Resumos

Abstracts

Sessão: Métodos Matemáticos em Probabilidade e Estatística

Session: Mathematical Methods in Probability and Statistics

Organizadores

Organizers

Victor Fossaluza - IME/USP
victor.ime@gmail.com

Rafael Izbicki - UFSCar
rafaelizbicki@gmail.com

I Congresso Brasileiro de
Jovens Pesquisadores em
Matemática Pura e Aplicada

The sticker collector's problem in the classroom

Adriano Polpo[*], M. Diniz[*], D. Lopes[*], L. Salasar[*]

[*]UFSCar, São Carlos, Brasil

Resumo

This note discusses how a generalization of the coupon collector's problem can be used in different undergraduate courses as a motivating example and an illustration of important results of probability theory.

Bayesian using modified Jeffreys prior for Weibull regression censored data

Al Omari Mohammed Ahmed*

*AlBaha University, Baha, Saudi Arabia.

Resumo

We have with regards to the Bayesian, developed an approach by using Jeffreys prior and modified Jeffreys priors with covariate obtained by using Gauss quadrature method. This is also done for maximum likelihood estimator to estimate the parameters of the covariate of the Weibull regression distribution given shape with right censored data. It has been seen that the estimators obtained are not available in closed forms, although they can be solved it for the given sample by using suitable numerical methods. The comparison criteria is the mean square error and the performance of these three estimates are assessed using simulation considering various sample size, several specific values of Weibull shape parameter. The results show that modified Jeffreys prior is better estimator compared to others.

Bayesian longitudinal item response modeling with restricted covariance pattern structures

Caio Lucidius N. Azevedo[*]

[*]UNICAMP

Resumo

Educational studies are often focused on growth in student performance and background variables that can explain developmental differences across examinees. To study educational progress, a flexible latent variable model is required to model individual differences in growth given longitudinal item response data, while accounting for time-heterogenous dependencies between measurements of student performance. Therefore, an item response theory model, to measure time-specific latent traits, is extended to model growth using the latent variable technology. Restricted covariance pattern models are proposed to model the variance-covariance structure of the student achievements. The main advantage of the extension is its ability to describe and explain the type of time-heterogenous dependency between student achievements. An efficient MCMC algorithm is given that can handle identification rules and restricted parametric covariance structures. A reparameterization technique is used, where unrestricted model parameters are sampled and transformed to obtain MCMC samples under the implied restrictions. The study is motivated by a large-scale longitudinal research program of the Brazilian Federal government to improve the teaching quality and general structure of schools for primary education. It is shown that the growth in math achievement can be accurately measured when accounting for complex dependencies over grades using time-heterogenous covariances structures.

Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes

Flávio Bambirra Gonçalves*

*UFMG, Belo Horizonte, Brasil

Resumo

In this paper we present a novel inference methodology to perform Bayesian inference for Cox processes in space and/or time where the intensity function depends on a multivariate Gaussian process. The novelty of the method lies on the fact that no discretization error is involved despite the non-tractability of the likelihood function and the infinite dimensionality of the problem. The method is based on a Markov chain Monte Carlo algorithm that samples from the joint posterior distribution of the parameters and latent variables of the model. A particular choice of the dominating measure to obtain the likelihood function corrects previous attempts to solve the problem in an exact framework. The models allow the use of covariates to explain the dynamics of the intensity function. Simulated examples illustrate the methodology and compare different alternatives for some of the MCMC steps.

Análise do padrão de pontos de óbitos por doença cerebrovascular no Rio de Janeiro com coeficientes variando espacialmente

Jony Arrais Pinto Junior*

*UFF, Rio de Janeiro, Brasil

Resumo

Este trabalho propõe um modelo para lidar com a heterogeneidade espacial presente no estudo de padrões geográficos de óbitos devido a doenças cerebrovasculares. A estrutura envolve a análise de um padrão de pontos com componentes que exibem uma variação espacial. Estudos preliminares mostram que a mortalidade por esta doença não apresenta uma distribuição geográfica uniforme, mesmo em países desenvolvidos. O modelo proposto é uma extensão do trabalho de Liang et al (2009), permitindo que o efeito das covariáveis associadas as unidades experimentais possam variar ao longo do espaço. Riscos relativos são obtidos para comparar diferentes níveis de covariáveis, diferentes localizações geográficas ou ambos. A metodologia é aplicada ao padrão de pontos de óbitos por doenças cerebrovasculares na cidade do Rio de Janeiro. Os resultados foram satisfatórios quando comparados com metodologias alternativas, incluindo o caso em que se considera fixo o efeito de todas as covariáveis. Nosso modelo é capaz de capturar e ressaltar importantes características dos dados que não seriam noticiadas em outras metodologias, provendo informações que são relevantes para a construção de políticas de saúde eficientes.

Multilevel binary regression with outcome uncertainty

Leonardo Bastos[*]

[*]Fiocruz, Rio de Janeiro, Brasil

Resumo

In this paper we present a multilevel binary model when the outcome is measured with uncertainty. We are interested in obtaining association measures, such as odds ratio, while taking into account specificity and sensitivity of the outcome. Posterior inference is implemented using Hamiltonian Monte Carlo and also integrated nested Laplace approximation (INLA). A simulation study is provided and the method is applied on Brazilian alcohol and drug abuse data sets.

Proposta de um Algoritmo para Indução de Árvores de Classificação para Dados Desbalanceados

Marcelo Lauretto[*] e Claudio Frizzarini

[*]EACH-USP, São Paulo, Brasil

Resumo

As técnicas de mineração de dados, e mais especificamente de aprendizado de máquina, têm se popularizado enormemente nos últimos anos, passando a incorporar os Sistemas de Informação para Apoio à Decisão, Previsão de Eventos e Análise de Dados. Por exemplo, sistemas de apoio à decisão na área médica e ambientes de Business Intelligence fazem uso intensivo dessas técnicas, envolvendo particularmente árvores de decisão. A mineração de informação e conhecimento a partir de grandes bases de dados tem sido reconhecida como tema chave de pesquisa em sistemas de banco de dados e aprendizado de máquina.

Concomitantemente a essa popularização, faz-se necessário o desenvolvimento de ferramentas de modelagem acessíveis, conceitualmente simples e com baixa necessidade de parametrização, que possam ser utilizadas (ao menos em análises mais simples) por profissionais que não sejam necessariamente especialistas nos métodos de modelagem subjacentes.

Algoritmos indutores de árvores de classificação, particularmente os algoritmos TDIDT (Top-Down Induction of Decision Trees), figuram entre as técnicas mais comuns de aprendizado supervisionado. A construção de uma árvore de decisão consiste em partições sucessivas do conjunto de treinamento original em subconjuntos menores. Uma das vantagens desses algoritmos em relação a outros é que, uma vez construída e validada, a árvore tende a ser interpretada com relativa facilidade, sem a necessidade de conhecimento prévio sobre o algoritmo de construção. Em um contexto de mineração de dados, mesmo que não sejam necessariamente utilizadas na classificação de novas instâncias, árvores de classificação podem ser construídas para fornecer descrições (na forma de regras de classificação) das características comuns aos membros de cada classe.

Todavia, são comuns problemas de classificação em que as frequências relativas das classes variam significativamente. Algoritmos baseados em minimização do erro global de classificação tendem a construir classificadores com baixos erros de classificação nas classes majoritárias e altos erros nas classes minoritárias. Esse fenômeno pode ser crítico quando as classes minoritárias representam eventos como a presença de uma doença grave (em um problema de diagnóstico médico) ou a inadimplência em um crédito concedido (em um problema de análise de crédito).

Diversos algoritmos TDIDT não possuem métodos adaptativos automáticos, demandando a calibração de parâmetros ad-hoc de custos ou, na ausência de tais parâmetros, a adoção de métodos de balanceamento dos dados. As duas abordagens não apenas introduzem uma maior complexidade no uso das ferramentas de mineração de dados para usuários menos experientes, como também nem sempre estão disponíveis.

Este trabalho apresenta uma descrição e alguns resultados empíricos de um algoritmo TDIDT em desenvolvimento, para construção de árvores na presença de dados desbalanceados. Esse algoritmo, denominado atualmente DDBT (Dynamic Discriminant Bounds Tree), utiliza um critério de partição de nós que, ao invés de se basear em frequências absolutas de classes, compara as proporções das classes nos nós com as proporções do conjunto de treinamento original, buscando formar subconjuntos com maior discriminação de classes em relação ao conjunto de dados original. Para a rotulação de nós terminais, o algoritmo atribui a classe com maior prevalência relativa no nó em relação à prevalência no conjunto original. Essas características fornecem ao algoritmo a flexibilidade para o tratamento de conjuntos de dados com desbalanceamento de classes, resultando em um maior equilíbrio entre as taxas de erro em classificação de objetos entre as classes.

On probability and subjectivism

Márcio Diniz*

*UFSCar, São Carlos, Brasil

Resumo

In 1900, at the II International Congress of Mathematicians held in Paris, David Hilbert presented ten (of a total of 23) unsolved (at the time) questions. Several became a hot topic, opening important research fields to solve them. In particular, the sixth problem (Mathematical treatment of the axioms of physics), was further explained as composed by two problems (i) axiomatic treatment of probability with limit theorems for foundation of statistical physics; and (ii) the rigorous theory of limiting processes "which lead from the atomistic view to the laws of motion of continua".

One may say that Kolmogorov's axiomatic approach of probability solved (i). Based on this approach, probability theory and statistics found powerful tools in measure theory to develop very useful results for applied and theoretical problems. Almost at the same time, another view, nowadays known as subjective, was also formalized by F. P. Ramsey and Bruno de Finetti but received much less attention.

This talk will show how this second school proved important results for mathematical statistics and its connection with other fields of mathematics such as functional analysis and set theory.

Análise Discrepante via Máxima Verossimilhança

Patrícia Viana da Silva*

*UFU, Uberlândia, Brasil

Resumo

Em medicina diagnóstica a condição do paciente em relação a determinada doença é avaliada a partir de sintomas, sinais indicativos ou de resultados de exames laboratoriais. Um procedimento diagnóstico utilizado nos últimos anos conhecido como Análise Discrepante propõe submeter todos os indivíduos a dois testes diagnósticos diferentes. A conclusão sobre o estado de saúde do paciente, doente ou não doente, é obtida pelos casos em que os testes concordam. No caso em que os resultados dos dois testes são discrepantes, discordantes, utiliza-se como referência um terceiro teste para critério de desempate. Com isso é criada uma estrutura de observações omissas, pois o resultado do último teste não é conhecido para todos os indivíduos.

Estudos que avaliam desempenho de testes diagnósticos usando a análise discrepante receberam críticas de superestimação das medidas de sensibilidade e especificidade usadas para avaliar o desempenho dos testes diagnósticos. Para melhorar o processo de estimação, esse trabalho propõe o uso do método de máxima verossimilhança com restrições sobre a probabilidade de omissão que garantem a identificabilidade do modelo.

Parameter estimation and identifiability in multivariate binary models with skewed link functions

Rafael Braz Azevedo Farias[*] e Marcia D'Elia Branco[**]

[*]DEMA-UFC, Fortaleza, Brasil
[**]IME-USP, São Paulo, Brasil

Resumo

Data sets with multivariate responses often appear in surveys where the data came from questionnaires. Opinion poll, sometimes simply referred to as a poll, are common examples of studies in which the responses are multivariate. One type poll that gain great prominence in Brazil in election years, is the survey of vote intent. However, despite the higher visibility of prognostic studies of election, opnion polls is a tool widely used to detect trends and positions of different social segments on various topics, be they political, social or governmental. We introduce in this work a class of multivariate regression models with asymmetric link functions to fit data sets with multivariate binary responses. The link functions here considered are quite flexible and robust, contemplating symmetrical link functions as special cases. Due to the complexity of the model, the issue of Bayesian identifiability in Multivariate binary models is discussed. It is important to note that the lack of identifiability may happen in differents ways, depending on whether the problem is in the prior, the likelihood function or the posterior distribution, different views on the issue of identifiability have been given in the literature. In our model the problem is on the the likelihood function. The Bayesian approach was considered and some Monte Carlo Markov Chain (MCMC) algorithms have been developed. Simulation studies have been developed with two objectives: i) verify the quality of the algorithms developed and ii) to verify the importance of choosing the link function.

The martingale approach for the occurrence of words in i.i.d. trials

Renato J. Gava[*]

[*]UFSCar, São Carlos, Brasil

Resumo

Consider a sequence of i.i.d. trials where each trial produces a letter from a finite alphabet. Given a collection of words, we look at this sequence till the moment t at which one of these patterns appears as a run. For example, if our alphabet is the set $\{H, T\}$ and we consider the word HTH then we get $\tau = 5$ in the realization $THHTH$. We show how the martingale approach introduced by Li can be employed to compute the mean and the generating function of τ and the probability that a pattern is the first one to appear.

On dynamic weighted entropies

Salimeh Yasaei Sekeh*, G. R. Mohtashami Borzadaranb**,
A. H. Rezaei Roknabadib**

*UFSCar, São Carlos, Brasil
**Ferdowsi University of Mashhad, Mashhad, Iran

Resumo

We study some properties of the weighted dynamic measures of entropy introduced in Di Crescenzo and Longobardi (2006). It is shown that the proposed measures could characterize distributions for some distributions in continuous and discrete cases. Moreover, a new stochastic order based on weighted dynamic entropies is presented and we provide some of its properties.

Statistical methods for preprocessing, integration, and quality assessment of synthetic lethality data

Samara Kiihl[*]

[*]UNICAMP, Campinas, Brasil

Resumo

Two genes are considered to be synthetically lethal (SL) when cells carrying a loss-offunction mutation in either of these genes are viable, but cells with loss-of-function mutation in both of the genes are not. In yeast, biologists have been observing the phenotype of mutant strains with two genes knocked out in search for SL, however, with over 12 million gene pairs to explore, only a small percentage of all possible interactions have been studied. We use data from the protocol called dSLAM (diploid-based synthetic lethality analysis by microarray) that measures growth of the mutants strains using microarrays. We describe some of the challenges associated with microarray data and propose a statistical model to improve prediction of SL pair. The model allows borrowing information across arrays and across genes to improve robustness and precision of the estimates. We use BioGRID, a curated database of genetic interactions, to demonstrate the advantages of our model by comparing it to naive approaches. Methods to speed up the search of SL pairs are crucial to identify all possible SL genes and may have an impact to overcome failure of cancer treatments targeting only one gene.

Multilevel binary regression with outcome uncertainty

Thaís C. O. Fonseca*, Marco A. R. Ferreira**

*UFRJ, Rio de Janeiro, Brasil
**Virginia Tech, Virginia, EUA

Resumo

We propose a new class of dynamic multiscale models for Poisson spatiotemporal processes. Specifically, we use a multiscale spatial Poisson factorization to decompose the Poisson process at each time point into spatiotemporal multiscale coefficients. We then connect these spatiotemporal multiscale coefficients through time with a novel Dirichlet evolution. Further, we propose a simulation-based full Bayesian posterior analysis. In particular, we develop filtering equations for updating of information forward in time and smoothing equations for integration of information backward in time, and use these equations to develop a forward filter backward sampler for the spatiotemporal multiscale coefficients. Because the multiscale coefficients are conditionally independent a posteriori, our full Bayesian posterior analysis is scalable, computationally efficient, and highly parallelizable. Moreover, the Dirichlet evolution of each spatiotemporal multiscale coefficient is parametrized by a discount factor that encodes the relevance of the temporal evolution of the spatiotemporal multiscale coefficient. Therefore, the analysis of discount factors provides a powerful way to identify regions with distinctive spatiotemporal dynamics. Finally, we illustrate the usefulness of our multiscale spatiotemporal Poisson methodology with an application to tornado reports in the American Midwest.

Imputation of multivariate continuous data with nonignorable missingness

Thais V. Paiva*

*Duke University, North Carolina, EUA

Resumo

Regular imputation methods have been used to deal with nonresponse in several types of survey data. However, in some of these studies, the assumption of missing at random is not valid since that the probability of missing depends on the response variable. We propose an imputation method for multivariate data sets when there is nonignorable missingness. A Dirichlet process mixture of multivariate normals is fit to the observed data under a Bayesian framework to provide flexibility. We provide some guidelines on how to alter the estimated distribution using the posterior samples of the mixture model and obtain imputed data under different scenarios. Lastly, we apply the method to a real data set.

I Congresso Brasileiro de
Jovens Pesquisadores em
Matemática Pura e Aplicada

Global estimation of Hidden Markov models using interval arithmetic

Tiago de Morais Montanher*

*IME-USP, São Paulo, Brasil

Resumo

Hidden Markov Models are important tools in statistics and applied mathematics, with applications in speech recognition, physics, mathematical finance and biology. The Hidden Markov Models we consider are formed by two discrete time and finite state stochastic process. The first process is a Markov chain and is not observable directly. Instead, we observe a second process which is driven by the hidden process. In order to extract conclusions from a Hidden Markov Model we must estimate the parameters defining it. In this article we present global optimization techniques to estimate these parameters by maximum likelihood and compare our estimates with the ones obtained by the local likelihood maximization methods already described in the literature. In order to evaluate the global maximum we provide an interval branch and bound algorithm based on interval Newton method and a symmetry breaking scheme. The algorithm starts with a local Baum-Welch method, which provides a warm lower bound for the problem. We also derive KKT conditions to obtain a new box elimination test. Our algorithm is able, in a successful execution, to find a box with prescribed width which rigorously contains at least one feasible point for the problem and such that the solution is an epsilon-global maximum. The objective function for this problem can be evaluated by the so called backward and forward recursions. In fact we can use only one of these recursions or we can combine both to evaluate function and its derivatives. These three formulations are equivalent using exact arithmetic. However they will usually be different in interval arithmetic due to the lack the distributivity law. In order to accelerate the convergence of upper bound of the global maximum we implement and compare interval extensions for the forward, backward and forward-backward equations and their respective derivatives. In order to make our interval bounds tight, we consider enclosures based on Taylor expansion of first and second orders and centered forms. We handle the underflow problems which arise frequently in the estimation problem for Hidden Markov models introducing a new scaling scheme which is not based on taking the log of the objective function. We present the results of numerical experiments illustrating the effectiveness of our approach.