Breaking the Zero Wall: Robust Poisson Hurdle Models

Conceição Amado, Instituto Superior Técnico and CEMAT

Many real-world problems involve count data with excess zeros, overdispersion and outliers, where standard Poisson or negative binomial models perform poorly. Typical examples arise in health-related outcomes (such as the number of adverse events per patient or per facility), in industrial quality monitoring (for instance, defect counts in steel manufacturing processes), and in the analysis of rare events per unit of time. Hurdle models address these challenges by separating the binary mechanism that governs whether any event occurs from the process that generates strictly positive counts. In this talk, I present a robust estimation approach for the Poisson hurdle model based on minimum distance methods. The procedure, built on distances such as the Hellinger and total variation distances, is applied separately to the binary and truncated Poisson components, preserving the interpretability of the two-part structure. A simulation study inspired by applications with many zeros and atypical observations shows that the minimum distance estimators maintain good efficiency in clean samples while substantially improving robustness relative to maximum likelihood and other existing estimators. The methodology will also be illustrated with practical examples, including industrial defect data, highlighting how robust hurdle modelling can support real-world decision-making in diverse domains.

Uncovering True Engagement: A Mixture Model for Active vs. Unattended Sessions

Marília Antunes, Faculdade de Ciências da Universidade de Lisboa and CEAUL

The CINDERELLA project aims to improve aesthetic outcomes and patient satisfaction following breast cancer surgery through an AI-based application that predicts post-surgical aesthetic results from patient data and medical images. App usage data provides an opportunity to study user behaviour; however, a software bug caused incorrect recording of session times. Recovering durations from login data may generate inflated values between sessions, while additional inflation may occur when users leave the app open without active interaction. Thus, observed usage times may arise from two latent processes: active and inactive periods.

A two-component finite mixture model is used to classify session times into exact or inflated categories. Log-normal and Gamma specifications were explored, incorporating socio-demographic, clinical, and user–application interaction covariates through mixtures of generalized linear models. Model estimation is performed using EM algorithms in a maximum likelihood framework and NUTS algorithms in a Bayesian framework.

From Data to Dialogue: Building Connections in Women-Centred Statistics and Data Science

Lisete Sousa, Faculdade de Ciências da Universidade de Lisboa and CEAUL

Statistics and data science are often presented as technical disciplines driven by methods, algorithms, and computational power. Yet, behind every dataset, there are people, stories, and communities that shape how knowledge is created and shared. In this talk, I reflect on my journey as a statistician and researcher working across interdisciplinary collaborations in health, life sciences, and bioinformatics, highlighting the role that dialogue, mentorship, and scientific generosity have played throughout that path.

Rather than focusing solely on methodological advances, this presentation explores how meaningful scientific work is frequently built through collaborative environments where diverse perspectives are valued and encouraged. I will discuss how women researchers and mentors have influenced my own scientific identity and how these experiences continue to shape my approach to teaching, supervision, and research collaboration.

The talk will also briefly present current work related to microbiome research and menopause-associated health risks, a project combining statistical learning, bioinformatics, and women's health data to better understand complex biological and social dimensions of ageing.

By sharing these experiences, I hope to contribute to a broader conversation about visibility, inclusion, and the importance of supportive scientific communities that empower future generations of women in statistics and data science.

Interpretable Machine Learning for Prognosis: When Sparsity Meets Biological Networks

Susana Vinga, INESC-ID, Instituto Superior Técnico and IDMEC

Many biomedical prediction tasks, such as tumor classification, relapse risk, and other time-to-event outcomes, are dominated by the "small n, large p" regime, where multiple models can fit the data equally well, and naïve feature selection becomes unstable. In this talk, I discuss how sparse learning and structured regularization help turn ill-posed problems into actionable and reproducible knowledge. I will introduce regularized generalized linear models and Cox regression enriched with network information, such as gene centrality, correlation structure and smoothness over biological graphs, to obtain compact and robust biomarker panels with clearer mechanistic interpretation. Drawing on cancer case studies, I will illustrate how network-aware penalties can improve generalization, reduce signature variability, and support clinically meaningful risk stratification. I will also emphasize practical validation lessons for high-dimensional biomedical data, including repeated resampling, stability analysis and transparent reporting of uncertainty. The goal is to bridge statistics, optimization and life sciences to build models that remain interpretable enough to support scientific insight and, ultimately, clinical decision support in precision medicine.