ILACSC

I Latin American Conference on Statistical Computing

BOOK OF ABSTRACTS

CONFERENCE THEME:

Statistical Computing for Data Science

July 22-24, 2016

Gramado, Brazil

Contents

KEYNOTE SPEAKERS

INVITED PAPER SESSIONS

CONTRIBUTED PAPER SESSIONS

CPS1: Statistical computing in the society

CPS2: Recent advances in distribution theory 1

CPS3: New challenges in statistical computing

CPS4: Recent advances in distribution theory 2

CPS5: New challenges in linear models

CONTRIBUTED POSTERS SESSIONS

Welcome to the 1st Latin American Conference on Statistical Computing (I LACSC 2016)

The I Latin American Conference on Statistical Computing (LACSC) will be held in Gramado, state of Rio Grande do Sul, Brazil, during July 22-24, 2016. The organization of this meeting is jointly carried out by the **International Association of Statistical Computing (IASC),** the **Brazilian Statistical Association (ABE)** and the **Department of Statistics at the Federal University of Rio Grande do Sul**.

The purpose of the I LACSC is gather researchers that use computational statistical methods, to share and discuss ways to improve the access to knowledge, and promote interdisciplinary collaborations. The scientific program is very appealing for most statisticians and other researchers interested in computational methods and data analysis. The scientific program includes invited paper and contributed paper sessions and also contributed posters.

Conference Theme:
Statistical Computing for Data Science

Conference's Aim:
The aim of this conference is to furnish a forum for the discussion and exchange of new ideas, concepts and recent methods regarding statistical computing for society information. Intellectual stimulation of research on statistical computing for Data Science is expected at this meeting.

We're looking forward to see you in Gramado.

Committees:

LOCAL ORGANIZING COMMITTEE
(Federal University of Rio Grande do Sul, Brazil)

Danilo Marcondes Filho - Chair
Sílvia Regina Costa Lopes - Chair
Cleber Bisognin
Eduardo de Oliveira Horta
Hudson da Silva Torrent
Márcia Helena Barbian

SCIENTIFIC PROGRAM COMMITTEE

Paulo Canas Rodrigues *(Chair, Federal University of Bahia, Brazil)*
Alba Martinez-Ruiz *(Universidad Católica de la Ssma. Concepción, Chile)*
Carlos Abanto *(Federal University of Rio de Janeiro, Brazil)*
David F. Muñoz *(Instituto Tecnológico Autónomo de México, Mexico)*
Francisco Louzada Neto *(University of São Paulo, Brazil)*
Graciela Boente *(University of Buenos Aires, Argentina)*
Javier Trejos *(Universidad de Costa Rica, Costa Rica)*
Miguel de Carvalho *(Pontificia Universidad Católica de Chile, Chile)*
Paula Brito (*University of Porto, Porto, Portugal)*
Ramsés Mena Chávez *(Universidad Nacional Autónoma de México, Mexico)*
Vera Tomazella *(Federal University of São Carlos, Brazil)*

TECHNICAL COMMITTEE

Bruno Ramos *(Federal University of Rio Grande do Sul, Brazil)*
Gabriel da Cunha *(Federal University of Rio Grande do Sul, Brazil)*
Luis Benites Sánchez *(University of São Paulo, Brazil)*
Jenifer C. Ribeiro *(Federal University of Rio Grande do Sul, Brazil)*

General Information:

1) SHUTTLE SERVICE IS AVAILABLE ACCORDING TO THE FOLLOWING SCHEDULE

Thursday, July 21
2:00pm from Salgado Filho International Airport, Porto Alegre to Laghetto Siena Hotel
6:00pm from Salgado Filho International Airport, Porto Alegre to Laghetto Siena Hotel

Friday, July 22
8:00am from Laghetto Siena Hotel to FAURGS Conference Center
8:45pm from FAURGS Conference Center to Laghetto Siena Hotel

Saturday, July 23
8:00am from Laghetto Siena Hotel to FAURGS Conference Center
6:00pm from FAURGS Conference Center to Laghetto Siena Hotel

Sunday, July 24
2:00pm from Laghetto Siena Hotel to Salgado Filho International Airport and Plaza São Rafael Hotel (SINAPE check-in).

2) CONFERENCE EXCURSION

On Sunday, **July 24th**, there will be an excursion tour to Caracol State Park *(http://www.parquedocaracol.com.br/)*. Shuttle service will be available, leaving at **8:30am** from the Laghetto Siena Hotel to Caracol State Park, and coming back at **11:15am**.
Fees: shuttle R$ 40.00 (payment to the local committee at any time during the conference, before July 24th), tickets R$ 18.00 (at the entrance of the park).
Important Notice: credit cards are not accepted.

3) PARTICIPATION CERTIFICATE

It will be made available on-line at the conference site *(http://www.redeabe.org.br/lacsc2016)*. Please notice: the certificates will be available only after the conference has ended, and it will be possible only for those who have completed their enrollment and paid the corresponding fees.

4) OFFICE HOURS

It will be open during the conference schedule.

Scientific Program (Schedule)

JULY 22, 2016 (FRIDAY)

08:00 – 09:30	Registration
09:30 – 09:45	Opening Ceremony (Locatelli Room)
09:45 – 10:45	**KS1**
10:45 – 11:15	Coffee break
11:15 – 12:45	**IPS1** (Locatelli Room) **IPS6** (Rembrandt Room)
12:45 – 15:00	Lunch
15:00 – 16:00	**IPS2** (Locatelli Room) **IPS3** (Rembrandt Room)
16:00 – 17:15	Open Meeting to discuss the creation of the Latin American Region of the IASC
17:15 – 17:45	Coffee break
17:45 – 19:00	Poster Session
19:30 - 20:30	Cocktail

KS1: Patrick Groenen, President of the IASC
KS2: Emilio Porcu
IPS1: Computational Methods in Bayesian Statistics
IPS2: Symbolic Data Analysis
IPS3: PLS Methods for Data Science: Theory and Computational Applications
IPS4: Statistical Challenges in Big Data
IPS5: Innovation and Statistics in Brazil
IPS6: Computational Methods in Survival Analysis
IPS7: Copula-based Computational Techniques
CPS1: Statistical computing in the society

JULY 23, 2016 (SATURDAY)

09:00 – 10:45	**IPS5** (Locatelli Room) **IPS7** (Rembrandt Room)
10:45 – 11:15	Coffee break
11:15 – 12:15	**CPS1** (Locatelli Room) **CPS2** (Rembrandt Room)
12:15 – 13:15	**CPS3** (Locatelli Room) **CPS4** (Rembrandt Room)
13:15 – 15:00	Lunch
15:00 – 16:00	**IPS4** (Locatelli Room) **CPS5** (Rembrandt Room)
16:00 – 16:30	Coffee break
16:30 – 17:30	**KS2**
17:30 – 17:45	Closing Ceremony
20:00 – 22:30	Gala Dinner

JULY 24, 2016 (SUNDAY)

08:00 – 12:00	Conference excursion
12:00 – 14:00	Lunch
14:00	Departure

CPS2: Recent advances in distribution theory 1
CPS3: New challenges in statistical computing
CPS4: Recent advances in distribution theory 2
CPS5: New challenges in linear models

Keynote Speakers

Majorization and Minorization Algorithms in Statistics

Patrick Groenen
(Erasmus University Roterdam, The Netherlands; President of the International Association for Statistical Computing)

Abstract: As an optimization method, majorization and minorization (MM) algorithms have been applied with success in a variety of models arising in the area of statistics. A key property of majorization algorithms is guaranteed descent, that is, the function value decreases in each step. In practical cases, the function is decreased until it has converged to a local minimum. The auxiliary function, the so-called majorizing function, is often quadratic so that an update can be obtained in one step. In this paper, we present an review of MM-algorithms. We present a classification of situations where iterative majorization can be used and show several inequalities that can be used for iterative majorization. We show how certain constraints can be easily imposed. Some guidelines are given for developing majorizing algorithms. We will present several examples of applications of MM algorithms in multidimensional scaling, logistic regression, support vector machines, weighted principal components analysis, and partial least squares.

Space-Time Covariance Functions for Planet Earth

Emilio Porcu
(University Federico Santa Maria, Chile)

Abstract: In this paper, we propose stationary covariance functions for processes that evolve temporally over a sphere, as well as cross-covariance functions for multivariate random fields defined over a sphere. For such processes, the great circle distance is the natural metric that should be used in order to describe spatial dependence. Given the mathematical difficulties for the construction of covariance functions for processes defined over spheres cross time, approximations of the state of nature have been proposed in the literature by using the Euclidean (based on map projections) and the chordal distances. We present several methods of construction based on the great circle distance and provide closed-form expressions for both spatio-temporal and multivariate cases. A simulation study assesses the discrepancy between the great circle distance, chordal distance and Euclidean distance based on a map projection both in terms of estimation and prediction in a space-time and a bivariate spatial setting, where the space is in this case the Earth. We revisit the analysis of Total Ozone Mapping Spectrometer (TOMS) data and investigate differences in terms of estimation and prediction between the aforementioned distance-based approaches. Both simulation and real data highlight sensible differences in terms of estimation of the spatial scale parameter. As far as prediction is concerned, the differences can be appreciated only when the interpoint distances are large, as demonstrated by an illustrative example.

Invited Paper Sessions

IPS1: Computational Methods in Bayesian Statistics

Organizer: David Fernando Muñoz

(Instituto Tecnológico Autónomo de México, Mexico)

A new Bayesian approach to one-parameter IRT models with large data sets

Jorge Luis Bazán

(University of São Paulo, Brazil)

Abstract: Rasch models are frequently used in large-scale assessments by considering several maximum likelihood (ML) estimation methods. In contrast, it is known that Bayesian estimation considering Markov chain Monte Carlo (MCMC) approaches are usually slow for large data sets. In this work, a new Bayesian approach considering integrated nested Laplace approximations (INLA) for one-parameter models (including the Rasch model) is proposed. The methodology is illustrated using simulated datasets and a large data set corresponding to a ENEM exam in Brazil. The main conclusion is that the new method provides a Bayesian alternative to analyze large-scale data sets which provide very similar results but faster than the traditional MCMC approach and perform similarly (or better) than ML estimation methods for parameter estimation.

Metropolis-adjusted Langevin algorithms for the estimation of a varying-dispersion Beta mean regression model

Luis Hilmar Valdivieso

(Pontificia Universidad Católica del Perú, Perú)

Abstract: In this article, we explore the use of alternative Metropolis-adjusted Langevin algorithms for Bayesian estimation of parameters from a new varying-dispersion Beta mean regression model for fractional response variables. This model extends the Beta mean regression model proposed in Bayes and Valdivieso (2016) by including a linear regression model for the precision parameter. In addition, we also study and compare these alternative Bayesian techniques to the classical likelihood approach that has been proposed to deal with this heteroscedastic model.

Solving the newsvendor problem under parametric uncertainty using simulation

David Fernando Muñoz
(Instituto Tecnológico Autónomo de México, Mexico)

Abstract: In this paper, we discuss the formulation and solution to the newsvendor problem under a Bayesian framework that allows the incorporation of uncertainty on the parameters of the demand model (introduced by the estimation process of these parameters). We present an application of this model with an analytical solution and we conduct experiments to compare the results under the proposed method and a classical approach. Furthermore, we illustrate the estimation of the optimal order size using stochastic simulation, when the complexity of the model does not allow the finding of a closed form expression for the solution.

IPS2: Symbolic Data Analysis

Organizer: Paula Brito
(University of Porto, Portugal)

Regression models for interval variables

Eufrásio Lima Neto
(Federal University of Paraíba, Brazil)

Abstract: Regression models are widely used to solve problems in many fields and the uses of inferential techniques play an important role in order to validate these models. An brief overview about the regression methods to model interval variables will be presented in this talk. Moreover, will be discussed a regression model for interval-valued variables based on copula theory that allows more flexibility for the model's random component. In this way, the main advance of the new approach is that is possible to consider inferential procedures over the parameters' estimates as well as goodness-of-fit measures and residual analysis based on general probabilistic background. A Monte Carlo simulation study demonstrates asymptotic properties for the maximum likelihood estimates obtained from the copula regression model. Applications to real data sets are also illustrate the useful of this new approach.

Clustering Interval Time Series

Paula Brito[1]
Joint work with Paulo Teles[1] and Elizabeth Ann Maharaj[2]
([1]University of Porto, Portugal; [2]Monash University, Australia)

Abstract: An interval time series (ITS) is a sequence of intervals observed in successive instants in time. We focus on the clustering of ITS where we examine and compare different approaches. One technique involves fitting space-time models to each of the ITS under consideration, and using the parameter estimates of the fitted models as inputs into clustering methods. Alternatively, we cluster the ITS on the basis of summary statistics of the midpoints and ranges of the observed intervals. Wavelet modelling is also applied and the obtained measures used for clustering. We use these approaches to cluster ITS on the sea level at different locations in Australia, based on daily interval data between 2023 and 2012.

IPS3: PLS Methods for Data Science: Theory and Computational Applications

Organizer: Alba Martinez-Ruiz
(Universidad Católica de la Ssma. Concepción, Chile)

A Study of Partial Least Squares and Ridge Regression and Their Relevance in Data Science

Luis Firinguetti
(University of Bío-Bío, Chile)

Abstract: Data Science is a multidisciplinary endeavor, in particular, a blend of statistical and computational thinking to extract knowledge from data. Important contributions from the domain of Statistics to Data Science come from Clustering, Data Reduction, Regression Modeling and Prediction, among others. In this work we focus on two Regression Modeling and Prediction techniques: Ridge Regression (RR), which is a regularization method, and Partial Least Squares (PLS), which is also a method of Data Reduction or data compression. In many realistic Data Science applications we may have to deal with high dimensional data matrices, where it is fairly common the presence of the so called multicolinnearity problem. To tackle this problem in the estimation of liner regression equations, a number of parameter estimation methods have been proposed, RR and PLS among others. In this presentation we compare through simulations, in the context of the Classical Linear Regression Model (CLRM), the properties of PLS and RR with Ordinary Least Squares (OLS). We consider varying degrees of multicollinearity, different numbers of observations and explanatory variables and different magnitudes of the error variance. It is found that both, PLS and RR, estimators produce significant reductions in the Prediction Mean Square Error (PMSE) and in the Total Mean Square Error (TMSE) over OLS. When comparing RR and PLS, the last estimator tend to produce better results when there are more variables in the model and they present greater collinearity. On the other hand RR tend to perform better when the level of noise is greater. We conclude this presentation with a discussion of some ongoing research on prediction confidence intervals for PLS and some variable selection techniques for RR, together with the advantages and disadvantages it may have in comparison with another well known regularization technique: the LASSO.

IPS4: Statistical challenges in Big Data

Organizer: Paulo Canas Rodrigues
(Federal University of Bahia, Brazil)

Statistics for Data Science

Paulo Canas Rodrigues
(Federal University of Bahia, Brazil)

Abstract: The increase in the amount and variety of data collected every day, in many disciplines, asks for new statistical strategies to handle it and to evaluate the research hypothesis timely. When dealing with high dimensional data and unstructured data, dimensionality reduction methods and visualization tools are of great interest. In this talk I will present a general overview of big data and how statistics can be of great usefulness to understand and to take advantage of this deluge.

Parameter Computation for Software Reability Growth Model

Rajesh Singh
(Sant Gadge Baba Amravati University, India)

Abstract: In the fastest growing world, the modern computers due to its ability of operating complex operations at great speed, accuracy, effective and huge storage capacity for the years, are being widely used all over for solving most of the complicated problems belonging to science, engineering, arts, education, business communication, construction technology, air traffic control and

maintance, nuclear reactors, automotive industry, hospitals and health care, etc. The software play very important role in performing all these works along with highly advanced computers. The reliability of the software is the most important characteristics and dynamic measure defined with the user's perspective related to operations of software and not the design of the program. The Software Reliability is defined as "the probability that the software will be functioning without failure under a given environmental condition during a specified period of time". Many more definitions of Software reliability are given by researchers. To compute the reliability of Software, the characterization of model is needed and then involved parameters are estimated. In this talk, the length biased exponential class model will be characterized as software reliability growth model although this model is well established for the computation of hardware reliability. The Bayesian method of estimation is having certain advantages over classical methods and thus, the Bayes estimators are proposed for the computation of parameters in case of non informative as well as informative priors. Lastly, these estimators will be compared with m.l.e.s to study their performance.

IPS5: Innovation and Statistics in Brazil

Organizer: Francisco Louzada Neto
(University of São Paulo - São Carlos, Brazil)

Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation: Why (not) to Randomize?

Julio Stern[1]
Joint work with Marcelo S. Lauretto[1], Fabio Nakano[1] and Carlos Alberto de Bragança Pereira[1]
([1]University of São Paulo, Brazil)

Abstract: This talk describes and comments some of the statistical models developed by the authors for a project lead by Datanexus, a Brazilian survey institute for media audience. The main goal of this project was to provide new alternatives for measuring television audience in São Paulo, Brazil. The goal of the statistical modeling was to help in the selection of a representative monitoring sample, that is, a small set of households which features (householders income, sex, age and scholarship) were similar to those in the population. We describ e an approach combining multiobjective programming and stochastic perturbations for providing monitoring samples that obey two performance criteria: optimality (similarity between features in the household sample and the population) and decoupling (low tendency to select the same households or pairs of households in different runs).

The core of our model is based on intentional sampling, a non-deterministic approach for yielding samples which meet specific prescribed criteria. The possibility of subsequently generalize statistically from such deterministic samples

to the general population has been the issue of long standing arguments and debates. Nevertheless, the intentional sampling technique presented in this talk explores pragmatic strategies for overcoming some of the real or perceived shortcomings and limitations of intentional sampling in practical applications.

Design of Clinical Trials: A Mixed Intentional-Randomized Sampling Approach

Marcelo S. Lauretto[1]

Joint work with Victor Fossaluza[1] *and Carlos Alberto de Bragança Pereira*[1]

([1]University of São Paulo, Brazil)

Abstract: Intentional sampling methods are non-randomized procedures that select a group of individuals for a sample with the purpose of meeting specific prescribed criteria. Such approach is intended for exploratory research or pilot studies where tight budget constraints or low availability of sample individuals preclude the use of traditional randomized representative sampling. In this paper we consider the problem of sequential allocation for clinical trials with few patients. The allocation problem consists in assigning each new patient to one, and only one, of two (or possibly more) alternative treatments (arms). A requisite usually stated by physicians is that the profiles in the alternative arms remain similar with respect to some relevant patients' characteristics (age, gender, disease, symptom severity and others). Fossaluza et al (2009) proposed a straightforward method for the sequential allocation problem, which measures the Aitchison distance between the compositional vectors under the two (or more) possible allocations, and chooses that allocation with the lowest final distance. Although this intentional sampling method provides arms with very similar profiles even in small trials, it is vulnerable to potential bias or human interferences in the patient's assignment. In this talk we show an extension of the sequential sampling approach proposed by Fossaluza et al (2009), adapting the stochastic perturbation introduced in Lauretto et al (2012). We also discuss how to conveniently set up perturbation parameters in order to yield a suitable balance between optimality (similarity among arms compositional vectors) and decoupling (absence or minimization of experimental bias in patients allocation).

i-Sports: A statistical tool for online talent detection

Anderson Ara[1]

Joint work with Alexandre Maiorano[2] *and Francisco Louzada*[1]

([1]University of São Paulo, Brazil; [2]University of São Paulo - São Carlos, Brazil)

Abstract: The main goal of this paper is to present an innovating online system, built through free software and statistics tools that allow a comparison of individuals in any sport modality. Particularly, this shown study is focused in the performance evaluation in soccer, using univariate and multivariate statistical methods. The univariate approach is given by the Z-CELAFISCS methodology and the multivariate is given through a novel construction of indicators using Principal Component Analysis, Factor Analysis and Copulas. The created system shows many dynamic online reports that allow us to observe the results of the subjects in the tests and who is more suitable for the practice of this sport.

On the Development of Statistical Products for Data Analysis

Francisco Louzada Neto

(University of São Paulo, Brazil)

Abstract: Nowadays there is a real need for statistical web based environments which facilitates data collection and survey applications, directly connecting data to real time informative reports, being accessible to anyone with only internet access. In this talk I focus on the presentation of some statistical web based products developed at the Centre for Mathematic Sciences Applied to Industry (CEPID-CeMEAI) for data analysis. Real success case applications are provided during the presentation.

IPS6: Computational Methods in Survival Analysis

Organizer: Vera Tomazella

(President of Associação Brasileira de Estatística, Federal University of São Carlos, Brazil)

Zero-inflated Cure Rate Regression Models for Time-to-Default Data

Francisco Louzada[1]

Joint work with Mauro R. de Oliveira Jr.[2] *and Fernando F. Moreira*[3]

([1]University of São Paulo, Brazil; [2]Federal University of São Carlos, Brazil; [3]University of Edinburgh Business School, UK)

Abstract: In this talk, I introduce a zero-inflated cure rate survival model, which accommodates fraud rate estimation in bank loan portfolios. Indeed, the approach enables the consideration of three different types of loan borrowers, namely, the fraudsters, those who are susceptible to default and those who are not susceptible to default. An advantage of our approach is to accommodate zero-inflated times, which is not possible in the standard cure rate models, usually considered in survival analysis. Estimation is based on the maximum likelihood. The proposed modeling is illustrated in a real dataset of loan survival

times, where focus is given on the zero-inflated Weibull cure rate model.

Generalized Extended Class of Defective Cure Rate Models

Ricardo Rocha
(Federal University of São Carlos, Brazil)

Abstract: The modeling of a cure fraction, also known as long-term survivors, is needed when there are observations in the data which are supposedly not susceptible to the event of interest. Such cases require special theoretical treatment, in a way that the modeling assumes the existence of such observations. A way to model cure rates is to use defective distributions. These distributions are characterized by having probability density functions which integrate to values less than one when the domain of some of their parameters is different from that usually defined. The defective models have the advantage of not need the assumption of the presence of immune individuals in the data set. In this work we investigate how to extend a baseline defective distributions using families of distributions. We will consider several well known families of distributions of the literature to exemplify the proposed methodology. The main purposes of this work are to introduce a result that can generate new defective distributions, given an extended family of distributions and a baseline defective distribution and to explore a variety of these families in order to show that they can properly be used to fit different kinds of data.

IPS7: Copula-based Computational Techniques

Organizer: Nikolai Kolev
(University of São Paulo, Brazil)

Dynamic Vine Copula Modelling

Flávio A. Ziegelmann
(Federal University of Rio Grande do Sul, Brazil)

Abstract: Vine copulas constitute a very flexible class of multivariate dependence models. In this paper, we allow the dependence parameters of the pair-copulas in a D-vine decomposition to be potentially time varying, bringing more flexibility to our model at the expense of a higher computational burden. The proposed model is explored in simulations and further evaluated with respect to the accuracy of Value-at-Risk (VaR) forecasts in crisis periods. Our findings in the empirical application illustrate that time variation is present in the dependence structure of multivariate financial returns. Moreover, the dynamic D-vine copula outperforms the static D-vine copula in terms of predictive accuracy for our data sets.

Estimating Multivariate Discrete Distributions using Bernstein Copulas

Victor Fossaluza[1]
Joint work with L. Esteves[1] and C. A. B. Pereira[1]
([1]University of São Paulo, Brazil)

Abstract: Measuring the dependence between random variables is one of the most fundamental problems in statistics, and therefore, determining the joint distribution of the relevant variables is crucial. Copulas have recently become an important tool for properly inferring the joint distribution of the variables of interest. Although many studies have addressed the case of continuous variables, few studies have focused on treating discrete variables. We present a nonparametric approach to the estimation of joint discrete distributions with bounded support using copulas and Bernstein polynomials. An application to real Obsessive-Compulsive-Disorder data will be discussed.

Energy Markets Modelling via Sibuya-type Copulas

Nikolai Kolev[1]
Joint work with Hugo Branco[1]
([1]University of São Paulo, Brazil)

Abstract: The dependence structure between prices for futures on crude oil and natural gas traded on NYMEX is analyzed using modern time-series and copula techniques. We model the log-returns on each commodity individually by Generalized Autoregressive Score (GAS) models and account for dependence between them by fitting various copulas to corresponding error terms. By performing a modified maximum likelihood approach and novel goodness-of-fit tests, we select the best copula, being member of the recently introduced Sibuya-copulas class.

Contributed Paper Sessions

CPS1: Statistical computing in the society

Traffic Accident Victims And After-Effects In Brazil: Disabled People

Paulo Tadeu Meira e Silva de Oliveira[1]
Joint work with Ana Paula Camargo Larocca, Aurenice da Cruz Figueira and Maria Izab El dos Santos

(¹University of São Paulo, Brazil)

Abstract: More than 1.5 million people die annually worldwide due to road traffic. Between 20-50 million suffer with serious road traffic injuries. Official data in Brazil indicate about 37,000 dead and 500,000 injured annually (WHO, 2013). The injured count, most permanently, creates a huge contingent of disabled people which dependent on social security and family support. Over a billion people live with some kind of disability (15% of world population) according to World Health Organization (WHO, 2010). This is lower than 2010 IBGE estimates, which indicates 45,606,048 people with at least one disability (almost 24% of Brazil's population). Disability still is considered a universal and social challenge and economic cost for individuals, families, communities and nations. The combination of factor such as age, sex and exposure to environmental risks, among others, increase the changes of a person to become disabling. According to Brazilian Traffic Accident Prevention Association (ABPAT), a large proportion of car accident victims end up as disabled person due to problems such as paralysis, amputations, burn injuries, and blindness among other types of health problems. This paper shows the problem in a quantified way by statistical terms and illustrated with data. It also indicate improvements so stakeholders that might relieve the problem, can better monitor it.

A Statistics computing solution in detecting procurement fraud

Oscar Centeno Mora
(University of Costa Rica, Costa Rica)

Abstract: Government purchases may lead to fraud. To prevent this, it is necessary to establish a statistics computing system allowing a timely and effective detection. The platform Fight Against Procurement Fraud is an analysis system which aims to detect purchases presenting high risk of fraud. This analysis system was created in 2015 by the Government Accountability Office of Costa Rica, and is constituted of five components. The computing component consists of the Extract-Transform-Load (ETL), clouds, analysis development needs, others. The market rules component is made of computer programs that establish indicators according to business criteria to detecting fraud cases. The analytics component performs statistical analysis. The reporting component is the link between the analysis and the visualization of the results. Finally, the management component link between the system and the analysis, and proposes improvements both in statistical and computing. The tools used in the analysis system are Oracle, QlikView, Redatam, SPSS and R. The main data analysis are descriptive (tables, charts, risk indicators), predictive analysis and detection anomalous cases (Data Mining techniques, Benford law, control charts, etc.). The results demonstrate the effectiveness of the platform to detect large number

of fraud cases. The level of visualization and computer development has made it possible to analyze all purchases in a few hours, under all possible scenarios.

A Hidden Markov Model-Based Methodology for Fault Detection In Continuous and Periodic Industrial Processes

Gustavo Matheus de Almeida[1]
Joint work with Song Won Park
([1]Federal University of Minas Gerais, Brazil)

Abstract: Continuous industrial processes as oil and gas and steelmaking plants have been also experiencing the actual data rich-world paradigm. A pure mathematical description of their operations constitutes in a hard task due to its complex nature. Both facts have led to the growing use of the direct data-driven approach with several purposes. This work focuses on the fault detection problem, an open question in practice yet. A fault consists in a deviation from the normal mode of operation, and so early detection is necessary to mitigate potential losses. Since the magnitude of faulty signals is generally incipient at the beginning, its recognition by control room operators by means of the commonly applied visual monitoring of key variables using trend plots is virtually impossible. Automatic and reliable fault detection systems may support this task with more rational decision making towards safer, cleaner, and more economical operations. This work proposes a methodology for developing fault detection systems for periodic and continuous industrial processes using the hidden Markov model (HMM) method, a statistical recognizer of sequential data. This study investigates mean vector and covariance matrix initial estimation, covariance matrix-type, model topology, and the use of overlapped data at the model identification step. The resulting HMM-based fault detection system is compared to classical multivariate statistical process control (MSPC) techniques, namely MEWMA, MCUSUM, PCA, and dynamic PCA. The case study is the DAMADICS actuator benchmark problem, commonly employed for development and comparison of monitoring systems. All possible faulty scenarios were investigated including abrupt faults of low magnitude and incipient faults, more difficult to detect in advance. Performance metrics as true detection rate showed the significant better performance given by the monitoring system derived from the proposed methodology by providing earlier fault detection given a common false alarm rate of 1%.

CPS2: Recent advances in distribution theory 1

Generalized Transmuted Exponential Distribution: Structural Properties and Applications

Maria do Carmo Soares de Lima[1]

Joint work with Gauss M. Cordeiro, Abraão D. C. Nascimento and Kassio F. Silva

([1]Federal University of Pernambuco, Brazil)

Abstract: We introduce and study the generalized transmuted exponential (GT-E) distribution that includes as specials cases the exponential and transmuted exponential distributions. Some mathematical properties of the new model such as the quantile and generating functions, ordinary and incomplete moments, mean deviations and Shannon entropy are investigated. The model parameters are estimated by maximum likelihood and the usefulness of the new distribution is illustrated by means of a simulation study and two real data sets. BesideS that, we perform a Monte Carlo simulation (with 1000 replications) to quantify some asymptotic properties of the MLEs of the GT-E parameters.

Extended Exponentiated Gamma Distribution: Properties and Applications

Juan Carlos Chaves Capella[1]

Joint work with Gauss M. Cordeiro and Maria do Carmo S. Lima

([1]Federal University of Pernambuco, Brazil)

Abstract: In this paper, we introduce and discuss a new four-parameter extented exponentiated gamma distribution. We present some of its mathematical properties, such as explicit expressions for the moments, quantile and generating functions and order statistics. We use the maximum likelihood method to estimate the model parameters and to determine the observed information matrix. We perform two applications to illustrate the performance of the new distribution and compared with other models. Besides that, we perform a Monte Carlo simulation (with 1000 replications) to quantify some asymptotic properties of the MLEs of the ELEG parameters.

The Gamma Gumbel Distribution: Properties and Applications

Wellington Marins Filho[1]

Joint work with Maria do Carmo S. Lima, Gauss M. Cordeiro, Juan Carlos Chaves Capella and Vítor L. Alves

([1]Federal University of Pernambuco, Brazil)

Abstract: The modeling and analysis of lifetimes is an important aspect of statistical work in a wide variety of scientific and technological fields. We introduce and study the gamma Gumbel distribution, which is obtained using the Gumbel distribution as a baseline for the gamma generator. Its flexibility, although it only depends on one additional parameter, is proved empirically. We derive some of its mathematical properties like explicit expressions for the moments, quantiles, generating function and Renyi entropy. The method of maximum likelihood is used for estimating the model parameters and the observed information matrix is determined. We provide two applications to compare the performance of the proposed distribution with other competitive models. Besides that, we perform a Monte Carlo simulation (with 1000 replications) to quantify some asymptotic properties of the MLEs of the GGu parameters.

CPS3: New challenges in statistical computing

Bootstrap Confidence Intervals for the Correlation Coefficient: A comparison of Execution Times with Parallel Implementations in R and C Languages

Ruben Carvajal Schiaffino[1]

Joint work with Luis Firinguetti

([1]University of Santiago, Chile)

Abstract: Computational Statistics is defined as “aiming at the design of algorithm for implementing statistical methods on computers, including the ones unthinkable before the computer age (e.g. bootstrap, simulation), as well as to cope with analytically intractable problems” (C. Lauro, 1996). R is an integrated suite of software facilities for data manipulation, calculation and graphical display (W. N. Venables, D. M. Smith and the R Core Team, 2016) that allow the implementation of many classical and modern statistical techniques. In spite its popularity, it is a known fact that its main deficiency is the time it takes to carry out a set of user defined instructions, because they are interpreted instructions. The R language provides means to increase velocity in the execution of programs, among them translating a byte code, interaction with compiled languages such as FORTRAN and C, and the possibility of parallel computing.

Selecting Items in a Computerized Adaptive Testing: New Criteria Using Bayesian Approach

Gilberto Pereira Sassi

(Fluminense Federal University, Brazil)

Abstract: With the advent of computer in the last decades, tests driven by computer has gradually replaced traditional paper-and-pencil tests. A key task in computerized testing is how to select an item given the updated examinee ability, called Item Selection Criterion. The goal is to select the most suitable item for the updated ability. In the Bayesian framework, this is accomplished selecting the item with minimum a posteriori variance. However, computing the a posteriori variance is a highly time consuming task. In the literature, the solution is to chose items minimizing the a posteriori variance in the beginning of test. After some items, when computing the a posteriori variance begins to become slow, the software shift to a different Item Selection Criterion which is faster but would not choose the best item for the updated ability. Here, we propose a two phase Item Selection Criterion: first, it creates an item pool with small to medium size using some Item Selection Criterion less precise and, then, in this item pool we select the item with minimum a posteriori variance. We enunciated this approach for the unidimensional logistic model of Item Response Theory and we compared them with the Bayesian criteria present in the literature by simulation study. We observed a lesser mean square error and bias. Additionally, these new criteria are faster than all Bayesian criteria which use the posteriori distribution in the Item Selection Criterion.

Recognition profiles on social networks: a case study in the analysis graph

Ana Gómez[1]

Joint work with Ricardo Barboza and Rodrigo Salas

([1]Escuela Colombiana de Ingeniería, Colombia)

Abstract: The interest in studying the behaviour of human beings in social networks is of great importance in government, political, commercial and educational institutions, among others. For these studies the application techniques, data analysis, such as Web Mining and Computational Statistics, can generate new knowledge. In this sense, from an academic approach it came also a particular interest in research on the theory and methods currently being used to quantify and process information relating to the participation of users in these popular digital spaces, in which they interact and communicate freely through a case study with real data. To this end, it was necessary first to find recent references related to the theoretical study on the social network analysis. Second, an investigation into what was needed is currently being done to apply the theory analysed and finally on the various types of software that can be found for to perform an analysis that applies algorithms based on the theory put forth by those references found. Established above, the application process where two sets of data analysed was developed extracted Facebook with the application of techniques Web Crawling later Gephi was applied to visualize necessary information through the generation of graphs, giving these numerical data structure, due that initially, all data were in a format made up ads containing textual information, multimedia, “Likes” and comments to announcements made by

other users in a similar format. The first set of data belong to a network of 58 nodes with edges 313, while the second consists of 414 nodes with edges 3441, each graph was designated G1 and G2 respectively. The nodes have a number identifier called ID, sex, age range of the user's account, the number of messages and code regional location, on the other hand edges have the source and destination nodes, the ID of the edge and weight, which is initially 1. For analysis tools R-project statistical program was used to G1 and G2 characterize graphs through the number of nodes, number of edges, medium grade, network diameter and density of the graph and then the existence of the Power Law was analysed in graphs through a linear regression model to recognize the type of trend posing, and measures centrality of the importance and influence of each node that is identified to determine groups or communities. Specifically, helps us understand who is most associated, closest to others, the most influential or which can serve as a bridge to take us to others.

CPS4: Recent advances in distribution theory 2

Comparison of data generation methods using GI0 distribution (for the single look case)

Débora Chan[1]

Joint work with Andrea Alejandra Rey, Julia Cassetti, Juliana Gambini and Alejandro Frery

([1]Universidad Tecnológica Nacional, Regional Buenos Aires, Argentina)

Abstract: The Synthetic Aperture Radar images are very useful, since interesting information about natural resources can be obtained from them, such as the control of ecological systems and the study of the consequences of human actions like deforestation. The reason why several statistical models have been introduced to analyse this kind of images is that they allow to detect many of their essential characteristics and also to interpret them. For instance, the family of GI0 distributions, which depends on three parameters related to texture, brightness and signal-noise relation (via the number of looks), can be applied to the study of intensity data of areas concerning different textures. In order to investigate the properties of the mentioned distribution, besides the quality of its parameter estimations, it is necessary to include simulation strategies. In this sense, a good random variable generator of computational low cost becomes an essential and powerful tool for the scientific research. Throughout this work, using the R platform, the authors propose four ways of generating data from a GI0 distribution when the number of looks is equal to one. For each of this generation strategies the advantages and drawbacks are considered analysing a wide range of the values of the distribution parameters. In particular, the authors have taken into account the computational cost as well as the degree of adjustment of the generated data to the original distribution by means of two well-known likelihood tests. Since the GI0 distribution verifies a power law in the single look case, the quality of the generated sample tail is of vital importance. Thus, the adjustment degree of the maximum sample distribution and the 75th and 90th sample percentiles are evaluated via confidence intervals.

The effect of misspecifying the random-effects distribution in beta mixed models

Olga Cecilia Usuga Manco[1]

Joint work with Eduwin Aguirre González and Freddy Hernández Barajas

([1]University of Antioquia, Colombia)

Abstract: Beta mixed models are useful methods for analyzing longitudinal and clustered data that assume values on the interval (a, b). Estimation is often based on maximum likelihood theory, which assumes that the underlying probability model is correctly specified. Although the hypothesis of normal distribution of random effects may be appropriate for the model, the study of other distributions and their effects on the estimates and the performance of the inferential procedures have not been studied so far. In the present work, we address the impact of the misspecification of the random-effects distribution on the estimates and on type I and type II errors of the test for the mean structure in beta mixed models. We study, via Monte Carlo, the misspecification of the random-effects structure under different scenarios. In this study, we adopted the approach in which we vary the true distribution for the generation of random effects and estimated the model parameters assuming normal random effects. The simulations results obtained show that the power and the estimates can be seriously affected, depending on the shape and the variance of the underlying random-effects distribution.

Alpha skew Gaussian Bayes classifiers

Anderson Ara[1]

Joint work with Francisco Louzada

([1]University of São Paulo - São Carlos, Brazil)

Abstract: The main goal of this paper is to introduce new procedures for Bayes classifiers, namely alpha skew Gaussian Bayes classifiers and sumarized by the alpha skew gaussian naïve Bayes classifier (ASGNB) and alpha skew gaussian k-dependence bayesian classifier (ASGKDB). The both methods are based on a novel generalization of the Gaussian distribution applied to continuous variables.

CPS5: New challenges in linear models

Repeated Responses in Misclassification Binary Regression

Magda Carvalho Pires[1]
Joint work with Roberto C. Quinino
([1]Federal University of Minas Gerais, Brazil)

Abstract: Binary regression models are commonly used in medical applications to identify risk factors for certain diseases. Currently, the analysis of binary regression models assumes that the response variable is measured perfectly. However, in some situations, the outcome is subject to misclassification: a success may be erroneously classified as a failure (known as a false negative in medical applications), or a failure may be misclassified as a success (false positive). This misclassification is an important problem in medical and environmental studies, as ignoring misclassification errors can produce biased covariate effect estimates. This study proposes to incorporate repeated responses in misclassification binary regression models. Repeated measures are not considered validation subsample because they are performed by the same imperfect test, i.e., there is no gold-standard test available. We use subjective prior distributions, as our conditional means prior (CMP), to evaluate and compare models. A data augmentation approach, Gibbs sampling and Adaptive Rejection Metropolis Sampling (ARMS) are used for posterior inferences, and a simulation using real data is presented. Satisfactory results for bias and efficiency were obtained using multiple repeated measures.

Improvements for external validity tests in fuzzy regression discontinuity designs

Alessandro Ribeiro de Carvalho Casalecchi[1]
Joint work with Sergio Firpo and Antônio Galvão
([1]Sao Paulo School of Economics - FGV, Brazil)

Abstract: One type of external validity in fuzzy regression discontinuity designs holds when the identified treatment effect (that for compliers) equals that for all compliance subpopulations at the cutoff of the design. Then the effect estimate for compliers is valid for the other compliance subpopulations. In the literature, tests for external validity as just defined usually test, effectively, for hypotheses involving means, and not conditional distributions. Besides, usually the tests are not performed under assumptions that allow researchers to draw conclusions about the joint distribution of potential outcomes within every compliance subpopulation, since such assumptions may be too restrictive. This paper has two purposes. First we show that, if the researcher combines a mild additional assumption and a specific testing strategy, external validity (as just defined) tests detect more alternatives and allow for more conclusions about the data generating process --- namely, the joint distribution of potential outcomes and the distribution of a common unobservable determinant of potential outcomes, both within every compliance subpopulation. The mild assumption is that the functional relationship between potential outcomes and a common unobservable determinant is bimeasurable. The specific testing strategy is that of using a test statistic that detects differences between two full (an not only functionals of) conditional distribution functions. While the null hypotheses effectively tested by some authors on functionals can be true even if external validity is false, under bimeasurability equality between full conditional distributions hold if and only if external validity is true. The second purpose of this paper is to use simulations to compare the performance of four different test statistics that detect differences between limiting conditional distribution functions.

Approach and solution of stochastic DEA model for longitudinal data

Jhon Jairo Vargas Sánchez[1]

Joint work with Edilberto Cepeda Cuervo and T. Gerard Olivar

([1]University of Magdalena, Colombia)

Abstract: For the purposes of measuring efficiency, DEA (Data Envelopment Analysis) technique is used. Two of the main research lines are stochastic DEA and DEA in a time horizon. The Chace Constrained Programming (CCP) is a classic stochastic DEA. DEA in a time horizon has two models known as Window Analysis and Malmquist index. The stochastic CCP model measures efficiency, but in an instant of time. DEA Models in a time horizon has some weaknesses 1) they do not explore the correlation structure in the output variables or input variables because variables are not considered as time series, 2) they do not consider the random error in the estimation of efficiency, 3) they are designed for very short periods of time (two periods in Malmquist), 4) they do not use a statistical technique to estimate efficiencies, 5) they assume independence in the efficiency calculation of one period to another and 6) neither DEA model known in literature is “stochastic” and “time horizon” at the same time. By using multivariate normal mixture models, specifically the random effect mixture model, we introduce a new DEA in a time horizon, it is stochastic too, and it is an extension of the CCP. Instead of considering univariate random variables, as CCP does, time series are considered in a new model. The research includes the study of the antedependence structure and parameter estimation by using AECM algorithm in the mixture. The new model we propose is the only DEA model which is stochastic and time horizon at the same time, also it is the only one that accepts time series as output variables. By using multivariate mixture

models the characteristics of covariance of the variables are captured. This new DEA model could be applied in the development of a new DEA methodology in real time.

Contributed Poster Sessions

Quadratic Programming Wavelength Selection for NIR spectra classification

Alessandro Kahmann[1]

Joint work with Michel José Anzanello, Bruna Martini Dalmoro and Matias Segelis Vieira

([1]Federal University of Rio Grande do Sul, Brazil)

Abstract: In the last few years, near infrared (NIR) spectroscopy has gained wide acceptance in many research fields. NIR is a simple, quickly and non-destructive technique, that allows to see the chemical composition of samples without previous preparation. Typically NIR provides a fast knowledge of a large number of absorbance values for a selected spectral range. Because it quantifies this characteristic of many chemical components, the resulting database usually contain many overlapped and noised variables, jeopardizing the prediction of a response variable. In this scenario, multivariate statistical techniques are needed to complement the analysis, being the variable selection, also called "frequency" or "wavelength" selection when applied to spectroscopic data, the most widespread.

This paper proposes a multiple class classification wavelength selection method, with filter and wrapper approaches. Using Quadratic Programming, the method aim to minimize the Mutual Information among the wavelengths of same class samples, at the same time it maximize the Mutual Information between the wavelengths and the response variable. The result of Quadratic Programming is used as a Wavelength Importance Index (WII). On filter approach the variables are retained according to a WII cutoff, while on wrapper approach the wavelengths are added to the model by a forward selection. Different classification techniques are used to verify the quality of the retained subset and to determine the generalizability of the method.

We applied our proposition to five NIR datasets, related to different products. These five datasets are characterized by different number of samples, classes and wavelengths, offering a solid basis for assessing the robustness of our propositions. On the wrapper method, the average classification accuracy raises 1.82% when compared to the full data set, retaining only 0.7967% of the original wavelengths. On the filter method, the average classification accuracy increased 0.95%, retaining 0.7962% of the original wavelengths.

Comparison between methods for calculating the number of degrees of freedom of unbalanced split plot experiments using simulations

Andréia Pereira Maria Hilário[1]
Joint work with César Gonçalves de Lima
([1]University of São Paulo, Brazil)

Abstract: In this study, we used the SAS system to perform the simulation with different experimental scenarios and analyze data sets in randomized blocks design with scheme of split plot and different levels of unbalanced. The methods Containment, Residual, Satterthwaite and Kenward-Roger, which are used to calculate the approximate number of degrees of freedom of the denominator of the statistical F, were compared using as comparison criteria the power test and the type I error rate.

Inference in the Generalized Log-Gamma Nonlinear Regression Models

Audrey Helen Mariz de Aquino Cysneiros[1]
Joint work with Priscila G. da Silva, Gauss M. Cordeiro and Francisco José A. Cysneiros
([1]Federal University of Pernambuco, Brazil)

Abstract: Generalized log-gamma linear regression models (Lawless, 1980) are defined when the response variable has the log--gamma distribution and the regression parameters are lin ear. We propose a new class of models called the generalized log--gamma nonlinear regression models (GLGNL) by considering a nonlinear structure for the regression parameters. We derive general matrix formulae for the second--order biases of the maximum likelihood estimators of these parameters by extending the results by Young and Bakir (1987). We use the general formula by Cox and Snell (1968) and the bootstrap technique (Efron,1979) to obtain the bias--corrected estimators. Simulation studies are performed for selected parameter values and some bootstrap confidence interval estimation. We also present an empirical application to real data.

Evaluation of Multiobjective Optimization Algorithms: an Application in Economical Design of Online Control for Attribute with Misclassification Errors

Augusto dos Reis Pereira[1]
Joint work with Lupércio França Bessegato
([1]Federal University of Juiz de Fora, Brazil)

Abstract: Taguchi et al (1989) proposed an economical procedure for monitoring online process control for attribute. It consists of inspecting a single item at every mth items produced. The process starts operating in-control (fraction of conformance is equal to p1). After the occurrence of a special cause (the process is out-of-control), the fraction of conformance shifts to p2. If the inspected item is non-conforming, the process is designated as out-of-control and production is stopped for possible adjustment; otherwise, production goes on. The problem involves finding the optimum sampling interval (m) that minimizes the average cost of the system of control. This study adopts a generalization of this model by considering variable design parameters and statistical constraints. The model also considers that the decision regarding the process is subject to misclassification errors. In order to determine the best monitoring and control strategy, the model requires the search for several values that minimize the expected unit cost of the system and meet the statistical restrictions. These restrictions are considered for making the model more robust to uncertainty of their input variables. The probabilistic model of the control system employs properties of an ergodic Markov chain. The procedure is illustrated by a numerical example in which multiobjective optimization is used to find Pareto optimal solutions of the problem. The performances of some multiobjective optimization strategies are evaluated and compared.

Bayesian Analysis of Finite Mixtures

Brian Alvarez Ribeiro de Melo[1]

Joint work with Lupércio França Bessegato

([1]University of São Paulo, Brazil)

Abstract: Finite mixtures are highly flexible parametric models capable to describe different data features and are considered in many contexts, especially in the analysis of heterogeneous data. In medical studies, there is a lot of heterogeneity, since patients are not alike, making finite mixtures a very reasonable option for analyzing this kind of data. Generally, in finite mixtures, all the components come from the same parametric family and are only distinguished by the associated parameter vector. In this paper, we examine the survival times of 1400 patients with heart disease from the Heart Institute of Sao Paulo University by considering a new finite mixture model with three components from different parametric families, the GLW model, composed of the densities from the Gama, Log-normal and Weibull distributions. The goal is to develop a Bayesian analysis of the GLW model, suitable to the presence of (right, left and interval) censored data, in the study of survival times. We develop an R package that allow us to generate samples from the posterior distribution of the GLW parameters using MCMC methods, to perform hypothesis tests for the regression coefficients, which are incorporated through the mean of the survival time using a logarithmic link function, to compute the Log Pseudo Marginal Likelihood (LPML) so we can choose among different models, to estimate some quantities of interest using the predictive distribution of the survival time among other functions. Results on the analysis of survival times show that patients with Chagas disease survive less time than patients with other heart conditions and the survival time of male patients is shorter than women's.

Spatiotemporal Analysis by using Autoregressive and Smoothing Models

Sílvia Regina Costa Lopes[1]

Joint work with Giovani Loiola da Silva and Cleonis Viater Figueira

([1]Federal University of Rio Grande do Sul, Brazil)

Abstract: This work deals with the analysis of spatiotemporal data by using autoregressive and smoothing models in the context of mixed generalized linear/additive models under a Bayesian approach. The spatial random effects component is modeled through both the intrinsic conditional autoregressive (CAR) and the proper conditional autoregressive (PCAR) priors. The temporal component present on the data is modeled mainly through first and second orders autoregressive models, as well as smoothing with cubic B-spline without the intercept and fixed knots. The autoregressive process is introduced into the model structure by two different procedures: an autoregressive model for the mean of the Gaussian prior distribution, and an autoregressive model imposed additively to the model. The methodology is applied namely to the number of infant deaths data (under one-year-age) per household in the Brazilian states between 1991 and 2013 (source MS/SVS/DASIS - Mortality Information System).

Statistical Computing applied to KDD in toxicogenomics dataset

Diego Carvalho do Nascimento[1]

Joint work with Felipe Prata Lima and Paulo Jorge Leitão Adeodato

([1]University of São Paulo - São Carlos, Brazil)

Abstract: The development of technologies such as microarray and next generation sequencing (Next-generation sequencing), has driven studies of gene expression and expanded the application possibilities of these increasingly technology. A recent field study using these technologies is the Genomic Toxicology (English, Toxicogenomics), which seeks to assess the effects of exposure to chemicals in living organisms, analyzing its effects at the molecular level, with changes in expression profiles gene in exposed organisms. With the large amount of data being produced by these technologies, a challenge is characterized by identifying and recognizing patterns in these data, and the interpretation thereof to useful knowledge production. This paper presents data mining (known also as Knowledge Discovery databases, or just KDD) under an optical knowledge extraction process big data as a tool for data analysis in genomics Toxicology and knowledge generation in the area. A study example is shown, and show a scenario where it gives the opportunity for the development of new applications, by integrating and computing statistical techniques.

SYMARMA: A new dynamic model for temporal data on conditional symmetric distribution

Francisco José de Azevêdo Cysneiros[1]

Joint work with Felipe Prata Lima and Paulo Jorge Leitão Adeodato

([1]Federal University of Pernambuco, Brazil)

Abstract: Gaussian models of time series, ARMA, have been widely used in the literature. Benjamin et al. (2003) extended these models to the exponential family distributions. Also in that direction, Rocha & Cribari-Neto (2009) proposed a time series model for the class of beta distributions. In this work, we develop an autoregressive and moving average symmetric model, named SYMARMA, which is a dynamic model for random variables belonging to the class of symmetric distributions including also a set of regressors. We discuss methods for parameter estimation, hypothesis testing and forecasting. In particular, we provide closed-form expressions for the score function and Fisher information matrix. Robust study is presented based on influence function. We conduct simulation studies to evaluate the consistency and asymptotic normality of the conditional maximum likelihood estimator for the model parameters. An application with real data is presented and discussed. The data and codes in R may be obtained from http://www.de.ufpe.br/\sim cysneiros/elliptical/time_series.html

Data Mining For Time Series: Automatic Forecasting For Health Statistical Data Using Holt Winters And Box-Jenkins Arima Models

Gabriel Cordeiro

(University of Costa Rica, Costa Rica)

Abstract: This research intends to develop a tool for automatic modeling of time series, which is capable of generating forecasts for several variables concerning information of the Caja Costarricense de Seguro Social (CCSS) (Costa Rican health system). The analysis focuses on public data for 14 time series obtained from the website of this institution, all of which refer to medical care provided by the institution related to outpatient (7 variables) and hospitalization (7 variables). Two techniques are evaluated: Holt-Winters and Box-Jenkins ARIMA models. The procedure considers each series and calibrates a model using both techniques. The results are compared to each other to determine which technique produces the most accurate estimate in terms of the error between the predicted value and the observed value. Several adjustment indexes are calculated in order to recommend a model: the relative error (ER), the root mean square error (RMSE), the mean absolute deviation (MAD) and finally, the mean absolute percentage error (MAPE). A summary index is created to evaluate these measures and propose a model. Using the recommended model, the study proceeds to generate forecasts and their confidence intervals. Since this tool is designed for automatic modeling, it assumes that the intervention of an expert is not required; however, one of the hypotheses to be validated, is the possibility that this is not practicable, or at least not entirely, because many times it cannot be ignored the judgment of an analyst for verification of the assumptions, specifically in the case of the approach of Box-Jenkins ARIMA models, besides the validation of the consistency of the results.

A Weak Version of Bivariate Lack of Memory Property and Applications in Finance Industry

Hugo Branco

(University of São Paulo, Brazil)

Abstract: We suggest a modification of the classical Marshall-Olkin’s bivariate exponential distribution allowing a positive mass concentrated along arbitrary line through the origin. It serves as a base of a new weaker version of the bivariate lack of memory property (being “”non-aging”” or “”aging”” depending on parameter values involved). The corresponding copula is obtained and we establish its disagreement with Lancaster´s phenomena. Characterizations, properties and finance industry applications of the novel bivariate memory-less notion will be discussed. Computational difficulties related to statistical inference will be outlined as well.

Maximum entropy distribution on a circular region under mean value constraints

J.C.S. de Miranda

(University of São Paulo, Brazil)

Abstract: Maximum entropy distributions are a valuable tool in simulation studies where, in some sense, besides the information we already have about a probability structure, we want to assume the least additional information about it. Using variational methods we determine the maximum entropy probability distribution with support on a circular region under mean value constraints. More precisely, we determine the probability density function, $f_{XY},$ of a random vector (X,Y) such that $\mathcal{I}m(X,Y)\subset \mathcal{D},$ where $\mathcal{D}=\{ (x,y)\in \mathbb{R}^2: x^2 +y^2\le1\},$ that maximizes the entropy functional $f\rightsquigarrow-\int_\mathcal{D}f\ln f\mathrm{d}\ell$ and satisfies the mean value constraints $\mathbb{E}X=\mu_X$ and $\mathbb{E} Y=\mu_Y,$ where μ_X and $\mu_Y,$ such that $(\mu_X,\mu_Y)\in\mathcal{D},$ are given.

Binaries classification applying a Simulated Annealing algorithm based on Neighborhood Data.

Luis Eduardo Amaya Briceño

(University of Bío-Bío, Chile)

Abstract: Classification methods consist of a series of strategies that seek to determine groups, under the general principle that objects (individuals) belonging to the same group, present characteristics more similar to each other (with respect to some selected criteria previously) compared with individuals who were assigned to other groups. Because partitioning methods are optimal local criterion to optimize, has sought to improve with the use of heuristic combinatorial optimization methods such as simulated annealing, tabu search, or acceptance thresholds among others, they have provided superior results to they obtained with classical methods such as k-means, hierarchical classification or dynamic clouds. In this paper we focus on the classification of binary data using an algorithm simulated annealing because metaheurísiticas implementing this objective has been very limited. based on the concept of neighborhood algorithm it was implemented using two aggregation criteria: the sum and L1. This algorithm was applied to binary data tables different sizes and different characteristics, the results are compared with traditional methods, this to demonstrate the effectiveness of the implemented algorithm.

Comparing interpolation methods for precipitation data

Miriam Rodrigues Silvestre[1]

Joint work with Edilson Ferreira Flores and Vinicius Carmello

([1]São Paulo State University, Brazil)

Abstract: The spatial interpolation is a technique used to estimate local values that have not been sampled, by using data values observed in known locations.

It is quite usual in climatology map-making and monitoring of climatic elements from data collected at weather stations. However, usually the elements are spatially correlated, and in these cases it may be feasible to use interpolation methods to consider the presence of correlation between observations, such as the Geostatistics. In this work are analyzed precipitation data for the Paraná branch of Paranapanema River Basin and compared the results with the application of various methods of interpolation between them, Geostatistics and deterministic methods: Inverse Distance Weighted (IDW) interpolation global Polynomial (GPI), Radial basis Function (RBF) and interpolation Local Polynomial (LPI). The results were evaluated by the statistics root mean square error (RMSE) and visual analysis of the produced map, basically if it has artifacts, which are unwanted features. They were compared isotropic and anisotropic models generated in Variowin® 2.21 software and ArcMap® 10.1, in Default and Optimized versions, and deterministic models for the dry crop year 1999/00 and the rainy crop year 2009/10. For the dry year, the Exponential Geostatistical model was more suitable and between deterministic methods highlighted the LPI method. As for the rainy year, the most suitable was the Gaussian Geostatistical model and between deterministic methods, the best was the IDW. Regarding the visual analysis of the results produced by the methods, it was observed that the maps obtained with the application of Geostatistics technique with the parameters set by the user in the Variowin software, provides more smoothed out, and without the artifacts that are found in other methods.

Dynamic Modelling of Covariance Matrices: a Cholesky Stochastic Volatility Model

Paloma Vaissman Uribe[1]

Joint work with Hedibert Freitas Lopes

([1]University of São Paulo, Brazil)

Abstract: The estimation of the covariance matrix, the simplest summary measure of dependence of several variables, can be a challenging task, specially if the number of variables is much greater than the number of observations of each variable. In many areas like neuroscience, finance and energy there has been a growing interest and great progress in developing computationally fast methods that can handle high-dimensional multivariate data. Nevertheless, estimation of a covariance matrix based on high-dimension data is still an open problem. Another major obstacle in modelling covariance matrices is the positive-definiteness constraint. Our goal is to estimate a bayesian dynamic model for covariance matrix based on the modified Cholesky decomposition, which results in a unique unit lower triangular matrix and unique diagonal matrix. The entries of the lower triangular and the log of the diagonal matrix are unconstrained and can be interpreted as regression coefficients and prediction variances when regressing each variable on its predecessors. Our model considers dynamics for the variances (stochastic volatility) and use shrinking priors for the regression coefficients such as the Normal-Gamma prior in order to induce sparsity. The full model is estimated via MCMC methods. Because of the conditional independence between the sequence of autoregressions, the model can be efficiently implemented by parallel computing.

An approach via MIDAS regression models: an application the growth rate of US GDP

Paulo Henrique Sales Guimarães
(Federal University of Lavras, Brazil)

Abstract: This work used a class of models called MIDAS (Mixed Data Sampling), which allows the independent and dependent variables in the study are in heterogeneous frequencies. In this approach, the explanatory variables (at higher frequency) are weighted by that functions with few parameters can achieve good flexibility. Thus, the MIDAS models can utilize the full potential of all the information contained in the data, so that estimates a lag parameter for each of the high frequency variable. For the MIDAS models used were found smaller forecast errors than in classes of ARIMA models for forecasting the time series for the growth rate of the gross domestic product of the United States (GNP).

A simulation study to compare the power of the F test in longitudinal data with different probabilities of drop out

Reginaldo Francisco Hilário[1]
Joint work with Clarice Garcia Borges Demétrio
([1]ESALQ - University of São Paulo, Brazil)

Abstract: In longitudinal experiments, often the expected response is not achieved due to the loss of observations that can occur unpredictably. The researcher can anticipate the expected loss of observations and plan an experiment that has more accuracy in the results. A useful computational method proposed by Verbeke and Lesaffre (1999), allows exploring different combinations of models with probabilities of drop out associated and compare them with models of real answers. To explore the method, were simulated observations with characteristics of longitudinal experiments with three different probability constant of losses of individuals.

Methodology for satellite seasonal monitoring of Costa Rican forests under climate change using the Normalized Difference Vegetation Index (NDVI)

Ricardo Alvarado

(University of Costa Rica, Costa Rica)

Abstract: The Intergovernmental Panel on Climate Change (IPCC) has declared that Central American countries (including Costa Rica) are located in a zone where the climate change is likely to produce particularly pronounced effects on the environment, on the economy and on the society in general. Monitoring the forests is a key issue for ecosystem comprehension and their ability to provide ecosystem services. Forest ecosystems provide climatic regulation, throughout carbon capturing and storing from the atmosphere, as well as erosion control and flood mitigation. In this work we propose a methodology to develop a satellite seasonal monitor system of Costa Rican forests that uses the Normalized Difference Vegetation Index (NDVI). This methodology can be extended to be used with other vegetation indexes. We present the methodological idea and the first results that allow an understanding of the behavior of the vegetation as a time series. We propose two types of comparisons of the time series in order to detect anomalies. The first one compares two specific dates while the second one compares one date against the rest of the series. The results show that there is a lot of variability at national level which indicates that an analysis for climatic sub-regions would be more useful. Comparisons among sub-regions and spatial correlation are important tasks to be performed.

Generalized Transmuted Weibull Distribution: Structural Properties and Applications

Thais Campos Lucas[1]

Joint work with Maria do Carmo S. Lima, Gauss M. Cordeiro and Juan Carlos C. Capella

([1]Federal University of Pernambuco, Brazil)

Abstract: This article introduces and study the generalized transmuted Weibull (GT-W) distribution. We study some mathematical properties as quantile function, moments, generating function, mean deviations and Shannon entropy. The model parameters are estimated by maximum likelihood and the usefulness of the new distribution is illustrated by means of some real data sets. Besides that, we perform a Monte Carlo simulation (with 1000 replications) to quantify some asymptotic properties of the MLEs of the GT-W parameters.

ORGANIZATION:

ABE

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

iasc

FINANCIAL SUPPORT:

Instituto de
MATEMÁTICA
E ESTATÍSTICA
UFRGS

UFRGS
PROREXT
PRÓ-REITORIA DE EXTENSÃO

pro pesq
Pró-Reitoria de Pesquisa - UFRGS

FAPERGS
Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul

CAPES

Instituto
de Física