Workshop on Statistical Modeling and Data Science UC 2024

Facultad de Matemáticas UC

August 12-13, 2024

This event aims to serve as a convergence point between the disciplines of statistics and data science, with significant relevance for education, training, collaboration, and knowledge promotion in this field. Its objective is to make an impact in the following areas:

- **Education and training:** The invitation of experts in statistical modeling provides a unique opportunity to educate and train students in programs related to statistics and data science. Given the substantial impact of these topics across various disciplines and sectors, we believe that a workshop like the one proposed can significantly enhance the skills and knowledge of the participants.
- **Knowledge update:** Statistics and data science are constantly evolving fields. Organizing a workshop of this nature allows participants to stay updated with the latest trends and tools in the field.
- **Exposure to real-world problem-solving:** Since the workshop will feature participation from distinguished national and international speakers working in various applications, real-world problems and challenges are expected to be presented. Participants can learn how these problems were tackled using the techniques they've learned.
- **Networking:** The workshop aims to bring together individuals who share professional and academic interests. This platform undoubtedly aids in the creation of valuable professional networks that may lead to employment opportunities, research collaborations, among others.
- **Knowledge dissemination:** Finally, the workshop aims to have a broader impact as participants share their knowledge and learnings with their colleagues and students outside of UC.

Background

The main activities to be conducted include an Inaugural Conference, in which a prestigious international guest will deliver a lecture of approximately 50 minutes on recent research work. Additionally, invitations are extended to seven national or international speakers to give a 30-minute talk (plus 10 minutes for questions and answers).

Schedule

Time	Monday 12th	Tuesday 13th
09:00 - 09:30	Registration	Registration
09:40 - 10:10	Garrit Page	Rolando de la Cruz
10:20 - 10:50		Christian Caamaño
11:00 - 11:30	Coffee break	Coffee break
11:40 - 12:10	Inés Varas	Pedro Ramos
12:20 - 12:50	Darlin Soto	Jonathan Acosta
13:00 - 13:30	Felipe Elorrieta	Closing Ceremony

Talks

Garritt L. Page

Joint Random Partition Models for Multivariate Change Point Analysis

Abstract

Change point analyses are concerned with identifying positions of an ordered stochastic process that undergo abrupt local changes of some underlying distribution. When multiple processes are observed, it is often the case that information regarding the change point positions is shared across the different processes. This work describes a method that takes advantage of this type of information. Since the number and position of change points can be described through a partition with contiguous clusters, our approach develops a joint model for these types of partitions. We describe computational strategies associated with our approach and illustrate improved performance in detecting change points through a small simulation study. We then apply our method to a financial data set of emerging markets in Latin America and highlight interesting insights discovered due to the correlation between change point locations among these economies.

Inés Varas

Predictive Validity through Missing Outcome Data

Abstract

Regression models are frequently used to analyze the impact of covariates on a particular outcome, with model coefficients quantifying these effects. In university predictive validity studies, coefficients are used to quantify the relationship strength between test scores and outcomes like academic performance. However, applying these models within selection processes introduces a significant statistical challenge. While test scores are available for all applicants, university performance data is only available for those admitted, leading to a missing outcome problem. This results in both the regression model and its parameters being non-identified. Although typical ignorable assumptions are often used to address missing data, they are not valid in the context of university selection tests. To address this issue, we propose a partial identification approach to better understand how missing data influences the relationship between selection criteria and academic success.

Darlin Soto

Spectral Density of X-ray Binary Systems

Abstract

The study of X-ray binary systems has sparked significant interest due to its crucial insights into physical phenomena, such as accretion flows around black holes. These studies are primarily conducted in the frequency domain through the power spectral density (PSD). Astronomers have observed that the PSD of binary systems can be explained by two components: noise and quasi-periodic oscillations (QPOs) and that QPOs can be fitted using Lorentzian functions. In this talk, we will see how ARMA time series models can also be used to describe the PSD of binary systems, demonstrating that the PSD of these models can also be decomposed as a sum of functions and that these functions can describe the QPOs.

Felipe Elorrieta

Progress in Irregularly Observed Autoregressive Models

Abstract

Time series analysis is generally performed on regularly observed data, assuming discrete time intervals. Thus, when observing a time sequence irregularly, it is often adjusted using a continuous time model or by transforming the sequence into an equally spaced time series through interpolation. In previous work, we proposed an alternative approach to model autoregressive processes observed irregularly in discrete time. This formulation provides greater flexibility in the model estimation procedure, allowing us to extend our models to negative autocorrelation, non-Gaussian, and/or multivariate data. The five models we have proposed so far following this approach are: iAR models, iAR-Gamma, iAR-T, CiAR, and BiAR. These models are implemented in a new package called iAR, available in both R and Python. The choice of each model depends on the type of time series being fitted. For instance, if the available data has a non-Gaussian distribution or is not positively autocorrelated, the appropriate model will vary. Our advancements in these models are mainly related to estimation methods. Among them, we aim to address the bias problem that conventional estimators (Maximum Likelihood and Least Squares) have in estimating autoregressive processes close to a unit root. To improve these estimation procedures, we developed a simulation-extrapolation (SIMEX) methodology in time series analysis. Results from simulations and real-world data show that this methodology significantly enhances the quality of estimates, reducing bias and increasing accuracy. Furthermore, we have proposed sequential estimation methods for the parameters of irregularly observed autoregressive models based on algorithms such as the Kalman Filter and Particle Filter. The results show that these online estimation methods are computationally efficient and capable of quickly adapting parameter estimates in response to changes in time series behavior, facilitating the detection of structural changes.

Rolando de la Cruz

Longitudinal Data Classification through Nonlinear Mixed Effects Models with Heterogeneity in the Random Effects in some Subpopulations

Abstract

A popular approach to analyzing and capturing the dynamics present in complex longitudinal data is frequently conducted with non-linear mixed effects (NLME) models under the regular assumptions of normality. The NLME models with normal distribution are commonly used for modeling complicated longitudinal trajectories, assuming that individuals come from homogeneous populations with normal errors. However, a homogenous population assumption may inappropriately ignore significant aspects related to between-individual and within-individual variability, leading to incorrect modeling outcomes. One way to address this situation is to propose NLME models that incorporate heterogeneity in the random effects, where we relax the assumption of a homogeneous subpopulation, which allows us to distinguish different parameters between several unobserved classes within a heterogeneous subpopulation. Following this idea, in the present work, we propose to fit NLME models with heterogeneity in the random effects using a stochastic version of the EM algorithm to estimate the model parameters and the probability of occurrence of an individual belonging to new classes observed within the subpopulations. To illustrate our proposal, we consider a study clinic related to the levels of the hormone β-HCG in pregnant women, a biomarker used to indicate changes during pregnancy. Our proposal provides a framework for modeling longitudinal trajectories, assuming individuals come from heterogeneous subpopulations. Finally, using the proposed models and considering the pregnant women's data, we class them into two groups, distinguishing between normal (baby birth) and abnormal (miscarriages) pregnancies and discretizing between the new possible subgroups or classes that emerge within the subpopulation of abnormal pregnancies by assuming heterogeneity.

Christian Caamaño

Discrete random fields

Abstract

Spatial discrete data are common in many fields and examples can be found in mining (counts of diamonds in kimberlite pipes, or of gold grains in alluvial placer deposits), forestry (counts of trees of a given species), ecology (sightings of wild animals), epidemiology (disease mapping based on reported infection cases), environmental sciences (radioactivity counts), criminology (thief counts). Binomial, negative binomial and Poisson distributions are very popular marginal model for independent and not identically distributed data. However in fitting regression model to spatial data, one needs to account for spatial dependence to ensure reliable inference for the regression coefficients. First, we propose a discrete random field by considering a sequence of independent copies of the Bernoulli random fields different sampling procedures lead to define random fields with binomial and negative binomial marginal distributions. Our approach can be viewed as a generalization of the spatial Bernoulli random fields proposed in Heagerty and Lele (1998). Second, we propose a random field with a Poisson marginal distribution considering a sequence of independent copies of a random field with an exponential marginal distribution as "inter-arrival times" in the counting renewal processes framework. Our proposal can be viewed as a spatial generalization of the Poisson counting process. For the proposed spatial random fields, analytic expressions for the covariance function and the bivariate distribution are provided. In an simulation study, we investigate the weighted pairwise likelihood as a method for estimating the Poisson random field parameters. Finally, the effectiveness of our methodology is illustrated by an analysis of reindeer pellet-group survey data, where a zero-inflated version of the proposed model is compared with zero-inflated Poisson Log-Gaussian and Poisson Gaussian copula models.

Pedro L. Ramos

Integrated Approaches in Statistical Modeling of Repairable Systems

Abstract

This work presents a suite of advanced statistical models for the analysis of repairable systems, focusing on Bayesian approach in scenarios of competitive risks. We develop hierarchical models for single repairable systems, where each model generalizes and expands previous concepts with power-law failure intensities and Bayesian statistical inference. The research advances in reliability modeling and risk assessment at various hierarchical levels of the system. Through Monte Carlo simulation studies, we evaluate the quality of our proposed approach and apply our methodologies in practical examples, such as the development of robotic units. These models provide a comprehensive and detailed insight into the analysis and management of repairable systems, demonstrating the applicability and effectiveness of these techniques in the context of complex system engineering.

Jonathan Acosta

Nonparametric Estimation of the Variogram and the Effective Sample Size

Abstract

This work introduces a novel approach for the nonparametric estimation of the variogram and the Effective Sample Size (ESS) in the context of spatial statistics. The variogram is crucial in modeling intrinsically stationary random fields, particularly in spatial prediction using kriging equations. Existing nonparametric variogram estimators often lack the guarantee of producing a conditionally negative definite function. To address this, we propose a new valid variogram estimator based on a linear combination of functions within a specified class, ensuring the satisfaction of critical properties. A penalty parameter that avoids overfitting is incorporated, thus eliminating spurious fluctuations in the estimated variogram function. Additionally, we extend the notion of the ESS, a crucial measure in spatial regression processes, to a nonparametric setting. The proposed nonparametric ESS relies on the reciprocal of the average correlation and is estimated using a plug-in approach. The proposed estimators' theoretical properties and consistency are discussed, and numerical experiments demonstrate their performance. This work not only enhances the robustness and flexibility of spatial statistical analyses through nonparametric methods but also extends the discussion to the application of these methods in spatiotemporal data, in both the nonparametric estimation of the variogram and the ESS, respectively, thereby broadening the potential impact of our research.