INTERNATIONAL WORKSHOP ON PROXIMITY DATA, MULTIVARIATE ANALYSIS AND CLASSIFICATION

FACULTAD DE CIENCIAS. UNIVERSIDAD DE GRANADA
October, 9-10, 2014
PROGRAMME

All talks will be held at the Conference Room

THURSDAY
OCTOBER 9

12.00 Welcoming remarks

12.05 Inaugural Talk: Cluster Analysis Through Model Selection
Elías Moreno, *Universidad de Granada*
Chair: José Fernando Vera, UGR

Clustering is an important and challenging statistical problem for which there is an extensive literature. Modelling approaches include mixture models and product partition models. Here we develop a product partition model and search algorithm driven by Bayes factors from intrinsic priors. The priors we develop for the partitions, and the number of clusters in the partition, lead to finding partitions with a smaller number of clusters, which does not happen if a uniform prior is used. We argue that this is desirable, since one reason for doing a cluster analysis is to find a small number of clusters that can help to understand underlying structure. However, we are also able to establish that our procedure is consistent, and hence will find the true underlying structure as the sample size increases. We illustrate our cluster algorithm with both simulated and real examples.

14.00 Lunch

16.00 Session 1
Chair: M^a del Mar Rueda. UGR

Jackknife for Randomized Response Techniques
Beatriz Cobo, *Universidad de Granada*

Randomized Response Technique introduced by Warner is a possible solution for protecting the anonymity of the respondent in a survey and is introduced to reduce the risk of escape or no response sensitive questions. This technique consists in the use of a random mechanism by which is selected one of two complementary questions: Do you belong to the group with the characteristic A? or Do you belong to the group that has not the characteristic A ?, where A is the sensitive characteristic. The respondent will answer yes or no and the interviewer does not have the ability to know what questions the respondent answered, thus protecting its confidentiality. A problem with these techniques is the variance estimate, since it depends on the random mechanism chosen and need to calculate the second-order inclusion probability of each pair of sample units. In some complex sampling designs, this is very complicated. An alternative is to obtain the variance using resampling techniques, such as jackknife.

Using multivariate auxiliary information to estimate discrete response variables in dual frame surveys

David Molina, *Universidad de Granada*

When conducting a survey, it is usual to collect information on multiple variables in addition to on variables of interest. These additional data are somehow related with response variables, they can be incorporated to the estimation process to improve accuracy of results obtained. Dual frame is a methodology recently arisen as an alternative to classic sampling theory. This new approach assumes that two frames are available for sampling and that, overall, they cover the entire target population. Extension of dual frame estimation techniques to the case of estimation of proportions when the variable of interest has discrete outcomes in the presence of multivariate auxiliary information is presented. Several estimators for the population proportions are calculated by using two different approaches: "single frame" and "dual frame". To check the efficiency of the proposed procedures in presence of different sets of auxiliary variables, some Monte Carlo experiments were carried out.

Nonparametric estimation in multiple frame surveys

Ismael Sánchez-Borrego, *Universidad de Granada*

Nonparametric regression methods have been used extensively for estimating the regression function in a wide range of fields. They allow the model to be correctly specified for much larger classes of functions, and have a great potential for application to a wide range

of problems. Kernel-based methods are well known for their good properties and for their adaptation to different settings. We consider the problem of estimating the population total in multiple frame surveys. In classic finite population sampling a basic hypothesis is the availability of a unique and complete list of units forming the target population to be used as a sampling frame. In practice frames that can be used for selecting the samples are generally incomplete or out of date. Sometimes a set of two or more lists is available for survey purposes. Multiple frame surveys have gained much attention and became largely used by statistical agencies and private organizations to decrease sampling costs or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. We propose a model-assisted estimator based on kernel regression that can handle continuous covariates. Asymptotic properties of the proposed method are derived and numerical experiments shows that it performs well in practical settings under different scenarios.

17.15 Coffee Break

17.45 Session 2

Chair: José Luis Zafra, Universidad de Granada

Default prediction in the construction industry

Juan Lara-Rubio, *Universidad de Granada*

At present, not discussed that the origin of the financial crisis in late 2007 came in the construction sector. The construction industry has been the main engine of the economy in many countries, especially in Spain. However, the economic recession has highlighted a number of consequences that can be considered to set actions in future similar crisis. This paper buids a non-parametric bankrupcy models based on the multilayer perceptron approach (MPL) and benchmarks their performance against other model which employ the tradicional logistic regression technique. Based on a sample of 25506 companies and constractors from construction industry, the results reveal that neural network models outperform the other technique both in terms of area under the receiver-operating characteristic curve (AUC) and as miscassifiaction cost.

Exploring long time effect on local government efficiency through a multivariate analysis

Gemma Pérez-López, *Universidad de Granada*

With the increase in public services offered at the local level and heightened budgetary and financial constraints on local government, new ways of managing local public services must be sought, in order to maximise the efficiency of resource management. Accordingly, local governments have introduced organisational changes, through contracting out, the creation of public agencies and increased cooperation with other municipalities and private companies. However, previous studies have only examined the relationship between contracting out and efficiency, and no in-depth survey has been made of the relationship between efficiency with other forms of management, nor of the joint impact of these forms of management on efficiency. The main aim of this study is to determine whether NPM delivery forms do in fact improve the efficiency of Spanish local governments. In this regard, we analyse the particular impact of the global recession. Our results suggest that in general terms the creation of agencies, contracting out and inter-municipal cooperation reduce cost-efficiency. However, these results also lead to the conclusion that during the global recession, some of these NPM delivery forms tended to become more efficient. Thus, empirical evidence suggests that the adoption of mixed firms contributes to higher levels of cost efficiency in the whole period considered, and thus it may be a suitable instrument in periods of crisis.

Multiple linear regression on factors determining online sustainability reporting

Francisco José Alcaraz, *Universidad de Granada*

In the public sector, diverse stakeholders are calling for the actions of public bodies to be more socially responsible and sustainable. This paper examines the real extent of sustainability information reporting by local governments in Spain, to identify and analyse the factors underlying and affecting this disclosure. As public administrations are making increasing use of the Internet to communicate with stakeholders, the methodology applied in this study is to analyse the websites of 55 major towns and cities, seeking 61 items that are recommended in the Global Reporting Initiative guidelines. By

applying multiple linear regression analysis, we identify the impact made by 13 factors on the sustainability reporting provided by these municipal authorities. The results show that social information is the most widespread, and that the dependent population has a positive impact on the sustainability disclosures in Spanish local governments. The major opportunities to improve sustainability practices lie in improving the disclosure of economic and environmental information, through the publication of formal sustainability reports and the enhanced coordination of information published individually by different departments.

Applying a two-step methodology to analyse the dynamic effect of contracting out in municipal solid waste efficiency

Diego Prior, *Universidad Autónoma de Barcelona*

This paper addresses an important gap in the literature of contracting out local public services that is the analysis of the cause-effect relationship between contracting out and cost efficiency in local government. For this purpose the study applies a two-step methodology which combines for the first time the application of the intertemporal frontier and the propensity score matching. The intertemporal frontier allows us to obtain a measure of the cost efficiency growth of the municipal solid waste (MSW) for a sample of XXX Spanish municipalities for the period 2002-2010, which is used in the second stage to determine whether the implementation of contracting out results in better levels of cost efficiency. In this sense, the propensity score matching technique compares the cost efficiency growth of municipalities that contract out the MSW with the cost efficiency growth of the non-contracting municipalities. The results suggest that the consequences of contracting out over cost efficiency are different depending on the implementation stage of this delivery form. In this respect, after a decrease of 1.4% in the initial phase of contracting out, this delivery forms favors the consecution of an increase of 3.2% in cost efficiency after three years, resulting in a cumulative efficiency growth of 1.8% for new contractors. In this sense, our results show up a learning effect over time in contracting out that may derive in a better contract management, overcoming the initial cost efficiency decrease of contracting out.

FRIDAY
OCTOBER 10

10.00 Invited Talk: Advances in Multimode-Clustering, Clustering and Dimensionality Reduction

Maurizio Vichi, *Sapienza University of Rome*

Big Data represented by data matrices X with a huge number of rows (objects, statistical units) and columns (variables) are generally analysed to synthesize the relevant information and to obtain a reduced data structure formed by prototype (mean profiles) objects and latent variables. This is achieved by the simultaneous grouping of rows and columns of X so that the results are informative and easy to interpret, because denote a compressed, but relevant representation of the big data, while trying to preserve most of the original information. The reduction could represent a strong synthesis to be directly interpreted in order to identify the most important characteristics in term of ideal objects and ideal variables. Alternatively, the reduction could be soft to obtain a light compression of the multivariate data in order to allow the successive application of other multivariate computationally complex methods on the compressed data matrix. In this presentation starting from an extension of standard K-means to simultaneously clustering of observations and features, namely Double K-Means (DKM), the model is developed in a probabilistic framework, an efficient coordinate ascent algorithm is proposed and the advantages of using this approach are discussed. DKM treats symmetrically the two modes of the data matrix (rows and columns) by producing a compressed set of prototype objects and prototype variables. The model is generalized to include an asymmetric treatment of the two modes so as to comprise both clustering and disjoint principal component analysis and its extension based on a probabilistic principal component analysis. A new coordinate ascent algorithm is developed and its performance is tested via simulation studies and real data sets. Finally, the results obtained on the real data are validated by building resampling confidence intervals for block centroids.

11.00 Coffee Break

11.30 Invited Talk: Robust Statistical Inference based on the Density Power Divergence Approach

Leandro Pardo, *Universidad Complutense de Madrid*

In any parametric inference problem, the robustness of the procedure is a real concern. A procedure which retains a high degree of efficiency under the model and simultaneously provides stable inference under data contamination is preferable in any practical situation over another procedure which achieves its efficiency at the cost of robustness or vice versa. The density power divergence family provides a flexible class of divergences where the adjustment between efficiency and robustness is controlled by a single parameter. Robust estimation based on the minimization of density power divergences has proved to be a useful alternative to the classical maximum likelihood based technique. The most popular hypothesis testing procedure, the likelihood ratio test, is known to be highly non-robust in many real situations. An alternative robust procedure of hypothesis testing based on the density power divergence is presented.

12.30 Session 3

Chair: Eva Boj, Universitat de Barcelona

Merging classes

Josep A. Martín-Fernández, *Universitat de Girona*

In model based clustering, any element from a given sample is assigned to a particular class according to its posterior probability to belong to that class. Similarly, in fuzzy clustering such posterior probability is substituted by the weight of belonging to that class. In this presentation, we are going to introduce a general method to explore how these probabilities or weights of belonging may allow us to combine classes and build a hierarchy from a set of classes. Our approach is based on the log-ratio methodology for compositional data, as the posteriors probabilities or weights vectors can be viewed as compositions. Previous known methods to build hierarchies over classes will be discussed as special cases of our approach, and improved by incorporating new strategies.

Logistic Biplots for Nominal and Ordinal Data

José Luis Vicente, *Universidad de Salamanca*

The Biplot method is a popular technique for analysing multivariate data. Recently, Logistic Biplots for binary data have been developed. In this paper we extend the Logistic Biplot to nominal and ordinal data. For nominal data the variables are represented as convex prediction regions rather than vectors while for ordinal data as directions divided into prediction segments. We study the geometry of such representations and construct computational algorithms for estimation of the parameters and representation of the prediction regions or directions. Two R packages developed for the new methods and an application to a survey on job satisfaction of doctorate holders in Spain are also presented.

Claim reserving with DB-GLM: extending the Chain-Ladder method

Eva Boj, *Universitat de Barcelona*

As is demonstrated in the bibliography, generalized linear models (GLM) can be considered as a stochastic version of the classical Chain-Ladder method of claim reserving in non-life insurance. We refer, e.g., to England (1999) and England and Verrall (2002) for a detailed description. In particular, the deterministic Chain-Ladder model is reproduced when a GLM is fitted to a run-off-triangle by assuming overdispersed Poisson error distribution and logarithmic link. In this presentation, we propose the use of distance-based generalized linear models (DB-GLM) in the claim reserving problem. We refer to Boj et al. (2012) where the main characteristics of the DB-GLM are studied. DB-GLM can be considered a generalization of the classical GLM to the distance-based analysis. The only information required to fit these models is a predictor distance matrix. DB-GLM can be fitted using the dbstats package for R (Boj et al., 2013). It is important to point out that DB-GLM contains as a particular instance ordinary GLM. Then it can be considered too as a stochastic Chain-Ladder claim reserving method. To complement the methodology and estimate reserve distributions and standard errors we develop a bootstrap technique adequate to the DB-GLM. We make an application with the well known run-of-triangle of Taylor and Ashe (1983). This research is part of the project: Semiparametric and distance-based methodologies with applications in bioinformatics, finance and risk management (grant MTM2010-17323).

14.00 Lunch

17.00 Session 4

Chair: José Miguel Angulo, Universidad de Granada

Rank based Solutions and monotone regression in non-metric MDS

Roxana Alemán. *Universidad Juan Misael Saracho, Bolivia*

When pairwise dissimilarities only satisfy the ordinal scale property in Multidimensional Scaling, the non-metric procedures comprise the appropriate choice. Usually in a least squares framework, non-metric MDS is based on the parameter estimation in an alternating algorithm that first look for a configuration and then find a monotone transformation to estimate the disparities using isotonic regression, to minimize the STRESS. Recently, a new procedure without resorting to use disparities is proposed such that the spatial configuration of the objects is updated by minimizing the sum of discrepancies between the rank orders of the dissimilarities and the estimated distancies. In this work a comparison of the performance of both procedures is showed both in terms of goodness of fit and of degenerate solutions.

Non-extensive approach to multifractal dependence assessment. Application to a real seismic series

Francisco Javier Esquivel, *Universidad de Granada*

This talk is focused on the assessment of dimensional dependence in terms of entropy measures in the multifractal domain. Limiting extensions of normalized dependence coefficients (see Furuichi 2006) are proposed based on a Tsallis-entropy formulation of generalized dimensions (Angulo and Esquivel 2014). The seismic activity associated with the volcanic eruption occurred during 2011 in El Hierro (Canary Islands, Spain) is analyzed under this approach and the non-extensive frequency-magnitude distribution.

Asymptotic extremal behaviour of spatially deformed stationary random fields

José Luis Romero Béjar, *Universidad de Granada*

In this contribution the complexity of assessment on extremal behaviour in real phenomena modeled by neither non-Gaussian nor non-stationary random fields is highlighted. A short review of the state of the art in areas related to this problem is performed. In this context, asymptotic extremal behaviour of spatially deformed stationary random fields is addressed. Under appropriate conditions for deformation these are harmonizable random fields. Some extensions and related results concerning the order of approximation are investigated for this class. Continuing research in this context is discussed

18.15 Coffe Break

18.45 Clossing Session

Chair: José Fernando Vera, Universidad de Granada

Clustering Unfolding models

José Fernando Vera, *Universidad de Granada*

In unfolding for two-way two-mode preference ratings data, the categorization of the set of individuals while the categories are represented in a low dimensional space may be an advisable procedure to facilitate their understanding, specifically for Big Data. In addition to considering groups of individuals of a similar preference pattern, homogeneous groups of objects are also considered, such that within each group there are clustered objects perceived to have similar attributes. In the context of latent class models and using Simulated Annealing, the cluster-unfolding model for two-way two-mode preference rating data has been shown to be superior to a two-step approach of first deriving the clusters and then unfolding the classes. However, the high computational cost makes the procedure only suitable for small or medium-sized data sets, and the hypothesis of independent and normally distributed preference data may also be too restrictive in many practical situations. Thus, an alternating least squares procedure may be also proposed, in which the individuals and the objects are partitioned into clusters, while at the same time the cluster centers are represented by unfolding. Real and artificial data sets are analyzed to illustrate the performance of the models.

Stability Analysis for the Nonstationary Spatial Covariance Structure Estimation using Multidimensional Scaling

Juan Antonio Roldán,” *Axesor, conocer para decidir” Consulting*

The well-known issue of non-stationary spatial covariance structure for problems of spatial interpolation has been widely treated in the literature, and several procedures have been proposed for the nonparametric approach to global estimation of the spatial covariance structure of a space-time random function. Among this, space deformation method based on Multidimensional Scaling have played an important role as first was showed by Sampson and Guttorp (1992), and several procedures have been proposed transforming the geographical space so that it exhibits isotropic spatial covariance structure. In this presentation, the study of the overall stability of the estimated spatial covariance by means of MDS is proposed to analyze the suitable choice of the MDS procedure as well as the dimensionality of the given solution for the estimation of the spatial covariance. Also, for the selected configuration, the leave-one-out cross validation procedure let us to study the stability of the statistical location of each station, and the influence of each station in the final configuration and therefore in the values of the derived spatial dispersion

Bootstrap confidence intervals and the Wald test

Josep Fortiana, *Universitat de Barcelona*

The Distance-Based Generalized Linear Model (DB-GLM) is a generalization of the classical GLM to the DB framework (Boj et al. 2012). The DB-GLM is non-linear on original predictors because its information is entered in the model by means of a squared distances matrix. In Boj et al. (2014) we proposed a definition of local influence coefficients for the DB-GLM depending on the nature of risk factors (numerical or categorical/binary). These coefficients measure the relative importance of each observed variable. In this presentation, we study how to adapt the Wald test of predictor significance to the DB-GLM environment. Firstly we apply the definition of influence coefficients and the bootstrap by pairs methodology to estimate the distribution of coefficients, as is given in Boj et al. (2014). In this way we are able to estimate the coefficients of the DB-GLM and its associated standard errors. Then, we propose a procedure to adapt the Wald statistic to the DB-GLM. We construct simple confidence intervals by using a standard normal distribution, and percentile t confidence intervals by using a bootstrap t* distribution in the sense of the reference MacKinnon (2006). The t* distribution of the percentile intervals follows the null hypothesis of the test in the bootstrap data generation process. In this way, the percentile t confidence intervals are useful to test the null hypothesis that a coefficient is equal to a fixed real value. We illustrate the calculation of percentile t* confidence intervals with the actuarial dataset of Hallin and Ingenbleek (1983). We estimate the related DB-GLM by using the dbglm function of the dbstats R package (Boj et al. 2013).

Functional analysis approach to canonical correlation

Carles M^{a} Cuadras, *Universitat de Barcelona*

We find some properties and eigen decompositions of two integral operators related to copulas. By using an inner product between two functions via an extension of the covariance, we study the countable set of eigen pairs, which is related to the set of canonical correlations and functions. Then a canonical analysis on the so-called Cuadras-Augé family of copulas is performed, showing the continuous dimensionality of this distribution. A diagonal expansion in terms of an integral is obtained. As a consequence, this continuous expansion allows us to generate a wide family of copulas.