Associação Brasileira de Estatística
21º SINAPE - Simpósio Nacional de Probabilidade e Estatística - 2014

Concurso de Melhor Trabalho de Iniciação Científica
Resumo dos Trabalhos Finalistas

Título: Correção de Bartlett bootstrap para a estatística da razão de verossimilhanças no modelo de regressão beta inflacionado

Autores: Laís Helen Loose; Fábio Mariano Bayer; Tarciana Liberal Pereira

Resumo: O modelo de regressão beta tem como objetivo permitir a modelagem de respostas que pertencem ao intervalo (0,1), como taxas ou proporções. No entanto, em situação práticas a presença de zeros e/ou uns é comumente observada. Para dados com essa característica o modelo de regressão beta inflacionado é adequado. Nestes modelos, os testes de hipóteses são frequentemente realizados baseados na estatística da razão de verossimilhanças. Os valores críticos são obtidos de aproximações assintóticas, o que pode conduzir a distorções de tamanho em amostras de tamanho finito. Neste sentido, o presente trabalho propõe a correção de Bartlett bootstrap para a estatística da razão de verossimilhanças no modelo de regressão beta inflacionado. Por meio de simulações de Monte Carlo é comparado o desempenho em amostras finitas da correção de Bartlett bootstrap com o teste da razão de verossimilhanças usual e com o ajuste de Skovgaard já proposto na literatura. Os resultados numéricos evidenciam o bom desempenho da correção de Bartlett bootstrap proposta. Ao final do trabalho também é apresentada uma aplicação a dados reais.

Palavras-Chave: *correção de Bartlett bootstrap; melhoramentos em pequenas amostras; regressão beta inflacionada; teste da razão de verossimilhanças.*

Título: Factor analysis with mixture modeling to evaluate coherent patterns in microarray data

Autores: João Daniel Nunes Duarte; Vinícius Diniz Mayrink

Resumo: The computational advances over the last decades have allowed the use of complex models to analyze large data sets. The development of simulated-based methods, such as the MCMC, has contributed to an increased interest in the Bayesian framework as an alternative to work with factor models. Many studies have applied the factor analysis to explore gene expression data with results often outperforming traditional methods for estimating and identifying patterns and meta-gene groups related to the underlying biology. In this work, we present a Sparse Latent Factor Model (SLFM) using a mixture prior (sparsity prior) to evaluate the signicance of each factor loading; when the loading is signicant the effect of the corresponding factor is detected through patterns displayed along the samples. The SLFM is applied to investigate simulated and real microarray data. The real data sets represent the gene expression for different types of cancer; this includes breast, brain, ovarian and lung tumors. The proposed model can indicate how strong is the observed expression pattern allowing the measurement of the evidence of presence/absence of the gene activity. Finally, we compare the SLFM with two simpler gene detection methods available in the literature. The results suggest that the SLFM outperforms the traditional methods.

Palavras-Chave: *Factor Model; Bayesian Inference; Gene Expression; Mixture; Sparsity.*

Título: Homogeneity tests for 2×2 contingency tables

Autores: Natália Lombardi de Oliveira; Adriano Polpo de Campos

Resumo: Using the likelihood ratio statistic, we develop a significance index, called P-value, to test the hypothesis of homogeneity in 2×2 contingency tables. The P-value does not depend on asymptotic distributions, and is based on the elimination of the nuisance parameter. Therefore, we obtain the exact distribution of the likelihood ratio statistic in a way that is, moreover, compatible with the likelihood principle. For a better understanding of significance indices to test homogeneity, we perform a study comparing the P-value with some frequentist indices (likelihood ratio test, chi-square test) and with the Full Bayesian Significance Test (FBST). This comparative study shows an interesting relation between all the analysed indices, Bayesian and frequentist.

Palavras-Chave: *Chi-square test; FBST; likelihood ratio test; P-value; significance indices.*

Título: Modelos de Mistura de Distribuições para Populações Heterogêneas

Autores: Carolina Valani Cavalcante; Kelly Cristina Mota Gonçalves

Resumo: Modelos de mistura de distribuições são de grande aplicabilidade em problemas de modelagem de fenômenos observados em populações que se comportam de maneira heterogênea, ou seja, são compostas por subpopulações. Durante esta monografia realiza-se um estudo acerca do ajuste desta classe de modelos, com base na abordagem Bayesiana, para dois casos distintos, o caso em que o n úmero de subpopulações é conhecido e quando este é desconhecido. Para tanto são apresentados os principais conceitos de Inferência Bayesiana úteis para o desenvolvimento deste trabalho. Em particular, o interesse concentra-se na inferência acerca de modelos de mistura Normais univariados. Neste caso, como a distribuição a posteriori do vetor paramétrico tem forma analítica desconhecida são necessários algoritmos de simulação estocástica, como os métodos de Monte Carlo via Cadeias de Markov. No entanto, no caso em que o número de subpopulações é desconhecido, tais métodos nãoo podem ser aplicados e uma opção é usar o algoritmo Monte Carlo via Cadeias de Markov com saltos reversíveis. Uma aplicação a dados artificiais é feita com o objetivo de comparar a performance das duas abordagens, ou seja, número de subpopulações conhecido ou não. Os resultados obtidos mostram que, como esperado, fixar este número no valor verdadeiro produz os resultados mais eficientes. Por outro lado, quando não se tem ideia sobre este número, considerá-lo também um parâmetro e estimá-lo é a melhor solução em termos de estimação e previsão, perdendo pouco para a primeira abordagem. Para a implementação destas técnicas foi utilizado o comando NMixMCMC presente no pacote mixAK do software R. Além disso, também aplicou-se ambas as técnicas a um conjunto de dados reais.

Palavras-Chave: *subpopulações; Inferência Bayesiana; mistura de distribuições; RJMCMC.*

Título: Métodos computacionais para realização de e ca inferência bayesiana em modelos dinâmicos e a lineares generalizados

Autores: Mariana Raniere Neves; Carlos Tadeu Pagani Zanini; Mariane Branco Alves

Resumo: A concentração de poluentes na atmosfera tem sido apontada, por vários estudos, como um fator que influencia na saúde e na qualidade de vida das pessoas. Quantificar o efeito de poluentes sobre desfechos epidemiológicos, bem como seu comportamento ao longo do tempo são questões de grande relevância, não apenas para

o estatístico, mas também para os órgãos públicos que administram a area de saúde. Neste âmbito, o presente trabalho se propõe a estudar o efeito da concentração de monóxido de carbono, conjuntamente com fatores climáticos, sobre óbitos diários de crianças com menos de 5 anos por doença respiratória na cidade de São Paulo ocorridas entre o dia Janeiro de 1994 e Dezembro de 1997. Para isso, utilizou-se a abordagem paramétrica bayesiana. Os modelos propostos pertencem à classe dos Modelos Dinâmicos Lineares Generalizados (MLDG), onde para a variável resposta, por se tratar de uma contagem, é assumida ter distribuição Poisson. Dada a falta de solução analítica para estimação dos parâmetros destes modelos, é necessário utilizar métodos aproximados para realização de inferência. Nesse sentido, escolheu-se utilizar os Métodos de Monte Carlo via Cadeias de Markov e o método Linear Bayes. Os resultados obtidos segundo as duas abordagens são comparados, ressaltando as vantagens e desvantagens decorrentes do uso de ambas.

Palavras-Chave: *Inferência Bayesiana; Linear Bayes; Métodos de Monte Carlo; Modelos Dinâmicos; Epidemiologia.*

Concurso de Melhor Dissertação de Mestrado
Resumo dos Trabalhos Finalistas

Título: Selecionador de Características para classificação de sinais de EEG e construção de Interfaces Cérebro-Máquina

Autores: Murilo Coutinho Silva; George Freitas von Borries

Resumo: A classificação de sinais de eletroencefalografia (EEG) vem sendo muito estudada recentemente para proporcionar aplicações como as Interfaces-Cérebro Máquina. Parte fundamental do processo de classificação é a chamada "extração de características" dos sinais de EEG. Na literatura, diversas técnicas de extração de características foram apresentadas e, entretanto, não existe uma técnica que supere as demais em todas as situações. Para solucionar este problema, este trabalho apresenta um novo algoritmo que seleciona automaticamente as melhores características obtidas por várias técnicas de extração simultaneamente, produzindo um conjunto ótimo e reduzido de características que não é necessariamente o mesmo para cada aplicação. Pelo uso do novo algoritmo, todas as técnicas já apresentadas na literatura e as técnicas futuras podem ser combinadas para produzir o melhor e mais poderoso conjunto de características gerando taxas de classificação excelentes. Neste trabalho, o selecionador de características é testado utilizando vários conjuntos de dados reais obtendo as melhores taxas de classificação quando comparado a outras técnicas de classificação de dados de EEG.

Palavras-Chave: *Eletroencefalografia; Máquina de suporte vetorial; Interface Cérebro-Máquina; Aprendizado estatístico de máquina.*

Título: Análise de dados com riscos semicompetitivos

Autores: Elizabeth González Patiño

Resumo: Em análise de sobrevivência, usualmente o interesse está em estudar o tempo até a ocorrência de um evento. Quando as observações estão sujeitas a mais de um tipo de evento (por exemplo, diferentes causas de óbito) e a ocorrência de um evento impede a ocorrência dos demais, tem-se uma estrutura de riscos competitivos. Em algumas situações, no entanto, o interesse está em estudar dois eventos, sendo que um deles (evento terminal) impede a ocorrência do outro (evento intermediário), mas não vice-versa. Essa estrutura é conhecida como riscos semicompetitivos e foi definida por Fine *et. al*. (2001). Neste trabalho são consideradas duas abordagens para análise de dados com essa estrutura. Uma delas é baseada na construção da função de sobrevivência bivariada por meio de cópulas da família Arquimediana e estimadores para funções de sobrevivência são obtidos. A segunda abordagem é baseada em um processo de três estados, conhecido como processo doença-morte, que pode ser especificado pelas funções de intensidade de transição ou funções de risco. Neste caso, considera-se a inclusão de covariáveis e a possível dependência entre os dois tempos observados é incorporada por meio de uma fragilidade compartilhada. Estas metodologias são aplicadas a dois conjuntos de dados reais: um de 137 pacientes com leucemia, observados no máximo de sete anos após transplante de medula óssea, e outro de 1253 pacientes com doença renal crônica submetidos a diálise, que foram observados entre os anos 2009-2011.

Palavras-Chave: *riscos semicompetitivos; fragildiade compartilhada; cópulas familia arquimediana; processo doença-morte.*

Título: Practical aspects of the estimation of mixture model via Dirichlet Process

Autores: Rosineide Ferando da Paz; Luís Aparecido Milan

Resumo: We review the Dirichlet process mixture model and investigate its performance as a classification method. The first aspect considered is its sensibility to the choice of location parameter of base distribution. The second aspect considers the performance of the model regarding the departure of the parameters of the component distributions. Simulation results with mixture of normal distributions indicate sensibility to location parameters choices and good performance even when component normal distributions differ only in variances. Finally, we apply the method to two data sets.

Palavras-Chave: *Dirichlet process; Mixture model; Density estimation; Non-parametric Bayesian; Gibbs sampling.*

Título: Modelos da Teoria de Resposta ao Item assimétricos de grupos múltiplos para respostas politômicas nominais e ordinais sob um enfoque bayesiano

Autores: Eduardo Vargas Ferreira; Caio Lucidius Naberezny Azevedo

Resumo: No presente trabalho propõem-se novos modelos da Teoria de Resposta ao Item para respostas politômicas nominais e ordinais (graduais), via dados aumentados, para grupos múltiplos. Para a modelagem das distribuições dos traços latentes de cada grupo, considera-se normais assimétricas centradas. Tal abordagem, além de acomodar a característica de assimetria aos dados, ajuda a garantir a identificabilidade dos modelos estudados, a qual é tratada tanto sob a ótica frequentista quanto bayesiana. Com relação aos métodos de estimação, desenvolveu-se procedimentos bayesianos através de algoritmos de Monte Carlo via cadeias de Markov (MCMC), utilizando o algoritmo de Gibbs (DAGS), com a verossimilhança aumentada (dados aumentados) e Metropolis-Hastings, considerando a verossimilhança original. As implementações computacionais foram escritas em linguagem C++, integradas ao ambiente computacional, gráfico e estatístico R, viabilizando rotinas gratuitas, de código aberto e alta velocidade no processamento, essenciais à difusão de tais metodologias. Para a seleção de modelos, utilizou-se o critério de informação deviance (DIC), os valores esperados do critério de informação de Akaike (EAIC) e o critério de informação bayesiano (EBIC). Em relação à verificação da qualidade do ajuste de modelos, explorou-se a checagem preditiva a posteriori, que fornece meios concretos de se avaliar a qualidade do instrumento de medida (prova, questionário etc), qualidade do ajuste do modelo de um modo global, além de indícios de violações de suposições específicas. Estudos de simulação, considerando diversas situações de interesse prático, indicam que os modelos e métodos de estimação produzem resultados bastante satisfatórios, com superioridade dos modelos assimétricos com relação ao simétrico (o qual assume simetria das distribuições das variáveis latentes). A análise de um conjunto de dados reais, referente à primeira fase do vestibular da UNICAMP de 2013, ilustra o potencial da tríade: modelagem, métodos de estimação e ferramentas de diagnósticos, desenvolvida neste trabalho.

Palavras-Chave: *Teoria da resposta ao item; Modelos politômicos; Distribuição normal assimétrica; Algoritmos MCMC; Dados aumentados.*

Título: Modelos Combinados AR-GARCH governados por distribuições estáveis

Autores: Thiago do Rêgo Sousa

Resumo: Estendemos a aplicação do modelo combinado AR-GARCH governado por distribuições GEV e apresentado por Zhao et. al. (2011) para um modelo governado por distribuições estáveis, já que estas distribuições podem ser utilizadas para modelar dados de finanças, incluindo os eventos extremos. Exploramos o método Bayesiano e de Máxima verossimilhança para estimação dos modelos com inovações GEV e estável. Posteriormente, investigamos as condições de estacionariedade de um modelo mais geral ARMA- power-GARCH com inovações estáveis proposto por Rachev et. al. (2002) e derivamos as condições de estacionariedade para um modelo assimétrico ARMA-APARCH com inovações estáveis. A estimação do modelo geral ARMA-APARCH com inovações GEV e estável foi mplementada em um novo pacote chamado GEVStableGarch disponível no CRAN do software R.

Palavras-Chave: *ARMA; GARCH; Distribuições estáveis; GEV; Estacionariedade.*