Book of Abstracts

COMPSTAT 2016

22nd International Conference on Computational Statistics

August 23-26, 2016

Satellite CRoNoS Workshop and Summer Course Functional Data Analysis

August 26-28, 2016

August 23-28, 2016

Auditorio Príncipe Felipe, Oviedo, Spain

POSTER SESSION III

CP001 **Room Ground Hall** **Chair: Francisco Torres-Ruiz**

CP0457: **Supervised classification using a distance-depth function**
Presenter: **Itziar Irigoien**, University Basque Country, Spain
Co-authors: Concepcion Arenas, Francesc Mestres

Supervised classification is used by researchers in a wide variety of fields as in taxonomic classification; in morphometric analysis for species identification; in ecological problems addressed to test the presence or absence of a particular species; in marine ecology to evaluate the similarity of distinct populations and to classify units of unknown origin to known populations; in genetic studies in order to summarize the genetic differentiation between groups or in the biomedical context, predicting the diagnostic category of a sample on the basis of its gene expression profile and some clinical features. A novel classifier rule is introduced based on an improvement of the distance-based discriminant (DB-discriminant), taking into account a depth function. This new model combines the DB-rule and the maximal depth classifier, obtaining a classifier that is often more accurate than both methods separately. To demonstrate its effectiveness the new classifier was compared with the DB-rule and the $k-$nearest neighbor classification method, using high-dimensional class-imbalanced cancer data sets, and evaluating the leave-one-out error rate, the generalized correlation coefficient, the sensitivity, the specificity and the positive predicted value for each class. The results show the good performance of the new classifier.

CP0467: **Estimation in the functional convolution model**
Presenter: **Tito Manrique**, UMR MISTEA - INRA Montpellier SUPAGRO, France
Co-authors: Christophe Crambes, Nadine Hilgert

An estimator is proposed for the unknown function in the Functional Convolution Model, which studies the relationship between a functional covariate $X(t)$ and a functional response $Y(t)$ through the following equation $Y(t) = \int_0^t \theta(s)X(t-s)ds + \varepsilon(t)$, where θ is the function to be estimated and ε is an additional functional noise. In this way we can study the influence of the history of X on $Y(t)$. We use the Continuous Fourier Transform to define an estimator of θ. The transformation of the convolution model results in the Functional Concurrent Model associated, in the frequency domain, namely $\mathcal{Y}(\xi) = \beta(\xi)\mathcal{X}(\xi) + \varepsilon(\xi)$. In order to estimate the unknown function β, we extended the classical ridge regression method to the functional data framework. We establish consistency properties of the proposed estimators and illustrate our results with some simulations.

CP0505: **A novel two-step iterative approach for clustering functional data**
Presenter: **Zuzana Rostakova**, Slovak Academy of Sciences, Slovakia
Co-authors: Roman Rosipal

An important task in functional data analysis is to divide a dataset into subgroups with similar profiles, or clusters. We address a problem in which classical functional data clustering techniques may fail when curve misalignment is present. Solutions in which registration or temporal alignment of the whole dataset precede the clustering step result in rapid distortions in the curve shapes when a dataset consists of many different curve profiles. Methods developed for simultaneous registration and clustering of curves mainly deal with linear transformation of time. This solution may also lead to unsatisfactory alignment when profiles of the curves are complex or the source of misalignment has a nonlinear character. We propose and validate a novel two-step approach, which iteratively combines clustering using a modified Dynamic Time Warping algorithm with the registration step applied separately to curves within estimated clusters. On generated and real functional data representing the sleep process we demonstrate the validity of the approach by measuring improvement in similarity between aligned curves in comparisons to: a) the case when clustering and registration steps are applied separately and b) other methods (e.g. k-means alignment, joined probabilistic curve clustering and alignment) for simultaneous curve registration and clustering.

CP0510: **Prediction of disease risk by high-dimensional genetic and environmental data**
Presenter: **Norbert Krautenbacher**, Technical University of Munich and Helmholtz Center Munich, Germany
Co-authors: Christiane Fuchs, Fabian Theis

The aim is to investigate the situation of having high-dimensional genetic and environmental data of individuals where the goal is to build a prediction model for the risk of suffering from the disease asthma. At the study one was also interested in the influence of the specific exposure variable farm-environment, so that a sample of the population should contain an appropriate number of observations with the combination farm/asthma. Since in the population both categories occur only rarely, a simple random sample would require a big sample size. In practice, however, it is not possible to take such a big sample, since collecting genomic data in terms of hundreds of thousands to millions of single-nucleotid polymorphisms (SNPs) is cost-intensive. Thus, a stratified random sample was taken from the population. Therefore, for analyzing the final sample two main issues occur: first, one has to correct for the arisen sample selection bias when learning and evaluating on biased training and test data sets. Second, the present genetic data containing 2.5 million SNPs have to be incorporated as features for dimension reduction and feature selection techniques which require special solutions.

CP0504: **Modified profile likelihood in complex models with many nuisance parameters**
Presenter: **Claudia Di Caterina**, University of Padova, Italy
Co-authors: Nicola Sartori

It is well known that usual frequentist inference procedures for a parameter of interest are generally highly inaccurate when dealing with statistical models where the number of nuisance parameters is large relative to the sample size. Among the alternative proposals put forward in the literature, the modified profile likelihood has proved to represent a valid solution to the problem. Specifically, the approximation to such pseudo-likelihood previously introduced allows us to overcome some difficulties related with its computation outside the class of exponential and group family models. Nevertheless, even this modification of the profile likelihood can be hard to obtain analytically under moderately complex scenarios. In order to further enlarge the domain of applicability of this technique, Monte Carlo simulation can be used to evaluate some expected values involved in the modified profile likelihood. It is shown how such an approach succeeds in providing a reliable inference on the parameter of interest in various frameworks, all considering a panel data structure: microeconometric fixed effects models with continuous or discrete response, models for datasets with missing values in the dependent variable or in the covariates, and parametric survival models for censored data.

CP0217: **Flexible Birnbaum-Saunders models**
Presenter: **Heleno Bolfarine**, University of Sao Paulo, Brazil

We introduce a new extension of the Birnbaum-Saunders distribution as a follow up to the family of skew-flexible-normal distributions. This extension produces a family of Birnbaum-Saunders distributions including densities that can be unimodal as well as bimodal. This flexibility is important in dealing with positive bimodal data, given the difficulties experienced by the use of mixtures of distributions when bimodality is present. Some basic properties of the new distribution are studied including moments. Parameter estimation is approached via the method of moments and also by maximum likelihood, including a derivation of the Fisher information matrix. Computational aspects of maximum likelihood implementation is discussed. Real data illustrations indicate satisfactory performance of the new model.

CP0235: **Influence of missing data on the estimation of the number of components of a PLS regression**
Presenter: **Frederic Bertrand**, Universite de Strasbourg, France
Co-authors: Nicolas Meyer, Myriam Maumy-Bertrand

 All rights reserved