SiDOR
Statistical Inference
Decision & Operations Research Group

Universidade Vigo

# Workshop on Statistics

## July 22nd, 2022

**Facultade de Ciencias Económicas e Empresariais**
**– Aula-Seminario 6 –**
**Universidade de Vigo**

## General Schedule

**11:00–11:35 A generalized additive model (GAM) approach to principal component analysis of geographic data.** Francisco de Asís López Álvarez (Universidade de Vigo)

**11:35–12:10 Clustering of nonparametric curves by the `clustcurv` package.** Marta Sestelo (Universidade de Vigo)

**12:10–12:45 Inverse probability weighted Cox regression to correct for ascertainment bias .** Mar Rodríguez-Girondo (Leiden University Medical Center, The Netherlands)

**12:45–13:20 The fundamental role of density functions in the binary classification problem.** Pablo Martínez-Camblor (Dartmouth-Hitchcock Medical Center and Geisel School of Medicine at Dartmouth, USA)

# Abstracts

## A generalized additive model (GAM) approach to principal component analysis of geographic data

**Francisco de Asís López Álvarez (Universidade de Vigo)**

*Abstract:* Geographically Weighted Principal Component Analysis (GWPCA) is an extension of classical PCA to deal with the spatial heterogeneity of geographical data. This heterogeneity results in a variance-covariance matrix that is not stationary but changes with the geographical location. Despite its usefulness, this method presents some unsolved issues, such as finding an appropriate bandwidth (size of the vicinity) as a function of the retained components. In this work, we address the problem of calculating principal components for geographical data from a new perspective that overcomes this problem. Specifically we propose a scale-location model which uses generalized additive models (GAMs) to calculate means for each variable and a variance-covariance matrix that relates the variables, both depending on the spatial location. This approach does not require to calculate an optimal bandwidth as a function of the number of components retained in the analysis. Instead, the covariance matrix is estimated using smooth functions adapted to the data, so the smoothness can be different for each element of the matrix.

## Clustering of nonparametric curves by the `clustcurv` package

**Marta Sestelo (Universidade de Vigo)**

*Abstract:* One basic but important goal in the statistics field is the comparison of curves between groups. Several nonparametric methods have been proposed in the literature to test for the equality of nonparametric curves. In this framework, when the null hypothesis of equality of curves is rejected, it can be interesting to ascertain whether curves can be grouped or if all these curves are different from each other. Software in the form of an R package (`clustcurv`) has been developed in order to allow determining groups with an automatic selection of their number. The package can be used for determining groups in multiple survival curves as well as for multiple regression curves. The applicability of the proposed methods is illustrated using real data.

### Inverse probability weighted Cox regression to correct for ascertainment bias

**Mar Rodríguez-Girondo (Leiden University Medical Center, The Netherlands)**

*Abstract:* Motivated by the study of genetic effect modifiers of cancer, we examine weighting approaches to correct for ascertainment bias of covariate effects in the context of Cox proportional hazards regression. (Family-based) outcome-dependent sampling is common in genetic epidemiology leading to study samples with too many events in comparison to the population and an overrepresentation of young, affected subjects. A usual approach for correcting for ascertainment bias in this setting is to use an inverse probability weighted Cox model, using weights based on external available population-based age-specific incidence rates of the type of cancer under investigation. However, the current approach relies on the assumption of oversampling of cases of all ages which is not realistic in relevant practical settings. Based on the same principle of weighting observations by their inverse probability of selection, we propose a new, more general approach. We compare the methods in simulations and illustrate the advantage of our new method with two real datasets. In both applications, the goal is to assess the association between common susceptibility loci identified in Genome-Wide Association Studies (GWAS) and cancer (colorectal and breast) using data collected through genetic testing in clinical genetics centers of the Netherlands.

### The fundamental role of density functions in the binary classification problem

**Pablo Martínez-Camblor (Dartmouth-Hitchcock Medical Center and Geisel School of Medicine at Dartmouth, USA)**

*Abstract:* In biomedical research, binary classification problems arise in a wide variety of problems, mainly involved in diagnostic and prognostic tasks but also have presence, for instance, in personalized medicine. The overall objective is to use the available information to correctly allocate subjects in groups. Frequently, this information implies high-dimensional data and the rationality behind the derived classification rules is difficult to understand. An adequate classification rule is frequently a trade-off between sensitivity and specificity. The ROC curve is a graphical tool which helps to understand, evaluate and compare the accuracy of diagnostic processes. We propose a procedure for estimating the optimal classification rules based on a penalized estimator of the underlying probability distribution functions in both the negative and positive populations. We study its asymptotic properties and its finite-sample behavior. Through Monte Carlo simulations, we compare our proposal with a support vector machine based ROC curve. We also illustrate its practical use on a real-world data problem. Results suggest that, despite some modern techniques promise to improve the results provided by other more traditional methods, in the binary classification problem, the limit is the actual relationship among the density functions. Statistical methods can provide a close approximation for the targeted quantity. Besides, to keep some rationality in the statistical analyses could result in much better classification accuracy than those based on computational power.