IX Workshop on Probabilistic and Statistical Methods

PROGRAM BOOK

Feb 9-11, 2022 - Virtual

Organized by Joint Graduate
Program in Statistics
PIPGES - USFCar/USP

Organization

ufscar PIPGES ICMC USP SÃO CARLOS 50 anos 1971 - 2021

IX Workshop on Probabilistic and Statistical Methods

February 9–11, 2022

ICMC-USP, São Carlos, SP, Brasil
http://wpsm.icmc.usp.br

PROGRAM

ICMC-USP and DEs/UFSCar

About the IX WPSM

ICMC-USP, São Carlos, SP, Brasil, February 9–11, 2022
`http://wpsm.icmc.usp.br`

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, Brasil.

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

This edition was held virtually through the Zoom platform.

Organizing Committee

Andressa Cerqueira, UFSCar (Chair)
Mário de Castro, ICMC-USP (Chair)
Michel Montoril, UFSCar
Vicente Garibay Gancho, ICMC-USP

Supporting Committee

ICMC-USP staff

Invited Speakers

Short courses

Marcelo Hilário - Universidade Federal de Minas Gerais
Peter Müller - The University of Texas at Austin
Valtencir Zucolotto - Universidade de São Paulo

Conferences

Inés Armendáriz - Universidad de Buenos Aires
Lurdes Inoue - University of Washington
Nalini Ravishanker - University of Connecticut

Mini conferences

Alejandro Róldan - Universidad de Antioquia
Aritra Halder - University of Virginia
Márcio Augusto Diniz - Cedars-Sinai Medical Center

Student talks

Jessica Suzana Barragan Alves - PIPGEs UFSCar/USP
Rodrigo Ferrari Lucas Lassance - PIPGEs UFSCar/USP
Vitor Amorim - PIPGEs UFSCar/USP

Satellite course

Daniel Takata - ENCE, IBGE

IX Workshop on Probabilistic and Statistical Methods

February 9–11, 2022

ICMC-USP, São Carlos, SP, Brasil

SCHEDULE

ICMC-USP and DEs/UFSCar

February 2–4, 2022

14h00 - 16h00 : **Satellite course** - Daniel Takata - ENCE, IBGE

February 9, 2022

10h00 - 12h00 : Valtencir Zucolotto - Universidade de São Paulo

12h00 - 14h00 : Lunch

14h00 - 14h30 : Alejandro Róldan - Universidad de Antioquia

14h30 - 15h00 : Jessica Suzana Barragan Alves - PIPGEs UFSCar/USP

15h00 - 16h00 : Lurdes Inoue - University of Washington

February 10, 2022

10h00 - 12h00 : Marcelo Hilário - Universidade Federal de Minas Gerais

12h00 - 14h00 : Lunch

14h00 - 15h00 : Nalini Ravishanker - University of Connecticut

15h00 - 15h30 : Aritra Halder - University of Virginia

15h30 - 16h00 : Vitor Amorim - PIPGEs UFSCar/USP

February 11, 2022

10h00 - 12h00 : Peter Müller - The University of Texas at Austin

12h00 - 14h00 : Lunch

14h00 - 15h00 : Inés Armendáriz - Universidad de Buenos Aires

15h00 - 15h30 : Márcio Augusto Diniz - Cedars-Sinai Medical Center

15h30 - 16h00 : Rodrigo Ferrari Lucas Lassance - PIPGEs UFSCar/USP

IX Workshop on Probabilistic and Statistical Methods

February 9–11, 2022

ICMC-USP, São Carlos, SP, Brasil

ABSTRACTS

ICMC-USP and DEs/UFSCar

Satellite course

Daniel Takata (ENCE - IBGE, Brasil)

Métodos estatísticos para a análise de dados do esporte

Abstract: Nos últimos anos, especialmente a partir da virada do século, a análise estatística de dados tem tido um papel cada vez mais relevante na ciência do esporte. Nos esportes coletivos, departamentos de análise de desempenho trabalham com modelos estatísticos em busca de padrões de comportamento de jogadores. Nos esportes individuais, modelos complexos são utilizados para a melhoria dos treinamentos e a otimização de desempenho. Como a revolução da análise de dados no esporte, os dias em que os treinamentos e as decisões eram baseados em tentativa e erro estão cada vez mais ficando para trás.

O minicurso será ministrado em três aulas. Na primeira, será feito um apanhado do que de mais relevante tem sido feito utilizando a estatística no esporte, incluindo análise de desempenho, tomada de decisões durante jogos, montagens de elencos etc. Na segunda, serão abordados métodos de análise de dados no futebol, que podem ser estendidos a outros esportes coletivos, como modelagem de placares de jogos e uso de métodos de aprendizagem de máquina para avaliação de desempenho. A última aula irá focar na análise de modelos para esportes individuais como atletismo e natação, incluindo a utilização de teoria de valores extremos para comparação de desempenhos.

Short courses

Marcelo Hilário (Universidade Federal de Minas Gerais, Brasil)
Multi-scale renormalization for percolation on random environments

Abstract: Multi-scale renormalization has been employed successfully in several areas of mathematics. Within probability theory, it is an important tool to study stochastic processes like random walks, interacting particle systems and percolation in random environments. In these lectures, I will explain the use of this technique to study the phase transition for percolation. I will concentrate on the example of a percolation model on a random stretched lattice studied recently in a joint work with Sá, Sachis and Teixeira (https://arxiv.org/pdf/1912.03320.pdf).

Peter Müller (The University of Texas at Austin, USA)
Nonparametric Bayesian data analysis

Abstract: All models are wrong, but some are useful. Many statisticians know and appreciate G.E.P. Box' comment on statistical modeling. Often the choice of the final inference model is a compromise of an accurate representation of the experimental conditions, a preference for parsimony and the need for a practicable implementation. The competing goals are not always honestly spelled out, and the resulting uncertainties are not fully described. Over the last 20 years a powerful inference approach that allows to mitigate some of these limitations has become increasingly popular. Bayesian nonparametric (BNP) inference allows to acknowledge uncertainy about an assumed sampling model while maintaining a practically feasible inference approach. We could take this feature as a pragmatic characterization of BNP as flexible prior probability models that generalize traditional models by allowing for positive prior probability for a very wide range of alternative models, while centering the prior around a parsimonious traditional model. A formal definition of BNP is as probability models on infinite dimensional parameter spaces. A typical application of BNP is to density estimation.

In this shortcourse we review some of the popular models, including Dirichlet process (DP) models, Polya tree models, DP mixtures and dependent DP (DDP) models. We will review some of the general modeling principles, including species sampling models, stick breaking priors, and normalized random measures with indpendent increments. We will briefly discuss some of the main computational algorithms and available software. The discussion will be illustrated by applications to problems in biostatistics and bioinformatics.

Topics covered:

- *Definition of BNP and introduction*

- *Density estimation: Dirichlet process (DP), Stick breaking, DP mixtures, DP clustering, DP mixtures: posterior simulation, Polya trees (PT)*
- *Regression: BNP survival regression, Dependent DP (DDP), Anova DDP, Weighted mixture of DP,*
- *Hierarchical priors: Hierarchical DP,*

Target audience & prerequisites: *Anyone with an appreciation for data analysis, and basic knowledge of Bayesian inference. At the level of, for example, Hoff (2009),* A first course in Bayesian statistical models. *Or any other basic text in Bayesian inference.*

References: *Items #2-4 are free on-line, #1 probably as free pdf in your library (same for Hoff (2009))*

1. *The course will follow the book: Müller, P., Quintana, F., Jara, A., and Hanson, T. (2015),* Bayesian Nonparametric Data Analysis*, Springer.*
2. *Maybe read (before the course) P. Müller and R. Mitra (2013), "Bayesian Nonparametric Inference – Why and How",* Bayesian Analysis*, 8, 269-302.*
3. *Lecture notes of a similar course: Müller, P. and Rodriguez, A., (2012)* Nonparametric Bayesian Inference*, IMS Lecture Notes, free at* `https://projecteuclid.org/euclid.cbms/1362163742`
4. *Excellent notes by Peter Orbantz, at* `http://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf`

Valtencir Zucolotto (Universidade de São Paulo, Brasil)
Formação de Pesquisadores e a Escrita de Artigos Científicos de Alto Impacto

Abstract: A necessidade crescente de comunicação científica tem motivado muitos pesquisadores a produzirem artigos científicos para publicações em revistas internacionais. Contudo, a escrita correta e eficiente de artigos científicos representa ainda uma grande barreira ao pleno desenvolvimento científico de muitos pesquisadores. Nesta palestra abordaremos tópicos relevantes em escrita científica como i) ideias e projetos de pesquisa; ii) principais seções de um artigo científico, iii) estilo e linguagem da escrita científica em inglês e iv) o processo editorial.

Conferences

Inés Armendáriz (Universidad de Buenos Aires, Argentina)
Group testing with nested pools

Abstract. The purpose of group testing is to identify the set of infected individuals in a large population in the most efficient way. In this talk we will review some group testing methods, and we will introduce a nested strategy. In the first stage, individual samples are divided in equally sized groups called pools and a single test is applied to each pool. Individuals whose samples belong to pools that test negative are declared healthy, while each pool that tests positive is divided into smaller, equally sized pools, which are tested in the next stage. We iterate this procedure k times. Finally, in the last stage, all remaining samples are tested. Given the probability of infection p, we determine the optimal number of stages and the optimal sequence of pool sizes, in order to minimize the mean number of tests per individual. Based on joint work with P.A. Ferrari, D. Fraiman, J.M. Martínez and S. Ponce Dawson.

Lurdes Inoue (University of Washington, USA)
Pragmatic Bayesian sequential decision-making

Abstract: In this talk we consider a surveillance program which utilizes biomarkers for treatment recommendations. Specifically, at each surveillance time point two decisions must be made, namely, whether the individuals will have their biomarker measured and, if so, whether they should initiate treatment. Traditionally, when solving such a decision problem, Bayesian sequential decision-making resorts to dynamic programming which is quickly challenged by the "curse of dimensionality". To address this issue, we propose a pragmatic solution that converts the multi-stage sequential decision making into replicates of two-stage decision problems. Accounting for the individuals' biomarker histories as well as benefits and harms incurred with testing and treatment, using a small-scale simulation study we show that the proposed method has comparable performance to that obtained when using dynamic programming. In addition, using a case-study, we show that the method lowers costs by recommending less frequent testing, but without compromising patients' expected survival.

Nalini Ravishanker (University of Connecticut, USA)
Bayesian models for time series of counts

Abstract: Modeling count time series is an important research area with applications in many diverse domains, as discussed in the 2016 CRC Handbook of Discrete-

valued Time Series. We sketch the use of Markov Chain Monte Carlo (MCMC) methods for Bayesian hierarchical dynamic modeling of vector time series of counts under a multivariate Poisson sampling distributional assumption, with an illustration to ecology data. However, this approach can be computationally demanding, especially in high dimensions. We describe an alternate flexible level correlated model (LCM) framework, which enables us to combine different marginal count distributions and to build hierarchical models for vector time series of counts, while accounting for association between components of the response vector. We employ the Integrated Nested Laplace Approximation for fast approximate Bayesian modeling, and illustrate the approach using an example on ride sourcing data in NYC using the R-INLA package.

Mini conferences

Alejandro Róldan (Universidad de Antioquia, Colombia)
Dispersion as a survival strategy in populations exposed to catastrophes.

Abstract: We consider stochastic growth models to represent population dynamics subject to catastrophes. We analyze the subject from different set ups considering or not spatial restrictions, whether dispersion is a good strategy to increase the population viability. We find out it strongly depends on the effect of a catastrophic event, the spatial constraints of the environment and the probability that each exposed individual survives when a disaster strikes. This is a joint work with V.V. Junior and F.P. Machado.

Aritra Halder (University of Virginia, USA)
Bayesian wombling for spatiotemporal Gaussian processes.

Abstract: With recent advances in geographic information systems, encountering spa- tiotemporally indexed data are abound. Spatiotemporal models envision a response surface possibly in $\mathbb{R}^d$, that evolves continuously over time, $t \in \mathbb{R}^+$. Such models generally begin by specification of a spatiotemporal stochastic process $Y(s,t)$, where $(s,t) \in \mathcal{S} \times \mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}^+$. The choice of response for such applications is generally motivated by the domain of application, they can be concentrations of pollutant or, particulate matter in air observed over a fixed spatial domain across time. Depending on the model, such processes can be directly assigned on the response or, modeled as an underlying latent geophysical process quantifying spatial and temporal variation. The granularity at which spatial indexing occurs varies between point-referencing (latitude-longitude) or, aggregated areal-referencing (census tracts, counties, zip-codes etc.). Temporal specification could occur at the level of days, months, year etc. Inference for spatiotemporal process models involve, (a) inference on model parameters associated with covariates responsible for variations in the trend surface (b) analyzing spatiotemporal variation in topological characteristics of the produced random surface from the specified model. This entails producing inferential frameworks for spatiotemporal curvilinear gradients, curvature, topological characterization of points, for example elliptic or saddle points, troughs peaks, which characterize change in response over the spatial reference domain across time. In this talk we outline a Bayesian hierarchical modeling framework that allows for exploration of such topological properties of Gaussian spatiotemporally indexed processes. We apply the developed framework to model temperatures in the northeastern United States.

Márcio Augusto Diniz (Cedars-Sinai Medical Center, USA)

Shrinkages priors for single cell experiments

Abstract: Recent advances in single-cell RNA-sequencing (scRNA-seq) allow large-scale quantitative transcriptional profiling of individual cells across a wide range of tissues such that clusters of different cell types are established, which can be seen as compositional data. Investigators are often interested in identifying clusters of cells that predict treatment response in clinical trials. However, the limited sample size of scRNA-seq experiments pose a challenge that require the use of shrinkage methods. Under a Bayesian approach, several shrinkage priors based on scale mixture of normal distributions have been discussed in the statistical literature. In this work, we study the operating characteristics of a selected set of priors (double exponential, horseshoe, logistic-normal and regularized horseshoe) with a Monte Carlo simulation and draw some conclusions about their use on scRNA-seq data.

Student talks

Jessica Suzana Barragan Alves (PIPGEs UFSCar/USP)
Flexible links to binomial regression models under Bayesian approach

Abstract. Several asymmetrical links with an extra parameter were proposed in the literature over last few years to deal with imbalanced data in binomial regression (when one of the classes is much smaller than the other). In this paper, we introduce flexible links for modeling binomial regression models which include an extra parameter associated with the link that explains some unbalancing for binomial outcomes. For all cases, the cloglog is a special case or the reciprocal version loglog link is obtained. A Bayesian MCMC inference approach is developed. A simulation study to evaluate the performance of the proposed algorithm is conducted. The results show a good performance of the proposed function links for imbalanced data.

Rodrigo Ferrari Lucas Lassance (PIPGEs UFSCar/USP)
Nonparametric pragmatic hypothesis testing

Abstract. Standard statistical tests have at least three major issues that have become more explicit in recent years: (i) outcomes that do not adhere to what a researcher wants to know, (ii) logical contradictions when applying multiple tests and (iii) rejection of precise hypotheses that are not relevant in a practical perspective. All of these problems are solved through the use of agnostic tests and pragmatic hypotheses. However, no study has yet been made to solve these issues in nonparametric tests, which is the objective of this paper. By expanding the theory in Coscrato et al. (2019), we delimit the different types of precise hypotheses of interest and the respective challenges each of them presents. Then, we provide an example of its application, showing how one can proceed to delimit a nonparametric pragmatic hypothesis and derive a Bayesian test that challenges (i)-(iii).

Vitor Amorim (PIPGEs UFSCar/USP)
Asymptotic Kac's lemma for hitting times

Abstract. We consider a sequence of symbols $A = a_0 \dots a_{n-1}$ of an alphabet A and ask for the first time that a stochastic process hits this sequence. This is the hitting time T_A and when the process starts by A, we call it the return time. The classic Kac's Lemma states for ergodic processes the intuitive fact that the expected return time to A is the inverse of the stationary measure of A. There is no similar formula for the expected hitting time. We present an exponential approximation for the hitting *time distribution, with an explicit parameter and a sharp error term, which allows us to derive an asymptotic version for the Kac's Lemma for all the moments of the hitting times. We also states some properties of the parameter of the exponential approximation that have consequences in the main result*

Organization