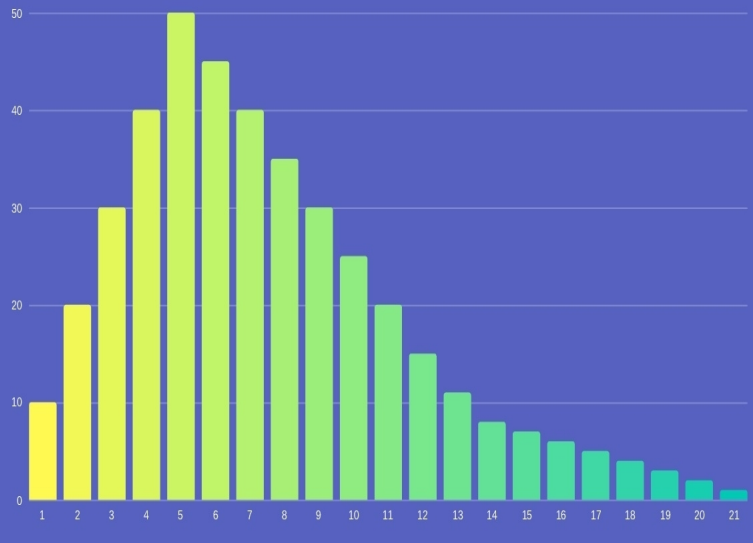
VII WORKSHOP ON PROBABILISTIC AND STATISTICAL METHODS

PROGRAM BOOK

Feb 13, 14, 15, 2019 - ICMC/USP, São Carlos/SP - Brazil

Meeting organized by the Joint Graduate Program in Statistics UFSCar/USP - PIPGEs



7th Workshop on Probabilistic and Statistical Methods

February 13–15, 2019

ICMC/USP, São Carlos, SP, Brazil

PROGRAM

ICMC/USP - DEs/UFSCar

About the 7WPSM

ICMC/USP, São Carlos, February 13–15, 2019

The Workshop on Probabilistic and Statistical Methods is an activity of the Joint Graduate Program in Statistics UFSCar/USP (PIPGEs), which brings together the research groups of probability and statistics working at ICMC-USP and UFSCar, in São Carlos/SP, Brazil.

The meeting intends to discuss new developments in statistics, probability and their applications. Activities include conferences and invited speaker sessions, contributed talks, poster sessions and a short course devoted to undergraduate/graduate students. The presentations of this new edition are related to probability and stochastic processes, statistical inference, regression models, survival analysis and related topics.

Organizing Committee

Daiane Zuanetti - UFSCar (Chair)
Katiane Conceição - ICMC/USP
Pablo Rodriguez - ICMC/USP
Ricardo Ehlers - ICMC/USP (Chair)
Sandro Gallo - UFSCar

Scientific Committee

Florencia Leonardi - IME/USP
Francisco Cribari - UFPE
Francisco Louzada Neto - ICMC/USP
Helio dos Santos Migon - UFRJ
Nancy Garcia - UNICAMP
Paulo Justiniano Ribeiro Junior - UFPR
Vera Tomazella - UFSCar

Support Committee

(students from PIPGEs)
Caio Moura Quina
Camila Sgarioni Ozelame
Deborah Bassi Stern
Gustavo Alexis Sabillón
Marcos Jardel Henriques
Victor Coscrato

Invited Speakers

Conferences

Alex Ramos - UFPE
Clarice Demétrio - ESALQ/USP
Juliana Cobre - ICMC/USP
Manuel Cabezas - Pontifícia Universidad Católica de Chile
Pedro Luis do Nascimento Silva - IBGE
Peter Müller - University of Texas at Austin
Rafael Izbicki - UFSCar
Robert Gramacy - Virginia Tech

Mini-Conferences

Anderson Ara - UFBA
José Augusto Fiorucci - UnB

Short Course

Anderson Castro Soares de Oliveira - UFMT

Special Sessions

Statistical Methods Applied to Genetic Data
(Chair: L. A. Milan)
Benilton de Sá Carvalho - UNICAMP
Júlia Maria Pavan Soler - IME/USP
Osvaldo Anacleto - ICMC/USP

Probability
(Chair: S. Gallo)
Alejandra Rada - CMCC/UFABC
Élcio Lebensztayn - IMECC/UNICAMP
Ludmila Rodrigues - IME/USP

7th Workshop on Probabilistic and Statistical Methods

February 13–15, 2019

ICMC/USP, São Carlos, SP, Brazil

SCHEDULE

ICMC/USP - DEs/UFSCar

WEDNESDAY 13 FEV

8h30 - 9h00 : Registration/Opening

9h00 - 10h00 : Peter Müller - University of Texas at Austin

10h00 - 10h30 : Coffee Break

10h30 - 11h30 : Rafael Izbicki - UFSCar

11h30 - 12h00 : José Augusto Fiorucci - UnB

12h00 - 14h00 : Lunch

14h00 - 15h00 : Juliana Cobre - ICMC/USP

15h00 - 16h00 : Oral Communications

- 15h00 - 15h20: Daisy Assmann Lima, Philipp Ehrl
  Universidade Católica de Brasília e Defensoria Pública da União
- 15h20 - 15h40: Demerson Andre Polli, Carlos Alberto de Ribeiro Diniz
  UFSCar/USP
- 15h40 - 16h00: Diego Nascimento, Osvaldo Anacleto, Lilia Costa, Taiza Santos, Francisco Louzada
  UFSCar/USP

16h00 - 17h00 : Coffee Break / Poster Session 1

17h00 - 18h00 : Clarice Demétrio - ESALQ/USP

THURSDAY 14 FEV

8h00 - 10h00 : Short Course by Anderson Castro Soares de Oliveira - UFMT

10h00 - 10h30 : Coffee Break

10h30 - 11h30 : Alex Ramos - UFPE

11h30 - 12h00 : Anderson Ara - UFBA

12h00 - 14h00 : Lunch

14h00 - 15h00 : Pedro Luis do Nascimento Silva - IBGE

15h00 - 16h00 : Oral Communications

- 15h00 - 15h20: Elizbeth Chipa Bedia, Vicente Garibay Cancho UFSCar/USP
- 15h20 - 15h40: Luciane Graziele Pereira, Sérgio Luiz Monteiro Salles-Filho UNICAMP
- 15h40 - 16h00: Victor Coscrato, Marco Inácio, Rafael Izbicki UFSCar/USP

16h00 - 17h00 : Coffee Break / Poster Session 2

17h00 - 18h00 : Manuel Cabezas - Pontifícia Universidad Católica de Chile

FRIDAY 15 FEV

8h00 - 10h00 : Short Course by Anderson Castro Soares de Oliveira - UFMT

10h00 - 10h30 : Coffee Break

10h30 - 11h30 : Robert Gramacy - Virginia Tech

11h30 - 12h00 : Closing

12h00 - 14h00 : Lunch

14h00 - 17h00 : Special Sessions

- *Statistical methods applied to genetic data (Chair: L. A. Milan):*
  - 14h00 - 15h00: Benilton de Sá Carvalho - UNICAMP
  - 15h00 - 16h00: Júlia Maria Pavan Soler - IME/USP
  - 16h00 - 17h00: Osvaldo Anacleto - ICMC/USP
- *Probability (Chair: S. Gallo):*
  - 14h00 - 15h00: Alejandra Rada - CMCC/UFABC
  - 15h00 - 16h00: Élcio Lebensztayn - IMECC/UNICAMP
  - 16h00 - 17h00: Ludmila Rodrigues - IME/USP

7th Workshop on Probabilistic and Statistical Methods

February 13–15, 2019

ICMC/USP, São Carlos, SP, Brazil

ABSTRACTS

ICMC/USP - DEs/UFSCar

Conferences

Alex Dias Ramos (Department of Statistics - UFPE)
Convergence time and phase transition in a non-monotonic family of probabilistic cellular automata

Abstract: In this talk, we will consider a one-dimensional probabilistic cellular automaton where their components assume two possible states, zero and one, and interact with their two nearest neighbors at each time step. Under the local interaction, if the component is in the same state as its two neighbors, it does not change its state. In the other cases, a component in state zero turns into a one with probability α, and a component in state one turns into a zero with probability $1-\beta$. For certain values of α and β, we show that the process will always converge weakly to δ_0, the measure concentrated on the configuration where all the components are zeros. Moreover, the mean time of this convergence is finite, and we describe an upper bound in this case, which is a linear function of the initial distribution. We also exhibit some results obtaines from mean-field approximation and Monte Carlo simulations, which show coexistence of three distinct behaviours for some values of parameters α and β. This work was developed joint with A. Leite.

Clarice Demétrio (ESALQ/USP)
Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data

Abstract: In the analysis of count data often the equidispersion assumption is not suitable, hence the Poisson regression model is inappropriate. As a generalization of the Poisson distribution the COM-Poisson distribution can deal with under-, equi- and overdispersed count data. It is a member of the exponential family of distributions and has the Poisson and geometric distributions as special cases, as well as the Bernoulli distribution as a limiting case. In spite of the nice properties of the COM-Poisson distribution, its location parameter does not correspond to the expectation, which complicates the interpretation of regression models specified using this distribution. In this paper, we propose a straightforward reparametrization of the COM-Poisson distribution based on an approximation to the expectation of this distribution. The main advantage of our new parametrization is the straightforward interpretation of the regression coeficients in terms of the expectation of the count response variable, as usual in the context of generalized linear models. Furthermore, the estimation and inference for the new COM-Poisson regression model can be done based on the likelihood paradigm. We carried out simulation studies to verify the finite sample properties of the maximum likelihood estimators. The results from our simula-

tion study show that the maximum likelihood estimators are unbiased and consistente for both regression and dispersion parameters. We observed that the empirical correlation between the regression and dispersion parameter estimators is close to zero, which suggests that these parameters are orthogonal. We illustrate the application of the proposed model through the analysis of three data sets with over-, under- and equidispersed count data. The study of distribution properties through a consideration of dispersion, zero-inflated and heavy tail indices, together with the results of data analysis show the exibility over standard approaches. Therefore, we encourage the application of the new parametrization for the analysis of count data in the context of COM-Poisson regression models. The computational routines for fitting the original and new version of the COM-Poisson regression model and the analyzed data sets are available.

Juliana Cobre (ICMC/USP, Brazil)
Why and how to measure the reliability of scientific co-authorship networks

Abstract: In this talk, we explain the association between a research group and a network and the different points of view that we can do it, more precisely what represents the nodes and the edges in such network. We justify why is important to measure statistically the reliability of scientific co-authorship networks, i. e., why it is not deterministic. In this study, we measure the reliability of networks by taking into account unreliable researchers and perfectly reliable edges. Some different inferential procedures presented and discussed are Bayesian inference using non-informative and informative priors.

This is joint work with Sandra Cristina de Oliveira (UNESP-Tupã)

Manuel Cabezas (Pontifícia Universidad Católica de Chile)
Hydrodynamic limit for the Atlas model

Abstract: The Atlas model is an interacting particle system where one starts with Poisson marks in $[0, \infty)$ which represent particles. As time runs, these particles perform independent Brownian motions, with the only exception being that, at any time, the leftmost particle has a drift to the right. Our main concern is to identify the possible behaviors of the leftmost particle (transience to the right, transience to the left, recurrence) as a function of the bias.

Pedro Luis do Nascimento Silva (IBGE)
Big data: potencial, paradoxos e a importância renovada do pensamento estatístico

Abstract: Vivemos numa era em que a disponibilidade e acessibilidade a dados não tem precedentes. 'Big data' é uma das tendências deste início do Milênio a confrontar o pensamento estatístico. Por um lado, há imenso potencial para aproveitar as novas fontes de informação que se tem tornado disponíveis, acessíveis e de baixo custo. Por outro lado, lacunas substanciais persistem e há imensos riscos de utilização inadequada dessas fontes pelos que desprezam as lições traduzidas nos principais fundamentos do pensamento e da metodologia estatística. Uma das falácias principais é a de que, com as imensas bases de dados disponíveis, não será mais preciso avaliar incerteza de estimativas, pois será possível 'conhecer' as quantidades de interesse a partir dos 'big data'. Apresenta-se o conceito de 'Índice de defeito dos dados' proposto por Meng (2018), e usa-se este conceito para mostrar que a qualidade de estimativas baseadas em pequenas amostras bem planejadas e executadas pode superar a de estimativas baseadas em conjuntos muito maiores provenientes de fontes orgânicas sujeitas a vieses de seleção. Penso que a metodologia estatística fornece a orientação essencial necessária para obter respostas atuais, relevantes, precisas e custo-efetivas às perguntas de interesse, mesmo na era do 'big data'. Apresentarei alguns exemplos para motivar a discussão dessas ideias.

Peter Müller (UT Austin, USA)
Bayesian Feature Allocation Models for Tumor Heterogeneity

Abstract: We characterize tumor variability by hypothetical latent cell types that are defined by the presence of some subset of recorded SNV's. (single nucleotide variants, that is, point mutations). Assuming that each sample is composed of some sample-specific proportions of these cell types we can then fit the observed proportions of SNV's for each sample. In other words, by fitting the observed proportions of SNV's in each sample we impute latent underlying cell types, essentially by a deconvolution of the observed proportions as a weighted average of binary indicators that define cell types by the presence or absence of different SNV's. In the first approach we use the generic feature allocation model of the Indian buffet process (IBP) as a prior for the latent cell subpopulations. In a second version of the proposed approach we make use of pairs of SNV's that are jointly recorded on the same reads, thereby contributing valuable haplotype information. Inference now requires feature allocation models beyond the binary IBP. We introduce a categorical extension of the IBP. Finally, in a third approach we replace the IBP by a prior based on a stylized model of a phylogenetic tree of cell subpopulations.

Lee, J., Müller, P., Ji, Y. and Gulukota, K. (2015) "A Bayesian Feature Allocation Model for Tumor Heterogeneity." Annals of Applied Statistics*, 9, 621-639.*

Zhou, T., Ji, Y., and Müller, P. (2017), TreeClone: Reconstruction of Tumor Subclone Phylogeny Based on Mutation Pairs using Next Generation Sequencing Data.

Zhou, T., Müller, P., Sengupta, S. and Ji, Y. (2016), PairClone: A Bayesian Subclone Caller Based on Mutation Pairs.

Xu Y, Müller P, Yuan Y, Gulukota K and Ji Y, (2015). "MAD Bayes for Tumor Heterogeneity – Feature Allocation with Exponential Family Sampling." Journal of the American Statistical Association*, 110, 503-514, PMID:26170513*

Lee, J., Müller, P., Sengupta, S., Gulukota, K. and Ji, Y. (2016), "Bayesian Inference for Intra-Tumor Heterogeneity in Mutations and Copy Number Variation", J. Royal Stat. Society C*, 65: 547?563. NIHMS #788717*

Rafael Izbicki (UFSCar, Brazil)

ABC-CDE: Towards Approximate Bayesian Computation with Complex High-Dimensional Data and Limited Simulations

Abstract: Approximate Bayesian Computation (ABC) is typically used when the likelihood is either unavailable or intractable but where data can be simulated under different parameter settings using a forward model. Despite the recent interest in ABC, high-dimensional data and costly simulations still remain a bottleneck. There is also no consensus as to how to best assess the performance of such methods. Here we show how a nonparametric conditional density estimation (CDE) framework can help address three key challenges in ABC, namely: (i) how to efficiently estimate the posterior distribution with limited simulations and different types of data, (ii) how to tune and compare the performance of ABC and related methods with CDE as a goal without knowing the true posterior, and (iii) how to efficiently choose among a very large set of summary statistics based on a CDE loss. We provide both theoretical and empirical evidence to justify the use of such procedures and describe settings where standard ABC may fail.

Robert Gramacy (Virginia Tech, USA)

Replication or exploration? Sequential design for stochastic simulation experiments

Abstract: In this paper we investigate the merits of replication, and provide methods that search for optimal designs (including replicates), in the context of noisy computer simulation experiments. We first show that replication offers the potential to be beneficial from both design and computational perspectives, in the context of Gaussian process surrogate modeling. We then develop a lookahead based sequential design scheme that can determine if a new run should be at an existing input location

(i.e., replicate) or at a new one (explore). When paired with a newly developed heteroskedastic Gaussian process model, our dynamic design scheme facilitates learning of signal and noise relationships which can vary throughout the input space. We show that it does so efficiently, on both computational and statistical grounds. In addition to illustrative synthetic examples, we demonstrate performance on two challenging real-data simulation experiments, from inventory management and epidemiology.

Mini-Conferences

Anderson Ara (UFBA, Brazil)
Redes Bayesianas: alguns métodos e aplicações

Abstract: Redes Bayesianas, também conhecidas como redes causais, redes de crenças ou redes probabilísticas de dependência, surgiram na década de 1980 e têm aplicadas em uma ampla variedade de atividades do mundo real. Em suma, são uma representação gráfica (grafo acíclico e direcionado) das variáveis e suas relações para um problema específico, sendo tal estrutura um elemento fundamental da rede. Nesta apresentação serão expostos alguns métodos de clássicos de construção da estrutura das redes e estimação de parâmetros, bem como aplicações recentes nas áreas financeira, biológica e educacional.

José Augusto Fiorucci (UnB, Brazil)
Time Series Forecasting and the Makridaks Competitions

Abstract: Accurate and robust forecasting methods for univariate time series are very important when the objective is to produce estimates for large numbers of time series. In this context, the Theta method's performance in the M3-Competition caught researchers' attention. The Theta method, as implemented in the monthly subset of the M3-Competition, decomposes the seasonally adjusted data into two "theta lines". The first theta line removes the curvature of the data in order to estimate the long-term trend component. The second theta line doubles the local curvatures of the series so as to approximate the short-term behaviour. We provide generalisations of the Theta method. The proposed Dynamic Optimised Theta Model is a state space model that selects the best short-term theta line optimally and revises the long-term theta line dynamically. The superior performance of this model is demonstrated through an empirical application. We relate special cases of this model to state space models for simple exponential smoothing with a drift.

Short-Course

Anderson Oliveira (UFMT, Brazil)
A utilização de redes sociais da internet para obtenção de dados

Abstract: À medida que as redes sociais da internet continuam se tornando integradas na vida cotidiana, os registros extensivos que estes sistemas arquivam como parte da operação normal prometem mudar os caminhos de pesquisa em várias areas de conhecimento. Assim, este minicurso discutirá o potencial da utilização destas redes sociais para levantamento de dados. Também será apresentado as limitações e dificuldades neste tipo de estudo. E por fim será apresentado alguns exemplos por meio da utilização de facebook e twitter.

Special Sessions

Statistical Methods Applied to Genetic Data

Benilton S Carvalho (UNICAMP, Brazil)
Identificação de Variantes Raras em Estudos Genômicos

Abstract: Ao longo da última década, com o desenvolvimento contínuo da tecnologia e metodologias analíticas, a geração de dados genômicos a partir de amostras biológicas teve seu custo reduzido em ordens de magnitude. Na década de 2000, o sequenciamento de um genoma completo custou aproximadamente 100 milhões de dólares. Atualmente, o custo é da ordem de 1.000 dólares, no mercado internacional. Como consequência deste processo, hoje é possível o sequenciamento de indivíduos em escalas maiores, chegando até o nível populacional. Assim, torna-se cada vez mais comum a realização de estudos para identificação de posições genômicas associadas com a ocorrência de fenótipos de interesse, como doenças complexas. O processo de análise de dados é composto por diversas etapas, que incluem filtros de diferentes tipos e modelos estatísticos para a avaliação efetiva de evidências de associação. Neste trabalho, apresentarei, da perspectiva analítica, as estratégias empregadas no Instituto Brasileiro de Neurociência e Neurotecnologia (BRAINN/FAPESP) nos estudos realizados com o intuito de identificar bases genéticas da epilepsia.

Júlia Maria Pavan Soler (IME-USP, Brazil)
Proteogenômica: a negação do one-size-fits-all

Abstract: A Proteogenômica inaugura uma nova fase de pesquisa multi-omics na Biologia Molecular, buscando integrar eficientemente grandes bancos de dados do genoma, transcriptoma e proteoma com informações clínicas. A promessa é identificar padrões específicos de pacientes e usar esse conhecimento na medicina personalizada e de precisão. Esta palestra tratará dos desafios interdisciplinares envolvidos, da abordagem e contribuição da estatística para essa área de pesquisa.

Osvaldo Anacleto (ICMC-USP, Brazil)
A stochastic transmission model to estimate social genetic effects in infectious diseases

Abstract: Current stochastic epidemic models ignore genetic heterogeneity in infectivity, which is the propensity of an infected individual to transmit diseases. Variation in this social interaction trait leads to the common superspreading phenomenon,

where a minority of highly infected hosts transmit the majority of infections. To date, is not known whether infectivity is genetically controlled. We present a novel stochastic transmission model which, by combining individual-level Poisson processes with bivariate random effects, can fully capture genetic variation in infectivity. We show that not only can this Bayesian model accurately estimate heritable variation in both infectivity and the propensity to be infected, but it also can identify parents more likely to generate offspring that are disease superspreaders. We also present a Bayesian analysis of a large-scale fish infection experiment which, for the very first time, shows that genetics does indeed contribute to variation in infectivity and therefore affects the spread of diseases.

Joint work with Andrea Doeschl-Wilson (U. of Edinburgh, Scotland) and Santiago Cabaleiro (CETGA, Spain)

Probability

Alejandra Rada (CMCC-UFABC, Brazil)
The Shortest Possible Return Time of β-Mixing Processes

Abstract: We consider a stochastic process and a given n-string. We study the shortest possible return time (or shortest return path) of the string over all the realizations of process starting from this string. For a β- mixing process having complete grammar, and for each size n of the strings, we approximate the distribution of this short return (properly re-scaled) by a non-degenerated distribution. Under mild conditions on the β coefficients, we prove the existence of the limit of this distribution to a non- degenerated distribution. We also prove that ergodicity is not enough to guaranty this convergence. Finally, we present a connection between the shortest return and the Shannon entropy, showing that maximum of the re-scaled variables grow as the matching function of Wyner and Ziv.

Élcio Lebensztayn (IMECC-UNICAMP, Brazil)
Phase transition for the frog model on trees

Abstract: The frog model is a stochastic epidemic model on a graph in which dormant particles begin to move and to infect other particles once they become infected. We study the frog model with geometric lifetimes on homogeneous and on biregular trees. With the help of branching processes, we obtain bounds for the critical parameter of the model.

Ludmila Rodrigues (IME-USP, Brasil)
Estimation of neuronal interaction graph from spike train data: method and application

Abstract: We address a basic question when analyzing experimental data in Neurobiology with respect to the the identification of the directed graph describing "synaptic coupling" between neurons. We present a novel estimator of effective connectivity, applying it to simulated and real data from a high quality multielectrode array recording dataset (Pouzat et al. 2015) from the first olfactory relay of the locust, Schistocerca americana. Our starting point is the procedure introduced in Duarte et al, 2016 and we present two novelties from the mathematical point of view: we propose a procedure allowing to deal with the small sample sizes met in actual datasets and we address the sensitive case of partially observed networks.

Oral Communications

Daisy Assmann Lima, Philipp Ehrl (Universidade Católica de Brasília e Defensoria Pública da União)
Individualistic Culture and Entrepreneurial Opportunities

Abstract: The present paper evaluates the effect of living in an individualistic society on entrepreneurial opportunities using cross-country data from the Global Entrepreneurship Index (GEI) in 2017.Individualism is a social trait that emphasizes freedom and rewards the personal achievements. We choose combine the fractional probit regression model with an instrumental variable approach through the conditional mixed-process (CMP) framework implemented by David Roodman's cmp command in the Stata. One of the reasons to use this method is that we can jointly estimate two or more equations with linkages among their error processes. And the individual equations need not be classical regressions with a continuous dependent variable. So instead of two linear OLS regressions in the usual 2SLS estimation, our IV-FPRM model performs a two-step procedure with two fractional probit estimations. Then we find that the number of opportunity startups (as opposed to necessity startups) is higher in individualistic countries. Part of this effect occurs because individualistic people perceive opportunities in a optimistic way, because they pretend to realize personal aims and because startups in individualistic countries are more innovative. These findings are robust to differences in institutions, religious affiliation, fertility, unemployment and education.

Demerson Andre Polli, Carlos Alberto de Ribeiro Diniz (UFSCar/USP)
Appraising products (or services) using a discrete scale: a generalization to the CUB model.

Abstract: The present paper evaluates the effect of living in an individualistic society on entrepreneurial opportunities using cross-country data from the Global Entrepreneurship Index (GEI) in 2017.Individualism is a social trait that emphasizes freedom and rewards the personal achievements. We choose combine the fractional probit regression model with an instrumental variable approach through the conditional mixed-process (CMP) framework implemented by David Roodman's cmp command in the Stata. One of the reasons to use this method is that we can jointly estimate two or more equations with linkages among their error processes. And the individual equations need not be classical regressions with a continuous dependent variable. So instead of two linear OLS regressions in the usual 2SLS estimation, our IV-FPRM model performs a two-step procedure with two fractional probit estimations. Then we find that the number of opportunity startups (as opposed to necessity startups) is higher in individualistic countries. Part of this effect occurs because individualis-

tic people perceive opportunities in a optimistic way, because they pretend to realize personal aims and because startups in individualistic countries are more innovative. These findings are robust to differences in institutions, religious affiliation, fertility, unemployment and education.

Diego Nascimento, Osvaldo Anacleto, Lilia Costa, Taiza Santos, Francisco Louzada (UFSCar/USP)
Multivariate time series high-dimensional shrinkage with dynamic graphical modeling

Abstract: Dynamic graphical models can be a good alternative to solve the challenge in analyzing/forecasting high-dimensional data. Represented as a graph, the estimation of network dynamics, through this class of models, aims to ensure stable inference and computationally feasible. Therefore, this work aims to compare the sparse Time Series Chain Graphical Model versus Multiregression Dynamic Model (MDM). As an exemplification, some extensions will be discussed regarding the empirical application which involves understanding causal mechanisms that underpin neural communication, using biosignals e.g. Electroencephalogram (EEG), regarding the safety of dose-response in electrical stimulation in a manipulation of human verticality.

Elizbeth Chipa Bedia, Vicente Garibay Cancho (UFSCar/USP)
Analysis of semi-competing risks data using Illness–Death processes with shared frailty inverse Gaussian: Application in colon cancer data.

Abstract: In semi-competing risks situation, which is a generalization of competing risks, only two events are generally considered, one terminal and one non-terminal. In this situation the terminal event (e.g. death) censors the non-terminal event (e.g. recurrence), while the occurrence of the non-terminal event does not prevent the terminal event from occurring. Usually, the two events are correlated. In this work, we study the modelling of semi-competing risks using the illness-death model with shared frailty. In this model the dependency between the terminal and non-terminal failure time is incorporated through of a shared frailty. We propose a shared frailty inverse Gaussian, as an alternative to the usually used Gamma. We introduce Weibull parametric models for the conditional transition rates. Maximum likelihood estimation is performed to fit the model to data set. First, a simulation study is provided to evaluate the performance of the maximum likelihood method in estimating parameters. Then the model is applied to colon cancer data sets.

Luciane Graziele Pereira, Sérgio Luiz Monteiro Salles-Filho (UNICAMP)
Data envelopment analysis (DEA) and multivariate analysis (MVA): integrating methods to analyze efficiency in the allocation of resources at UNICAMP

Abstract: Data Envelopment Analysis (DEA) is a non-parametric technique for comparing input and output data aiming to measure the efficiency of each decision making units (DMU). Through this analysis, it is possible to define an efficiency frontier, which can be used as a benchmarking for others DMUs. On the other hand, statistical multivariate analysis (MVA) scale models highlight important aspects of the information contained in the data, as with Multidimensional Scaling. These models work on the basis of proximity measures between pairs of objects. In Factorial Analysis (FA) and Principal Component Analysis (PCA), correlation coefficients are usually used, and scale models may use different measures of proximity. The aim of our research is integrate the Data Envelopment Analysis and statistical multivariate analysis to measure the efficiency of the different teaching and research units that compose UNICAMP. The concept of efficiency considers the observation of such units in the formation of people and in scientific production, from the use of human and financial resources. Utilizing our analysis, the policymakers can set goals for efficiency improvements, since efficiency in teaching and research are among the university's strategic objectives

Victor Coscrato, Marco Inácio, Rafael Izbicki (UFSCar/USP)
The NN-stacking: feature weighted linear stacking through neural networks

Abstract: Ensemble methods are known to be great boosters for prediction performance of regression methods. A simple way to ensemble regression estimators is by combining them linearly, as done by Breiman. Even though such approach is useful from an interpretative perspective, it often does not lead to great predictive power. In this work, we propose a nove lmethod called NN-Stacking, a linear ensembler that flexibilizes Breiman's method by allowing non static coefficients. The methods uses neural networks to optimize a squared loss function for obtaining optimal estimates for the stacking coefficients. We show that while our approach keeps the interpretative features of Breiman's method, it leads to better predictive power.

Poster Session 1

Alex Rodrigo dos Santos Sousa, Nancy Lopes Garcia, Brani Vidakovic (Universidade Estadual de Campinas)
Encolhimento bayesiano sob briori Beta de coeficientes de ondaletas em modelos com erros gaussianos e positivos

Abstract: Considere o problema de regressão não paramétrica de estimação de curva por meio da observação de pontos desta curva. Métodos de encolhimento de coeficientes de ondaletas são aplicados aos dados no domínio das ondaletas para redução de ruído para que a função possa ser estimada por expansão em bases de ondaletas. O presente trabalho propõe uma abordagem bayesiana de encolhimento de coeficientes de ondaletas com a utilização da distribuição beta com suporte em (-m,m) como distribuição a priori para os coeficientes das ondaletas em modelos com erros aleatórios aditivos gaussianos e positivos. Fórmulas explícitas para casos particulares das regras de encolhimento são obtidas, propriedades estatísticas como viés, risco clássico e bayesiano das regras são analisadas e performances das regras propostas são obtidas em estudos de simulações envolvendo as funções testes de Donoho-Johnstone. Aplicações em conjuntos de dados reais nas áreas de Espectrometria e Spike Sorting são feitas.

Amanda Morales Eudes D'Andrea, Vera Lucia Damasceno Tomazella (UFSCar/USP)
Fragilidade não paramétrica aplicada para dados de múltiplos sistemas reparáveis

Abstract: Em sistemas reparáveis, é muito importante considerar a escolha do modelo apropriado para modelar o tempo até a ocorrência do evento de interesse. Em muitas situações, o processo de Poisson não-homogêneo modela muito bem os dados. Mas a heterogeneidade entre os sistemas deve ser incluída, isso pode ser feito pelo modelo de fragilidade, isto é, a incorporação de uma variável aleatória não observada multiplicando a taxa de ocorrência de falhas. A proposta deste trabalho é considerar a fragilidade de forma não paramétrica e apresentar uma aplicação com um conjunto de dados real.

Ana Paula Jorge do Espirito Santo, Vicente Garibay Cancho (UFSCar/USP)
A new survival model for lifetime with long-term survivors and unobserved heterogeneity

Abstract: The Frailty models are often used to modeling unobserved dependence

and heterogeneity in the individual in survival data, wich can be modeled as an unobserved random variable acting multiplicatively on the baseline hazard function. In some situations, it may be appropriate to consider discrete frailty distributions, where individuals with long-term survival times include zero frailty (immune or cured). In this work, we develop a new survival model induced by discrete frailty with Katz distribution which can account for overdispersion, equidispersion and underdispersion. The new model encompasses as special case the mixture cure rate model, promotion cure rate model and cure rate model with dispersion, besides has the proportional hazards structure when covariates are modeled through θ that allows a study of the risk ratio independ on time t. We construct a regression model to evaluate the effects of covariates in the cured fraction, we discussed inference aspects for the proposed model in a classical approach, an EM algorithm is then developed for determining the maximum likelihood estimates of the parameters. Finally, the modeling is fully illustrated on a data set on colorectal cancerin which we observe that the therapy with Levamisole+5-FU increases the lifetime of patients and cured fraction.

Átila Prates Correia (ICMC-USP)
Kolmogorov extended axioms and negative probabilities

Abstract: As an attempt to extend the concept of probability to negative numbers, the content to be exposed in this poster makes use of measure theory in order to achieve this goal. More precisely, given a measurable space (Ω, Σ), we associate to it a finite signed measure P under convenient restrictions (axioms) which turn feasible the concept of negative probability. Although we do not provide an interpretation for the negative probabilities, we propose the structure such concept should fit in.

Breno Gabriel da Silva , Yana Miranda Borges , Brian A. Ribeiro de Melo (Universidade Estadual Maringá)
Ajuste de extensões da distribuição weibull para dados meteorológicos do estado de Goiás

Abstract: Quando o objetivo é descrever situações cujo interesse é analisar o tempo até a ocorrência de um evento, o ponto de partida é escolher uma distribuição de probabilidade adequada ao problema, como nos casos de analisar o tempo até a falha de um componente eletrônico, crescimento de uma determinada bactéria, evolução de alguma doença, entre outros. Se a distribuição candidata não se ajustar bem aos dados, temos então que as estimativas serão imprecisas. Este trabalho tem como objetivo avaliar a capacidade de ganho real no ajuste de onze distribuições de probabilidade derivadas da distribuição Weibull, uma vez que nos últimos tempos foram

sendo criadas inúmeras extensões desta. Em relação aos materiais, utilizaremos um conjunto de dados de precipitação, somente em que esta é maior que zero, do estado de Goiás no período de maio de 2017 a maio de 2018. Em relação aos métodos, utilizaremos o teste não-paramétrico de Kolmogorov-Smirnov (KS), que verifica se determinada distribuição teórica ajusta-se bem à distribuição empírica dos dados e o critério de informação de Akaike (AIC) para a seleção do modelo adequado. Assim, com o modelo selecionado pretendemos em trabalhos futuros verificar a influência de covariáveis no índice de precipitação de uma determinada região.

Camila Ozelame, Anderson Ara, Francisco Louzada (UFSCar/USP)
Redes bayesianas para classificação: diferentes métodos e comparação

Abstract: Em problemas de classificação, inúmeras técnicas estatísticas e computacionais podem ser aplicadas, uma possibilidade é dada pela metodologia de Redes Bayesianas (RBs) com estruturas específicas, portanto, também conhecidas como classificadores Bayesianos. As RBs unem as teoria de grafos e de probabilidades atuando com maestria em dois dos objetivos que dominam a análise de dados, a descrição e a predição. As RBs conseguem identificar a estrutura, mais provável, das variáveis analisadas, bem como a maneira que interagem umas com as outras. Utilizando a noção gráfico-teórica de d-separação são capazes de identificar grupos de possuam suposição de independência condicional e fornecem a distribuição a posteriori de uma variável resposta dado um conjunto de variáveis explicativas. Diversos classificadores são apresentados na literatura, cada uma com uma abordagem diferente a respeito das probabilidades conjunta e condicional entre as variáveis estudadas. Dentre elas destacam-se: Naïve Bayes (NB); Tree Augmented Naive Bayes (TAN); Bayesian Network Augumented Naive Bayes (BAN); General Bayesian Network (GBN); K-Dependence Bayesian Network (KDB); Averaged One-Dependence Estimator (AODE). Neste trabaho, realizamos uma comparação da capacidade preditiva de tais classificores utilizando benchmarkings de dados reais e artificias. Para quantificar a qualidade do das estruturas propostas são utilizadas quatro medidas de performance de desempenho: sensibilidade, especificidade, acurácia e o coeficiente de correlação de Matthew.

Carolina Grejo, Pablo Martin Rodriguez (UFSCar/USP)
A general Maki-Thompson model with directed inter-group interactions

Abstract: In this work we considered a new extension for the Maki-Thompson model which incorpores inter-group directed contacts. We obtain limit theorems for the remaining proportion of ignorants and discuss some examples and possible applications.

Cristel Ecaterin Vera Tapia (UFSCar/USP)
Estimation of the number of communities in the stochastic block model

Abstract: The Stochastic Block Model was introduced by Holland et all. (1983) and falls in the general class of random graphs. In this model, the nodes are classied in groups or communities, such that, considering to each node in the graph a associated latent discrete random variable, describing its community label, the probability of a edge connecting two nodes, depends only on the values of the latent variables. In this context, Andressa Cerqueira and Florencia Leonardi (2018) proposed the Krichevsky-Tromov estimator for the number of communities in the stochastic block model. Based on these results, we consider in this work a simple extension, to include the possibility of existence of a nite number of edges between pairs of nodes.

Danilo Sarti, Carlos Tadeu dos Santos Dias (USP - LCE)
Fitting probabilistic models to data and risk exposure

Abstract: This paper aims to show how a computational methodology can be applied to fitting different probabilistic models to data. Simulations of random variables are used to generate datasets. Cullen and Frey graphical methodology is explored as a guideline for the candidate distributions to model the data. In order to build such graphics higher order moments such as kurtosis and symmetry are used and discussed. Parametric distributions are fitting to the data using likelihood, method of moments as long as other methods focusing on fitting optimization. Several goodness of fitness measures are applied and compared for discussion of their uses and implications. An example of how such method can be used for assessing exposure to risk is made via a R dataset from packages avaiable at CRAN website.

George Lucas Moraes Pezzott, José Galvão Leite, Luis Ernesto Bueno Salasar (UFSCar/USP)
Spatial capture-recapture model for open population

Abstract: In this work we propose a spatial capture-recapture model to estimate the number of animals in an open population. The statistical model conform to data obtained through individual tag capture-recapture sampling performed in different areas within the habitat, taking into account the rates of births and deaths during the study period, the geographical locations of the catches and the movement of the animals between the sampling regions. Therefore, we was able to estimate the number of the live animals in each occasion and sampling region as well as the size of the area of movement of the species. We propose a Bayesian approach to the inferential process

and derive a MCMC algorithm from simple computational implementation, from the use of augmented data techniques. Our proposed methodology is illustrated in a real capture-recapture data set of arachnids in a cave.

Gilson Yuuji Shimizu, Rafael Izbicki (UFSCar/USP)
Conformal Prediction via Densidade Condicional

Abstract: Os métodos de machine learning tem como principal finalidade a predição de uma variável resposta para um dado conjunto de variáveis explicativas. Geralmente, estamos interessados apenas na esperança condicional desta variável resposta. Contudo, quando a distribuição da variável resposta é complexa, apenas a esperança condicional pode ser insuficiente para descrever a incerteza desta predição. Neste trabalho, nós propomos um método para estimar uma banda de predição, num contexto de regressão, onde nenhuma suposição é feita sobre a distribuição dos dados e qualquer método de machine learning pode ser utilizado. O método conformal prediction é combinado com a estimação de densidade condicional para a construção desta banda. Exemplos com dados simulados são apresentados para demonstrar a eficácia do método. Palavras chave: machine learning, conformal prediction, banda de predição e densidade condicional.

Glauber Márcio Silveira Pereira, Carlos Alberto Ribeiro Diniz (UFSCar)
Distribuição COMPoisson generalizada parcialmente correlacionada

Abstract: Desenvolvemos a distribuição COM-Poisson generalizado parcialmente correlacionado (CPGPC), é uma generalização da distribuição Poisson generalizada parcialmente correlacionada desenvolvida por (Luceño, 1995) e COM-Poisson desenvolvida por Shmueli, G Minka & Boatwringht (2005). Desenvolvemos a estimação dos parâmetros com uma abordagem clássica. Os estimadores de máxima verossimilhança e estimadores de momentos dos parâmetros são determinados. Intervalos de confiança normal e bootstrap são construídos para os parâmetros da distribuição. Fazemos simulação para os casos citados.

Isis F. Mascarin, Katiane S. Conceição (UFSCar/USP)
Distribuições discretas zero-modificadas para modelar dados de contagem zeros faltantes

Abstract: Em diversas situações práticas onde o conjunto de dados constitui-se de observações de contagem, é possível notar a ausência de observações zero. A ocor-

rência deste cenário pode ser devido à probabilidade nula de ocorrência de tal valor ou de algum problema de amostragem. No caso em que a probabilidade de observação zero é positiva, deve-se assumir uma distribuição zero-deflacionada em vez de uma zero-truncada, para obter de forma correta as estimativas dos parâmetros da distribuição que explica adequadamente o comportamento destes dados, bem como a frequência esperada de zeros. Neste trabalho, o interesse é fazer tais estimações por meio do algoritmo EM, utilizando distribuições zero-modificadas adequadas para o estudo. O conteúdo inclui introdução às famílias de distribuições Série de Potência (PS) e Série de Potência Zero-Modificada (ZMPS), incluindo sua versão hurdle, a partir da qual foram desenvolvidos os cálculos para obtenção dos estimadores de máxima verossimilhança dos parâmetros da distribuição ZMPS. Além disso, apresenta-se alguns resultados iniciais obtidos de simulação aplicada à distribuição Poisson, que pertence à família PS.

Jaime Phasquinel Lopes Cavalcante, Luciana Moura Reinaldo (Universidade Federal do Ceará - UFC)

Modelos lineares generalizados (MLGs) e sua aplicação em Ciências Atuariais

Abstract: A utilização de métodos estatísticos na rotina da Ciência Atuarial tem desenvolvido, historicamente, um papel central, tanto nos assuntos teóricos quanto práticos. Nesse sentido, corroborando com estudos outrora publicados, o presente estudo possui como objetivo principal demonstrar a aplicação da metodologia dos Modelos Lineares Generalizados com foco em uma problemática atuarial. A justificativa para o estudo surge do fato de que há uma vasta área de aplicações para o MLG (HABERMAN; RENSHAW, 1996), mas pouco exploradas, especialmente no Brasil. Diante do exposto, considerou-se a base contida em Kaas et al. (2008), que reflete a experiência anual de um portfólio de seguros de automóveis. Com isso, buscou-se relacionar a frequência de sinistros aos fatores de risco: sexo, região, tipo do carro, situação laboral. Portanto, ajustou-se um MLG com relação média-variância do tipo Poisson com função de ligação canônica. Ademais, são apresentados os diagnósticos do modelo que confirmam o bom ajuste da modelagem e a análise dos desvios. Finalmente, foi possível determinar que, em média, um motorista com os piores fatores, em comparação com aquele que tem a melhor combinação, realiza um sinistro a cada 6,6 anos.

João Vitor Magri da Silva, Elisangela Aparecida da Silva Lizzi (Universidade Tecnológica Federal do Paraná)

Modelagem bayesiana espacial aplicada à taxa de homicídios por violência nas unidades federativas brasileiras.

Abstract: Introdução: A taxa de homicídios e índices de violência no Brasil são elevados, quando comparados a taxas internacionais, em regiões e estratos específicos da população estas taxas são superiores à países latinos-americanos envolvidos em guerrilhas internas e conflitos militares. Objetivo: Modelar a relação entre taxas de homicídios por violência, IDH e o efetivo de policiamento militar distribuídos em todas as 27 unidades federativas do Brasil. Métodos: Os dados relativos à taxa de homicídios por violência, usou o filtro de diferentes causas, como: violência por gênero, mortes violentas por causa indeterminada e óbitos por armas de fogo. Neste cenário, foi proposta uma modelagem bayesiana espacial, usando-se modelo espacial com efeito aleatório BYM, onde modelou-se a dependência espacial com matriz de adjacência de ordem 1 que representa a localização de cada UF e seus respectivos vizinhos de primeira ordem. O número de policiais por 100 mil habitantes e o IDH entrou como preditores. O modelo foi implementado com apoio computacional do software R e utilizou métodos de aproximação de Laplace (INLA). Resultados: Não foi possível estabelecer uma relação entra a taxa de homicídios e o policiamento, porém em relação ao IDH quanto menor o índice maior a taxa de violência naquela região.

Jonathan Kevin Jordan Vasquez, Josemar Rodrigues (USP)
Modelo de Sobrevivência em dois estágios com taxa de cura

Abstract: Este projeto foi motivado pela análise dos dados de sinusite em pacientes infetados pelo HIV (GIOLO; COLOSIMO, 2006). A estimativa de Kaplan-Meier da função de sobrevivência sugeriu a existência de pacientes imunes a sinusite tornando-se inadequado a utilização dos modelos tradicionais em Análise de Sobrevivência. Para verificar se a infecção pelo HIV aumenta o risco da ocorrência da sinusite, propomos neste projeto o modelo Exponencial e Weibull em dois-estágios (Rodrigues, J. Teoria Unificada de Análise de Sobrevivência, 2009) com taxa de cura.

Katy Rocio, Alex Mota, Vera Tomazella, Vicente Calsavara (UFSCar/USP)
Modelo de fragilidade discreta zero inflacionado poisson

Abstract: Os modelos de fragilidade são utilizados para a modelagem de heterogeneidade na análise de dados de sobrevivência. Na analise desses dados a distribuição da fragilidade, em geral, é assumida contínua. Entretanto, existem algumas situações nas quais é apropriado considerar a fragilidade distribuída discretamente, por exemplo, quando a heterogeneidade dos tempos de sobrevivência surge por causa da presença de um número aleatório de falhas por unidade ou pela causa da exposição a danos em um número aleatório de ocasiões. Neste trabalho estendemos os modelos

de fragilidade de riscos proporcionais permitindo distribuições discretas, em particular a distribuição Zero Inflacionado Poisson (ZIP). Neste contexto podemos observar a possibilidade de indivíduos com fragilidade zero que corresponde a um modelo de falha limitado que contém uma proporção de unidades que nunca falham (sobreviventes de longa duração ou modelo com fração de cura).

Poster Session 2

Lucas Leite Cavalaro, Gustavo Henrique de Araujo Pereira (UFSCar/USP)
Comparação de métodos de seleção de variáveis em MLGD com covariáveis correlacionadas

Abstract: Os modelos lineares generalizados duplos (MLGD), diferentemente dos modelos lineares generalizados (MLG), permitem o ajuste do parâmetro de dispersão da variável resposta em função de variáveis preditoras, aperfeiçoando a forma de modelar fenômenos. Desse modo, os mesmos são uma possível solução quando a suposição de que o parâmetro de dispersão constante não é razoável e a variável resposta tem distribuição que pertence à família exponencial. Considerando nosso interesse em seleção de variáveis nesta classe de modelos, estudamos o esquema de seleção de variáveis em dois passos proposto por Bayer e Cribari-Neto (2015) e, com base neste método, desenvolvemos um procedimento para seleção de variáveis em até ""k"" passos. Para avaliar o desempenho do nosso procedimento em dados que apresentam comportamento próximo ao observado na prática, realizamos estudos de simulação de Monte Carlo em MLGD considerando covariávis com diferentes correlações. Os resultados obtidos indicam que o nosso procedimento para seleção de variáveis apresenta, em geral, performance semelhante ou superior à das demais metodologias estudadas sem necessitar de um grande custo computacional.

Luís Felipe Barbosa Fernandes,Evandro Marcos Saidel Ribeiro (USP)
Modelagem de crédito com a técnica de redes bayesianas aplicadas ao segmento de indústrias de alimentos e bebidas

Abstract: O trabalho consiste na elaboração de um modelo Bayesiano capaz de prever o risco de inadimplência de uma empresa com base em um conjunto de oito indicadores financeiros. Para isso, foi usado um banco de dados da Serasa Experian formado por um conjunto de 368 empresas, com balanços patrimoniais para três períodos consecutivos. O trabalho divide-se em duas etapas: modelagem e construção do aplicativo. A modelagem foi realizada através das técnicas de padronização dos dados, categorização e seleção do modelo. Após essa etapa, a Rede Bayesiana resultante foi elaborada através de um aplicativo. Esse aplicativo, desenvolvido através do pacote Shiny, do RStudio, permite que um usuário obtenha o risco de uma empresa, a partir da atribuição de valores para os oito indicadores financeiros selecionados. O usuário deve digitar os valores dos oito indicadores financeiros da empresa em questão. Após digitar os valores, o aplicativo exibe os valores digitados padronizados e categorizados em uma tabela visível ao usuário. Os valores categorizados são submetidos à rede, que calcula a probabilidade da empresa pertencer á cada uma das 22 classes de risco. A classe com maior probabilidade é a classe de risco da empresa.

Luiz Carlos Medeiros Damasceno, Luís Aparecido Milan (IFNMG/UFSCar/USP)
Modelos de mistura aplicados a dados com dependência espacial

Abstract: O objetivo desse poster é apresentar instrumentos para inferência para modelos de misturas com dependência espacial entre as variáveis usando metodologia de M.V. e/ou bayesiana. Entre os aspectos a serem tratados esta a seleção de modelos, estimação via MCMC, bem como verificação de performance. Serão realizados estudos de simulação para verificar a performance dos métodos propostos e comparação com métodos propostos na literatura. Os métodos propostos também serão aplicados a dados reais.

Luiz Gabriel Fernandes Cotrim, Daiane Aparecida Zuanetti (UFSCar/USP)
Inferência Bayesiana para Modelos de Mistura de Regressão: uma aplicação em dados educacionais brasileiros

Abstract: É comum que a variável resposta não se relacione com as covariaveis de forma homogênia para toda a populacão. Uma maneira interessante de modelar esses dados é considerar que a heterogeneidade observada surgiu devido ao fato de que a população é composta por K subpopulações e que as variaveis a serem analisadas se relacionam de maneira diferente em cada subpopulação. Se as subpopulações são desconhecidas, podemos assumir que a distribuição da variável resposta condicionada as covariaveis é um modelo de mistura. Denominamos esta classe de modelos de modelos de mistura de regressão. Apresentamos metodos de estimação e seleção Bayesianos para o caso em que temos uma mistura de regressões normais, comparamos os resultados ao algoritmo EM e aplicamos a metodologia em dados educacionais brasileiros, considerando o índice de desenvolvimento da educação básica (IDEB) como nossa variavel de interesse.

Marcelo da Silva, Jorge Bazán (UFSCar/USP)
Validação da matriz Q em modelos da teoria da resposta ao item multidimensionais

Abstract: A matriz Q é um componente bastante simples e intuitivo utilizado originalmente por uma nova classe de modelos de variáveis latentes multidimensionais, conhecida como modelos de diagnóstico cognitivo (MDC) com o objetivo de especificar a relação item-traço em um instrumento de medição. Recentemente, a matriz Q foi incorporada nos modelos da teoria da resposta ao item multidimensionais (TRIM). A construção da matriz Q é tipicamente feita por especialistas no tema dos itens, ou seja, é um processo subjetivo que pode implicar em equívocos e, consequentemente, resultar em importantes implicações práticas. Assim, a verificação da exatidão da

matriz Q faz-se necessário tanto em MDC como nos modelos da TRIM. Baseando-se no método para validar teoricamente as especificações da matriz Q apresentado por Jimmy de la Torre em MDC, propomos um método de validação da matriz Q em modelos da TRIM. Para isso, definimos e comparamos o comportamento de alguns critérios para buscar uma matriz Q adequada aos dados, adaptamos um algoritmo de busca de matriz chamado Algoritmo de Troca por Ponto e realizamos um estudo de simulação para avaliar o desempenho do método proposto.

Naiara Caroline Aparecido dos Santos, Breno Gabriel da Silva, Yana Miranda Borges, Brian A. Ribeiro de Melo (Universidade Estadual de Maringá)

Modelo de regressão quasi-poisson para dados de tentativas de suicídio por intoxicação exógena do estado de São Paulo

Abstract: Segundo a Organização Mundial da Saúde (OMS), a intoxicação exógena é um dos três principais motivos de tentativas de suicídios no mundo. Assim, objetiva-se verificar qual grupo (sexo e faixa etária) possui a maior taxa de tentativas de suicídio por intoxicação exógena. Desta forma, foi realizado um estudo dos registros de casos do Estado de São Paulo, Brasil. Os dados foram extraídos do Sistema de Informações sobre Mortalidade (DATASUS) e características populacionais do Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2017. As análises estatísticas foram realizadas no software R versão 3.5.0, em que utilizou-se o modelo de Regressão Poisson, também conhecido como Modelo Log-Linear de Poisson, o qual pertence a família de Modelos Lineares Generalizados. Inicialmente ajustou-se um modelo log-linear de Poisson, que apresentou fortes evidências de superdispersão, indicando que o modelo não era o mais apropriado. Sendo assim, ajustou-se o modelo quasi-Poisson, de modo a acomodar a variabilidade presente, mostrando-se adequado. Verificou-se que homens cometem suicídio 0.376 vez mais que as mulheres, e que a maior taxa de tentativa de suicídio por intoxicação ocorre entre 20 e 39 anos. Por outro lado, a menor taxa ocorre em indivíduos acima de 80 anos.

Oilson Alberto Gonzatto Junior, Marcos Jardel Henriques, Camila Sgarioni Ozelame, Anderson Ara, Mariana Curi, Francisco Louzada Neto (UFSCar/USP)

Um Modelo de regressão com respostas Beta Inflacionada e componentes principais

Abstract: A distribuição Beta é comumente utilizada para descrever o comportamento de uma variável aleatória com suporte em um intervalo aberto e, em particular, quando a problemática envolve uma proporção, o intervalo (0,1). Essa distribuição

pode ser estendida com a adição apropriada de pontos de interesse ao seu suporte, o que caracteriza a distribuição Beta Inflacionada. Se o objetivo é examinar a existência e quantificar a influência de fatores que atuam sobre o comportamento de uma proporção de interesse, o modelo de Regressão Beta Inflacionado é uma alternativa interessante. Se houver uma alta quantidade de covariáveis possivelmente correlacionadas, pode-se associar o modelo de regressão com a técnica das componentes principais. Para ilustrar a aplicação de tal metodologia, um conjunto de dados reais foi utilizado.

Oscar Holguín Villamil, Jonas Rafael Dos Santos (Universidade Estadual Paulista "Júlio de Mesquita Filho")

IRAMUTEQ y ATLAS.TI en la descomposición genética como herramientas para el análisis de contenido de la práctica docente y de políticas públicas: aplicación del análisis de correspondencia.

Abstract: El trabajo describe el proceso de categorización, clasificación y la implementación de la técnica de estadística multivariada de análisis de correspondencia en el desarrollo del análisis de primer orden que arrojan los datos obtenidos en una investigación de tipo bibliográfico cuyo objetivo por una parte es la caracterización de la práctica educativa de docentes universitarios en la experiencia de la especialización en educación en tecnología de la UFScar y los componentes del concepto EaD apropiados por un experto investigador del campo. Y por otra la identificación de patrones de comparación entre programas de política educativa de primera infancia implementadas en Brasil y en Colombia. Dada la naturaleza de rastreo documental y sistematización de experiencias; se desarrolla un planteamiento inferencial de descomposición genética del fenómeno lingüístico y comunicativo que soporta los dos estudios y desde los cuales es posible el trabajo analítico de la estadística; para ello se implementan los programas ATLAS.ti e IRAMUTEQ, los cuales permiten almacenar todos los datos, la codificación, categorización y el análisis inferencial de los resultados obtenidos a partir de la descomposición genética como opción de análisis de contenido y modelación en investigación cualitativa.

Patty Mercedes Arce Flores (Pontificia Universidad Católica del Perú)

Estimativas para la probabilidad de eventos sorpresivos en una cadena de Markov.

Abstract: En este trabajo estudiaremos el siguiente aspecto sobre cadenas de Markov con espacios de estados finitos S. Denotaremos por $\tau(y)$ el tiempo que le toma a la cadena en llegar por primera vez a un estado y. Fijado un tiempo t y un estado inicial x, buscamos estimar la probabilidad de que la cadena llegue por primera

vez al estado "y" en "t" pasos, es decir buscamos estimar el valor de $Px(\tau(y) = t)$. *En este trabajo enunciaremos cotas superiores para esta probabilidad, que dependen del tamaño del espacio de estado y el tiempo "t" y no tanto de las probabilidades de transición que tiene la cadena. Estas estimaivas son Obtenidas por J. Norris et y nuestro trabajo de tesis se basa en ese artículo. Las estimativas que presentaremos en este trabajo son las siguientes: (a) Para una cadena de Markov con n estados,* $Px(\tau(y) = t) \leq n * t$. *(b) En una cadena de Markov reversible con n estados,* $Px(\tau(y) = t) \leq (2n * t)^{(1/2)}$ *para* $t \geq 4n + 4$ *(para un tiempo "t" suficientemente grande). (c) Para un camino aleatorio sobre un grafo simple con* $n \geq 2$ *vertices,* $Px(\tau(y) = t) \leq 4e(logn) * t$. *En dicho artículo se construyen ejemplos que muestran que estas cotas están muy cerca de ser óptimas.*

Roberta de Souza, Carlos A. R. Diniz (UFSCar/USP)
Modelo de regressão geométrico de ordem k: uma aplicação em operações de crédito.

Abstract: Atrasos em operações de crédito, como no pagamento de consecutivas parcelas de empréstimo, faz com que o cliente se torne inadimplente e possa ser transferido do sistema bancário para uma empresa de cobranças. Algumas políticas internas bancárias consideram 90 dias em atraso o tempo para um cliente se tornar inadimplente, o que corresponde a três parcelas consecutivas não pagas. A sequência de parcelas pagas em dia e em atraso, até que três parcelas consecutivas em atraso ocorram, pode ser observada numa sequência de respostas binárias e a quantidade de respostas desta sequência representa uma variável com distribuição geométrica de ordem k (Philippou & Muwafi, 1980). Neste trabalho, um modelo de regressão geométrico de ordem k foi proposto para a avaliação de covariáveis na probabilidade de atraso da parcela de crédito, como também no tempo médio até a inadimplência, em clientes inadimplentes de um banco brasileiro (neste caso, k=3). Os parâmetros foram estimados por método Bayesiano, neste caso as distribuições a posteriori dos parâmetros foram obtidas por método de simulação estocástica de MCMC utilizando o algoritmo de Metropolis-Hastings (Gamerman & Lopes, 2006). O diagnóstico bayesiano para ajuste do modelo e verificação de observações influentes foi analisado por resíduos quantílicos aleatorizados (Jiang et al., 2013) e medidas de divergência (Peng & Dey, 1995).

Solange Ferreira Silvino, Yana Miranda Borges, Breno Gabriel da Silva, Naiara Caroline Aparecido dos Santos, Brian Alvarez Ribeiro de Melo. (Instituto Federal de Roraima - IFRR)
Estudo de modelos de previsão para demanda de processos distribuídos em uma unidade judiciária do Estado de Roraima

Abstract: A previsão de demanda em unidades judiciárias é um importante fator de impacto no desempenho de tribunais, onde flutuações de demanda não previstas causam altos impactos nas unidades judiciárias, gerando sobrecarga de trabalho aos servidores e, consequentemente, demora em tramitação de processos. Neste contexto, este artigo tem como objetivo analisar o comportamento da série de processos distribuídos no período de 2000 a 2015 na unidade judiciária 2ª Vara Cível do Tribunal de Justiça do Estado de Roraima, sob o olhar da previsão de demanda baseada em métodos de séries temporais comparando modelos de previsão. Neste trabalho, aplicar-se-á vários testes estatísticos que direcionam para os melhores modelos de previsão a serem utilizados.

Thiago Gottardi, Rosana Teresinha Vaccare Braga (ICMC-USP)
Providing past and future awareness to adaptive systems using stochastic processes

Abstract: In the context of self-adaptive systems, software behavior may vary at real-time. These systems depend on self-awareness in order to self-improve and self-adapt. According to several road-maps for these systems, many authors have pointed that this unpredictability requires specific measures to monitor the system status. In this work, we discuss how these systems can be modelled as stochastic processes. We present an analytic proven metric based on Markov chain to assess behavior by comparing their execution to past successes and failures. It is expected that the metric and associated models can be used to improve the management of those systems, as well as providing a basis for risk analysis of their execution. After presenting a case study, we also discuss how our proposal is feasible to improve the monitoring of these systems. In conclusion, this study indicates how stochastic methods could eventually become essential for monitoring self-adaptive systems by enabling past and future awareness. We also discuss how this monitoring can be effective for dev-ops development techniques, which are composed by distributed teams generating frequent releases. These teams require intensive usage of monitoring tools to cope with unpredictable risks that may arise throughout development.

Vanessa Helena Pereira, Theodore Gyle Lewis, Leonardo Tomazeli Duarte (UNICAMP)
Criteria importance through intercriteria correlation method: a mathematical implementation for NBA players efficiency analysis

Abstract: Usually, solving problems involves determining and choosing criteria,

the consequences of which are critical in a decision-making process. The determination of subjective preferences is always difficult, especially when it comes to involving multiple criteria decisions. An interesting mathematical method, the CRITIC (Critical Importance Through Intercriteria Correlation) involves important tools of Descriptive Statistics and Principal Component Analysis (PCA). The objective of this work is to implement the CRITIC method in open computing environment for scientific applications, the SciLab, in a classification study of players of NBA (National Basketball League). This type of method allows to classify mathematically and by multiple criterias the players according to their performances. The data are selected from the General List of Official Leading Players in the Regular Season 2017-2018. Our approach points to new types of statistical methods of analyzing athlete in addition to standard methods for efficiency analysis.

Walkiria M. Oliveira Macerau, Luis A. Milan (UFSCar/USP)
Inferência bayesiana de modelos de misturas de distribuições assimétricas aplicados à dados de criptomoedas.

Abstract: Resumo: Criptomoedas são moedas virtuais descentralizadas que tem recebido muita atenção recentemente, tendo sido apontada por alguns analistas como sendo o futuro das transações financeiras. Uma característica que esse mercado tem apresentado é alta volatilidade. Bitcoin é a pioneira dentre as criptomoedas, mas desde seu lançamento em 2009 um grande número de criptomoedas tem sido lançadas e esse número já ultrapassa 2100 [1]. O volume de recursos aplicados em criptomoedas também aumenta a cada dia o faz desse fenômeno algo a ser estudado e melhor entendido. Nesse trabalho testamos algumas alternativas de modelos que expliquem o comportamento de algumas características, em particular o retorno, de algumas criptomoedas. Utilizaremos as distribuições normal e t de Student assimétricas e também misturas dessas distribuições. Utilizaremos a abordagem bayesiana para ajustar os modelos combinada com os métodos MCMC (Monte Carlo Markov Chain). Métodos de seleção de modelos serão empregados para identificar quais modelos explicam melhor o comportamento dessas criptomoedas, com ênfase para os critérios EAIC, EBIC e DIC. A implementação dos métodos será feita utilizando o software R.

Palavras-chave:. Modelos de mistura, distribuições assimétricas, criptomoedas, distribuição normal, distribuição t de Student, abordagem bayesiana.

Referência: [1] Site: https://coinmarketcap.com/all/views/all/#BRL

Yana Miranda Borges, Breno Gabriel da Silva, Naiara Caroline Aparecido dos Santos, Brian Alvarez Ribeiro de Melo (Universidade Estadual de Maringá)
Uma análise de modelos de regressão Poisson para dados de doenças isquêmicas do coração no Brasil em 2016

Abstract: Ao estudarmos situações cuja variável dependente descreve dados de contagem, uma abordagem possível é estimar um Modelo Linear Generalizado com distribuição de Poisson. No entanto, devido às especificidades da distribuição Poisson, em que média e variância são iguais, há casos em que a ausência de equidispersão pode subestimar ou superestimar as estimativas, havendo necessidade de ajustar o modelo ou de procurar modelos alternativos mais adequados aos dados. O presente estudo avaliou ajustes pelo modelo Poisson, quasi-Poisson e binomial negativa para analisar a taxa de doenças isquêmicas do coração, considerando sexo, faixa etária e região. Dos modelos verificados neste estudo, o modelo quasi-Poisson apresentou um melhor ajuste quando comparado com o modelo log-linear Poisson e binomial negativa. Como resultado, verificou-se que entre os homens a taxa de ocorrência de morte por doenças isquêmicas do coração é 1.83 maior que entre as mulheres. Quanto à faixa etária, a taxa apresenta-se crescente, tomando-se como referência o grupo de menores que vinte anos. As regiões não apresentam grandes diferenças, no entanto, a região Norte é a que apresenta a menor taxa de mortalidade, quando comparada à região Centro-Oeste.

Yury Rojas Benites, prof. Vicente Garibay Cancho (UFSCar/USP)
Um novo modelo de regressão para taxas e proporções

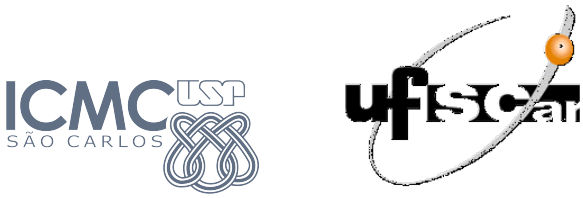
Abstract: Neste trabalho é apresentado um novo modelo estatístico para modelar dados no intervalo continuo (0,1). O modelo proposto é baseado na distribuição SB de Johnson, onde consideramos como transformação o quantil da distribuição valor extremo generalizado ao invés do quantil da distribuição logística. A nova família é estendida para modelos de regressão, onde fazendo uma reparametrização do modelo na mediana da variável resposta modelamos a mediana e o parâmetro de dispersão conjuntamente. A estimação dos parâmetros é baseada no método de máxima verossimilhança. Estudos de simulação são feitos para avaliar o desempenho das estimativas de máxima verossimilhança. Uma aplicação com dados reais de câncer colorretal é apresentado para ver a utilidade do modelo.

Participants

Adriano Kamimura Suzuki, ICMC-USP
Afonso Celso Penze Nunes da Cunha, ICMC-USP
Alejandra Rada, CMCC-UFABC
Alex de la Cruz Huayanay, UFSCar/USP
Alex Leal Mota, UFSCar/USP
Alex Ramos, UFPE
Alex Rodrigo dos Santos Sousa, Universidade Estadual de Campinas
Alfredo Ribeiro de Freitas
Amanda Morales Eudes D'Andrea, UFSCar/USP
Ana Carolina Freitas Xavier, Instituto Agronômico de Campinas
Ana Paula Jorge do Espirito Santo, UFSCar/USP
Anderson Ara, UFBA
Anderson Castro Soares de Oliveira, UFMT
André Luis Moutinho Teizen, USP
André Luiz Tirollo dos Santos, UFSCar
Anna Caroline Felix Santos de Jesus, USP
Átila Prates Correia, ICMC-USP
Benilton de Sá Carvalho, UNICAMP
Breno Gabriel da Silva, UEM
Bruna Ambrozim Silveira, Data Science Academy
Bruna Luiza de Faria Rezende, UFSCar/USP
Caio Moura Quina, UFSCar/USP
Camila Sgarioni Ozelame, UFSCar/USP
Carlos Alberto Oliveira de Matos, Unesp
Carlos Alonso, ICMC-USP
Carlos Franklin, UFSCar/USP
Carolina Grejo, UFSCar/USP
Cherlynn Daniela da Silva Arce, UNESP
Clarice Demétrio, ESALQ/USP
Cleide Mayra Menezes Lima, Universidade Federal do Piauí
Cristel Ecaterin Vera Tapia, UFSCar/USP
Cristine Lemos Corrêa Amaral, Faculdade Pitágoras
Daiane Aparecida Zuanetti, UFSCar
Daiane de Ascenção Cardoso, ENCE
Daiane de Souza Santos, UFSCar/USP
Daisy Assmann Lima, Universidade Católica de Brasília
Daniel Simionato, UFSCar
Danielle Lopes, ICMC-USP
Danillo Magalhães Xavier Assunção, UFSCar/USP
Danilo Augusto Sarti, USP
Deborah Bassi Stern, UFSCar/USP
Demerson Andre Polli, UFSCar/USP
Djidenou Hans Amos Montcho, Universidade Federal do Rio Grande
Élcio Lebensztayn, IMECC/UNICAMP
Elizbeth Chipa Bedia, UFSCar/USP
Enio Junior Seidel, UFSM
Felipe Aleshinsky, UFSCar
Felipe de Moura Ferreira, ICMC-USP
Francisco Antonio Rojas Rojas, UFSCar
Francisco José de Almeida Fernandes, IME-USP
Gabriel Avila Casalecchi, UFSCar
Gabriel Gomes Ferreira, USP
Gabriel Ianhez Pereira dos Santos, UFSCar
George Lucas Moraes Pezzott, UFSCar
Guilherme Antonio Alves de Lima, UFSCar
Guilherme Martins Lopes, IFSC
Gustavo Alexis Sabillón Lee, USP
Henrique Trivelato de Angelo, UFSCar
Ianní Muliterno, UFPE
Isaac Cortés Olmos, UFSCar/USP
Isis Fernanda Mascarin, UFSCar/USP
Jaime Enrique Lincovil Curivil, USP
Jaime Phasquinel Lopes Cavalcante, Universidade Federal do Ceará
Joabe Alves Carneiro, UFSCar
João Victor Zuanazzi Leme, UFSCar/USP
João Vitor Magri da Silva, UTFPR - CP
Jonathan Kevin Jordan Vasquez, Universidade de São Paulo
José Augusto Fiorucci, UnB
Josimara Tatiane da Silva, UFSCar/USP
Júlia Maria Pavan Soler, IME-USP
Juliana Cobre, ICMC-USP
Julio Cesar Pereira, UFSCar
Katiane Silva Conceição, ICMC-USP
Katy Rocio Cruz Molina, USP
Laís Sebastiany de Souza Santos, UFSCar
Leandro Resende Mundim, ICMC-USP
Lorena Yanet Cáceres Tomaya, UFSCar/USP
Lucas Leite Cavalaro, UFSCar/USP
Luciana Moura Reinaldo, UFC
Luciane Graziele Pereira, UNICAMP
Ludmila Rodrigues, IME-USP
Luis Aparecido Milan, UFSCar
Luis Ernesto Bueno Salasar, UFSCar
Luís Felipe Barbosa Fernandes, USP
Luis Felipe Borges de Messis, ICMC-USP
Luis Gustavo Sabino, Unicamp
Luiz Carlos Damasceno, IFNMG/USP/UFSCar
Luiz Gustavo Simão Pereira, UFSCar
Luiz Otávio de Oliveira Pala, Universidade Federal de Alfenas
Manuel Cabezas, Pontifícia Universidad Católica

de Chile
Marcelo Andrade da Silva, UFSCar/USP
Marcos Jardel Henriques, UFSCar/USP
Maria Lígia Chuerubim, Universidade Federal de Uberlândia
Mariane Romildo dos Santos, USP
Marina Gandolfi, UFSCar/USP
Marina Gonzaga de Oliveira, UFSCar/USP
Mário de Castro, USP
Matheus dos Santos Barbosa da Silva, Instituto de Física de São Carlos
Milena Nascimento Lima, UFSCar/USP
Milton Miranda Neto, UFSCar/USP
Murilo Henrique Soave, ICMC-USP
Naiara Caroline Aparecido dos Santos, UEM
Nancira Riberio Madi, UFSCar
Nancy Garcia, Unicamp
Nayara Fernandes de Mendonça, UTFPR
Oilson Alberto Gonzatto Junior, UFSCar/USP
Oscar Holguin Villamil, Universidade Estadual Paulista Júlio de Mesquita Filho
Osvaldo Anacleto, ICMC-USP
Pablo Rodriguez, ICMC-USP
Patty Mercedes Arce Flores, Pontificia Universidad Católica del Perú
Paulo Freitas Gomes, Unicamp
Paulo Oliveira, IME-USP
Pedro Ferreira Filho, DEs-UFSCar
Pedro Floriano Ribeiro, UFSCar
Pedro Luis do Nascimento Silva, IBGE
Peter Müller, University of Texas at Austin
Rafael Izbicki, UFSCar
Rafaela Cristina de Camargo, UFSCar
Renan Douglas Floriano Scavazzini, UEM
Ricardo Gonçalves da Silva, ICMC-USP
Ricardo Sandes Ehlers, ICMC-USP
Robert Gramacy, Virginia Tech
Roberta de Souza, UFSCar/USP
Rodrigo Barrem, FIA Business School
Samirian, UNICEP
Sandra Cristina de Oliveira, Faculdade de Ciências e Engenharia/UNESP
Sandro Gallo, UFSCar
Sandro Gonçalves, USP
Sandro Martinelli Reia, IFSC
Solange Ferreira Silvino, IFRR
Steve Ataucuri Cruz, UFSCar
Tainá Santana Caldas, UFSCar/USP
Taís Roberta Ribeiro, UFSCar/USP
Talita Zara Crevelim, UEM
Thiago Gottardi, ICMC-USP
Thiago Roberto do Prado, UNESP
Vanessa Helena Pereira, UNICAMP
Vera Tomazella, UFSCar
Victor Azevedo Coscrato, UFSCar-USP
Vinicius Hideki Yamada Santiago, UFSCar
Vitor Gustavo de Amorim, UFSCar/USP
Waldomiro Barioni Júnior, Embrapa
Walkiria Maria de Oliveira Macerau, UFSCar/USP
Yana Miranda Borges, UEM
Yury Rojas Benites, UFSCar/USP

Organizers



Support