem of selection a priori of variables. So, we refer some goodness-of-fit methods for the probabilistic model considered; we obtain homogeneous

groups of variables through the EM algorithm; we propose a "dual discriminant analysis", which enables us to affect new variables into one of the previously defined groups of variables and finally, we present an application of this new approach.

A Bayesian Approach to the Vectorization of Objects Boundaries from Digital Images

Francesco Finazzi
University of Bergamo, Department of Information Technology and Mathematical Methods

Extracting objects boundaries from digital images is a core task in many image analysis techniques for industrial applications. Although many algorithms can provide raw, low-level descriptions of the boundaries, little has been done in order to extract concise geometrical descriptions of the objects shapes. This kind of geometrical description is useful when the object shape is simple, namely it can be described by enumeration of a limited number of geometrical entities. This work presents a new Bayesian approach for extracting 2D objects boundaries as minimal closed sequences of segments and arcs, called mixed polygons. The sequence is minimal in the sense that it is able to describe all the geometrical properties of the shape without being redundant. Based on geometrical measures evaluated on the object shape, a prior distribution is introduced in order to favour mixed polygon with good geometrical properties, avoiding short sides, collinearity between segments and so on. On the other hand, the prior distribution is not too restrictive since it allows for complex mixed polygons with any number of sides. The posterior distribution is obtained by combining the prior distribution with a likelihood function based on a simple image model. Object and background are assumed to be uniform in their colour and to be observed with random noise. The estimation process is based on a two stage procedure combining reversible jump and classic MCMC methods. The reversible jump MCMC (RJMCMC) method is viewed as a model selection technique and it is used here to estimate the correct number of sides of the mixed polygon. A classic MCMC algorithm is then implemented in order to obtain spatial probability distributions on the mixed polygon parameters, namely vertices position and arcs radius. The spatial distributions are used to provide spatial accuracy information on the extracted geometrical description. A convergence criterion for the RJMCMC method is provided and it is used to switch from the RJMCMC to the MCMC stage. This approach is applied to both synthetic and real images in order to evaluate the correctness of the estimates, the robustness to the image noise and the computational burden. The approach is attractive since it does not rely on the output of low-level image processing methods and it allows for incorporating useful geometrical a priori information of the object shape. It can find application in reverse engineering, object recognition, quality control and remote measurement.

Applying the Dynamic Coregionalization Model to Particulate Matters Mapping Using Satellite Data

Alessandro Fassò, ***Francesco Finazzi***
University of Bergamo, Department of Information Technology and Mathematical Methods

The multivariate dynamic coregionalization model has been recently introduced in environmental spatio-temporal statistics because of its flexibility and easy estimate. In this talk we consider the above model for mapping airborne particulate matters in the "Padano-Veneto" region, North Italy, including the Alps. This is done by using airborne particulate matters measured by an irregularly spaced ground level network and regularly spaced satellite measurements of aerosol optical thickness (AOT). A first point favoring the multivariate dynamic coregionalization model is maximum likelihood estimation. This is done by a quasi-closed form EM algorithm which is preferred to Newton-Raphson type algorithms because of its stability for the model considered, even for fairly large parameter sets. Stability is especially important in spatio-temporal applications because of the computational complexity of the above optimization. In our application, for a single day we have data from 107 PM10 stations and 54×21 AOT measurements giving a variance-covariance matrix which is 1241×1241 in size. Parameter uncertainty is covered by bootstrap, for assessing asymptotic normality, and by standard deviations, which are based on the marginal likelihood Hessian, for confidence intervals. Moreover, thanks to the same marginal likelihood computations, likelihood ratio tests can be used at the identification and model selection stages, which are finalized by BIC and crossvalidation. The latter job is computer intensive, as the leave-one-out crossvalidation entails 107 replications of above EM estimation. As well known, AOT measurements are not available under cloudy conditions so that we have to manage a large amount of missing data. This is a natural task thanks to the state space representation of the model and to the above mentioned EM algorithm. In order to check the sensitivity to missingness, an extensive simulation campaign is performed with missing rate ranging from 0% to 90%, showing the reliability of the method for the case under study. Eventually, mapping is suitably performed thanks to dynamic Kriging formulas which are obtained by the plug-in approach and by integrate maps over time. This improves the memoryless approach of repeating the traditional Kriging every day.

The Power of Simplicity

Nicholas Fisher
University of Sydney

One of the delights of statistical research is that the solution to a problem can often be found by identifying a simple, key idea that unlocks the door to

understanding and solving the problem. This talk illustrates two simple devices for this purpose that have wide applicability, using a number of consulting applications that range from the micro to the macro, and from Lesser Statistics to Greater Statistics (in the language of John Chambers).

Comparison of Classical and Bayesian Approaches for Intervention Analysis

Glaura Franco
UFMG
(Joint work with Dani Gamerman and Thiago R. Santos)

Intervention analysis has been recently the subject of several studies, mainly because real time series present a wide variety of phenomena that are caused by external and/or unexpected events. In this work, transfer functions are used to model different forms of intervention to the mean level of a time series. This is performed in the framework of state-space models. Two canonical forms of intervention are considered: pulse and step functions. Static and dynamic explanation of the intervention effects, normal and non-normal time series, detection of intervention and study of the effect of outliers are also discussed. The performance of the two approaches is compared in terms of point and interval estimation through Monte Carlo simulation. The methodology was applied to financial time series and showed satisfactory results for the intervention models used.

On the Combined Estimating Functions Method with Applications in Finance

Melody Ghahramani
University of Winnipeg

Aerambamoorthy Thavaneswaran
University of Manitoba

Volatility plays an important role in financial forecasting and in option pricing. Estimating function theory is well suited to financial data. Various volatility models such as GARCH and Stochastic Volatility models are such that the first two conditional moments of the observed process depend on the parameter of interest. When there are two estimating functions for the same parameter, a more informative estimating function may be obtained by combining them as in Ghahramani and Thavaneswaran (2009). The combined estimating functions method had been studied by Naik-Nimbalkar and Rajarshi (1995) and also by Thompson and Thavaneswaran (1999) in the Bayesian context (See also McLeish and Small (1988) and Heyde (1997)). In this talk, I will discuss some

recent applications of the combined estimating functions method for some financial time series. Applications include hypothesis tests for ARMA models with GARCH errors, GARCH model identification and, nonlinear recursive forecasting.

yNon-Parametric and Structured Graduation of Mortality Rates

Víctor M. Guerrero
Instituto Nacional de Estadística y Geografía (INEGI)

Eliud Silva
Universidad Carlos III de Madrid

In this work, we present a non-parametric method to estimate trends in mortality rates. We first realize that the wrong recording of deaths misrepresent the phenomenon under study leading to an increase (or decrease) of its intensity and timing at a certain age, in detriment of another. This situation can affect timely decision making and policy creation, both in the public and the private sectors. Therefore, graduating data come up as an alternative to solve this problem. Our graduation method combines goodness of fit and smoothness of the non-parametric approach with information from a given structured mortality rate. So, the user is able to control both smoothness and structure in the resulting estimated mortality. One main goal of our proposal is to allow analysts to compare mortality trends, because we can fix at the outset equal percentages of smoothness and structure for different datasets. This is important because we emphasize that trends and graduated data are comparable only when they have the same amounts of smoothness. We apply the Hodrick-Prescott Multi-Variate filter, usually employed in Economics, as a tool to estimate unobserved variables, including relevant information of the phenomenon under study, as well as smoothness. Thus, we take into account the random errors resulting from a demographic relation that involves the unobserved variable. The most important contribution of our work is the definition of two indexes, one of smoothness and the other of structure, which are employed to select the smoothing constants required by the filter. Two perspectives of the methodology are emphasized. On the one hand, the proper fit and smoothness and, on the other, the combination of two sources of information, thus giving the analyst the possibility of choosing which one offers the greatest credibility. The calculations are supported by Kalman filtering, so that they are relatively simple and the usefulness of our approach is shown via empirical examples that use different mortality indicators. In the first example, we propose a 2010 goal for the year 2000 male population in the United Kingdom, so that it has the same mortality experience as Japan in 2006. The second example proposes a goal for Chile's female population in 2010, so that its annual mortality indicator is the same as the Japanese women experience in 2006. The third example makes use of United States mortality for the male population, as seen from a longitudinal (by cohort)

approach and by period. The final example shows how this methodology can be used by different specialists, such as anthropologists, demographers or statisticians, since it combines paleodemographic data of the XIX century for Mexico coming from a cemetery and from a parish.

Divide and Recombine for the Analysis of Very Large Datatsets

Saptarshi Guha**, Jin Xia, Bowei Xi, **William Cleveland
Purdue University

Divide and recombine (D&R) is a framework for the analysis of very large datasets, ubiquitous today in science, engineering, business, and government. The data are divided into subsets, an analysis method is applied to each subset or to each subset in a sample, and the subset outputs of the method are recombined. The goal of data analysis, whether the dataset is very large or very small, should be comprehensive analysis that does not miss important information in the data. The 1000s of analysis methods of statistics and machine learning can be divided into two groups. Mathematical methods, which result in numerical output, enable automated learning by the computer. Visualization methods, which result in visual output, enable human guidance to the process of automated learning. Both mathematical methods and visualization methods are critical to comprehensive analysis. The computing of D&R is embarrassingly parallel. Recent development of very effective distributed software environments that exploit this, have resulted in feasible computation. This provides a mechanism for comprehensive analysis of very large datasets because it enables both mathematical and visualization methods. In a D&R analysis, mathematical methods are typically applied to all subsets, and visualization methods are typically applied to a representative sample guided by variables from mathematical methods. To achieve its potential, D&R requires much further research in all areas that are involved in the analysis of data: computational environments, mathematical methods, visualization methods, and theory. The goal of the research is to discover methods of division and recombination that provide optimal results from the analysis methods, given that the data must be divided.

R and Hadoop: Analyzing Massive Data Sets using RHIPE

Saptarshi Guha
Department of Statistics, Purdue University

The statistician of today is confronted with the proliferation of massive data sets. Well known examples are the Netflix film recommendations, network security packet traces, and the social network graphs of Twitter and Facebook.

The ready availability of cheap and powerful computing resources allow us to tackle some of the computational and analytical challenges posed by the complexity and size of massive data sets. In this presention, I will discuss how one can address these problems within the R environment by making use of the RHIPE (R and Hadoop Integrated Programming Environment) package. RHIPE is a software system that integrates the R open source project for statistical computing and visualization with the Apache Hadoop Distributed File System (HDFS) and the Apache MapReduce software framework for the distributed processing of massive data sets across a cluster. Distributed programming with massive data sets is by nature complex – issues such as data storage, scheduling and fault tolerance must all be handled. RHIPE uses its tight integration with the HDFS to store data across the cluster. Similarly, it takes advantage of MapReduce to efficiently utilize all the processing cores of the cluster. Vital, but difficult to implement details, such as task scheduling, bandwidth optimization and recovery from failing computers are handled by Hadoop MapReduce. Most importantly, RHIPE hides these details from the R user, by providing an idiomatic R interface to Hadoop and HDFS cluster. The design of RHIPE strives for a balance between conceptual simplicity, ease of use and flexibility. Algorithms, designed for the MapReduce programming model, can be implemented using the R language, executed from R's REPL (read-eval-print-loop) and the results are directly returned to the user. In this talk, I will first outline the design of RHIPE along with some simple examples of MapReduce-based solutions implemented in RHIPE. I will then discuss two real-world examples of using RHIPE. In the first application, we monitor over a hundred gigabytes of internet traffic data, partition them as individual connections and use RHIPE to assist in the development of intrusion detection systems. This application will also highlight how HDFS can be used as a query-able database of millions of R objects. In the second application, we demonstrate the use of RHIPE in the storage, processing and modeling of Voice over IP data. While RHIPE presents an intuitive and powerful extentsion to the R programming environment, a number of problems remain to be addressed. I will present an overview of future improvements to the package and end with a brief discussion of some classes of problems that can and cannot be tackled within the environment described here.

Generalized Fiducial Inference for Wavelet Regression

Jan Hannig
University of North Carolina at Chapel Hill

Thomas Lee
University of California at Davis

In this article we apply Fisher's fiducial idea to conduct statistical inference for wavelet regression. We first develop a general methodology for handling model selection problems within the fiducial framework. With this new methodology we then propose fiduical based methods for performing wavelet curve

estimation, as well as constructing both pointwise and curvewise confidence intervals. It is shown that, under some mild regularity conditions, both the new fiducial based pointwise and curvewise confidence intervals have asymptotically correct coverage. Furthermore, simulation results show that these new fiducial based methods, especially for constructing pointwise confidence intervals, also possess promising empirical properties. To the best of our knowledge, this is the first time that the fiducial idea has been applied to a nonparametric estimation problem.

Random Generation for Constructing Supersaturated Designs

Hiroki Hashiguchi
Saitama University

Supersaturated designs are fractional factorial designs in which the number of columns for potential effect is greater than the number of runs. They are helpful when the number of factors to be assigned in an experiment is large, for example, when screening to find a few pivotal factors from many candidates in the primary stage of a scientific investigation and product innovation. Constructing supersaturated designs is one of the important issues as well as the data analysis using supersaturated designs. There are several theoretical and computational approaches to construct supersaturated designs, such as random balance design, application of BIBD, algorithmic approach including permutation of rows and columns and so forth. This talk concerns with a random generation method to construct mixed-level supersaturated designs. This method includes two stages mainly. The aim of the first stage is to attempt to minimize the sum of orthogonality between the two columns by uniformly random generation. The other stage is to adjust the coincidence numbers of rows so that they can close to their average. This approach is compared with the previous studies for a construction by numerical examples. In the numerical examples, the lower bound for the sum of orthogonality can be used as a benchmark of design optimality.

Projecting Future Building Water Losses From Climate Scenarios

Ola Haug
Norwegian Computing Center

The anticipation of substantial future climate change gives rise to increased focus on weather related risks from the insurance industry. The vulnerability of life as well as non-life products is affected, and knowledge of future risks is

valuable. Most apparently, premiums may be updated accordingly, but communicating dedicated loss-preventive measures to clients, building contractors and regulators could help save costs and personal inconvenience. We have established statistical claims models for the coherence between externally inflicted water damage to private buildings in Norway and selected meteorological variables. As a spin-off, spatial variations in vulnerability are recognized as well. Based on these claims models and downscaled climate model data with different climate models and CO2 emissions scenarios, projected loss levels of a future scenario period are compared to those of a control period. Inherently involved in the loss projections are uncertainties introduced from imprecise climate model scenario data, claims models misspecification, and fitting the claims models to limited amounts of data. Among these error terms, the claims models estimation uncertainty is the only component that is quantifiable. Disregarding the first two components, our analyses identify an incontestable, but moderate increase in the losses. Along the western coast of Norway, the claim estimates depend significantly on which climate model has been invoked, whereas the choice of CO2 emissions scenario turns out to be irrelevant almost everywhere.

The Influence of Industry Structure on the Presence of Small Firms: The Case of Philippine Manufacturing Industries

Lourdes Homecillo
National Statistics Office - Philippines

This study focuses mainly on the presence of small firms in the Philippine manufacturing sector and identifies the dimensions of industry structure that influence entrepreneurs to discover and exploit business opportunities in certain industries than in others.
In the context of entrepreneurial research, numerous studies have been undertaken that examined the influence of industry structure on the decision to exploit business opportunities. Since it has been established that industry-level differences exist in the exploitation of entrepreneurial opportunities, Shane (2003) identified six dimensions of industry structure that influence firm formation, as follows: industry profitability, cost of inputs, capital intensity, advertising, industry concentration, and average firm size. Empirical studies of Caves (1998), Acs and Audretsch (1989), Dean and Meyer (1998), including the of findings of Khemani and Shapiro (1987) concluded that market concentration was generally predict to exert a negative influence on firm entry and exit. The findings of their studies suggest that high concentration ratio in an industry inhibits other smaller firms to exploit business opportunities in the industry as the market is monopolized by only a few large players that have specialized production and marketing processes. Likewise, studies point out that new firm formation is also more common in industries with lower average firm (employment) size. This is because these types of firms require less capital requirements and lower cost inputs than larger firms.

Methodology and Results
The study made use of economic data from the 2005 Annual Survey and 2006 Census of Philippine Business and Industry for manufacturing that were undertaken by the Philippine NSO. The 3-digit PSIC industry group is used as the level of analysis. However, only 24 industry groups where the number of small establishments is considered significant were included. The assumption is that these industry groups are considered representative of industries in the manufacturing sector where many entrepreneurs exploit business opportunities as evident by the presence of large number of small and medium-sized establishments. Certain indicators compiled from the survey were used to represent the dimensions/factors of industry structure.

Regression analysis was used as the method of determining which among the dimensions of industry structure yielded significant influence on the exploitation of business opportunities in certain manufacturing industries. The results of the analysis reveal that only two dimensions of industry structure have significant negative influence on entrepreneurs' decision to exploit business opportunities, as follows: concentration ratio and average firm size. The results of this study are consistent with those of earlier empirical researches, some of which were cited in the literature portion.

Statistical Models on Risk Management

D. S. Hooda
Jaypee Institute of Engineering & Technology

Every investor wants to maximize his profits by selecting proper strategy for investment. There are investments like governments and bank securities, real estate, mutual funds and blue chips stocks that have low return but relatively safe because of a proven record of non-volatility in price fluctuations. On the other hand, there are investments which being high return, but may be prone to great deal of risk and the investor makes loss in case the investment goes sour. It is very interesting to some statistical models to invest his funds for maximum return and at the same time his risk of losing his capital is minimized. In the present communication we discuss a deterministic model due to Markowitz who gave the concept of mean variance efficient frontier to find all efficient portfolios that maximize the expected returns and minimize the variance. Risk aversion index and Pareto-optimal sharing of risk sharing are explained. Some measures of portfolio analysis based on entropy mean-variance frontier and maximum entropy model in risk sharing are proposed and studied.

Testing of Stochastic Dependence of Attributes in Contingency Table

D.S. Hooda
Jaypee Institute of Engineering &Technology

In the present paper we derive a new information theoretic model for testing and measurement of stochastic 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.

An Integrated Probability-Based Approach for Multiple Response Surface Optimization

Okay Isik, *Resit Unal, Ghaith Rabadi*
Old Dominion University

The purpose of this paper is to suggest strategies in order to improve the quality of processes and products with multiple quality characteristics. An integrated probability based approach will be followed in the modeling and optimization of the problem, which will utilize strengths of probability-based and desirability approaches. Conformance probability metric is the commonly used optimization criterion for probability-based approaches and it will be shown that particularly when conformance probability is high and one-sided response types are studied; due to the bathtub shape of the solution surface it can prematurely stop the search process and can give biased solutions in mean response values. Another concern is when the number of responses grows high; a feasible solution set may not exist due to the response bounds. Therefore, penalization of infeasible solutions can help to identify near feasible solutions, and also help decision makers articulate their preference information efficiently in order to find compromising solutions. The utility of the proposed method will be illustrated with two examples.

Three-Dimensional Quantile Plots and Animated Quantile Plots of the Prediction Variance for Response Surface Designs

Dae-Heung Jang
Pukyong National University

Giovannitti-Jensen and Myers (1989) proposed variance dispersion graphs to evaluate the overall prediction capacity of response surface designs. This

variance dispersion graph uses the maximum, the average, and the minimum prediction variances on concentric spheres inside region of the interest. Khuri et al. (1996) suggested quantile plots of the prediction variance for response surface designs. This quantile plot uses the quantiles instead of the maximum, the average, and the minimum prediction variances on concentric spheres inside region of the interest. We must select several values of radii inside region of the interest and draw several corresponding quantile plots for evaluating the overall prediction capacity of response surface designs. Simply we can the same task using three-dimensional quantile plot as a plot. Robinson and Khuri (2003) proposed the quantile dispersion graphs. We can extend the idea of the three-dimensional quantile plots to the quantile dispersion graphs. As an another extension of quantile plots, we can suggest animated quantile plots. Through animated quantile plots with automatic sequantial change of radii inside region of the interest, we can evaluate and compare the overall prediction capacity of response surface designs.

Monitoring Structural Change in Time Series Models

Mardi Jankowitz
University of South Africa

Structural stability is very important in time series since future estimation is based on it. If unstable relationships are used for estimation, the forecasts can be biased, inaccurate, and not meaningful at all. This applies especially in econometrics. Structural changes can occur as outliers (spikes), a single shift or fluctuations. In statistical literature it is identified as change points, break points, step changes, or jumps. Most research has been done on statistical quality control. In the field of engineering it is known as edge detection and edge preservation, and applied in the areas of signal processing, image processing, computer graphics, pattern recognition, geology, etc. The literature shows research where linear and nonlinear smoothers were compared to monitor changes. Algorithms using local linear smoothing and adaptive ridging, and some based on kernel smoothing were developed. For nonlinear smoothers, the running median removes outliers and preserves shifts, but it was found that the repeated median removes outliers from signal with trend. Lower-upper-lower-upper (LULU) smoother is a class of nonlinear smoothers introduced by Rohwer in 1989. These smoothers are compositions of the extreme selectors, the minima and the maxima. They also have very attractive mathematical properties in particular their way of dealing with impulsive noise in the form of block pulses. In this study LULU smoothers were applied to smooth time series with trend. Their performance is compared to thos of other smoothers in detecting structural change. The results will be discussed.

Selection of Variables That Are Relevant to Multivariate Process Monitoring Goals

Luan Jaupi
CNAM

Multivariate process control charts have been increasingly popular to monitor many different industrial processes. There are many reasons for this, but the main one is the recent advances that have occurred in multivariate quality control methods. For processes where huge amounts of multidimensional data are available, multivariate projection methods, such as principal component analysis (PCA) and partial least squares (PLS), have received much attention by a variety of industries. This paper introduces a new technique that can be used in the early process monitoring design stage to select from among the set of quality characteristics or process parameters a smaller set that is adequate to process control and ensure a product satisfies yield specifications. The critical subset of variables needed to ensure suitable process monitoring are identified by using Rao (1964)'s principal component analysis of instrumental variables. The effectiveness of the proposed procedure is demonstrated through a real example.

Exploration of the Recommended Model of Decision Tree to Predict a Hard-to Measure Measurement in Anthropometric Invited

JongHoo Choi, ***Jae-Poong Jeong***
Korea University

Anthropometric survey is important as a basis for human engineering fields. According to our experiences, there are difficulties in obtaining the measurements of some body parts because respondents are reluctant to expose. In order to overcome these difficulties, we propose a method for estimating such hard-to-measure measurements by using easy-to-measure measurements those are closely related to them. This study aims to explore a recommended model of decision tree to predict a hard-to-measure measurement in anthropometric survey. We carry out an experiment on cross validation study to obtain a recommended model of decision tree. We use three split rules of decision tree, those ard CHAID, Exhaustive CHAID, and CART. CART result is the best one in real world data. The method we propose will be illustrated with real data from the 1992 Korea national anthropometric survey.

Sequential Fixed Width Confidence Intervals for the Offset between two Network Clocks

Jun Li, ***Daniel Jeske***
University of California, Riverside

Estimation of the offset between two network clocks has received a lot of attention in the literature with the motivating force being data networking applications that require synchronous communication protocols. Statistical modeling techniques have been used to develop improved estimation algorithms, with a recent development being the construction of a confidence interval based on a fixed sample size. Lacking in the fixed sample size confidence interval procedures is a useable relationship between sample size and the width of the resulting confidence interval. Were that available, an optimum sample size could be determined to achieve a specified level of precision in the estimator and thereby improve the efficiency of the estimation procedure by reducing unnecessary overhead in the network that is associated with collecting the data used by the estimation schemes. A fixed sample size confidence interval that has a prescribed width is not available for this problem. However, in this paper we develop and compare alternative sequential intervals with fixed width and demonstrate that an effective solution is available.

Data Mining Application for Estimation of Customer's Wallet Size

Seohoon Jin, Seok-Won Oh
Korea University

Differentiated communication strategy for customer is needed for effective CRM. Gathering and analyzing various information of customers are very important to understand and differentiate customers according to their value and status. Especially, in credit card market of Korea with its cut-throat competition, it is more critical to analyze customer information and apply it to CRM. Generally a customer has several credit cards of different companies. To find share of wallet (SOW) of each customer is very useful to build customer communication strategy for credit card company. High SOW customer should be retained and low SOW customer can be a market to attack. In this study, we built the estimation model for each customer's wallet size(WS). Information of customers who have more than 4 companies' credit card is shared to credit card companies. This information was used for building up estimation model for WS of customers who have less than 3 companies' credit card. Several data mining techniques were applied and resulting models were verified with validation data set.

A Study on the Datamining Score Model Management Process

Jiyong Jung, ***Seohoon Jin***
Korea University

Today, most of leading companies are introducing CRM (customer relationship management) for their business. Since customer acquisition brings more cost and effort than customer retention, management of relationship with customers is very important. In order to improve their CRM capability companies use some sophisticated tools. Data mining is one of effective tools for CRM. This study is about data mining application for effective CRM. We focus on the development of propensity model for Cross-Sell and it's management. This study deals with a credit card company case. Credit card company has a lot of customer information including customer contacts and credit card usage histories. These information can make companies to understand customers. When companies sell a product to customers, the response probability is different for each customer because customers have different needs along with their status. Therefore it is necessary to find right customer who will buy a product. Supervised learning technique of data mining can be used for getting response score of each customer. Customers and market change as time goes by, so score models cannot maintain the performance. Therefore we need to improve the performance by adjusting the parameters for scoring or rebuilding the score model. In this study, we propose the management process of score model. By keeping up with the process, companies can maintain the performance of the score model.

Probabilistic Structural Equation with Bayesian Networks

Lione Jouffe
Bayesia

Bayesian networks, also called Bayesian Belief Networks (BBN), are probabilistic graphical models that allow representing non-deterministic knowledge. Qualitatively, each random variable has a corresponding node in the graph, and arcs link the variables that are directly dependent. Quantitatively, each node has an associated probability distribution that is usually represented with tables. Those distributions are marginal distributions for root nodes, and conditional distributions for nodes with parents to quantify the probabilistic relationships between the linked variables. BBN then allow compactly encoding the Joint Probability Distribution. Efficient inference algorithms can be used to rigorously take into account any piece of evidence (hard or soft) on an arbitrary subset of variables for updating the probability distributions of all the unobserved variables. As all the variables can be input or output variables, inference can be used for simulation and/or diagnosis. Originally exclusively built by hand to model expert knowledge, it is now

possible to automatically learn BBN through data analysis. BBN then offer a single and unique theoretical framework to carry out a large set of data mining and data analysis tasks: unsupervised learning for discovering all the direct relationships that hold between the variables, supervised learning for finding the probabilistic profile of a target variable, data segmentation for clustering individuals, and variable segmentation for inducing hidden concepts. Furthermore, thanks to their probabilistic nature, BBN naturally handle missing values, dynamically during learning. And last but not least, from a communication point of view, their graphical representation makes them highly readable and understandable, which eases the knowledge discovery and dissemination process. By combining unsupervised learning, variable segmentation and data clustering, we use a real example to describe how to exploit BBN for building Probabilistic Structural Equations (PSE) as an efficient alternative to traditional Structural Equation Modeling and PLS approaches. The obtained PSE graphically represents the relations between the key variables (the targets), the identified concepts (factors), and the manifest variables. Bayesian inference is then used for what-if scenario generation, sensibility analysis, and realistic driver analyses that take into account costs and difficulty to apply the induced action policies.

Statistical Issues in the Comparison of Multi-dimensional Profiles

Karen Kafadar
Indiana University

Several common problems arise in a collection or large database of profiles: (1) Determine the number of features are needed to characterize a multi-dimensional profile; (2) Estimate the "false match" probability (and its uncertainty) without resorting to a comparison of all pairs of profiles; (3) Design a sequential sample to achieve (1) and (2) that accommodates an increasing database of profiles. These issues arise in fraud detection (identify behavior that differs substantially from a customer's typical profile), pattern comparisons (e.g., fingerprints, bite marks) and genetic studies (e.g., microarray experiments, spectra from proteomics experiments to identify proteins, DNA profiles). In this talk, I will describe scenarios in which these issues arise, and propose some possibilities for addressing them.

Can one Extract Causal Information from High-Dimensional Observational Data?

Markus Kalisch, Peter Bühlmann, Marloes Maathuis
ETH Zürich

Understanding cause-effect relationships between variables is of interest in many fields of science. It is a well-established scientific principle to determine the total causal effect of one variable on another via randomized controlled intervention experiments. Sometimes, however, experiments are too time consuming, expensive or unethical. We discuss an approach that aims at extracting bounds on causal effects by using observational data only. We outline the underlying theory and discuss strengths and limitations of the approach. Furthermore, we present the R-package "pcalg" for using the discussed method.

Dynamics of Abstinence and Condom Use Among Unmarried Youth Aged 15–24 in Uganda

Brian Kanzira
Makerere University, Uganda

Worldwide, young people's sexual health is a major concern especially in the context of HIV/AIDS, Sexually Transmitted Diseases/Infections (STD/Is) and unplanned pregnancies. Two-thirds of all people infected with HIV live in sub-Saharan Africa, although this region contains little more than 10% of the world's population. AIDS has caused immense human suffering in the continent. The most obvious effect of this crisis has been illness and death, but the impact of the epidemic has certainly not been confined to the health sector; households, schools, workplaces and economies have also been badly affected. It is estimated over 1.4 million adults and children die as a result of AIDS in sub-Saharan Africa. Methods have been tried in Africa to try and reduce on this epidemic of which Abstinence and Condom use have proven to tackle the prevalence rate. This though has come with some dynamics. This thesis establishes the current levels of, and changes in, sexual abstinence and condom use among young unmarried people and determines factors associated with them. Abstinence is considered in two ways: primary abstinence refers to delayed onset of sexual intercourse, whilst secondary abstinence refers to periods during which sexual intercourse did not take place amongst those who had already been sexually active. The study analyses quantitative data from 186 respondents in a stratified cluster survey and qualitative data from focus group discussions and key informants. The survey uses both a questionnaire and an Event History Calendar (EHC). Logistic regression is used to analyze the probabilities of abstaining and condom use while Cox's proportional hazards and piecewise constant hazards models are used to analyze age at first sex and duration of secondary abstinence. Consistency of condom use at both first

and latest events is analyzed using Multinomial Logistic Regression. Multilevel modeling is used to explore cluster level variation. The results show that young people are more likely to delay first sex if they reside in Kabale, are aged 15–16, avoid parties/clubs, avoid alcohol consumption, are not indecently assaulted and have a positive attitude towards abstinence. Being in age group 15–16, not taking alcohol and avoiding parties/clubs are associated with secondary abstinence. Residence in Mukono, being female, being in school, attainment of secondary education, listening to radio, positive attitude to condom use, higher age at first sex and having had two or more relationships are associated with condom use. Condom use level varies by sexual event and relationship while the trend of hazard of initiating sex varies by age cohort. Consistency of condom use is associated with residence in Mukono, secondary education and higher age at first sex. There is a significant random variation in both primary and secondary abstinence at village level. In conclusion, there appear to be dynamics of condom use and abstinence among young people in Uganda. An EHC can be reliably used to collect data on sexual abstinence episodes. Some patterns of condom use and sexual abstinence are different from results reported in other studies.

New Approach to the Analysis of Signal-to-ratio-Noise Ratio in Robust Parameter Design

Toshihiko Kawamura
Institute of Statistical Mathematics

Kazuo Tatebayashi
Fuji Xerox Co., ltd

Hiroe Tsubaki
Institute of Statistical Mathematics

Taguchi's dynamic robust parameter design is an engineering methodology for improving the quality of products or processes in the automotive industry. However, this important tool in statistical quality control lacks a theory of robust parameter design. Our purpose is to provide such a basis as approach to the analysis of signal-to-noise (SN) ratio. Taguchi proposed various performance measures for evaluating the performance of signal-response systems. We do not use Taguchi's SN ratios for two reasons. First, Taguchi's approach uses regression analysis (least squares) to minimize the variation in data, regardless of how the variation of functions is evaluated.Theis important the basic concept of robust parameter design is to evaluate not measurement errors but fluctuations in functionality. We directly evaluated the input/output ratio or the difference between the logarithmic input and output relative to an ideal or target signal-response relationship. This approach is close to the original idea of the Taguchi methods, which is to directly evaluate variations in functionality or sensitivity. Second, our approach makes it possible to evaluate the robustness of data with collapsed orthogonality where it is not necessarily

possible to design the experiments strictly. The proposed methodology is illustrated with real data on an automotive suspension system.

EEG Classification Models in Driver's Microsleep Prevention

Jan Klaschka
Institute of Computer Science, Academy of Sciences, Prague, Czech Republic

Present work is a part of a multicentric project aimed at prevention of traffic accidents caused by driver's microsleeps. The data analytical part of the project is focused at development of EEG classifiers capable of distinguishing somnolence (sleepiness) from other brain states. In the target application, some sorts of drivers (e. g. truck drivers on long distances) should be, while driving, EEG-monitored. The EEG signals, registered with electrodes mounted in a special cap, headband or glasses, should be subject to real-time analysis and an alarm should be activated when somnolence is detected. It would be perfect to have a universal classifier applicable to any driver without laboratory testing of his/her EEG. However, due to the well known fact that the EEG patterns typical for different brain states are highly individual, such a universal classifier would inevitably be very inaccurate. Thus, good classifiers have to be tailored for specific subjects and trained using EEG data from laboratory testing of the prospective users. This does not, nevertheless, mean that a good classifier for a person must be "purely individual", i. e. derived exclusively from the person's own data. We can consider "mixed" classifiers, too: A mixed classifier for a person results from combining a "purely individual" classification model with a model trained on the data of other well selected subjects. Mixed models might not only increase the classification accuracy, but also allow for less extensive testing of the drivers. This work adresses the following questions: (1) May the mixed classifiers outperform the "purely individual" ones? (2) How should the mixed models be constructed? (How to select the data contributing to the model? How to combine the models?) The presentation will concentrate on the results of extensive computational experiments with EEG data from the Joint Laboratory of System Reliability, Faculty of Transportation Sciences, Czech Technical University in Prague. Various strategies of mixed model construction will be compared, and superiority of some of them over the "purely individual" models will be demonstrated.

Model of Optimization of Trade on Internal and a Foreign Market

Yana Kulishova
Donetck National University

The income of enterprises is multiplied on 0.4% due to multiplying realization at the internal market, it is related to the small volume of internal market, unattractiveness for enterprises. An index at the parameter of time (t) testifies to the decline in a dynamics arrived at the observance of the folded proportions of realization to the internal and oversea market. If to take into account correlation of internal market in a model, export and income, got metallurgical enterprises in 2009year, crises tendencies began in which to show up, the values of coefficients of elasticity will assume a completely another.

Ukraine belongs to number of developing countries. Foreign trade activities of Ukraine are realised by export of the raw goods, import of the equipment and the technologies, insufficient volume of attraction of foreign investments. In the modern terms of globalization of the world system, subsequent liberalization of economy of Ukraine and its entry in WOT requirements rise to the management in relation to providing of competitiveness of enterprises. The modern model of development of foreign economic activity of enterprises is characterized the considerable degree of openness. In the total in 2008 a the foreign trade turnover turn was exceeded by GDP in 1,017 times, export relatively to GDP was 46,8%, import – 54,9%. In the structure of export in 2008 41,2% was on ungracious metals, including on the metal products – only 15,4%. the import metals occupies about 7%. A low competitiveness of metallurgical enterprises, low level of the labour and payment of workers productivity, wearing out of capital assets, low innovative activity and technological lag, the developed countries is the retentive factors of development and increase of efficiency of their activity at the oversea markets, that conditioned by absence of effective mechanism of management development of enterprises with external economic potential. Therefore raw industries should be focused on requirements of home market. The author the model is offered describes dependence between the income of enterprises income and activity at the oversea market and realization on national. It is offered to the metallurgical enterprises to extend an internal market in realization of products. It is impossible to promote an income due to realization of products on an export and certainly necessity of expansion realization to the internal market, that it is expected after the model of dynamic sedate dependence income and export activity and realization at the internal market. Model parameters values of coefficients of elasticity. The income of metallurgical enterprises is multiplied on 1,73% at multiplying an export on 1%. Products have the most paying concerns to the oversea market.

Control Charts for Incident Rates in the Construction Industry

Alan Lee
University of Auckland

Nicholas Fisher
University of Sydney

Ross Sparks
CSIRO Mathematics, Informatics and Statistics

Monitoring incident rates in industrial enterprises is important part of an overall strategy to achieve greater safety in the workplace. In particular, it is important to detect rapidly any upward change in incident rates, so that remedial action can be taken. Control charting techniques have been in use for many years to detect changes in incident rates in industrial and more recently in medicine and public health, for example to detect changes in adverse event rates, hospital admissions and disease incidence, but do not seem to have been used in incident monitoring. In this paper we describe some control charting techniques that can be used to give timely warning of changes in incident rates. Unlike industrial applications, a key metric in comparing different charting procedures is the steady-state average run length. We discuss methods for computing this metric in the case where the exposure to risk varies from observation to observation and apply these to the design of a plan in the construction industry.

Auc Based Variable Selection for Data Mining Score Model

Won Ho Lee, ***Gyeong-Min Lee***
Korea University

AUC based variable selection for data mining score model All financial transactions seek profit but avoid the risk. One of the risks that we want to sidestep is bankruptcy. In order to predict bankruptcy it should be utilized variety of variables such as company's debts, BIS ratio, management policy, market conditions. In this paper, among the many variables that influence, how to select necessary variables to the bankruptcy prediction model was applied. Typically, for bankruptcy predictions, we had to decide the model, and then to select significance variables with proper variable selection method. If we considered all possible combinations of variables, we can find the best model. However, the model can contain unnecessary variables. Moreover, unnecessary many independent variables can make multicollinearity problems. Therefore, it is important to effectively choose the significant independent variables. This study inquires into appropriate variable selection based on AUC. In this paper we utilize the logistic regression model that is usually used for classifying target variable. A variable selection method select with variables are based on a change rate of AUC.

Modified Shewhart Control Charts for Conditionally Heteroskedastic Models

Esmeralda Gonçalves, Nazaré Mendes-Lopes
University of Coimbra

Joana Leite
Polytechnic Institute of Coimbra

In recent years the applicability of control schemes, like the Shewhart chart, has been extended from independent processes to time series, namely, with the introduction of modified charts which incorporate the time series structure into its design. Control charts allow monitoring whether an observed process diverts from a supposed target process, by issuing out-of-control alerts. So, when presented with more than one control chart to monitor a time series, raises the question of evaluating the best design to detect a deviation from target as soon as possible. The average run length (ARL) is widely used as a performance measure for control charts and, when dealing with time series, is defined as the average number of instants that must go by before one indicates an out-of-control condition. Severin and Schmid (1999) introduced the first control charts for conditionally heteroskedastic processes, specifically for GARCH processes, and established theoretical bounds for the in-control ARL of modified Shewhart charts for pure ARCH processes. This lower bound was further developed by Pawlak and Schmid (2001). The relevance of these schemes in finance is well documented in Severin and Schmid (1999), who illustrated the implementation of a control chart to the daily returns of a stock, and in Schipper and Schmid (2001), who showed how a trading system for a financial series can be construed using information given by a control chart. In this study, we revisit the work developed for the ARCH model, suggest a new lower bound for the ARL and show how it improves the bound derived from Pawlak and Schmid (2001). For TARCH processes (Rabemananjara and Zakoian, 1993), we define the modified Shewhart chart and, taking into account the results presented in Gonçalves and Mendes-Lopes (2007), determine bounds for the in-control ARL of modified Shewhart charts. In addition, we present a simulation study to assess the quality of the bounds calculated for the ARL by comparing them with the estimated ARL. This simulation study also allows making some considerations about the behaviour of the ARL, which, in turn, provides some hints for future work.

Optimal Discrete Choice Experiments Design Under Model Uncertainty

William Li
University of Minnesota

Ke Wang
Fudan University

In applications of discrete choice experiments (DCE's), researchers may not be certain about the specific model forms for describing the possible interactions among attributes involved in the DCE's. In this case, it is important to construct the DCE optimally at the design stage so that the resulting design is robust against misspecification of the underlying model. We propose a general framework for constructing model robust DCE's based on a design criterion called Bayesian Information Capacity. The performance of the proposed designs is assessed with respect to relative efficiency and minimal level overlap properties. We also demonstrate several applications in the quality area.

Heavy Metal Migration During Electroremediation of Fly Ash From Different Wastes – Modelling

Ana T. Lima
Utrecht University

Paulo C. Rodrigues*, João T. Mexia*
New University of Lisbon

Fly ash is an airborne material which is considered hazardous waste due to its enrichment on heavy metals. Depending on the waste from which they are originated, fly ash may be further valorised, e.g. as soil amendment or concrete and ceramics adjuvant, or landfilled, when defined as hazardous material. In any case, heavy metal content has to be decreased either for fly ash valorisation or for complying with landfill criteria. The electrodialytic (EDR) process is a remediation technique based on the principle of electrokinetics and dialysis, having the aim to remove heavy metals from contaminated solid media. EDR was here applied to fly ashes from the combustion of straw (ST), from the incineration of municipal solid waste (DK and PT) and from the co-combustion of wood (CW). A statistical study, using F tests, Bonferroni multiple comparison method and a categorical regression, was carried out to determine which variables ("Ash type", "Duration", "Initial pH", "Final pH", "Acidification" and "Dissolution") were the most significant for EDR efficiency. After establishing these, the selected variables were then used to characterize some kinetic parameters, from metals migration during EDR, using a biregressional design. Cd, Cr, Cu, Ca and Zn migration velocity and acceleration to the electrodes (anode and cathode) were then considered. Cd and Cu migration to the cathode were found to be significantly influenced by "Ash type", "Duration", "Final pH" and "Dissolution".

DD-Classifiers: New Nonparametric Classification Procedures

Juan Cuesta-Albertos
University of Cantabria, Spain

Jun Li
University of California, Riverside

Regina Liu
Rutgers University

Most existing classification algorithms assume either certain parametric distributions for the data or certain forms of separating surfaces. Either assumption can greatly limit the potential applicability of the algorithm. We introduce a new nonparametric classification algorithm using the so-called DD-plot. This algorithm, DD-classifier, is completely nonparametric, requiring no prior knowledge of the underlying distributions or of the form of the separating surface. Thus it can be applied to a wide range of classification problems. The DD-classifier can be easily implemented and its classification outcome can be clearly visualized on a two-dimensional plot regardless of the dimension of the data. Furthermore, it can also be robust against outliers or contamination. The asymptotic properties of the proposed classifier and its misclassification rate are discussed. The DD-classifier is shown to be asymptotically equivalent to the Bayes rule under suitable conditions. The performance of the classifier is also examined using simulated and real data sets. Overall, the DD-classifier performs well across a broad range of settings, and compares favorably with most existing nonparametric classifiers.

Equivalent Poisson and iid Sampling Models: A Bayesian View

Albert Lo
Hong Kong University of Science and Technology

Conditional on the total counts, the sites of a non-homogeneous Poisson process are independent and identically distributed random variables from a distribution that is the normalized intensity function. This suggests that the estimation of the intensity measure of a non-homogeneous Poisson process and the estimation of a distribution from an iid sample are similar problems. The Bayesian approach to the inference problem requires the additional manipulation of Gamma typed process prior distributions. It is shown that the result that Dirichlet processes form a conjugate family of priors for iid sampling from an unknown distribution is equivalent to the result that Gamma processes form a conjugate family of priors when sampling a path of a non-homogeneous Poisson process. In addition, it is also shown that these conjugate prior properties are also equivalent to an absolute continuity property of Gamma processes, as well as to the latter's Palm distribution theory.

Selecting Designs for a Computer Experiments

Jason Loeppky
UBC Okanagan

William Welch
UBC

Computer experiments are now widely used in many areas of science to model the output of a physical process. Due to the computational complexity of the code it is often necessary to build a fast emulator that can be used for exploration of the underlying response surface. For example the approximations are often used for optimization of the code output, understanding the sensitivities of the inputs to the code output or gaining a better understanding of the code to improve the overall fit of the response surface. The literature surrounding the approximation of a computer code is rich with examples of using orthogonal arrays or space filling designs to approximate the code output. In this talk we will investigate when standard orthogonal arrays are useful, reliable and convenient in the context of modelling a computer experiments.

A Class Comparison Method With Filtering-Enhanced Variable Selection For High-Dimensional Data Sets

Lara Lusa
Department of Medical Informatics, University of Ljubljana, Slovenia

Edward Korn
National Cancer Institute (NIH)

Lisa McShane
National Cancer Institute (NIH)

High-throughput molecular analysis technologies can produce thousands of measurements for each of the assayed samples. A common scientific question is to identify the variables whose distributions differ between some pre-specified classes (i.e. are differentially expressed). The statistical cost of examining thousands of variables is related to the risk of identifying many variables that truly are not differentially expressed, and many different multiple testing strategies have been used for the analysis of high-dimensional data sets to control the number of these false positives.
An approach that is often used in practice to reduce the multiple comparisons problem is to lessen the number of comparisons being performed by filtering out variables that are considered non-informative 'before' the analysis. However, deciding which and how many variables should be filtered out can be highly arbitrary, and different filtering strategies can result in different variables being identified as differentially expressed. We propose the filtering-enhanced variable selection (FEVS) method, a new multiple testing strategy

for identifying differentially expressed variables. This method identifies differentially expressed variables by combining the results obtained using a variety of filtering methods, instead of using a pre-specified filtering method or trying to identify an optimal filtering of the variables prior to class comparison analysis. FEVS can be extended and used to combine different analyses performed on the same data set, for example combining the results obtained using different normalization methods, or using different methods for variable identification. We proved that the FEVS method probabilistically controls the number of false discoveries.
In this talk we will present the FEVS method and we will show with a set of simulations and with an example from the literature that FEVS can be useful for gaining sensitivity for the detection of truly differentially expressed variables. We will also present the FEVS R package that we developed.

A Control Chart Based on Sample Means and Sample Ranges for Monitoring Bivariate Processes

***Marcela Machado**, Antonio Costa*
UNESP

Control charts are often used to observe whether a process is in control or not. When there is only one quality characteristic Shewhart control charts are usually applied to detect process shifts. However, there are many situations in which it is necessary to control two or more related quality characteristics simultaneously. The traditional tool used for monitoring the mean vector of multivariate processes is the T2 chart; however, the T2 statistic is not easy to calculate, once the user is not familiar with the computation of vectors. Moreover, recent studies have shown that if compared with the use of simultaneous Xbar charts, the T2 chart is not always faster in signaling process disturbances. Similarly to the mean vector, the covariance matrix is also subject to changes. The generalized variance |S| chart is the usual tool for monitoring the covariance matrix of multivariate processes. The |S| statistic has the same drawback of the T2 statistic, that is, it is difficult to calculate, once the user is not familiar with the computation of matrixes too. Recent studies have proposed simpler statistics for monitoring the covariance matrix, for example, the use of sample ranges or sample variances, which are more familiar for the users. In general, these charts are faster in signaling than the |S| chart. There are a few recent papers dealing with the joint control of the mean vector and the covariance matrix of multivariate processes. In this article, we propose a single chart based on the synthetic procedure for monitoring the mean vector and the covariance matrix of bivariate processes, named as the synthetic MRMAX chart. The sample points correspond to the maximum among four values: the standardized sample means and sample ranges of two quality characteristics. The main reason to propose the MRMAX chart lies on the fact that it is easier to deal with it if compared with the use of the joint T2 and |S| charts. The proposed chart only requires the computation of sample means and sample ranges, reminding univariate control charts. The

properties of the proposed chart were obtained by mathematical expressions. We investigated the performance of the synthetic MRMAX considering the usual sample sizes of 4 and 5, three levels of correlation and shifts in the mean vector and/or in the covariance matrix. The synthetic MRMAX chart is faster in signaling than the joint T2 and |S| charts, except when the correlation between the two quality characteristics is high. When the correlation is high, the joint T2 and |S| charts are, in general, faster in signaling assignable causes that only affect the mean and/or the variability of one of the two quality characteristics.

Gini Index in Czech Republic

Luboš Marek, Michal Vrabec
University of Economics Prague

The authors compute the values of Gini index in Czech Republic over period of years 1995–2009. They analyze the trend of Gini index over time and show that there is a linear dependency between this index and the time. Gini index is computed over the same period separately for men and woman, too. Authors compare the values of index for both sex and show that there is the significant difference between values for men and woman. The values for men are greater and the scissors between men and women are more and more opened over time. Futher is this index computed for three age group (till 30, 30–50, over 50). Very similar values are reached for two "older" groups, for the youngest group the values are dramatically smaller. The authors compute the values of Gini index for each region in Czech Republic in the next part of article. The time period for analysis is shorter because the new defininition of regions strarted in year 2000. The situation is very similar for all of regions excepting capitol Prague. The values for Prague are considerable and they are comparable with some advanced European countries. The table of Gini index values for some chosen countries are published in the last part of article. The authors compare the values for Czech Republic with values for these countries. The construction of Gini index is based on database around 3 million records with income values in Czech Republic. For the analysis is used the base definition of Lorenz curve and definition of Gini index. There is very detailed interval distribution of incomes for purposes of analysis. The intervals have the length 500 Czech crowns (apro-ximatelly 20 Euro) and data had the form: income interval frequency.

Estimating the Quality of Packet Transmission in Peer-to-Peer Applications

Natalia Markovich
Institute of Control Sciences, Russian Academy of Sciences

A main problem in the teletraffic theory concerns the transmission of the information with minimal loss and delay of the delivery. The information is transmitted by packets that may have constant or random lengths. The inter-arrival times (IATs) between packets are constant or random depending on the application. We consider the packet traffic in peer-to-peer (P2P) applications like Skype and IPTV where the packet lengths (PLs) and IATs are both random. The peers in a P2P overlay network may randomly join or leave the structure. Leaving peers cause the loss of all stored data at those peers and hence, downgrade the quality. We study the loss and delay at the packet layer. A peer failure or silence periods may cause the lack of packets in an aggregated stream of information sent to a peer and hence, a delay. The study of maxima of IATs is important since the IATs reflect the lack of packets. The main idea is that the packet loss may be caused not only by silence periods or peer failure but also by exceedances of the rate of transmission over the threshold that can be interpreted as a channel capacity. Such exceedances occur mostly in clusters of packets caused by their statistical dependence. Such clusters or conglomerates of frequently coming packets generate large rates which are approximated by the ratio Rate = PL/IAT. We determine the cluster as a set of packets between two consecutive packets such that their corresponding rates do not exceed the channel capacity. Hence, the loss or lack of packets lead to the delays which reflect on the restoration of images or voice samples. We focus on the estimation of the means and high quantiles of the general delay caused by different reasons and the delay in the clusters, the lossless time, the maximal IAT and the byte loss in the clusters. They are considered as quality indices of the packet transmission. For this purpose, we apply Wald\'s equation and nonparametric high quantile estimators. The data at our disposal contain the PLs and IATs between packets observed during a limited time interval. The quality indices (the delay, lossless time and byte loss) are generated for some capacity value using the PLs and IATs. All samples have moderate sizes. The estimation problems arise from the dependence and heaviness of tails of the PLs, IATs and the random quality indices. For example, the delay in the clusters is shown to be a dependent heavy-tailed random variable with infinite first two moments. A declustering of the data beforehand is proposed. This implies that the data are separated into independent blocks and one can deal with block maxima (or other representatives of blocks) as with independent data. The high quantiles of quality indices are estimated by means of quantiles of corresponding block maxima using the estimate of the extremal index. The proposed methodology of the quality control can be applied to any multimedia packet stream in P2P applications.

Optimal Supersaturated Designs Under Measures of Multicollinearity

***Chris Marley**, Dave Woods, Sue Lewis*
University of Southampton, UK

Suppose a scientist or engineer is interested in establishing which inputs are most important in a scientific or manufacturing system. These inputs may be, for example, controllable process variables or chemical constituents of a fluid. An experiment can be used to establish which of these inputs, or 'factors\', have substantial effects on the response(s) of interest. Such an experiment consists of several design points, made up of different combinations of the levels, or values, of the factors. Supersaturated designs may be defined as experimental plans with at least as many factors as runs. Such designs can prove useful in screening situations, where the experimenter only wishes to investigate which are the few dominant factors, as opposed to fitting a detailed statistical model. Supersaturated experiments are attractive to industry because they enable experimenters to investigate a large number of factors in a relatively small number of runs. Thus costs can be reduced when experiments are very expensive to perform. Most existing criteria for generating or assessing supersaturated designs focus only on dependencies between pairs of factors. We propose a new class of criteria for supersaturated designs based on measures of multicollinearity among subsets of the factors. Unlike some existing criteria, this new class can be used to design experiments where factor levels cannot be set independently. Although there is much attention given to supersaturated designs where the factors can be independently set at one of two levels (high or low), there is little work on cases where the combinations of factor levels cannot be set independently. This may occur when the experimenter has a list of compounds which have different properties. In this case, the properties are the factors in the experiment and one compound must be chosen for each run of the experiment. Clearly, choosing the compound fixes the levels of all the factors for this run. We apply the new criteria to such an experiment with a large list of compounds to choose from. This example is used this to demonstrate the benefits of the new methodology and to illustrate some desirable properties of the resulting designs.

FDA for Tree-Structured Data Objects

J. S. Marron
University of North Carolina at Chapel Hill

The field of FDA has made a lot of progress on the statistical analysis of the variation in a population of curves. A particularly challenging extension of this set of ideas, is to populations of tree- structured objects. Deep challenges arise, which involve a marriage of ideas from statistics, geometry, and numerical analysis, because the space of trees is strongly non-Euclidean in nature. These challenges, together with some first approaches to addressing

them, are illustrated using a real data example, where each data point is the tree of blood vessels in one person's brain.

Object Oriented Data Analysis

J. S. Marron
University of North Carolina at Chapel Hill

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Recent developments in medical image analysis motivate the statistical analysis of populations of more complex data objects which are elements of mildly non-Euclidean spaces, such as Lie Groups and Symmetric Spaces, or of strongly non-Euclidean spaces, such as spaces of tree-structured data objects. These new contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics. Even in situations where Euclidean analysis makes sense, there are statistical challenges because of the High Dimension Low Sample Size problem, which motivates a new type of asymptotics leading to non-standard mathematical statistics.

The Rise and Fall of Quality in the Introductory Statistics Course for North American Business Students

John McKenzie
Babson College

In 1986 the first Making Statistics More Effective in Schools of Business (MSMESB) conference was held at the University of Chicago. The organizers listed a number of ways that the introductory statistics course needed to change to better educate its future business and industrial leaders in undergraduate and MBA programs (Easton, Roberts, and Tiao, 1988). Subsequent annual MSMESB conferences up to 2002, including two with W. Edward Deming present, and articles from these conferences, led to many positive improvements in schools of business. Among these is the use of statistical software to analyze real-world data and to determine critical values, instead of hand calculations of naked numbers and table look-ups. Other recommendations included an emphasis on time series instead of cross-sectional data and less emphasis on hypothesis testing, have not been heeded. Increased topical coverage to include control charts and other quality tools is the focus of this presentation. These techniques began to appear in a number of courses after articles such as Levine's 1992 "Business Statistics Curricula

Lacks Quality" appeared. In recent years these topics have almost disappeared from these required business school courses. This presentation provides evidence for this ascend and descend by summarizing surveys such as Rose, Machak, and Spivey (1988); Strasser and Ozgur (1985); McKenzie and Taylor (1996); McAlevey, Everett, and Sullivan (2001); and Haskin and Krehbiel (under review), and discusses the consequences for business and industry. These changes are apparent in a number of successful North American textbooks such as those written by Anderson, Sweeney and Williams, and Berenson and Levine over the past 25 years. The talk explains some reasons for this disconnect between academics and practitioners. It concludes with some suggestions on how to reverse this slide. Asking business students to critique a Pareto chart is an excellent way to build upon their previous knowledge of other displays and helps better prepare them for occupations in business and industry. Another way is consider tests for special causes other than the familiar test to locate one point more than 3s from center line. Explaining the rationale for some of these tests via simulation is a superb way to show many of the students how they will solve problems in their future. Bringing back in-class active-learning exercises such as the Red Bed Experiment is outstanding way to reinforce important statistical concepts in a business context. Providing more real-world examples of quality management data from service industries instead of the more frequently presented manufacturing industries is especially important in North America.

Automatic Crack Detection Algorithm for Vibrothermography Sequence-of-Images Data

William Meeker, *Ming Li, Steve Holland*
Iowa State University

Vibrothermography is a technique for finding cracks through frictional heat given off in response to vibration. Vibrothermography provides a sequence of infrared images as output of the inspection process. A fast and accurate automatic crack-detection algorithm for the sequence-of-images data will greatly increase the productivity of vibrothermography method. Matched filtering is a technique widely used in signal detection, and it is the optimal linear filter to maximize the signal-to-noise ratio in the presence of additive uncorrelated stochastic noise. Based on key features from images of known cracks, we can construct a three-dimensional matched filter to detect cracks from the vibrothermography data. In this paper, we evaluate the matched filter developed from a vibrothermography inspection sequence-of-images. We then use the output of the matched filter to develop a detection rule. The probability of detection for the matched filter detection algorithm is then compared with the probability of detection for a simpler detection algorithm that is based on a scalar measure of the amount of heat generated in an inspection. Our results show the matched filter algorithm provides improved detection capability when a flaw signature is known approximately.

Warranty Prediction Based on Auxiliary Use-rate Information

William Meeker
Iowa State University

Yili Hong
Virginia Tech

Usually the warranty data response used to make predictions of future failures is the number of weeks (or another unit of real time) in service. Use-rate information usually is not available (automobile warranty data are an exception, where both weeks in service and number of miles driven are available for units returned for warranty repair). With new technology, however, sensors and smart chips are being installed in many modern products ranging from computers and printers to automobiles and aircraft engines. Thus the coming generations of field data for many products will provide information on how the product has been used and the environment in which it was used. This paper was motivated by the need to predict warranty returns for a product with multiple failure modes. For this product, cycles-to-failure/use-rate information was available for those units that were connected to the network. We show how to use a cycles-to-failure model to compute predictions and prediction intervals for the number of warranty returns. We also present prediction methods for units not connected to the network. In order to provide insight into the reasons that use-rate models provide better predictions, we also present a comparison of asymptotic variances comparing the cycles-to-failure and time-to-failure models.

Model Evaluation in Presence of Rare Classes

Giovanna Menardi*, Nicola Torelli*
University of Trieste

When dealing with a classification task, evaluating the accuracy of the classifier plays a role that is at least as important as the model estimation. Indeed, both the choice of the best classification rule among alternative ones and the extent to which a classification rule may be operatively applied to real problems for labelling new unobserved examples, depend on our ability in measuring the classification accuracy. The issue becomes of even more critical significance when classification is performed in a binary imbalanced framework, that is when one class is rare. It has been widely reported that such imbalance heavily compromises the process of learning, because the classifier is overwhelmed by the majority (negative) examples, and ignore the minority (positive) ones. Although interest in learning with imbalanced data has recently grown, the issue of model assessing has not received yet as much attention as the one focusing on model training. In fact, even if an effective classification rule was trained on the data, the class imbalance would still lead to non-

negligible consequences when evaluating the model accuracy. A first problem concerns the choice of the evaluation metric to be used for performance measurement. The use of common measures, such as the error rate, yields to misleading results because they depend on the class distribution. More appropriate metrics are based on different propensity towards false negatives and false positives (e.g. precision, recall or ROC curves). Although these evaluation metrics share some drawbacks, the research activity focusing on this issue is very fruitful and several advances have been made. In fact, the evaluation of the accuracy of a classifier in unbalanced learning is subject to a more serious problem than the choice of an adequate error metric, concerning the estimate of such accuracy: whatever evaluation metric is chosen, its expression depends on the unknown probability distribution underlying the data, and hence estimation of this quantity has to be considered. In most of the literature about imbalanced classification, the empirical analysis consists in estimating the classifier over a training set and assessing its accuracy on a test set. However, in real data problems, there are not enough examples from the rare class for both training and testing the classifier and the scarcity of data leads to high variance estimates of the error. It stands to reason that poor estimates of the classifier's performance may lead to erroneous conclusions about the quality of the classifier and proposing more and more sophisticated learning methods becomes a wild-goose chase if we are not able to evaluate their accuracy. In this work we show how smoothed bootstrap techniques may be effectively used in imbalance learning to improve the quality of the estimates of the accuracy.

A Semi-Theoretician's Mid-Day Confession: The True Meaning of i.i.d. in (Applied) Statistics

Xiao-Li Meng
Department of Statistics, Harvard University

Doing good (applied) statistics is inherently – and increasingly – difficult. The size of the data and the complexity of their structure are increasing, as are the depth and specificity of the investigation goals. Yet the available time for conducting the study is decreasing due to intensive competition, especially for funding. Statistical consultants are therefore increasingly asked to perform magic, such as providing scientifically valid causal conclusions for a deeply stratified subgroup based on a handful of weighted samples with wildly varying weights. And by the way, the analysis must be done in one week and the methods must be implementable by (and explainable to) analysts whose statistical experience might come largely from reading output from SAS/SPSS/Stata. This is not a cynical observation, but rather a real challenge that we as statisticians must face in order for our profession to remain at the core of quantitative scientific investigation. In this talk, I will report my own experiences both crying and smiling as a member of a team of statistical consultants for the National Latino and Asian American Study (NLAAS), a recent survey of psychiatric epidemiology, which measured over 5000

variables and embedded experiments on different survey instruments. In particular, I will report on the success and failure of using Bayesian modeling and multiple imputation to deal with the respondents' untruthful self-reporting regarding health service use, as detected by the embedded experiments.
The true meaning of i.i.d. will become clear only by the end of my talk; unless, of course, you have already deciphered it from this abstract...

Network–WIDEE Statistical Modeling and Prediction of Computer Traffic

George Michailidis
University of Michigan

Computer network use is becoming increasingly widespread, both in terms of number of users and variety of applications. In order to provide consistently high quality service, network engineers and other professionals must monitor several aspects of the network, including the traffic intensity on the links that comprise the network. As networks grow, this type of monitoring has potential to become burdensome in terms of resources required. Motivated by the prospect of monitoring only a small subset of links, we explore the problem of using observed traffic measurements on selected links to predict the traffic on other, unobserved links. The characteristics of such unobserved links are learned through auxiliary data. Although more expensive to obtain, this extra data set provides the necessary information to represent important structure in the network, and can significantly improve the results of prediction as compared with more naive approaches. In addition, we introduce an adjusted control chart methodology that shows possible applications of our prediction results in situations where all links may be observed.

The Mathematical Model of an Assembly Link Checking Stand

Jiří Michálek
Institute of Information Theory and Automationof the ASCR, Prague, Czech Republic

The mathematical model of an assembly link checking stand Jiří Michálek Institute of Information Theory and Automation Czech Academy of Science The construction of the proposed model is motivated by a real case of a checking stand where valves coming from the assembly link are tested and repaired respectively. The assembly link presents one-flow production of valves that are coming to the test station. When a valve passes the test the valve is marked OK and released to expedition. In case a valve does not pass the test the first attempt of repair follows. After repairing the valve goes through the test again

and if the test is passed the valve is released. But when the test is not fulfilled the second attempt of repaire is starting and so on until the valve is OK. In the real situation the number of possible repaires is not limited but in the model this number is limited by three attempts for simplicity. These valves that could not be repaired during three attempts are declared as scraps and excluded from the next operations. The mathematical model is based on Markov and semi-Markov chains. The embedded Markov chain has 6 states in a simplier form considered in this contribution. These states are: state T for testing, state OK after passing the test, three states R1, R2 and R3 for successive repaires. State S is for scraps. The values of transition probabilities for the transition matrix were determined by estimates obtained from the automatic collection of data performed at the checking stand. The embedded Markov chain is nonperiodic and irreducible having the unique class of ergodic states. In accordance with the real situation a random amount of time spent in every state must be considered. Possible probability distributions are like log-normal or Weibull, also in special situations normal distribution could be acceptable. The matrix of transition probabilities together with probability distributions of random times define in a unique way the corresponding semi-Markov chain. This semi-Markov chain is in fact a regenerative process because after every state T the probabilistic structure of the process is the same. The most interesting characteristics are the mean length of time interval between two states OK and simmilarly the mean time between two scraps. Further interesting characteristics are the probability of accurence of a scrap or number of scraps within a time period, e.g. within a shift. The model can serve as a suitable tool for observing changes in its characterictics under changes in input transition probabilities and parameters of time distributions.

Quasi-Quantile Regression

Ivan Mizera
University of Alberta

Quantile regression, arguably "one of the most important development in statistics in recent years" (according to the new edition of the Ramsay's and Silverman's Functional Data Analysis book), is becoming an increasingly popular data-analytic tool. After reviewing several areas of applications of this methodology, illustrating the type of achievable insights on data examples and emphasizing the involvement of modern methods of convex optimization, we concentrate on one of its "twilight zones" (in the terminology of the Koenker's monograph on the subject): fitting of parametric models of quantile regression in the circumstances typically leading, in the mean-regression context, to what is generally referred to as generalized linear models. In particular, we focus on quantile regression of counts, where the existing methodology still exhibits somewhat peculiar aspects; we try to adapt the ideology of quasi-likelihood to quantile regression there, in the hope of obtaining more conventional, but still valid fitting strategies.

Multi-Treatment Regression Approach for the Time Evolution of Heavy Metals Concentration in a Electrodialtic Removal Method

***Elsa Moreira**, João Tiago Mexia*
Research Center in Mathematics and Applications, Faculty of Sciences and Technology, Nova University of Lisbon

The electrodialytic removal of heavy metals from waste materials impregnated with chemicals is a remediation process that promotes the re-use. The electrodialytic removal of Cu, Cr and As was tested in wood chips treated with chromated copper arsenate (CCA). The method uses a low-level direct current as the cleaning agent, in the presence of an extracting solution. Several experiments were conducted using different extracting solutions and initial current intensity, which will be considered as different treatments. A polynomial model was fitted to the time evolution of each metal concentration in the electrolytes. Based on this modeling and aiming to choose the best treatment in jointly removing the three metals, five experiments were selected in order to be analyzed under multi-treatment regression approach. In this approach, instead of a sample for each treatment, there is a linear regression in the same variables, both controlled and dependent. Then, instead of the action of the treatments on the sample mean values, the action on the regression coefficients is studied. ANOVA algorithms and multiple comparison methods are adapted aiming to perform the comparison between the coefficients of different regressions. Data was unbalanced, that is, the model matrix $\boldsymbol{X}$ for the values of the controlled variables was not the same in all regressions. Thus, a special covariance matrix (diagonal by blocks) had to be considered in order to perform the hypotheses testing. The results point to the choice of the oxalic acid 2.5% and a current ranging from 20 to 40 mA.

Probabilities of Extreme Half-Spaces

Stephan Morgenthaler
Ecole polytechnique fédérale de Lausanne

Debbie Dupuis
HEC, Université de Montréal

In this talk we will present a method based on the generalized Pareto distribution for estimating probabilities of extremal half-spaces involving multivariate observations. If the multivariate distribution is elliptical we can use affine transformations between half-spaces. This allows us to combine estimators based on a variety of one-dimensional projections of the multivariate data set, rather than using a single projection. More reliable estimators result from this practice. Furthermore, the technique suggests some natural estimates of the sampling variance. An application to risk assessment in financial data will be given.

Risk Management: A Differential Diagnosis

John Morrison
Asymptotix

I am leaving the Large and Complex Data Set as the silent elephant in the room for the purposes of this paper. Instead, I am looking along a differential perspective at the nature of the end-client of the Large and Complex Data Set. I want to look at the key Large and Complex Data set (as an abstract) in the real world today; the Data Warehouse in a Financial Institution. Basic Economics tells you that the data warehouse (or whatever it may be called today) is a supply partner to the client applications which demand its data. In the real world in Risk Management in Financial Institutions, these client applications are for the most part either appliance model or Model Development Environments. These are the applications which actually consume extractions of information from massive data sets. This paper is an examination of the context and management of these two types of Complex Data Set clients in the real world of banking as we know it today. I provide a further examination of the demands being placed on the client layer for the Complex Data Set. Consequently I can depict a supply chain of data management requirements in banking today. Finally, I draw some conclusions about the future of the large and complex data The key concepts which require to be addressed in this panel of the conference are; 1) Large and Complex Data Sets, 2) Extractions of information from massive data sets & 3) Illustration of how the extraction of information from large and complex data sets is of critical concern to an end-user. This paper (for which this document is only the abstract) is the foundation for a presentation addressing these criteria. I will do that in the context of the Credit Crisis. I do that by zooming in on the underlying technical aspects of the paper. The key issue I want to share is in the questions surrounding true transparency in Banking and Financial Markets. As I have said, Banking Transparency is a genuinely scary very large scale complex data set management problem of our time. It could be strongly argued that there is an overwhelming political need to get it done, both in the US and Europe. The question is is it a social externality i.e. is it such a challenge that it is 'beyond' the private sector market system to solve and is it such an important social need that the state in both the US and Europe is required to step in and support the development of demonstrably transparent data management in Banking and Financial Services more widely?, in banking and securities industries.

On a Discrete Laplacian Based Method for Outliers Detection in Phase I of Profiles Control Charts

Francisco Moura Neto
Polytechnic Institute/State University of Rio de Janeiro (IPRJ/UERJ), Brazil

Maysa De Magalhães
National School of Statistical Sciences/Brazilian Institute of Geography and Statistics (ENCE/IBGE), Brazil

Monitoring profiles is useful in maintaining production processes output according to previously set standards, thus guaranteeing high quality products (Kang, L. and Albin, S.L. (2000), Zhang H. and Albin, S. (2009)). One relevant aspect of the monitoring refers to Phase I where the standard is defined. This depends on historical data retrieved from the process and the distinction between in-control behavior and out-of-control behavior of the process (outliers). We consider here an effective method to detect outliers based on a discrete Laplacian operator from conveniently defined scalar products of historical data profiles, and the associated Fiedler vector in order to determine the outliers and, in this way, improving the estimator of the in-control profile model (Archip, N et al, 2005). Simulation results imply that the methodology is promising. We compare the results with those obtained by Zhang H. and Albin, S. (2009).

Control Charts for Monitoring Linear Profiles

Francisc Moura Neto, *Viviany Fernandes*
Polytechnic Institute/State University of Rio de Janeiro (IPRJ/UERJ), Brazil

Maysa De Magalhães
National School of Statistical Sciences/Brazilian Institute of Geography and Statistics (ENCE/IBGE), Brazil

Control charts are useful in monitoring production processes, usually by keeping track of a few key quality characteristics of the outcome of the production process. The monitoring can be accomplished by means of univariate or multivariate charts. In some cases, variables that reflect main characteristics of the production process can be represented by some profile, that is, they stand as linear or nonlinear relationships. Here, we propose the monitoring of a linear profile employing a flexible model for the quality characteristic of the production process. The model uses the Kalman filter methodology coupled with a chi-square control chart to monitor the process in phase II operations. By simulation, we conclude that the chart based on Kalman filter out-performs the chart based on the ordinary least squares and the competing charts based on exponentially weighted moving averages.

Why Fill Space?

Werner Mueller
JKU Linz

Space-filling, particularly equidistant designs have become the paradigmatic approach for setting up large-scale computer simulation experiments. Since output from these experiments are predominantly modeled by correlated random fields, we will investigate two approaches that design the experiments to allow for obtaining precise predictions over the whole experimental domain. Both take the uncertainty of the estimated parameters of the correlation structure of the random field into account. The first one corresponds to a compound D-optimality criterion for both the trend and covariance parameters. The second one relies on an approximation of the mean squared prediction error already proposed in the literature. It will be shown, that typically these two approaches yield similar optimal designs, but such that can be rather far from the ubiquitous equidistant space-filling ones. As a side issue we will also show that a Kiefer-Wolfowitz type equivalence relation can typically not be achieved in the correlated error setup. The talk is based upon joint work with L.Pronzato, M.Stehlík and H.Waldl.

Influence of Macroeconomics Indicators in Power Prices: the Spanish Case

Maria Pilar Muñoz
Universitat Politècnica de Catalunya (UPC)

David A. Dickey
Department of Statistics, North Carolina State University

Several are the causes that have influenced the volatility of power prices in developed countries. In general, the current volatility makes it difficult to predict price behavior, even more so in the case of power markets. In addition, the very nature of components used in power generation such as gas, coal and petroleum are very volatile. So, to know and determine the causes of variations in power prices are very important, even more so in the case of Spain, which has to import the majority of its commodities, of which the prices are quoted in dollars. The objective of this work is twofold. First of all, from an econometric point of view, to examine the relationship between Spanish electricity spot prices, the US dollar/Euro Exchange rate and oil prices in order to better understand the evolution of the former. The findings in this part are very interesting. Results have shown that there is a long-run relationship between the three variables under study. Regarding short-run relationships, both Spanish electricity prices and the US Dollar/Euro exchange rate are influenced by the evolutions of oil prices although Spanish electricity prices are not affected in the short run by changes in the evolution of the US Dollar/Euro exchange rate. Finally, and considering the transmission of

volatility between the three variables, the three volatilities are related, in the sense that the volatility of both indicators (US Dollar/Euro exchange rate and oil prices) affects one of the Spanish electricity prices (Muñoz and Dickey, 2009). The estimation of the multivariate volatility has been carried out by mean of a VAR(4)-GARCH(1,1)-BEKK model. Secondly, extreme value theory has been applied for modelling the extreme events of those variables and detecting asymptotic dependency between them. The findings in this part are that there is weak asymptotic dependence in two of the three possible relationships between variables; specifically, the extreme events between Spanish electricity prices and the exchange rate are asymptotic related, with a similar result for the Spanish Electricity and oil prices. Meanwhile the extreme events of the two macroeconomic indicators (the exchange rate and oil prices) present asymptotic independence. The main conclusion for this work is that the Spanish electricity market is truly depending on the macroeconomics indicators that are used; therefore, the fluctuations in the exchange rate could pose a possible risk in the prices of fossil fuels. If so, that would introduce a high risk in the Spanish electricity market. The corollary of this conclusion is that for countries like Spain, which import fossil fuels, the use of renewable energy could significantly increase their security.

Statistics of High Frequency Financial Data: Background and New Developments

Per Mykland
University of Chicago and University of Oxford

Recent years have seen an explosion of the literature in the area of estimating parameters (such as volatility) on the basis of high frequency financial data. We give some background for this kind of inference, and then discuss challenges and recent innovations in the area. The talk is particularly focused on issues involving market microstructure and local likelihood.

Inference in Multi-state Semi-Markov Models with Interval-Censored Data

Vijay Nair
University of Michigan

Multi-state models arise naturally in survival, reliability, and risk analysis applications. Inference in such models in the presence of censored data presents interesting challenges. This is especially true of interval censoring which is common. The talk will begin with a review of inference for Markov models. Difficulties in doing inference with semi-Markov models are then

described. The problems are more computational than conceptual in nature. Some methods for likelihood and Bayesian inference in parametric semi-Markov models, using data augmentation and stochastic approximation techniques, are described. Extensions to situations with covariates, non-parametric and other situations will be outlines as time permits. This talk is based on joint work with Yang Yang.

Ruin Probability in Data Transmission for Computer Network by Symmetric Alpha Stable Distributions Approach

Hassan Naseri
Ministry of Energy

Helda Abdshah
Ministry of Agriculture

In this paper will be discussed on a problem of data transmission in computer network by using of stable distributional hypothesis in a ruin strategy. Concepts are illustrated for the case of number of data packets arrives at a node of the network in each time-slot and at least one packet is waiting and each time one packet can be transmitted .We can assume that the number of packets arriving are random variables with symmetric stable distribution set, first we define a Risk Crowd Process and then we try to find ruin behavior in this model as follows: first we will introduce a nice case where this is clearly visible stems from queuing theory by stable distributions then we suppose a sequence of risks which are assumed to be independent and identically distributed in a discrete time domain, hence let at each time $n \in N$ a random number of data packets arrives at a node of network; let in time-slot $[n, n+1)$ one packet can be transmitted provided that at the beginning of this slot at least one packet is waiting, therefore if Xn be the number of packets arriving at n then $X1, X2, \ldots$ would be independent and we can suppose they are identically distributed with a symmetric stable distribution .Suppose that just before time n=0 there L0 is a random , but independent of $X1, X2, \ldots$. After one unit of time the number of packets decreases by 1. But at the same time X1 new packets arrive and so $(L0 + X1 - 1) +$ (where $x+ = \max\{0, x\}$) packets are waiting just before the beginning of the next slot .In general , the number Ln of packets just before the beginning of the $(n+1)-$ the slot fulfils the recurrence relation $Ln = (Ln-1 + Xn-1)+$, Ln is called the queue length at time n. we can write Ln in the other following style: $Ln = \max\{0, Yn, Yn-1 + Yn, \ldots, L0 + Y1 + \ldots + Yn\}$, where $Yi = Xi-1$. We can prove that if $Y1, Y1, \ldots$ be independent and identically distributed. Then, $(Yn, Yn-1 + Yn, \ldots, Y1 + \ldots + Yn)$ same distribution as $(Y1, Y1 + Y2, \ldots, Y1 + \ldots + Yn)$ for all $n = 1, 2, \ldots$ Now suppose that the sequence $\{Ln, n \in N\}$ describes the evolution of the packets which are waited for transmitting in the network.By an equation for definition on Ln, Ln can be called Risk Crowd Process and it can be positive for some $n \in N$. This event $\{L1 > 0\} U \{L2 > 0\} U \ldots$ is called the (technical) Ruin of the Network.

Estimation of R=P(Y<X) for Two-Parameter Exponential Distribution

Parviz Nasiri
University Payame Noor

In this paper we consider the estimation of the stress-strength parameter $R = P(Y < X)$, when X and Y are independent and both are two-parameter exponential distributions with the common location parameters but different scale parameters. It is observed that the maximum likelihood estimators do not exist in this case, and we propose a modified maximum likelihood estimator, and also an approximate modified maximum likelihood estimator of R. Analyses of two data sets have also been presented for illustrative purposes.

Statistical Inference for the Right Endpoint of a Light Tailed Distribution

Cláudia Neves
University of Aveiro

In Extreme Value statistics we often encounter testing procedures for assessing the presence of the Gumbel domain, attached to the simple null hypothesis of a shape parameter equal to zero. The problem of assessing for light-tailed distributions with finite or infinite right endpoint is seldom referred. In this talk, we present two testing procedures which enable us to distinguish light-tailed distribution functions with finite right endpoint from those with infinite endpoint lying in the Gumbel domain. Illustrative examples are also provided.

The Effect of PLS Regression in PLS Path Model Estimation When Multicollinearity is Present

Rikke Nielsen*, Kai Kristensen, Jacob Eskildsen*
Aarhus School of Business, Aarhus University

PLS path modelling has previously been found to be robust to multicollinearity both between latent variables and between manifest variables of a common latent variable (see e.g. Cassel et al. (1999), Kristensen, Eskildsen (2005), Westlund et al. (2008)). However, most of the studies investigate models with relatively few variables and very simple dependence structures compared to the models that are often estimated in practical settings. A recent study by Nielsen et al. (2009) found that when model structure is more complex, PLS path modelling is not as robust to multicollinearity between latent variables as previously assumed. A difference in the standard error of path coefficients of

as much as 83 % was found between moderate and severe levels of multicollinearity. Large differences were found not only for large path coefficients, but also for small path coefficients and in some cases the difference could lead to a change in conclusions about variable importance. A possible remedy of multicollinearity in PLS path models is to replace the OLS regressions by PLS regressions. PLS regression is then used to estimate the path coefficients of the structural model in the case of multicollinearity between latent variables. More recently, PLS regression has also been proposed as a 'Mode PLS' alternative to Modes A and B for estimating the inner weights when the measurement model has a formative specification and multicollinearity is present between manifest variables (Esposito Vinzi (2009)). The present study investigates the effect of using PLS regression compared to standard OLS regression in the presence of multicollinearity in both the structural and measurement models of a PLS path model. This is done by means of a simulation study that is designed to reflect model complexity and multicollinearity as experienced in practice.

Integration of SPC and APC in Semiconductor Manufacturing Process

Ken Nishina
Nagoya Institute of Technology

APC (Automatic Process Control), especially the feedback control, is frequently used in semiconductor manufacturing processes. Scheming an integration of SPC (Statistical Process Control) and APC in the process, the viewpoints to apply SPC effectively to the manufacturing process with APC can be indicated as follows: (1) APC reinforces SPC. (2) SPC complements APC. In this study, how to utilize control charts based on the viewpoints is discussed. Box and Kramer (1992) and others have proposed the integration of SPC and APC. Their main proposals are control charts based on the time series model. In contrast with their proposals, our first viewpoint is a scheme of the effective integration of SPC and APC. In the feedback control systems the manipulated variable can be regarded as an input variable of the manufacturing process such as treatment time. In the process there are some factors which have the interaction with the manipulated variable. Therefore, the manufacturing processes itself can be controlled effectively by monitoring the relationship between the monitoring variable of being process output and the manipulated variables of being process input. Traditionally, process control by control chart has been a passive approach by monitoring the output variable only; however, it can be developed into an active approach by monitoring the relationship between the input variable and the output variable. Concretely, the ratio of output to input (the process rate) is recommended as the monitoring variable. It can be shown that APC reinforces SPC. Feedback control is aimed at reduction of the between–batch variation not the within–batch variation. A main action to reduce the within–batch variation is the maintenance, for example the removal of by–products and the exchange of parts of the

manufacturing equipment. APC is insufficient for the reduction of the within–batch variation. Then, SPC is useful to control the within–batch variation. This is our second viewpoint. R charts is widely used to control the within–batch variation, however, R charts are statistical tools to monitor the amount of the within–batch variation only. In many cases, some systematic variations can be seen in the within–batch variation; moreover, the systematic variations are varied over the time series between the maintenances. It can be visualized by principal component analysis. Therefore, the systematic patterns and the departure from the systematic patterns should be monitored by using another statistical tool instead of R charts. It is proposed that T2 – Q charts, which was proposed by Jackson (1979), can be applied to monitor the within–batch variation. It can be shown that SPC complements APC.

Bridging the Way Between Academia and Practice: the ENBIS Way

Irena Ograjenšek
University of Ljubljana, Faculty of Economics

Shirley Coleman
Newcastle University, ISRU

Bart De Ketelaere
Katholieke Universiteit Leuven, Faculty of Bioscience Engineering

The ENBIS Session at ISBIS 2010 will take a form of a panel entitled "Bridging the way between academia and practice: the ENBIS way" and will address the following issues: (1) How to establish cooperation between academia and businesses: alternative approaches. (2) Real-life showcases of success and failure in projects based on cooperation between academia and businesses. (3) EU-funding of cooperation between academia and businesses. (4) Dissemination of experiences from cooperation between academia and businesses.

Composite Likelihood for Random Fields: Rainfall Applications

Simone Padoan
Ecole Polytechnique Federale de Lausanne

Moreno Bevilacqua
Università Ca' Foscari di Venezia,

In the last few decades random fields have emerged as a common tool for the statistical modeling of spatial analysis. A common objective of spatial analysis

is to quantify and characterize the behavior of environmental phenomena such as precipitation levels, wind speed, or daily temperatures. These analyses are therefore useful devices for understanding and predicting environmental events. The maximum likelihood method is considered the best choice when estimating random fields, but it can be cumbersome in some cases. For instance, assuming Gaussianity, it is impractical when dealing with large dataset for computational reasons. Furthermore, random fields for spatial extremes often involve intractable multivariate densities so that likelihood inference is not feasible. We show that a good strategy is to approximate the likelihood of spatial models using the composite likelihood. The estimation procedure based on this approach is computationally more efficient than that of the standard likelihood. Also with complex models the composite likelihood provides a flexible framework for inference, that is otherwise difficult to supply. We illustrate with an analysis of precipitation data the reliability and versatility of the proposed method.

Extrinsic Analysis on Manifolds is Computationally Faster than Intrinsic Analysis, with Examples from Shape and Image Analysis

***Vic Patrangenaru**, Xiuwen Liu, Leif Ellingson*
Florida State University

Rabi Bhattacharya
University of Arizona

Samanmalee Sughatadasa
Texas Tech University

In our technological era, non-Euclidean data abounds, especially due to advances in digital imaging. Patrangenaru (1998) introduced extrinsic and intrinsic means on manifolds, as location parameters for non-Euclidean data. Once large sample and nonparametric bootstrap analysis was set in place around Y2K by Bhattacharya and Patrangenaru, a flurry of papers in computer vision, statistical learning, pattern recognition, medical imaging and other computational intensive applied areas using these concepts followed. When pursuing such location parameters in various instances of data analysis on manifolds, scientists are using intrinsic means, almost without exception. In this paper, we debunk the intrinsic analysis myth, by using John Nash's celebrated isometric embedding theorem and an its equivariant version. We show that for each intrinsic analysis there is an extrinsic counterpart, that is computationally faster, and give some concrete examples in shape and image analysis. The authors is grateful to NSF-DMS-0805977 and to NSA-MSP-H98230-08-1-0058 for their support. This is joint work with R. N. Bhattacharya, X. Liu, L. Ellingson and S. Sughatadasa.

A Methodology for 3D Scene Reconstruction from Digital Camera Images

Michael Crane, Wei Liu
Florida State University, USA

Xavier Descombes
INRIA - Sophia Antipolis, France

*Wei Liu**, Vic Patrangenaru,** Xiuwen Liu*
Florida State University

This paper concerns a methodology for a two steps reconstruction of a 3D scene, including texture, in absence of occlusions, from arbitrary partial views. Patrangenaru and Patrangenaru (2004) derived a projective mean shape based reconstruction for planar scenes, using image fusion around landmark configurations of arbitrary views. In this paper we first analyze the advantages and limitations of such a reconstruction of a a close to planar remote scene from its partial aerial views, by specializing this algorithm to affine transformations. Furthermore, using a projective shape reconstruction of a 3D finite configuration from its uncalibrated camera views, as developped in Patrangenaru, Liu and Sughatadasa (2010), we reconstruct projectively a locally polyhedral 3D scene. This research was supported by the National Science Foundation, by the National Security Agency and by the INRIA/Florida State University Associated Team SHAPES grant.

Empirical Retrospective Power: Background, Planning and Use

Fortunato Pesarin, Monjed Samuh
Padua University

Until very recently, many authors, especially Psychologists, start using what so called \textsl{retrospective power} (also called a posteriori power, post hoc power, observed power, achieved power, or conditional power) in response to the demand of some scientific journals and editors especially when the outcome of the test is not significant or slightly significant. This paper is raised at the time when misunderstandings and misconceptions abounded concerning retrospective power; it has been noticed that some authors disagree to calculate the retrospective power in the sense is unhelpful in the presence of the crude p-value and some others advocate the use of retrospective power in the sense that has another interpretation than what we have from the crude p-value. This study tries to discover the nature problem according to this concept, tries to summarize what is available in the literature and to dispel some confusion concerning this concept. Retrospective power within permutation approach is developed, its convergence to prospective power is investigated as well as the connection between prospective and retrospective

power is studied. Finally, real data applications and simulation studies are considered.

On Multi-Aspect Permutation Testing

***Fortunato Pesarin**, Luig Salmaso*
University of Padova

Often in the experimental context the action of a treatment is on more than one aspect of a distribution. In the sense that it is relatively rare that responses are merely shifted of a fixed quantity and is quite more frequent that each individual react to treatment differently from others: some individuals may have null, others small and others large effects. From the one hand the effects rarely can be assumed fixed; from the other the resulting distribution "stochastically dominates" the null one. Moreover, it is generally difficult to justify that random effects are "independent" of null responses. That is, denoting with X the null response, with d the random effect, and with $X1$ the treated response, we may write $X1 = X + d$ is larger in distribution than X and so two variables X and d may not be independent. In particular they are not independent if assumed normally distributed. This issue may produce specific testing problems. For example, when testing for equality of two distributions in a two-sample problem where the treatment is presumed to act on several different aspects of a distribution, e.g. on first and second moment, several different tests, each specific to one of the aspects of interest, should be applied leading to a multi-aspect testing issue. If the global testing is of interest, a combination of several dependent partial tests then becomes a necessity. This problem is tackled and solved within the nonparametric combination (NPC) of dependent permutation tests and related theory according to Pesarin F. (2001): Multivariate Permutation Tests, with Applications in Biostatistics, Wiley, Chichester. Moreover, several application examples are also presented: 1) based on the property that if two or more tests are unbiased and at least one of them is consistent their NPC is consistent, concerns the consistency and robustness of any multi-aspect test; 2) one is concerning the exact solution of the well-known Behrens-Fisher problem when, based on the notion of randomization of individuals to treatments, data exchangeability is assumed in the null hypothesis; 3) one more concerns testing problems when the treatment effect is positive on some subjects and negative on some others, as it occurs with drugs having genetic interaction, giving rise to multi-sided tests; 4) a further one is related to testing for the so-called multivariate stochastic ordering.

Accuracy in Cell Image Segmentation Algorithms

***Adele Peskin**, Alden Dima, Joe Chalfoun*
National Institute of Standards and Technology, Boulder

We have performed image segmentations on a large number of images, using five imaging conditions and two mammalian cell lines, in order to study trends in the segmentation results and make predictions about segmentation accuracy. Comparing results from approximately 40000 cells, we find a linear relationship between the highest segmentation accuracy seen for a given cell and the fraction of pixels in the neighborhood of the edge of that cell, the fraction at risk for error using any method when cells are segmented. We call the ratio of the size of this neighborhood to the size of the cell the "extended edge neighborhood" which we can use in defining a quality measure of cell image segmentation. Such a measure is important for accurate estimation of cell types, cell sizes, and the effects on cells with environmental changes.

Optimizing Execution Runtimes of R Programs

***Sascha Plazar**, Peter Marwedel, Jörg Rahnenführer*
TU-Dortmund

The GNU R language is very popular in the domain of statistics. Its functional character supports the rapid development of statistical algorithms and analyses. Statisticians around the world profit from the immense R package archive CRAN where researchers offer their algorithms in form of R programs for free usage. One of the disadvantages of R is that programs have to be evaluated and processed by a runtime interpreter. Such an interpretation requires a lot of time and delays the execution. Thus, a lot of computing power is wasted compared to imperative languages like ANSI C, which can be automatically optimized and translated to machine code by a sophisticated compiler. This abstract proposes an approach which exploits various optimizations and the workflow of toolchains for imperative languages to accelerate R programs. To this end, we are proposing a toolchain which is divided into four phases. Phase 1 applies source level optimizations on R. Phase 2 transforms such optimized R code and libraries to C code. In the next phase, the generated C in turn can be optimized, employing existing and newly developed optimization techniques. In the final phase, a standard compiler will translate the C code into machine code for a fast execution on a host machine. Our goal is to speed up R programs automatically on average by a factor of 50 or better. In a case study, we manually applied the optimizations common subexpression elimination (CSE) and dead code elimination (DCE) to R programs to evaluate their positive impact on the programs' execution times. CSE replaces multiple occurrences of the same expressions by a single variable holding the same value. Applied to strMCMC, a function for estimating graphical models with a Markov chain Monte Carlo approach, CSE was able to remove eight expressions which otherwise would have to be recomputed

several times. DCE removes code which would be executed on no account. By applying DCE to the same program, three if-statements inside the commonly used which() function could be removed which always evaluate to false. These optimizations reduced the overall execution time by 10 % and 5 %, respectively. In order to demonstrate the advantages of avoiding a time consuming interpretation of R programs to achieve high performance, we exemplarily translated pieces of R code into C. For this purpose, we evaluated the hot spot of the frequently used R package rda for Regularized Discriminant Analysis. By translating a single for loop of rda's apply() function and compiling it with the GCC compiler, we were able to speed up this function by a factor of 90. This led to a total reduction of 71 % concerning the overall runtime of the rda package. These excellent results attest that our envisioned toolchain will be highly effective for accelerating R programs. The results also show that a speedup by a factor of 50 is achievable by optimizing R programs and translating them into an imperative language in order to generate efficient machine code.

A Cooperative Game Theory Approach to Classification Analysis

Laura Pontiggia
University of the Sciences in Philadelphia

Traditional methods used for classification analysis, such as cluster analysis, principal components, and discriminant analysis, are difficult to use when the predictors are not nicely distributed, complex interactions and patterns exists in the data, or there is a large number of predictors. This talk presents a cooperative game theory approach to classification analysis, based on the use of the Shapley Value imputation. This approach to classification problems considers each variable as a player, and it considers each subset as a way of building coalitions among players to maximize their payoff. The Shapley Value can be used to evaluate the worth of the variables (players) that form a particular subset (coalitions).

The Risk-Return Tradeoff and Leverage Effect in a Stochastic Volatility-In-Mean Model

Jesper Christensen
CREATES and School of Economics and Management University of Aarhus

Petra Posedel
Faculty of Economics and Business Zagreb, University of Zagreb

The relation between risk and return is central to financial economics. Asset pricing theory predicts a positive risk-return tradeoff. Empirically, conditional first and second moments of asset returns are time-varying, and this must be accounted for when testing the risk-return relation. Empirical research also documents a strong leverage effect that potentially makes it more difficult to identify the risk-return relation. Thus, according to Fisher Black's explanation of the leverage effect, a price drop increases the debt-equity ratio and hence expected risk. The increase in risk would in turn increase expected returns in case of a positive risk-return relation. Depending on whether the empirical researcher associates the increase in risk with the initial price drop (negative return) or manages to link it to the higher subsequent (expected) returns, the apparent empirical risk-return relation may be of either sign. This highlights the identification issue, and may suggest why the empirical literature tends to show mixed results on the significance and sign of the risk-return relation. Evidently, further analysis should be carried out in a model that in addition to the risk-return relation explicitly accommodates a separate leverage effect.

Spatial Extremes: Application to Precipitation Data in North of Portugal

Dora Prata Gomes
FCT-UNL

Manuela Neves
ISA-UTL

Extremely rainy winters can drastically reduce agricultural production, increase the prices of the necessity goods and lead to hazardous health conditions. The interesting question is: how heavy can the heaviest precipitation level in 100 years be? Understanding and predicting the spatial rainfall, temporal variability and trends of extreme weather events is crucial to prevent catastrophes. Tools for statistical modeling of univariate extremes are well-developed. However, extending these tools to model spatial extreme data is an active area of environmental research. Spatial data sets are now easy to obtain. The most natural way for continuous space specification of extremes is provided by the theory of max-stable processes which can be seen as an extreme value analogy of Gaussian processes. Different characterizations of max-stable processes with unit Fréchet margins have been proposed in Smith (1991), Schlather (2002) and in more recent work of Schlather and others.

Estimation is performed by approximating the full likelihood, inadequate for max-stable processes, by the pairwise likelihood of Cox and Reid (2004). Here, after an idea of what has been done to overcoming these difficulties an application to annual maximum series of precipitation levels in the North of Portugal at different sites will be studied. Our goal is to try to understand how extreme precipitation is distributed over the North of Portugal. Exploratory analysis based in several techniques such as functional Boxplots for complex space-time data visualization is applied, Genton(2010). Estimate return levels is also done using several models for the spatial max-stable processes.

Profiles as Outputs in Supervised Data Analysis

Marco Reis
University of Coimbra

Images are a particular case of profiles, a term used to accommodate data structures that retain information about a given entity (arrays of data), indexed by time and/or space. For instance, grey images are examples of 2D profiles whereas multi-channel images fall in the category of 3D profiles. By extending image analysis to the analysis of profiles, one is able to generalize, in a more efficient way, the developments achieved to other non-related applications but falling under the same general abstract umbrella of "a profile".
There is currently a large body of knowledge accumulated in the Image Analysis and related fields (regarding 2D and 3D profiles), where image features are extracted in order to accomplish a given data analysis task. The goals can be as varied as the geometrical assessments of objects and/or their distribution, 1 measurements of displacements in bodies subject to environmental changes, 2 quality characterization and assessment of goods, 3–6 quantification of the presence of compounds 7 or objects, 8 etc.
This heterogeneity of goals founds also a parallel in the variety of approaches used to address them, and a plethora of methodologies are now available to perform image enhancement, segmentation, 9 texture characterization 10 and features extraction. 11 All these developments have in common the fact that image features are used as inputs or predictors in the analysis, being the key elements to infer objects or systems properties or behaviour.
However, there are also important applications in practice where the interest is centred on the prediction of image features, or, rephrasing in more general terms, where the goal is to predict the effect of a number of input variables on profiles, which play now the role of outputs in such supervised analysis tasks. Noting that very little attention has been devoted to this topic, we present and discuss in this communication several applications where this problem emerges, and address methods to handle them. The first application regards the control of 1D profiles of fibre orientation in the paper and paperboard industry, in order to control the existence of curl in the final product. The second is relative to the prediction of colour in candles, after treatment with decolorizing agents at different conditions.

Seasonal Fractional Processes with Heteroscedastic Innovations

Valderio Reisen
UFES, Brazil

This paper discusses the possibility that the daily average Particulate Matter (PM_10) concentration is a seasonal fractionally integrated process with time-dependent variance (volatility). In this context, one convenient extension is to consider the SARFIMA model with GARCH type innovations. The model proposed here is a SARFIMA process with two fractional parameters. The model is theoretically justified and its usefulness is corroborated with the application to PM_10 contaminant. The model shows it is able to capture the dynamics in the series and to improve the out-of-sample forecast of PM_10 pollutant. The results obtained by a SARFIMA model are compared to the results obtained by a SARMA model, both using heteroscedastic errors.

Analysis and Forecasting of the NYSE Composite Index

Paulo Rodrigues
CMA – Research Center for Mathematics and Applications, Nova University of Lisbon, Portugal

The New York Stock Exchange (NYSE) composite index is a market capitalization weighted index covering price movements of common stock (ordinary shares) of some 2300 firms listed on NYSE. Base index value is $50, and all shares are related to an aggregate market capitalization value as of December 31, 1965 adjusted for capitalization changes. The financial crisis of 2007–2010, which resulted in the collapse of large financial institutions and downturns in stock markets around the world, had given a higher level of visibility to NYSE and other stock markets all around the world. Singular Spectrum Analysis (SSA) arises as a natural generalization of principal components methods, suited for time series analysis. Since its introduction, SSA has been successfully applied in a broad number of fields of research such as climatology, geophysics and meteorology. Applications to econometrics and finance are still at their early days being the method unknown over those fields of research. SSA leaves two decisions to the analyst, namely the choices of the window length l, and the number of leading eigentriples for conducting the reconstruction m. In practice the window length is usually chosen to be close to half of the length of the series and divisible by the period of expected periodicity. For the choice of the number of leading eigentriples, SSA shares the same problem of model identification as PCA and many proposed methods for PCA can be applied in SSA. In this paper we examine the NYSE composite index. First a general analysis of the index is made and then the SSA is used to effectively disentangle the original time series in components of trend, cycle and seasonality, for suitable values of l and m. Then, an algorithm is presented

in order to produce out-of-sample forecasts for the NYSE. The data set used consists of the NYSE composite index, ranging from December, 1965 to December, 2009 (529 monthly observations).

Transforming the Data Deluge into Data-Driven Insights

Robert Rodriguez
SAS Institute

This presentation will describe emerging opportunities for statisticians in business environments where customer value and competitive strategy are driven by solving statistical problems. These problems are characterized by large volumes of data and often require software solutions that integrate data management, analytical computing, and delivery of results. I will survey the opportunities from my perspective as a developer of statistical software, using examples from various business sectors, including retail and financial services. The presentation will highlight approaches to software development and high-performance statistical computing that play a growing role in business analytics, along with technical skills and professional training that can equip statisticians to contribute more effectively.

Set and Boundary Estimation Under Shape Restrictions

Alberto Rodriguez Casal
Universidad de Santiago de Compostela

The problem of estimating a set S from a random sample of points arises in connection with some applications in statistical quality control, clustering, image analysis and statistical learning. This problem can be established in a more formal way as the problem of estimating the support of an absolutely continuous probability measure P from n independent observations drawn from P. So, the goal here is to estimate a set, not a parameter or a function. Assuming that the set of interest belongs to some family of sets can be useful in order to find an efficient estimation method. The case where S is assumed to be convex has received a special attention. If we assume that S is the support of the distribution which generates the sample points, there is a quite obvious estimator: the convex hull of the sample, that is, the smallest convex set which contains the sample. However, if S is not convex, the convex hull of the sample can be a bad choice. In this talk support estimation under the assumption that the set satisfies a much more flexible shape restriction, which is named alpha-convexity, will be presented. It will be showed that the new estimator can achieve, in a much more general setting, the same convergence rate as the convex hull. Support estimation is also connected to another interesting problem: the estimation of certain geometric characteristics of the

set such as the volume or the surface area. It seems natural to think that the volume or the surface area of a good support estimator should provide good approximations of these geometrical quantities. Here we analyze the problem of boundary length estimation when it is assumed that the set is alpha-convex.

Liquidity, Risk and Return: Specifying an Objective Function for the Management of Foreign Reserves

Yuliya Romanyuk
Bank of Canada

An objective function is a key component of a strategic portfolio management model used to determine optimal allocations of assets and possibly their associated liabilities over some investment horizon. This paper discusses perspectives and investment philosophies for the management of foreign reserves, and investigates how to translate the three common policy objectives for reserves (liquidity, safety, and return) into objective functions for strategic reserve management. The paper identifies stochastic programming as a practically advantageous modelling framework to capture the objectives of foreign reserves management, and concludes with an illustration of a strategic reserve management model that trades off expected net returns with costs and liquidity issues related to a potential liquidation of a portion of the portfolio.

Non-Metric Partial Least Squares Methods

Giorgio Russolillo
CNAM, Paris

In 1966, Herman Wold proposed the estimation of principal components and related models by means of a Non-linear Iterative PArtial Least Squares (NIPALS) procedure. At the beginning, NIPALS was "just" an algorithm able to implement a PCA through a suite of simple OLS regression so as to avoid the inversion of the correlation matrix. This property allows handling missing data without a priori imputation. Successively, NIPALS algorithm was developed in order to implement a variety of Multidimensional data analysis methods. Moreover, two new methods, called in literature PLS Regression (PLS-R) and PLS Path Modeling (PLS-PM), instead, were devised to perform respectively regularized regression and Structural Equation Models in a soft modeling framework. Nowadays, among the open issues that currently represent the most important and promising research challenges in PLS-PM and PLS-R, there is the specific treatment of non-metric (nominal and ordinal) variables and specific treatment of non-linearity. This work shows how NIPALS-type algorithms (commonly known as PLS algorithms), properly adjusted, can work as Optimal Scaling algorithms. This new feature of PLS allows us to devise

three new PLS methods: Non-Metric NIPALS, Non-Metric PLS Regression and Non-Metric PLS Path Modeling. The Non-Metric PLS (NM-PLS) methods can be used with different aims: – To analyze at the same time variables observed on different measurement scales; – To investigate non linearity; – To discard the hard assumption of linearity in favor of a milder assumption of monotonicity. Non-Metric NIPALS, Non-Metric PLS Regression and Non-Metric PLS Path Modeling generalize the respective standard methods in order to handle variables observed on a variety of measurement scales, as well as to cope with non linearity problems. Three new algorithms are proposed to implement NM-PLS methods. All these algorithms provide at the same time specific PLS model parameters as well as scaling values for variables to be scaled. Scaling values provided by these algorithms have been proved to be optimal, in the sense that they optimize the same criterion of the model in which they are involved. Moreover, they are suitable, since they respect the constraints depending on which properties of the original measurement scale we want to preserve.

Practical Limitations When Using Excel with Large Data Sets

William Rybolt
Babson College

As large data sets become more common, those who have used Excel to edit and analyze small data sets need to become aware of practical considerations which will impact their future usage. The physical limits of Excel workbooks have been well documented. ("So what is new in Excel 2007?" http: //visio.mvps.org/Excel_2007.htm; "The official blog of the Microsoft Excel product team" http://blogs.msdn.com/excel/default.aspx Microsoft Excel 2010) These limits include file size, workbook size, and the number of worksheet rows and columns. The limits are determined by the release of Excel (2003, 2007, and 2010) being used. Practical limits include not only the physical limits, but also the amount of time to complete common operations on large data sets. When dealing with small data sets, the time for Excel to complete operations is easily ignored. As the data sets become larger, the amount of time to complete computations becomes more noticeable. This presentation examines the effect of seven factors on the time to complete worksheet operations. These seven factors are worksheet size, Excel release, operating system, and four PC hardware configuration factors (available memory, processor speed, number of processors, and number of bits). The naïve assumption is that the time to complete a task is a linear function of the factors above. This presentation summarizes the results of running simple operations such as Copy, Paste, Delete, Find, Replace, Sort, Average, and Standard Deviation while varying the factors above. The major emphasis is on how the size of the data set affects the time to complete tasks. The focus is on the conditions under which the scaling becomes nonlinear. The goal is, first, to provide simple guidelines to facilitate decisions on how to allocate resources among the seven factors mentioned above when using Excel with large datasets. The second goal is make users aware that as the size of the datasets

become larger, the time to complete the computations becomes excessive. This forces the user to seek more powerful computing resources, migrate to other software, or modify the datasets to make the computations more manageable.

Classification of Pinot Noir Wines from Casablanca Valley, Chile: a Chemometrical Approach for Designation of Origins

***Jorge Saavedra Torrico**, Waldo Quiroz, Manuel Bravo*
Pontificia Universidad Católica de Valparaiso

Gyeorgy Lukaicsy
Corvinus University of Budapest

José Miguel Carot Carot
Universidad Politécnica de Valencia

Today consumers have a greater awareness and demand ever more information about the foods they consume, especially with regard to its composition, nutritional properties and origin. In this sense, the European Union has developed since 2002 a whole series of regulations regarding to food safety and traceability, which considers the rights of consumers and producers in terms of food adulteration, fraud or deceptive practices.
One of the Chilean most important export products is wine. Chilean wine industry has presented a dramatic growth during the last two decades. The export volume increased 12-fold from 1990 to 2008 reaching 6,2 million Hl with the value of $US 1,3 billion, leading that today Chile as the fifth wine exporter of the World.
In this context, Casablanca Valley subregion is one of the youngest wine producing areas of Chile. This sub region is in a special point of interest of national and international wine producers and consumers because of its unique climate affected by the Pacific Ocean. Furthermore, Casablanca Valley is nominated as the first cool climate wine region of Chile producing mainly Chardonnay, Sauvignon blanc and Pinot noir.
In the present work the classification of wines of Casablanca Valley and the comparison of different chemometrical approaches were evaluated for the first time. The characterization of Pinot noir wines was achieved by the determination of UV-VIS spectra, anthocyanin profile and elemental content. Multivariate tools evaluated in this study were Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Soft Independent Modelling of Class Analogy (SIMCA) and Partial Least Square Discriminant Analysis (PLS-DA). No considerable differences were detected among the applied statistical procedures.
The results suggest at least three well separated groups: two subzones in Casablanca Valley and the third clearly corresponds to the Loncomilla Valley (known outlier). This information could be so many important for future Regulatory Councils conformation and Designation of Origin like a formal method to authentication and traceability of wine. Further investigations are

recommended, concerning environmental factors, namely the climate and the soil properties.

Three-Way PLS Approach to Analyze and Monitoring a Soft-Drinks Bottling Process

***Jorge Saavedra Torrico**, Jaime Gomez*
Pontificia Universidad Católica de Valparaiso

Luis Puente D.
Tecnología Química Universidad de Chile

José Miguel Carot Carot
Universidad Politécnica de Valencia

The Statistical Process Control (SPC) has been extensive and successful utilized in the quality improvement process in manufacturing and service industries, including pharmaceutical, foodstuffs, polymer, semiconductor and biochemical industries. An important subgroup of SPC is the multivariate statistical process control (MSPC), specially the approach based in the work developed by Nomikos and MacGregor: multiway principal component analysis (MPCA) and multiway partial least squares (MPLS). In the last years, many attentions has been placed to process transitions, startups and restarts, since these abnormalities often arrives to the loss of production time, undesirable and not detectable variability increasing, production of off-grade materials and, to inconsistent reproducibility of product grades. The present research deal with the modeling an analysis of a continuous soft-drinks bottling process by a multivariate approach based on the 3-Way PLS. For this purpose the process behavior was studied, related to a set of off-processing, transitions grades and/or abnormal star-ups, detected in four month process period. Since the process had been controlled only with a univariate SPC approach, the research aim was to investigate and explore the magnitude of variability increasing and the relation with non yet detected process shift and out of control process. A total of 1938 cases were investigated, measuring 8 process variables. Finally, the process was modeled, filtering the high grade of autocorrelation and cross-correlation within the studied variables. Thereby, the responsible variables for that behavior were detected; by way of detect the disturbance grade over the normal variability pattern in the final product.

Quantile Based Reliability Analysis

***Paduthol Sankaran**, Unnikrishnan Nair*
Cochin University of Science & Technology

Quantile functions are equivalent alternatives to distribution functions in modelling and analysis of statistical data. The present paper discusses the role

of quantile functions in reliability studies. We present the hazard, mean residual, variance residual and percentile residual quantile functions, their mutual relationships and expressions for the quantile functions in terms of these functions. Further some theoretical results relating to the Hankin and Lee (2006) lambda distribution are discussed. We develop nonparametric estimator of hazard quantile function.

Semiparametric Approaches for Modelling Air Pollution Processes

Olha Bodnar, ***Wolfgang Schmid***
European University

In the last years spatio-temporal processes have been intensively discussed in literature. They have turned out to be extremely useful for modelling environmental processes. Nowadays we find a variety of applications to different types of processes like, e.g., atmospheric pollutant concentrations, precipitation fields and surface winds. Fassò and Cameletti (2009) introduced a very general spatio-temporal process for modeling the concentration of PM10. It can be presented as a state-space model. In this model it is assumed that there is a linear relationship between the PM10 concentration and the geographical and meteorological covariates. In Bodnar and Schmid (2010) a similar model is considered and it is used to calculate the locally weighted scatterplot smoothing (LOESS) kriging predictor. Their approach is based on the idea to find a balance between a local and a more global method. This means that necessarily not all available measurement stations are used to interpolate the process at an arbitrary position. Criteria for finding an optimal set of included stations are discussed in the paper. In the present paper we introduce a semiparametric spatio-temporal model. It can be considered as an extension of the models discussed in Fassò and Cameletti (2009) and Bodnar and Schmid (2010). This model is more flexible because it is not assumed that there is a linear relationship between the concentration and the geographical and meteorological covariates. It has further desirable properties. For instance, it is guaranteed that the concentration process is always positive and the present concentration depends directly on previous values. The price of the generalization is a process which is more difficult to handle. Because both the state and the space equations are non-linear, the standard Kalman filter cannot be used for estimating the model parameters. For that reason we make use of the unscented Kalman Filter suggested by Julier et al. (2000). This method is combined with the EM algorithm and a Nadaraya-Watson type estimator for the nonparametric component to get estimators of the process parameters. Furthermore we consider the problem of interpolating the process at arbitrary locations. This is done in a similar way as described in Bodnar and Schmid (2010). Our results are applied to model the PM10 concentration of the Berlin-Brandenburg region in Germany.

Robust Xbar Control Charts

Marit Schoonhoven, Ronald Does
IBIS UvA

Muhammad Riaz
Quaid-i-Azam University

This paper concerns the design of the Xbar control chart when the in-control process mean and process standard deviation are unknown and therefore have to be estimated in Phase I. We consider the situation that the Phase I data are uncontaminated and various situations in which the Phase I data are contaminated. A comparison of the different Phase I estimators of the process mean and the process standard deviation presented in literature is made, and their effect on control chart performance is studied. The Phase I estimators of the mean that are considered are the mean, the median, the 25% trimmed mean of means (Rocke, 1989), the Trimean (Tukey, 1977) the mean rank (Jones-Farmer et al., 2009) and the Hodges-Lehmann Estimator (Alloway and Raghavachari, 1991). The Phase I estimators of the process standard deviation are the average absolute deviation from the median, the median absolute deviation from the median and the robust estimator of Tatum (1997). Furthermore, we derive correction factors for the control limits in order to obtain accurate control limits when limited data are available for parameter estimation.

The Study of Extended Multivariate Time Series Through Independent Component Analysis

Fernando Sebastiao
Department of Mathematics, School of Technology and Management, Polytechnic Institute of Leiria, and CM-UTAD
Irene Oliveira
Dpt of Mathematics, University of Trás-os-Montes and Alto Douro, and CM-UTAD

In multivariate data analysis, the study of time series has been used in various techniques, including the Multi-channel Singular Spectrum Analysis (MSSA). This technique is Principal Component Analysis (PCA) (Jolliffe, 2002) of extended matrix of initial lagged series, hence also designated in the climatological context as Extended Empirical Orthogonal Function (EEOF) Analysis (von Storch and Zwiers, 1999). We present Independent Component Analysis (ICA) (Hyvärinen et al., 2001) to study the extended matrix of time series, as an alternative to the method MSSA for climate data. Although ICA is a statistical computational technique widely used in several areas such as image processing, biomedical signals, telecommunications, economic field, but it is not yet widely applied in climate research. As it is known that the atmospheric system is very complex, and ICA can play an important role in cases where the classical PCA does not extract all the essential information underlying a data set in space and time, since ICA involves higher order statistics while PCA only uses the second order statistics conditioned to non

correlated Principal Components (PCs). Another aspect to consider in these techniques is the maximization of the variability retained in the first components, although in PCA the PCs are ranked in descending order of variance, in ICA the Independent Components (ICs) are not necessarily sorted out in a specific order. In the literature there are different methods of ordering ICs, since performing ICA through a large number of existing algorithms provides the extraction of ICs and these are not sorted out according to any criterion. Therefore we present some existing methods of ordering ICs and we propose a new one. We present an example of time series for meteorological data and some comparative results between the techniques under study, and we analyse the quality of the reconstructions of the original data through the sum of square errors.

Tails of Correlation Mixtures of Elliptical Copulas

Hans Manner
Maastricht University

Johan Segers
Université Catholique de Louvain

Correlation mixtures of elliptical copulas arise when the correlation parameter is driven itself by a latent random process. For such copulas, both penultimate and asymptotic tail dependence are much larger than for ordinary elliptical copulas with the same unconditional correlation. Furthermore, for Gaussian and Student t-copulas, tail dependence at sub-asymptotic levels is generally larger than in the limit, which can have serious consequences for estimation and evaluation of extreme risk. Finally, although correlation mixtures of Gaussian copulas inherit the property of asymptotic independence, at the same time they fall in the newly defined category of near asymptotic dependence. The consequences of these findings for modeling are assessed by means of a simulation study and a case study involving financial time series. Paper available at http://arxiv.org/abs/0912.3516.

The Equity of Financial Service Distribution

Nicoleta Serban*, Sungil Kim*
Georgia Institute of Technology

Research in service distribution equity has emerged as economic and social equity advocates recognized that where people live influences their opportunities for economic development, access to quality healthcare, and political participation. In this research presentation, service distribution equity is concerned with where and when services have been and are accessed by

different population groups. Leveraging new statistical methods for modeling spatial-temporal data with a hierarchical structure, this paper estimates demographic association patterns to service accessibility varying over a large geographic area (Georgia) and over a period of 13 years. The focus of this study is on financial services but it generally applies to any other service operation. The underlying modeling is a (hierarchical) varying coefficient model where the coefficients vary both in time and space. The model includes both separable space-time varying coefficients and space-time interaction terms. We introduce an inference procedure for assessing the shape of the varying regression coefficients using confidence bands.

An Introduction to Morphological Industrial Control via Case Studies

Jean Serra
ESIEE, University of Paris-Est, France

Computer vision is used for several types of industrial controls. The relevance of such an approach involves a few criteria. Five of them are investigated here, namely the mode of representation, the possible presence of individual objects, translation invariance, measurements, and speed. The above criteria are introduced via examples of typical situations, where one of these criteria is emphasised more than the others, namely individual analysis and rice quality, robust metrology and antibiograms, translation invariance and motor inspection, and speed and envelope reading. The basic concepts of mathematical morphology are introduced by means of the previous situations. The two notions of filtering and of segmentation are illustrated by openings, granulometries and by watershed techniques respectively.

Historical Overview of Image Analysis and Mathematical Morphology

Jean Serra
ESIEE, University of Paris-Est, France

In twentieth century, the evolution of image analysis followed the rhythm of the decades. It appeared at the beginning of the sixties, in association with the first images for land observation (USA) and with optical microscopy (Europe). The latter experienced a boom in the seventies with the introduction of video sensors, mainly in the two domains of biology (cytology and histology) and material sciences where, for the first time the structures under study were quantified in an automatic way. The next change came at the beginning of the eighties, as a consequence of the oil crisis. Automatic control of the industrial processes became an economical issue. In parallel but independently, new

sensor technologies such as tomography or ultra-sound were introduced in medicine, that generated digital images to be processed. Another change commenced at the end of the eighties with the need for image and video compression, and with the boom of Internet a few years later. Finally, in our current decade, the emphasis is on maps. The resolution of the satellites for earth observation has been drastically reduced by a factor of 10 (from 5 m to 50 cm), which induces new problems and new products (e.g. Google Earth). All these themes, although they appeared successively, are still evolving with varying degrees of rapidity depending on the domain, and on the development of methods such as mathematical morphology or wavelets.

Visual Detection of Change Points

***Sackmone Sirisack**, Anders Grimvall*
Division of Statistics, Department of Computer Science, Linköping University

The rapid growth of systems for automatic data collection has increased the need for algorithms that can efficiently reveal important features of large datasets. For example, it is often of great interest to examine the presence of abrupt changes in long bivariate or multivariate time series of data. A number of numerical algorithms and statistical estimators are widely used to detect sudden changes in the mean or other parameters of univariate probability distributions. However, changes in twodimensional or multidimensional distributions of random variables are often best detected by visual inspection. Here, we present a set of user-friendly animation tools that can be utilized to detect temporal changes in the bivariate distribution of one or more pairs of random variables. The entire dataset or specific features of this dataset are used as a static background, while a sequence of user-selected subsets is highlighted. In particular, we show how animated bubble-charts can reveal temporal changes in the frequency of outliers or the conditional expectation or variance of one variable given another. Other tools, based on animated bar charts, enable visual detection of changes over time in the synchrony of increases and decreases in multiple time series of data. Our visualization tools have been implemented as VBA macros for Excel, and advantages and disadvantages of different techniques for animating graphs in this software will be discussed. In particular, it will be shown how datasets comprising tens of thousands of cases can be handled. Time series of daily climate and share price records will be used to illustrate the power of visual tools for change-point detection. Our analysis of climate data will focus on sudden shifts in the mean and seasonal patterns, whereas our study of share price data will focus on temporal changes in the synchrony of share prices representing the same or different branches.

Customer Satisfaction and Loyalty in the Polish Banking Sector. Empirical Studies Perspective from Years 2007–2009.

Lukasz Skowron, Stanislaw Skowron
Technical University of Lublin

In the current age of dynamic market changes, all companies, regardless of their size and sectors in which they are operating, to ensure survival on the market and further development, are forced to efficiently manage the relations on the company-customer plane. Customer satisfaction and loyalty models, can be perceived as effective tools thanks to which companies can regularly and objectively measure the effectiveness of their actions with regard to the specified group of clients, compare their results with the results of the competition and choose a strategy of improving relationships with customers, which subsequently will contribute to better financial performance of given companies. Therefore, it can be said that such models are becoming the foundation of the customer relation management process of any modern organizations with the special emphasis on the banking sector. In the article authors will concentrate on analytical issues concerning satisfaction and loyalty of Polish customers in the banking sector. The research of Polish banking sector was primary, carried out among the group of 1884 people – clients of banks – during three years time period – 2007–2009. For the methodology procedure authors decided to employed a model based on the traditional European EPSI (the main differences between polish model and the basic EPSI are the number of questions which describes each of the areas of the model and implementation of the three quality areas instead of two used in original EPSI framework). The process of estimating particular structural relations occurring between particular areas of the applied model was carried out by means of the SmartPLS program. Data collected during the research allowed authors to make the following analyses: statistical frequencies, path coefficients, level of adjustment of the model and the particular measures of the total influence of analysed areas (both direct and indirect – through other dependent areas). Obtained empirical results shows that the financial crises of the year 2008 has changed the character of the process of building customer satisfaction and loyalty in the Polish banking sector by strengthening the influence of the "image" area. Additionally it is evident from the detailed analysis that the perception among Polish customers to an individual bank now varies much more than before. This means that the spread between satisfied and dissatisfied customers was higher in 2009 than during previous years (2007 and 2008).

Explicit Martingale Estimating Functions for Diffusions With Jumps

Michael Sorensen
University of Copenhagen

A diffusion with jumps is a stochastic process given by a stochastic differential equation driven not only by a Wiener process, but also by another stochastic mechanism that causes the process to make jumps. This other mechanism can be a Lévy process, or more generally, a random measure on a suitable space. Diffusions with jumps are often use as models for financial time series. When the data are continuous time observations, likelihood inference for diffusions with jumps has long been well understood; see e.g. Sorensen (1991). However, continuous time observations are not available in practice, and for discrete time observations the likelihood function is not explicitly known and usually extremely difficult to calculate numerically. Therefore alternatives like estimating functions are even more useful for jump diffusions than for classical diffusions. We present a highly flexible class of diffusions with jumps for which explicit optimal martingale estimating functions of the type introduced by Kessler and Sorensen (1999) are available. These are based on eigenfunctions of the generator of the diffusion. The class of Pearson diffusions, investigated in Forman and Sorensen (2008), has the property that the generator maps polynomials into polynomials. Therfore it is easy to find polynomial eigenfunctions. Here we generalize these ideas and consider a class of diffusions with jumps for which the generator has the same property using ideas from Zhou (2003). The generator of a diffusion with jumps is considerably more complicated that that for a classical diffusion: It is a differential-integral operator. However, it turns out that a simple condition on the compensator of the jump measure is enough to ensure that explicit optimal martingale estimating functions can be found. We illustrate the general theory by concrete examples. The talk is based on joint work with Mathias Schmidt.

Spatio-temporal Surveillance of Road Crashes

Ross Sparks*, Ellis Patrick*
CSIRO Australia

Spatio-temporal surveillance methods for detecting outbreaks of disease are fairly common in the literature with the SCAN statistic setting the benchmark. If the shape and size of the outbreaks are known in advance, then the SCAN statistic can be trained to efficiently detect these, however this is seldom true. Therefore we want to devise plans that are efficient at detecting a number of outbreaks that vary in size and shape. These outbreaks could be spatially dispersed or very local in nature, and methods ideally should have robust properties in situations. The approach is based on dividing the entire geographical region under surveillance into a lattice based on quantiles, and

then applying CUSUM and moving average methods. These approaches are compared using simulation, and the results of these will be reported. The more impressive surveillance methods based on the outcomes of these simulations will be applied to a road crashes application.

Exact Testing for Small Sampled Censored and Missing Data

Milan Stehlik
Johannes Kepler University Linz

In reliability engineering, the inference problem for the complete samples and large data sets are commonly rare events. Typically, missing data are present and censoring has been applied. Moreover, samples are frequently small because of many reasons (e.g. expensive observations or rare event structure or the failure process). During the talk we will discuss recent results on the exact likelihood ratio tests of scale and homogeneity hypotheses when samples are from exponential, Erlang, gamma, Weibull and generalized gamma distributions. The asymptotical tests are typically oversized and thus inappropriate for small samples. We will focus on the reliability prediction when some data is missing or is censored. The reliability prediction when some data is missing plays a major role in many reliability programs (e.g. for a variety of reasons over 90 % of the data in the Reliability Analysis Center does not have the individual failure times recorded). Testing for Type I and progressively Type II censored small samples data from exponential distribution will be provided (see [2]). We will provide also recent results for exact testing with missing data (see [4,9]). The competing risk model, often modeled by the mixture, will be also mentioned and difference between the upper and lower contamination for the two component mixture alternative will be explained. The real data examples will illustrate the topics discussed. Gamma distribution parameter estimation for field reliability data with missing failure times.

Risk Adjusted Survival Time Monitoring

Stefan Steiner
University of Waterloo

Mark Jones
CRISP

Monitoring medical outcomes is desirable to help quickly detect performance changes. Previous applications have focused mostly on binary outcomes, such as 30-day mortality after surgery. However, in many applications survival time data are routinely collected. In this article we propose an updating exponentially weighted moving average (EWMA) control chart to monitor risk-

adjusted survival times. The updating EWMA (uEWMA) operates in continuous time so scores for each patient always reflect the most up-to-date information. The uEWMA can be implemented based on a variety of survival time models and can be set up to provide an ongoing estimate of a clinically interpretable average patient score. The efficiency of the uEWMA is shown to compare favorably to competing methods.

Parameter Estimation for Fisher-Snedecor Diffusion

Nenad Šuvak
Department of Mathematics, University of Osijek

Ergodic diffusion with an invariant Fisher-Snedecor distribution is studied. In particular, the problem of parameter estimation is treated and the quality of estimators is illustrated by the simulation results. Classical approach to the asymptotic analysis of estimators is implied by some properties of the Fisher-Snedecor diffusion, such as the ergodicity, stationarity and exponentially decaying alpha-mixing property. The unknown parameters which are treated here are the autocorrelation parameter and parameters which coincide with the shape parameters of the invariant Fisher-Snedecor distribution. Estimation procedure for all parameters is based on the discrete observations form the process of interest. Estimation of the autocorrelation parameter and estimation of shape parameters of invariant distribution are treated separately. In particular, the autocorrelation parameter is estimated under the assumption that the shape parameters of the invariant distribution are known. The corresponding estimator is determined by the generalized method of moments based on the consistent estimator for the autocorrelation function – Pearson's sample autocorrelation function. Furthermore, the shape parameters of the invariant distribution are estimated under the assumption that the autocorrelation parameter is known. The bivariate estimator of the shape parameters is determined by the method of moments based on the empirical counterparts of the first and the second moment of the invariant distribution. Ergodicity, stationarity and exponentially decaying alpha-mixing property of the Fisher-Snedecor diffusion are crucial for the analysis of asymptotic properties of these estimators. Consistency of estimators is implied by the ergodic theorem for stationary sequences and the continuous mapping theorem. Asymptotic normality of the bivariate estimator of shape parameters of invariant distribution follows from the central limit theorem for alpha-mixing sequences and the standard delta method. Furthermore, the explicit form of the limiting covariance matrix of the bivariate estimator of shape parameters is calculated according to the method based on the closed form expression for the spectral representation of transition density and important properties of the finite system of orthogonal polynomial eigenfunctions of the infinitesimal generator of Fisher-Snedecor diffusion. The estimation procedure based on the martingale estimation equations due to Forman and Sorensen (2008) is briefly discussed. The discrete observations from the Fisher-Snedecor diffusion are

simulated in statistical software R and used for further analysis of the quality of presented estimators.

The Bayesian Theory of Games

Jimmy Teng
Academia Sinica

The current prevailing non cooperative games theory is without statistical decision theoretic foundation. This fundamental problem started with Nash equilibrium and pervades the whole current games theory, causing it to be plagued with issues of multiple equilibriums and insensible equilibriums. Consequently, there is need for myriad refinements. This paper presents the Bayesian theory of games. It reconstructs the basic structure and solution concept of non cooperative games theory by using Bayesian statistical decision theory. The fundamental solution concept is no longer Nash equilibrium but Bayesian rational prior equilibrium. Bayesian rational prior equilibrium requires agent to make rational statistical predictions and decisions, starting with first order non informative prior and keeps updating it with statistical decision theoretic and game theoretic reasoning until a convergence of conjectures is achieved. The Bayesian theory of games methodically studies the formation and updating of conjectures on strategies which are done only in an ad hoc manner in the current games theory, starting with the first order uninformative prior. The Bayesian rational prior equilibrium approach differs from the current games theory in the following ways: I. It analyzes a larger set of games, including noisy games, games with unstable equilibrium and games with double or multiple sided incomplete information games which are not analyzed or hardly analyzed under the current games theory. II. For the set of games analyzed by the current games theory, it functions as an equilibrium selection criterion and selects the most common sensible and statistically sound equilibrium among equilibriums and eliminates insensible and statistically unsound equilibrium. III. It differentiates between simultaneous move and imperfect information. In current games theory, imperfect information is treated as identical to simultaneous move. In contrast, the Bayesian theory of games treats sequential move with imperfect information as a special case of sequential move with observational noise term. When the variance of the noise term approaches its maximum such that the observation contains no informational value, there is imperfect information (with sequential move). IV. It treats games with complete and perfect information as special cases of games with incomplete information and noisy observation whereby the variance of the prior distribution function on type and the variance of the observation noise term tend to zero. Consequently, there is the issue of indeterminacy in statistical inference and decision making in these games whereby the equilibrium solution depends on which variances tends to zero first. It therefore identifies equilibriums in these games that have so far eluded the classical theory of games. The issue of indeterminacy also arises in games with games with imperfect information and maximum incomplete information.

The GARCH Structural Credit Risk Model

Enrique Ter Horst
Euromed Management

Samuel Malone
Universidad de los Andes

Abel Rodriguez
University of California at Santa Cruz

Recently documented evidence on the "credit spread puzzle\" points toward the explicit modeling of stochastic asset volatility in structural credit risk models as a necessary and potentially important innovation (Huang and Huang, 2003). To that end, we adapt the GARCH option pricing model of Heston and Nandi (2000) to the valuation of risky corporate debt in a Merton (1974) setting. In light of recent findings that maximum likelihood estimation methods may be superior to traditional calibration approaches (Ericson and Renneby, 2005), we develop a new Expectation Maximization (EM) algorithm for a GARCH structural credit risk model and we conduct a simulation study to sort out the relationship between model implementation method and model mis-specification. We find that maximum likelihood estimation of the Merton model is superior in the absence of model mis-specification, but that Merton's (1974) calibration approach can exhibit superior performance when asset volatility is stochastic. Finally, we apply the GARCH model to a cross section of US banks and financial institutions on two dates of interest: December 14th, 2007, and August 29th, 2008.

Clustering of Stations in Monitoring Networks

*Andrea Pastore, Roberto Pastres**, Stefano Tonellato***
Università Ca' Foscari Venezia

We will consider the general problem of site classification in monitoring networks, where measurements are collected repeatedly in time at a fix set of stations. An application to the water quality monitoring network in Venice Lagoon will be provided by combining functional data analysis and probabilistic clustering. In particular, we will consider the classification of monitoring stations in terms of some trophic variables.

Variable Selection in Partial Least Squares Methods: Overview and Recent Developments

Laura Trinchera
SUPELEC, Department of Signal Processing and Electronic Systems, Gif-sur-Yvette, France

Recent developments in technology enable collecting a large amount of data from various sources. Moreover, many real world applications require studying relations among several groups of variables. The analysis of landscape matrices, i.e. matrices having more columns (variables, p) than rows (observations, n), is a challenging task in several domains. Two different kinds of problems arise when dealing with high dimensional data sets characterized by landscape matrices. The first refers to computational and numerical problems. The second deals with the difficulty in assessing and understanding the results. Dimension reduction seems to be a solution to solve both problems. We should distinguish between feature selection and feature extraction. The first refers to variable selection, while feature extraction aims to transform the data from high-dimensional space to low-dimensional space. Partial Least Squares (PLS) methods are classical feature extraction tools that work in the case of high-dimensional data sets. Since PLS methods do not require matrices inversion or diagonalization, they allow us to solve computational problems. However, results interpretation is still a hard problem when facing with very high-dimensional data sets. Moreover, recently Chun & Keles (2010) showed that asymptotic consistency of PLS regression estimator for the univariate case does not hold with the very large p and small n paradigm. Nowadays interest is increasing in developing new PLS methods able to be, at the same time, a feature extraction tool and a feature selection method. The first attempt to perform variable selection in univariate PLS Regression framework was presented by Bastien et al. in 2005. More recently Le Cao et al. (2008) and Chun & Keles (2010) proposed two different approaches to include variable selection in PLS Regression, based on L1 penalization (Tibshirani, 1996). In our work, we will investigate all these approaches and discuss the pros and cons. Moreover, a new version of PLS Path Modeling algorithm including variable selection will be presented.

A "Reason for Unreason": Returns-Based Beliefs in Game Theory

Chander Velu, ***Sriya Iyer***, *Jonathan Gair*
University of Cambridge

A sizeable experimental literature shows that there is a dissonance between the theory and empirics of games in a strategic setting where defection is predicted to be the optimal outcome over cooperation. In particular, players in experimental settings appear to cooperate more than the theory would predict.

In this paper we provide a new method termed the "returns-based beliefs" approach of forming subjective probabilities that is based upon the expected returns of a particular strategy, in proportion to the total expected returns of all strategies. Our approach combines a decision analytic solution concept and Luce's (1959) probabilistic choice model. We show how our approach provides a closer description of empirical observations in both the Prisoner's Dilemma and the Traveler's Dilemma. We test the closeness of fit of our model using data from Selten and Chmura (2008) for constant sum 2X2 games and develop an extension of our model using the concept of second order beliefs.

Dominance and Innovation

Chander Velu**, **Sirya Iyer
University of Cambridge

Do dominant or less dominant firms innovate more? Theoretically it has been shown that within an asymmetric mixed strategy game of a patent race, the less dominant firm invests more than the dominant firm. But the empirical data on patent races is divided. In this paper, we argue that the decisions that concern strategic choice in innovation may be influenced by expected relative returns. Our approach, which we call the returns-based beliefs approach, is based upon subjective probabilities. It combines a decision analytic solution concept and Luce's (1959) probabilistic choice model. In particular, we show how the use of the returns-based beliefs approach provides support for the thesis that dominant firms invest more in R&D within an asymmetric mixed strategy game. Consequently, we argue that the returns-based beliefs approach is more in line with recent empirical studies of innovation. We also provide empirical evidence using UK R&D data across a range of industries from 2001–2006 that shows that firms' spending on R&D is related more to their own profitability than that of their competitors, which is consistent with the returns-based beliefs approach. We discuss the managerial implications of our theoretical approach and the empirical findings.

Model Selection in Mixture-Process Experiments

***Antonio Vieira**, Luiz Henrique Dal Bello*
PUC-Rio, Brazil

We present a model selection procedure for use in Mixture-Process Experiments. Certain combinations of restrictions on the proportions of the mixture components can result in a very constrained experimental region. This results in collinearity among the covariates of the model, which can make it difficult to fit the model using the traditional method based on the significance

of the coefficients. For this reason, for model selection, a methodology based on information criteria will be proposed. Any Scheffé model for mixtures can be combined with the model for the process variables. The full Mixture-Process model can be obtained by additive combination of the Scheffé mixture model and the model for the process variables. Alternative models have been suggested, with the introduction of "cross" terms in mixture models and process models, resulting in the full cross model. Initially, the full cross model is fitted using the traditional step-by-step method (stepwise, forward or backward). This first model fitted is called base model. Next, one can try to obtain a better model using information criteria, and taking into account equivalent terms and the base model terms. A case study is presented to exemplify this model selection procedure. Initially, a model (Model M1) was fitted by backward selection of variables. Another model (Model M2) was fitted using the modified AIC (Akaike information criterion). For each model selected, the proportions of the components and the levels of the process variables were determined that give the desired response and minimize the width of the prediction interval for the response. Model M2 was selected because it provided a better value of the information criterion and of other statistics (PRESS and MSE) as well.

Outlier Detection in Small Samples of Same-Quantity Ratios: A Simulation Study Based on Business Indicators of Healthcare Quality

Gaj Vidmar
University Rehabilitation Institute, Republic of Slovenia

Rok Blagus
Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana

Healthcare quality monitoring within the Ministry of Health in Slovenia includes over 100 business indicators of economy, efficiency and funding allocation that are annually analysed for over 20 hospitals. The essence of associated statistical analyses is identification of outliers, in which we aimed at a compromise between state-of-the-art and wide understandability. Our exploratory approach, which the present study aims to scrutinise, combines three types of methods: common outlier detection tests useful in small samples (Grubbs, Dean&Dixon, and Nalimov test), all based on normality assumption and hence applied conditionally upon results of normality tests; the Tukey boxplot rule; and control charts. The indicators are same-quantity ratios (thus bound between 0 and 1) that are not appropriately treated either as proportions or as fixed-denominator ratios - they are random-denominator ratios with highly correlated numerator and denominator. Funnel plots – a standard for monitoring hospital and/or physician performance indicators – are therefore not appropriate control charts for them since virtually all points would get labelled as outliers because of huge denominators (thousands of

square meters, millions of Euro) yielding uselessly narrow confidence intervals for the average proportion. Hence, we chose the Double Square Root (Shewart) Chart (DSQRTC) plotting square-root of numerator vs. square-root of difference between denominator and numerator, whereby we replaced the traditional control limits (based on the underlying assumption of binomial distribution like in funnel plots, therefore much too narrow) by newly defined ones, obtained using linear regression through the origin (since no costs can be incurred without income etc.) and estimating control limits using 95% confidence interval for prediction. We studied performance and agreement of the chosen methods through a large simulation study on realistic data. Two types of ratios were generated: the small ones belonged to the $[0, 0.2]$ interval and the large ones to $[0.5, 1]$. Samples of size 5, 10, 20, 25 and 30 were drawn from three distributions previously found to best fit the empirical data: Pert (bounded), Burr and 3-parameter-loglogistic (both non-negative). Zero, one or two simulated outliers were included in the samples. The outliers were generated by increasing the relevant parameter (mode, scale parameter and mean for Pert, Burr and 3PLGL, respectively) by 50%, 100%, 150% and 500% while holding other parameters (related to dispersion and shape) fixed. The simulation was performed with R code using rejection approach. Results showed that performance of the methods varies greatly across the conditions. As expected, small-sample tests become less usable as sample size increases. Among the formal tests, Dixon's performed the worse overall. The simple boxplot method performed the most variedly, but it was the only useful for tiny samples. Our variant of DSQRTC proved too conservative in tiny samples and too liberal under no-outliers condition with $N >= 20$ (both holding also for Nalimov test, and for boxplot for small ratios), but it appeared by far the most useful (though still far from perfect) to detect actual outliers with N getting larger.

Google Street View: Overview & Computational Challenges

Luc Vincent
Google, Inc.

Unveiled in May 2007, Google Street View is the result of a substantial research and engineering effort by a team of software engineers, mechanical engineers, UI designers, computer vision scientists, operations experts, and scores of others. The initial vision for Street View was provided by Google co-founder Larry Page, who personally collected street scene videos from his moving car in order to bootstrap research in this area. Turning this initial vision into a successful product required developing major new pieces of technology, including robust data collection platforms (vans, cars, tricycles, snowmobiles, etc.), systems for computing accurate pose from imperfect sensors, various software components to stitch, blend, color correct and warp collected imagery, a number of systems to address privacy issues, processing pipelines to ingest and process a Gargantuan flow of data on a daily basis, various front-ends, and a lot more. This presentation will give an overview and brief history

of the Street View project, and highlight some of the unique algorithmic, computational and scale-related challenges the Street View R&D team is addressing.

A Bayes Approach to Goodness-Of-Fit Tests for Lifetime Models

Petr Volf
Institute of Information Theory and Automationof the ASCR, Prague, Czech Republic

In the framework of intensity models for lifetime data, the residual process (martingale residuals) is defined as a difference between estimated cumulated intensity and observed counting process. Hence, residual process is constructed from observed data, its properties depend on properties of estimator of cumulated hazard rate. While in a case without regression, as well as in Aalen's additive regression model, residual processes are the martingales, in some other cases (Cox's model, AFT model) behaviour of estimates, and therefore of residuals, is more complicated. In such instances, Bayes approach can offer a reasonable alternative. To demonstrate it, we shall first present a Bayes procedure of estimation in semi-parametric intensity models. Further, Bayes construction of residual processes and model goodness-of-fit testing will be proposed. Bayes analysis is often connected with MCMC methods. They are used for obtaining approximate representation of posterior distribution. In cases considered here we deal with nonparametric baseline hazard rate. In Bayes solution, its representation will be made from piecewise-constant functions (or from splines or from other functional bases). Alternatively, baseline distribution can be obtained as a (Dirichlet) mixture of Gauss distributions. Once a posterior sample of hazard rate (i.e. representation of its aposteriori distribution obtained by MCMC procedure) is available, we can construct a sample of cumulated intensities and corresponding residuals. Their use for assessing a model fit will be illustrated on examples with Cox's and AFT regression models.

Income Distributions in Czech Republic

Michal Vrabec, Luboš Marek
University of Economics Prague, Czech Republic

This article describes the characteristics income distribution in Czech Republic over the years 1995–2009 (some 2000–2009. We have data over the years 1995–2009 for all characteristics, but the region organization was changed in 2000 so that the data are not comparable. We know the basic statistics (average, median and selected quintiles) and some trimmed means. We

compare these statistics one to other (so, we compare the single regions), and the region statistics against the overall Czech Republic statistics. A general approach to distribution of earning modeling under curent heterogenity conditions don't permit to fit by some chosen distribution function or probably density function. All of the component distributions of this mixture model correspond to an employee's group with greater homogeneity of earnings.

Analyzing Computer Experiments: What Matters

William Welch
University of British Columbia

We compare strategies for the analysis of an initial computer experiment. A prior treating the real-valued output as a realization of a Gaussian stochastic process (GaSP) is fairly common now, but the user is faced with some operational decisions. First, the mean of the GaSP may be just a constant or could be a regression model in the input variables. Secondly, there are several possible families for the (assumed stationary) covariance function of the process. We take an evidence-based approach to evaluating such modelling options. While the main thrust of the talk is analysis, it turns out that design of an experiment has some important and perhaps surprising implications. Our findings and recommendations will be based on results from simple illustrative functions, real codes, and simulations. Joint work with Jason Loeppky (University of British Columbia, Okanagan) and Jerome Sacks (National Institute of Statistical Sciences)

Statisticians meet Optimizers: Visualizing Sampling Variability in Plots and Finding Outliers

Rajiv Menjoge, ***Roy Welsch***
MIT

Tri-Dzung Nguyen
University of Illinois

Some recent results show the importance of taking advantage of modern optimization ideas and algorithms to advance statistical theory and practice. Our first example proposes a general method for providing a description of the sampling variability of a plot of data. The motivation behind this development is that a single plot of a sample of data without a description of its sampling variability can be uninformative and misleading in the same way that a sample mean without a confidence interval can be. The method works by using bootstrap methods to generate a large number of plots that might have arisen from different samples from the population, and then conveying the

information given in the collection of plots by methodically selecting a few representative plots. The distance between two scatter plots, in general, is the solution to the "Assignment Problem" that assigns each point in one scatter plot to a point in the other, such that the assignments have minimal cost. The "Assignment Problem" is a network flow problem which can be solved efficiently by, for instance, the Hungarian Algorithm. The collection of plots is ordered so that each plot is as close as possible to neighboring plots using a traveling salesman type algorithm. The second example addresses outliers in regression problems. Least trimmed squares (LTS) regression has desirable properties and forms the basic building block for many robust regression algorithms, but is very computationally expensive due to its combinatorial nature since it tries to minimize the sum of the smallest p squared residuals. LTS is equivalent to a concave minimization problem under a simple linear constraint set. We propose maximum trimmed squares (MTS), an "almost complementary" problem that maximizes the sum of the q smallest squared residuals, in direct pursuit of the set of outliers rather than the set of clean points as in LTS. MTS can be formulated as a semi-definite programming problem that can be solved efficiently in polynomial time using interior point methods. In addition, under reasonable assumptions, MTS is guaranteed to identify outliers, no mater how extreme they are.

FASTAT: A Second Opinion Statistical Analyzer

Lee Wilkinson
University of Illinois

Statistical expert systems were the focus of intense research more than a decade ago, but commercial and open-source statistics packages have generally failed to exploit the results of that research. FASTAT is an attempt to combine an expert statistical system with a new graphical user interface (GUI) in order to provide non-expert users with a second opinion on data analyses they have performed with a conventional statistical package. The expert system includes extensive testing for anomalies, diagnostics of model residuals, and remedial recommendations. The GUI takes advantage of contemporary developments in interface design. This GUI – tested on both novice and expert users – has no keyboard input, user manual, or dialogs. All interaction is designed for a mouse or touch-screen interface. FASTAT output is a browser-based document with exportable graphics, technical references, and links to Web statistical resources.

The Construction and Use of Crossover Designs in Medical Research

Emlyn Williams
Australian National University

Crossover designs are common in clinical trials. They are characterized by the application of sequences of treatments to subjects. For example the efficacy and safety of an inhaled drug might be evaluated by presenting the drug and a placebo (in some order) to a number of patients.
Often subjects are used to evaluate more than two treatments and then the construction of optimal experimental arrangements becomes important. The design and analysis of crossover experiments need to be able to accommodate the possibility of carry-over effects from the successive application of treatments to subjects.
Different model specifications have been proposed for crossover designs in clinical trials; these include:

1. The additive model – carry-over effects are included as an additive component.
2. The interactive model – a full set of interactions between direct and first-order carry-over effects are included.
3. The self-adjacency model – a separate set of parameters are included when a treatment carries over to itself.
4. The placebo model – no carry-over is included for placebos.
5. The proportionality model – the carry-over effect is assumed to be additive and proportional to the direct treatment effect.

The construction of suitable crossover designs needs to take into account the proposed model for analysis.
The software package CycDesigN (www.cycdesign.co.nz) has facilities for the optimal construction of crossover designs for a wide range of model types including those listed above. This talk will discuss the types of crossover designs, their properties, implementation and optimal construction using CycDesigN. Some examples of the use of crossover designs in clinical trials will be presented.

An Adaptive Threshold Method Applied to Public Health Surveillance

William Woodall*, John Szarka, Linmin Gan*
Virginia Tech

Daily data from hundreds of hospitals and public health departments from across the country are collected and analyzed through the U.S. Centers for Disease Control and Prevention's (CDC) BioSense program in order to detect health-related outbreaks of disease or illness. The statistical methods used to decide if there is an outbreak are incorporated into the Early Aberration

Reporting System (EARS). The EARS methods allow one to monitor syndrome counts from each source and/or the proportions of counts of a particular syndrome relative to the total number of facility visits. A modification of the adaptive threshold monitoring approach of Lambert and Liu (2006, JASA) is introduced and compared to the use of the EARS methods both in terms of their design when there is no outbreak, using an empiric recurrence interval metric, and their power to detect outbreaks. Generally, the adaptive threshold methods perform better than the corresponding EARS methods, especially in the case of monitoring rates.

Spatially Adapted Experimental Design for Computer Experiments

Henry Wynn
London School of Economics

Hugo Maruri-Aguilar
Queen Mary University of London

Noha Youssef
London School of Economics

Two methods can be combined to create experimental design procedures which are truly adaptive in that they exploit learning about the local smoothness of the underlying spatial process. The first method is to spatially adapt the covariance of a Gaussian Process (GP) by an elementary expansion into wavelets, which can express local dependence. Haar function wavelets are a starting point in this regard. The second method is to vary the spatial "density" of a space-filling design and in this a newly developed type of spatial adapted Sobol' sequence is favoured. A motivating theorem shows that the latter are D-optimal, in the classical sense, if the local density of the design matches, in a certain way, the local prevalance of Haar wavelets. One analogue of D-optimality for sampling from processes is maximum entropy sampling (MES) and this is used to develop an adaptive version of the result, particularly for application to computer experiments.

DOE in Computer Simulation for Automobile Design

Shu Yamada
University of Tsukuba

In the recent technology development, effective utilization of information technology is the essential for the success in many fields. In manufacturing sector, digital engineering has been a popular term that implies effective application of information technology where the computer simulation and CAD

system has located at the central of digital engineering. Therefore how to implement quality management with computer simulation is an essential part to actualize high quality and customer satisfaction. This paper discusses application of statistical tool in the field of digital engineering under the scheme of quality management. In particular, application of DOE in computer simulation in Automobile Design is discussed. In some cases, the tools can be applied as well as the traditional approach, while some modifications on the traditional approach of the DOE are required to adapt the computer simulation. Practical guidelines are provided for the modifications of the approaches. The approach of this paper is utilization of the examples in order to derive general guideline. One example is to apply design of experiments for computer simulation that is a case to be required modifications.

Visualization for Datamining

Yoshiro Yamamoto
Tokai University, Japan

In the data mining field, visualization plays the very important role for Exploratory Data Analysis (EDA). Visualization is necessary for the interpretation which is as a result of the cluster analyses and the correspondence analysis and so on. 1. introduce about 2. interactive representations, interactive clustering plot and interactive choropleth maps. Dendrogram is usually used to express a result of the hierarchical cluster analysis. The dissimilarity between the objects or clusters can be expressed by the dendrogram. On the other hand, the objects can be visualized in Euclidean space by using multidimensional scaling (MDS) or principal component analysis (PCA). Objects coordinates in the space can express not only the dissimilarities between objects but also other information, such as factor loadings plot. Visualization of a cluster analyses has to be able to compare many results interactively. I introduce about the way. For hierarchical cluster analysis, the cluster is represented as closed domains in Euclidean space, in which the objects are represented as points. Each domain is drawn according to the clustering process or user's request. A decision of the number of clusters is also a big problem for cluster analysis. In addition to visualization, I propose the way to give a candidate for the number of clusters. On the regional analysis, clustering analysis does a very important role to interpret the relation between various areas. On clustering for regional data, types of data to be analyzed are given in various forms. For instance, there are rectangular data of continuous variables, dissimilarities, similarities and so on. It is easy to understand the cluster analysis result of regional data from visualization using the map. In this report, I introduce some methods of visualizing the result of clustering result onto map. Some of representation methods are the varieties of choropleth maps. Moreover, we illustrate interactive graphics to achieve various type of expression of choropleth map on the Web GIS. We describe some graphical representation of asymmetric type of data.

Estimating Covariation: Epps Effect, Microstructure Noise

Lan Zhang
University of Oxford and University of Illinois at Chicago

This paper is about how to estimate the integrated covariance <X,Y> of two assets over a fixed time horizon [0,T], when the observations of X and Y are "contaminated" and when such noisy observations are at discrete, but not synchronized, times. We show that the usual previous-tick covariance estimator is biased, and the size of the bias is more pronounced for less liquid assets. This is an analytic characterization of the Epps effect. We also provide optimal sampling frequency which balances the tradeoff between the bias and various sources of stochastic error terms, including nonsynchronous trading, microstructure noise, and time discretization. Finally, a two-scales covariance estimator is provided which simultaneously cancels (to first order) the Epps effect and the effect of microstructure noise. The gain is demonstrated in data.

Simultaneous Variable Selection for Heteroscedastic Regression Models

Zhongzhan Zhang
College of Allied Sciences, Beijing University of Technology

Darong Wang
Beijing University of Technology

The simultaneous variable selection for mean model and variance model in heteroscedastic linear models is discussed in this paper. We propose a criterion named PICa based on the adjusted profile log-likelihood function, which can be employed to jointly select regression variables in the mean model and variance model. The efficiency of the proposed criterion is compared with the naive AIC and BIC through a Monte Carlo simulation, and it is shown that PICa outperforms AIC, and is comparable with BIC. In addition, when the sample size is not large, it performs the best.

INDEX of names of other registered authors for papers with multiple authorship

Title: **ISBIS-2010**
Type of Publication: Book of Abstracts

Submitted: Authors, Co-Authors
Format: A5
Number of Pages: 113 pages
Year of Issue: 2010

Published: Zeithamlová Milena, Ing. - Agentura Action M
Vršovická 68
CZ - 101 00 Praha 10
actionm@action-m.com
http://www.action-m.com

Printed: Reprostředisko UK MFF
Sokolovská 83
CZ - 186 75 Praha 8

No editorial and stylistic revision.