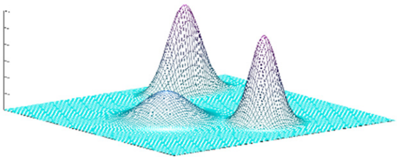
III International Workshop on Proximity Data,
Multivariate Analysis and Classification
S E
I O

Book of Abstracts

III International Workshop on Proximity Data, Multivariate Analysis and Classification

October 26-27, 2017

Valladolid (Spain)



Universidad de Valladolid

Editors:

José Fernando Vera, Universidad de Granada.

Eva Boj del Val, Universidad de Barcelona.

Luis Ángel García Escudero, Universidad de Valladolid.

Agustín Mayo Iscar, Universidad de Valladolid.

Edited by the Publications Service of the University of Barcelona.

ISBN: 978-84-697-7418-2

Organizing Committee

Eva Boj del Val, UB (evaboj@ub.edu)

José Fernando Vera, UGR (jfvera@ugr.es)

Luis Ángel García, UVA (lagarcia@eio.uva.es)

Agustín Mayo, UVA (agustinm@eio.uva.es)

Scientific Committee

Carles Mª Cuadras, UB (ccuadras@ub.edu)

José Fernando Vera, UGR (jfvera@ugr.es)

Eva Boj del Val, UB (evaboj@ub.edu)

Albert Satorra, UPF (albert.satorra@upf.edu)

José Luis Vicente, USAL (villardon@usal.es)

José A. Martín, UDG (josepantoni.martin@udg.edu)

Luis Ángel García, UVA (lagarcia@eio.uva.es)

Alfonso Gordaliza, UVA (alfonsog@eio.uva.es)

Eustasio del Barrio, UVA (tasio@eio.uva.es)

Miguel A. Fernández, UVA (miguelaf@eio.uva.es)

Important Dates

October, 16th: Deadline for abstract submission

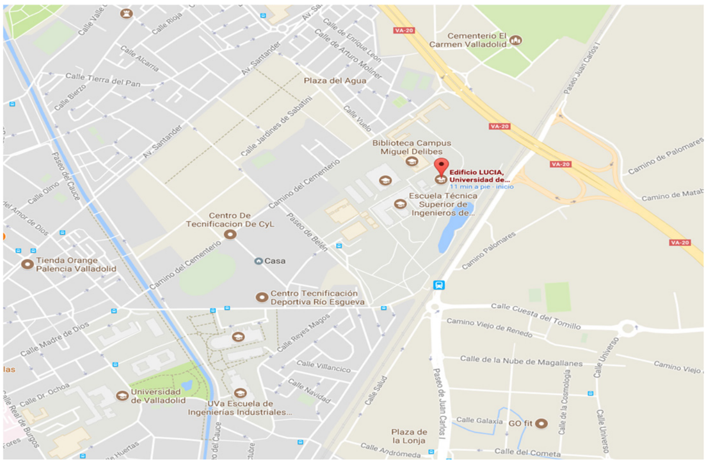
October, 19th: Notification of acceptance

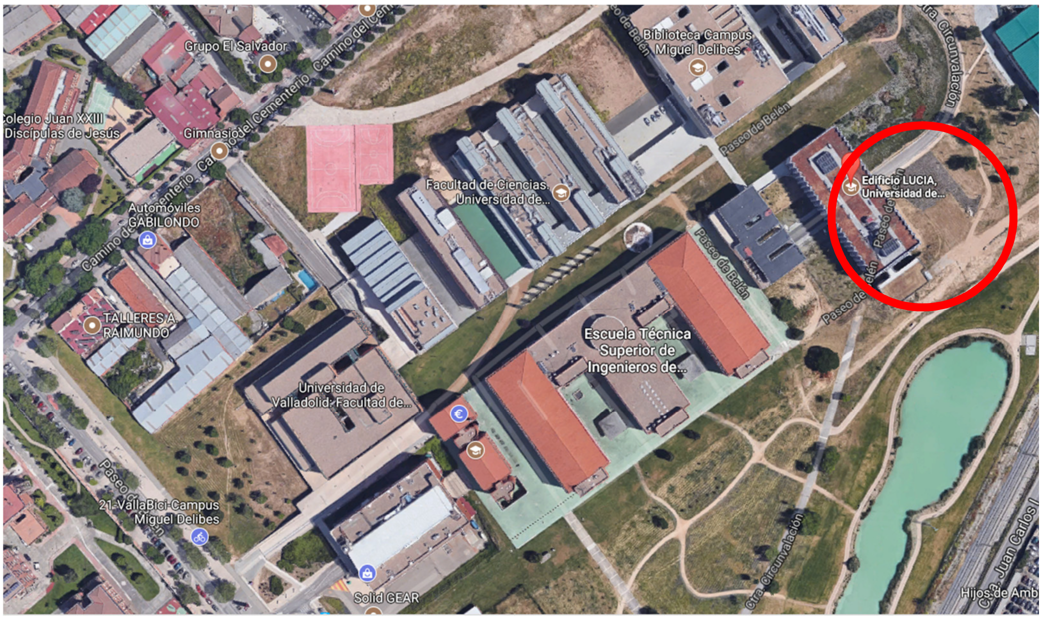
October, 20th: Deadline for registration

Location:

Edificio LUCIA (Dependencias del IMUVA), Universidad de Valladolid

Paseo de Belén, 19, 47011 Valladolid (Campus Miguel Delibes)


Cementerio El Carmen Valladolid
Plaza del Agua
Biblioteca Campus Miguel Delibes
Edificio LUCIA, Universidad de...
Escuela Técnica Superior de Ingenieros de...
Centro De Tecnificacion De CyL
Tienda Orange Palencia Valladolid
Casa
Centro Tecnificación Deportiva Río Esgueva
Universidad de Valladolid
UVa Escuela de Ingenierías Industriales...
Plaza de la Lonja
GO fit
Calle Cuesta del Tornillo
Camino Viejo de Renedo
Calle de la Nube de Magallanes
Camino del Cementerio
Paseo de Belén
Paseo Juan Carlos I
Camino de Palomares
VA-20


Grupo El Salvador
Camino del Cementerio
Biblioteca Campus Miguel Delibes
Colegio Juan XXIII Discípulas de Jesús
Gimnasio
Paseo de Belén
Edificio LUCIA, Universidad de...
Automóviles GABILONDO
Facultad de Ciencias Universidad de...
Ronda Circunvalación
TALLERES A RAIMUNDO
Escuela Técnica Superior de Ingenieros de...
Universidad de Valladolid Facultad de...
21 VallaBici Campus Miguel Delibes
Solid GEAR
Juan Carlos I
Hijos de Amb

Sponsors:

S E I O Sociedad de Estadística e Investigación Operativa

http://www.seio.es/

Universidad de Valladolid

http:/www.uva.es

dE IO Dpto. de Estadística e I.O.
Universidad de Valladolid

http://www.eio.uva.es/

imUVa

http://www.imuva.uva.es/

esdUVa
Escuela de Doctorado **Universidad** de **Valladolid**

http://escueladoctorado.uva.es/

Annual meeting of the SEIO Working Group on Multivariate Analysis and Classification (AMyC), Valladolid, October 26 - 27, 2017:

The III International Workshop on Proximity Data, Multivariate Analysis and Classification has taken place during October, 26-27, 2017 in Valladolid (Spain):

http://www.eio.uva.es/wamyc/

It has been organized by the Multivariate Analysis and Classification Spanish SEIO Group AMyC:

http://amyc.seio.es/

The Spanish Group of Multivariate Analysis and Classification is a Working Group of more than 50 researchers from all the Spanish universities. Every year, the Working Group organizes a meeting to promote the communication between its members and between them and other researchers, and to contribute to the development of the Multivariate Analysis and Classification field and related problems and applications. The first International Workshop on Proximity Data, Multivariate Analysis and Classification took place in Granada, October 2014 (http://www.ugr.es/~amyc/EVENTS/WAMYC/) and the second in Barcelona, October 2016 (http://www.ub.edu/wamyc/).

The topics of interest comprise any related problem to Multivariate Analysis and Classification both from a theoretical or a computational point of view, and their applications. It also includes problems related to unsupervised or supervised statistical learning related to big data analysis.

Reduced Program:

Date:	Thursday 26th	Friday 27th
08:30-09:00	**Registration**	
09:00-10:00	**Plenary Invited Talk:** *Chair: C. Matrán* Exploration of the variability of variable selection based on distances between bootstrap sample results Hennig, Christian	**Plenary Invited Talk:** *Chair: A. Gordaliza* Big Data world: wide consensus in estimation using parallelized inference Cuesta-Albertos, Juan Antonio
10.00-11:00	**Session 1:** *Chair: E. Boj del Val* Distance-based logistic classification and prediction error Boj del Val, Eva Minimum density power divergence estimators for polytomous logistic regression models Castilla González, Elena M. Visualizing dynamic proximities and frequencies by means of mathematical optimization Guerrero, Vanesa	**Invited Talks (and proposal for the next workshop):** *Chair: J.L. Vicente Villardón* A proposal of cluster distance-based regression procedure Boj del Val, Eva Two-way mixture distance-association model Vera, José Fernando Logistic biplots for binary data revisited Vicente Villardón, José Luis
11:00-11:30	**Coffee break**	**Coffee break**
11:30-12:30	**Session 2:** *Chair: M.A. Fernández Temprano* Classification via sparse cost-sensitive Support Vector Machines Benítez Peña, Sandra Enhancing the Naïve Bayes classifier by adding sparsity and performance constraints Sillero Denamiel, M. de los Remedios Novel order restricted boosting and backfitting classification procedures with an application to diagnosis of industrial motors Conde del Rio, David	**Session 4:** *Chair: E. del Barrio Tellado* Meta-analysis of clustering procedures based on k-barycenters in the Wasserstein space Inouzhe Valdés, Hristo Unsupervised classification of functional data: an application to classification with air navigation data Gordaliza Pastor, Paula Robust functional clustering via trimming and constraints García Escudero, Luis Ángel
12:30-13:30	**Plenary Invited Talk:** *Chair: A. Mayo-Iscar* To get the best, tame the beast: constrained ML estimation for mixture models	**AMyC Group Meeting**

	Greselin, Francesca	

Date:	Thursday 26th	Friday 27th
13:30-16:00	**Lunch**	**AMyC Group Meeting/Closing/Lunch**
16:00-17:00	**Session 3:** *Chair: M. Comas Cufí* A new approach for clustering in the context of qualitatives scales González del Pozo, Raquel Estimating the parameters of a logistic-normal-multinomial distribution Comas-Cufí, Marc	
17:00-18:00	**Ateneo IMUVA:** *(Sala de Grados I, Facultad de Ciencias)* Location theory and some basic physical and social principles in a nutshell Puerto Albandoz, Justo	

Program:

Thursday 26th

08:30-09:00 Registration

09:00-10:00 Plenary Invited Talk (Chair: C. Matrán)

Christian Hennig. University College of London (UCL)

http://www.homepages.ucl.ac.uk/~ucakche/

Exploration of the variability of variable selection based on distances between bootstrap sample results

Christian Hennig and Willi Sauerbrei

Abstract: It is well known that variable selection in multiple (linear, generalised linear or nonlinear) regression can be unstable and that the model uncertainty can be considerable. The model uncertainty can be quantified and explored by bootstrap resampling, see Sauerbrei et al. (2015). In this presentation we will present some approaches that use the results of bootstrap replications of the variable selection process to obtain more detailed information about the data.

Analyses will be based on distances between the results of the analyses of different bootstrap samples. Distances could be computed between the vector of predictions for all observations from the different analyses, or between the lists of selected variables. The distances can be used to map the bootstrap results by mutidimensional scaling and to cluster them. Clusters are of interest because they could point to substantially different interpretations of the data that could arise from different selections of variables supported by different bootstrap samples. The distances also allow to pinpoint influential and atypical observations by quantifying the atypicality of a bootstrap result and evaluating for each observation how atypical the bootstrap samples are to which it contributes.

These and further issues will be illustrated by some data examples including the study on ozone effects in children analysed in Sauerbrei et al. (2015).

References:

Sauerbrei W., Buchholz A., Boulesteix A.-L., Binder H. (2015) On stability issues in deriving multivariable regression models. Biometrical Journal 57, 531-555.

10.00-11:00 Session 1 (Chair: E. Boj del Val)

Distance-based logistic classification and prediction error

Eva Boj del Val and Mª Teresa Costa Cor

Abstract: In Boj *et al*. (2017) it is showed a procedure for the estimation of prediction error, PE, the square root of mean squared error, MSE, for distance-based generalized linear models (Boj *et al*., 2016). Expressions are developed when the general cases of power families of error distributions and of links are used. As a first step, MSE is approximated with the sum of the process variance and of the estimation variance. The estimation variance can be estimated by applying the delta method and/or by using bootstrap. When using bootstrap one is able to obtain an estimation of the distribution of each predicted value. To help us in the knowledge of the randomness of the new predicted values, confidence intervals can be calculated by taking into account the bootstrapped distributions. Now, it is showed the expression of PE for the generalized linear model with Binomial error distribution and logit link function, the logistic regression. Its calculus and their usefulness are illustrated to solve the problem of Credit Scoring, where policyholders are classified into defaulters and non-defaulters. Two sets of real credit risk data are analyzed and probabilities of default are estimated. Distance-based logistic regression models are fitted using the *dbglm* function of the *dbstats* package for R (Boj et al., 2014).

References:

Boj, E., Caballé, A., Delicado, P., Fortiana, J. (2014). dbstats: distance-based statistics (dbstats). R package version 1.0.4.

Boj, E., Delicado, P., Fortiana, J., Esteve A., Caballé, A. (2016). Global and local distance-based generalized linear models. TEST 25, 170–195.

Boj, E., Costa, T., Fortiana, J. (2017). Prediction error in distance-based generalized linear models. In: Palumbo, F., Montanari, A. and M. Vichi (eds.). Data science. Innovative developments in data analysis and clustering. Series: Studies in classification, data analysis, and knowledge organization. Volume 1, pp. 191–204. Springer International Publishing.

Acknowledgements: Multivariate Statistics and Classification Group of the Spanish Society of Statistics and Operations Research; Spanish Ministerio de Educación y Ciencia (grant MTM2014-56535-R) and Generalitat de Catalunya, AGAUR (2017SGR152).

Minimum Density Power Divergence Estimators for Polytomous Logistic Regression Models

Elena Castilla, Abhik Ghosh, Nirian Martin and Leandro Pardo

Abstract: The polytomous logistic regression model (PLRM) is widely used in health and life sciences as well as in other different areas where we need to analyze a nominal qualitative response taking values in a set of unordered categories. Examples of such response variables frequently occurring in medical and other applied sciences include disease symptoms that have been classified by subjects being absent, mild, moderate, or severe; or the invasiveness of a tumor classified as in situ, locally invasive, ormetastatic, etc.

The most common existing way to estimate the parameters in this PLRM is the maximum likelihood estimator (MLE), which is the main base of most of the existing literature on logistic models. However, this estimator is clearly known to be non-robust with respect to the possible outliers in data. Although there also exist some alternative estimation procedures for the PLRM other than the MLE, the important issue of robustness against outliers in data was ignored in all these methods.

In this talk we will present robust estimators under the PLRM based on the minimum divergence approach with the density power divergences. We will study their asymptotic distribution and robustness properties, as well as define Wald-type tests statistics for linear hypotheses. Simulation studies provide further confirmation of the validity of the theoretical results established before. An approach for the data-driven selection of the tuning parameter is also proposed with empirical justifications.

Visualizing dynamic proximities and frequencies by means of Mathematical Optimization.

Emilio Carrizosa, **Vanesa Guerrero** and Dolores Romero Morales

Abstract: In order to visualize dynamic proximities and frequencies, we develop a visualization framework which extends the standard Multidimensional Scaling and has a global optimization model at its heart. Difference of Convex functions and Nonconvex Quadratic Binary Optimization techniques are combined as a solution approach. Our methodology is illustrated using a dynamic linguistic real-world dataset.

11:00-11:30 Coffee break

11:30-12:30 Session 2 (Chair: M.A. Fernández Temprano)

Classification via sparse cost-sensitive Support Vector Machines

Sandra Benítez-Peña, Rafael Blanquero, Emilio Carrizosa and Pepa Ramírez-Cobo

Abstract: Support Vector Machine (SVM) is a powerful tool to solve binary classification problems. Many real world classification problems, such as those found in credit-scoring or fraud prediction, involve misclassification costs which may be different in the different classes. Providing precise values for such misclassification costs may be hard for the user, whereas it may be much easier to identify acceptable misclassification rates values. Hence, we propose here a novel SVM model in which misclassification costs are considered by incorporating performance constraints in the problem formulation. In particular, our target is to seek the hyperplane with maximal margin yielding misclassification rates below given threshold values.

This novel model is extended by performing Feature Selection (FS), which is a crucial task in Data Science, making thus the classification procedures more interpretable and more effective.

The reported numerical experience demonstrates that our model gives the user control on the misclassification rates in addition to the usefulness of the proposed FS procedure. Indeed, our results on benchmark data sets show that a substantial decrease of the number of features is obtained, whilst the desired trade-off between false positive and false negative rates is achieved.

Enhancing the Naïve Bayes classifier by adding sparsity and performance constraints

Rafael Blanquero, Emilio Carrizosa, Pepa Ramírez-Cobo and **M. Remedios Sillero-Denamiel**

Abstract: Naïve Bayes is a tractable and efficient approach to classification learning. However, as it is common in real classification contexts, datasets are often characterized by a large number of features which may complicate the interpretation of the results as well as slow down the method's execution. In addition, the consequences of misclassifications may be rather different for different classes. Hence, it is crucial to control misclassification rates in the most critical cases, possibly at the expense of higher misclassification rates in less problematic classes. In this work we propose a sparse version of the Naïve Bayes in which a variable reduction approach, that takes into account the dependencies among features, is embedded into the classification algorithm. Moreover, a number of constraints over the performance

measures of interest are embedded into the optimization problem. Unlike typical approaches in the literature modifying standard classification methods, the achievement in the different individual performance measures under consideration is controlled. Our findings show that, under a reasonable computational cost, the number of variables is significantly reduced obtaining competitive estimates of the performance measures.

Novel order restricted boosting and backfitting classification procedures with an application to diagnosis of industrial motors

David Conde, Miguel A. Fernandez, Cristina Rueda and Bonifacio Salvador

Abstract: The classical problem of classifying observations in one of *k* classes has received a lot of attention in the last decades due to its applicability in a very wide range of problems in different scientific fields. As a consequence, many methods and techniques to build classification rules have been developed, from the classic linear or quadratic rules to the more recent k-nearest neighbors, support vector machines or decision trees. Even more recently, methods based on so-called weak classifiers have been developed and a family of procedures called boosting procedures has been defined.

However, none of these procedures is able to take advantage of an issue that frequently appears in applications. This issue is the existence of additional information on the problem that can be expressed as order restrictions among the classes of interest. In a series of papers, we have developed restricted classification rules that incorporate this additional information to linear and quadratic rules and studied their good behavior. In this work, we define new restricted classification procedures that allow us to incorporate these order restrictions to boosting and backfitting procedures that fit the additive logistic model, in order to increase the accuracy of classification problems with two or more classes.

When the number of classes *k* is greater than two, the one-against-all strategy allows to reduce the problem to *k* binary problems but this methodology does not allow to incorporate the additional information. In this work, we propose a procedure that allows to incorporate the additional information reducing the problem to *k – 1* binary problems.

We show the good performance of the new procedures in a simulation study, comparing their performance with that of several well-known classification rules. Finally, we apply these procedures to a problem appearing in industrial practice, the diagnostic of an industrial electrical motor, also showing the good behavior of our procedures. This is a problem that is being extensively considered in engineering literature as it is very important from the organizational and economical point of view. A motor is monitored and its signals are processed. From these signals the motor is diagnosed so that if it is decided that it is in a state close to suffering a break, the motor is stopped and repaired before the break occurs. The consequences of misclassification are clear as stopping a motor when it is not needed can be very expensive and not stopping it and suffering a break can be even worse.

12:30-13:30 Plenary Invited Talk (Chair: A. Mayo-Iscar)

Francesca Greselin. Università degli Studi di Milano-Bicocca

www.economia.unimib.it/GRESELIN

To get the best, tame the beast: constrained ML estimation for mixture models

Francesca Greselin, Luis-Angel García-Escudero, Alfonso Gordaliza, Salvatore Ingrassia and Agustín Mayo-Iscar

Abstract: This paper presents a review about the usage of eigenvalues restrictions for constrained parameter estimation in mixtures of elliptical distributions according to the likelihood approach. The restrictions serve a twofold purpose: to avoid convergence to degenerate solutions and to reduce the onset of non- interesting (spurious) local maximizers, related to complex likelihood surfaces. The likelihood function may present local spurious maxima when a fitted component has a very small variance or generalized variance (i.e., the determinant of the covariance matrix), compared to the others (Day, 1969). Such a component usually corresponds to a cluster containing few data points either relatively close together or almost lying in a lower-dimensional subspace, in the case of multivariate data. These solutions are of little practical use or real world interpretation (McLachlan and Peel, 2000). To get the best from the EM algorithm widely used for model estimation, a constrained estimation of the covariance matrices can help in driving the maximum likelihood approach toward sensible solutions and far from the ”wild”, non-useful ones.

We begin presenting the strongest constraints, i.e. equality among spherical covariance matrices, and the well-known homoscedastic assumption. Afterwards, some authors reinterpreted the safe, above mentioned methods, looking for milder conditions, like the lightest assumption of equality of covariance determinants (McLachlan and Peel, 2000, Section 3.9.1). Gaussian parsimonious mixture models have been proposed by Banfield and Raftery (1993), to get intermediate component covariance matrices lying between homoscedasticity and heteroscedasticity. A broad family of contributions in the literature arise from Hathaway’s seminal paper and deals with setting constraints on the ratio between the maximum and the minimum eigenvalue (among many others, Ingrassia, 2004 and García-Escudero et al. 2008). The same approach, simultaneously with impartial trimming, can be adopted in the context of robust statistical methods. Unfortunately these proposals are not equivariant, and this remark motivated Gallegos and Ritter (2009) and Rocci et al. (2017) to introduce affine equivariant constraints. The methods herein described have been extended to other mixtures of elliptical models.

In our survey, for each position recalled so far, we discuss the algorithms needed for their exact or approximate implementation through the EM and recall theoretical results on the obtained estimator (whenever available). When “taming the beast”, the constrained maximization of the likelihood provides stability to the obtained solutions.

References:

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803–821.

Day, N. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56(3), 463–474.

McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. John Wiley & Sons, New York. Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. Statistical Methods & Applications, 13(4), 151–166.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. The Annals of Statistics, 36(3), 1324–1345.

Gallegos, M. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixture. Sankhya (Ser. A), 71, 164–220.

Rocci, R., Gattone, S. A., and Di Mari, R. (2017). A data driven equivariant approach to constrained Gaussian mixture modeling. Advances in Data Analysis and Classification, 1-26.

13:30-16:00 Lunch

16:00-17:00 Session 3 (Chair: M. Comas Cufí)

A new approach for clustering in the context of qualitative scales

Raquel González del Pozo, José Luis García-Lapresta and David Pérez-Román

Abstract: Linguistic appraisals of presidential candidates given by voters play an essential role during electorate campaigns. In recent years, some voting systems based on linguistic appraisals have arisen. One example is the Majority Judgment voting system introduced by Balinski and Laraki (2007, 2011), where voters assign a linguistic term to each candidate. In the political context, cluster analysis has hardly considered the appraisals given by voters, being mainly applied to identify demographically relationships among voters, voting tendencies or for classifying Republican and Democratic candidates in United States (U.S.)

taking into account the multiple donor networks (see Seabrook 2009, Pearson and Cooper 2012 and Dowdle et al. 2013, among others).

The main purpose of this contribution is to illustrate an agglomerative hierarchical clustering procedure which takes into account linguistic appraisals (see González del Pozo et al. 2017). To do this, we considered the opinions given by a random representative sample of adults living in U.S. from the January 2016 Political Survey conducted by the Pew Research Center (2017). In that survey, voters used the following five terms qualitative scale: 'terrible president', 'poor president', 'average president', 'good president', 'great president' for evaluating the presidential candidates.

Usually, qualitative scales used in surveys are based on the implicit assumption that they are uniform. However, sometimes individuals can perceive that the psychological proximity between linguistic terms is not always the same. In the setting of non-uniform scales, we can apply the concept of ordinal proximity measure (see García-Lapresta and Pérez-Román 2015) which allows us to measure the consensus in a group of agents when a set of alternatives is evaluated through non-necessarily uniform qualitative scales.

The proposed clustering procedure uses similarity functions based on consensus measures to merge clusters, in such a way that presidential candidates are classified into different clusters when the degree of consensus among voters is maximized.

References:

Balinski, M. and Laraki, R.: A theory of measuring, electing and ranking. Proceedings of the National Academy of Sciences of the United States of America 104, pp. 8720-8725, 2007.

Balinski, M. and Laraki, R.: Majority Judgment. Measuring, Ranking, and Electing. The MIT Press, Cambridge MA, 2011.

Dowdle, A., Limbocker, S., Yang, S., Sebold, K. and Stewart, P.: The Invisible Hands of Political Parties in Presidential Elections: Party Activists and Political Aggregation from 2004 to 2012, Palgrave Macmillan US, 2013.

García-Lapresta, J.L. and Pérez-Román, D.: Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering. Applied Soft Computing 35, pp. 864-872, 2015.

González del Pozo, R., García-Lapresta, J.L. and Pérez-Román, D.: Clustering U.S. 2016 presidential candidates through linguistic appraisals. Advances in Fuzzy Logic and Technology 2017. Proceedings of EUSFLAT 2017. The 10th Conference of the European Society for Fuzzy Logic and Technology, pp .143-153, 2017.

Pearson, P.T. and Cooper, C.I.: Using Self Organizing Maps to Analyze Demographics and Swing State Voting in the 2008 U.S. Presidential Election. Lecture Notes in Artificial Intelligence, Vol. 7477, Springer, 2012.

Pew Research Center. http://www.pewresearch.org, 2017.

Seabrook, N.R.: The Obama effect: Patterns of geographic clustering in the 2004 and 2008 presidential elections. The Forum 7, 2009. doi:10.2202/1540-8884.1308.

Estimating the parameters of a logistic-normal-multinomial distribution

Marc Comas-Cufí, J.A. Martín-Fernández, G. Mateu-Figueras and J. Palarea-Albaladejo

Abstract: The logistic-normal-multinomial distribution is the compounding probability distribution resulting from considering the multivariate logistic-normal as the distribution for the probability parameter vector of the multinomial distribution. This distribution can be used to model multivariate count data when only the relative relations between parts are of interest. It may be considered as an alternative to the Dirichlet-multinomial distribution and it can be used to deal with counting zeros.

Billheimer and others (2001) proposed a Bayesian approach to estimate its parameters by means of a Markov chain Monte Carlo simulation. In contrast, Xia and others (2013) proposed a Monte Carlo EM algorithm to fit the parameters where the expected values are calculated using a Metropolis-Hasting algorithm. In the univariate case, Hughes and other (1998), considered to solve the problem using numerical integration. In all these works, the proposed approach gives satisfactory results when it is applied to a particular case. However, when it is applied in more general settings, problems may appear. In consequence, for general scenarios, a more comprehensive proposal is required.

In this work we consider different alternatives to calculate the expected value in the Monte Carlo EM algorithm and compare their performance. In particular, for the method proposed by Xia and others (2013), we substitute the Metropolis-Hasting algorithm by a standard Monte Carlo algorithm using different variance reduction techniques.

References:

Billheimer, D., Guttorp, P. and Fagan, W.F. (2001). Statistical Interpretation of Species Composition. Journal of the American Statistical Association, 96(456), 1205--1214.

Hughes, G., Munkvold, G.P. and Samita, S. (1998). Application of the logistic-normal-binomial distribution to the analysis of Eutypa dieback disease incidence. International Journal of Pest Management, 44(1), 35--42.

Xia, M., Chen, J., Fung K.F. and Li H. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. Biometrics, 69, 1053--1063.

17:00-18:00 Ateneo IMUVA (Sala de Grados I, Facultad de Ciencias)

Location theory and some basic physical and social principles in a nutshell

Justo Puerto Albandoz

Friday 27th

09:00-10:00 Plenary Invited Talk (A. Gordaliza)

Juan Antonio Cuesta Albertos. Universidad de Cantabria

http://personales.unican.es/cuestaj/

Big Data world: wide consensus in estimation using parallelized inference

Juan A. Cuesta-Albertos

Abstract: Let us assume that we want to estimate the k components of a mixture on ^{d}, ,…, , and that we have *m* units (hospitals, computation or research centers,…) involved in the estimation. Assume that each unit U_j, j = 1,… , m, processes a separate sample, and that as a result it sends us an estimation of the k components: .

Our task is to find k probability distributions: to summarize those estimations. But, just in case there exist some very discrepant 's we should look for a kind of robust consensus between the estimations the units have sent us.

Our proposal is to consider the k x m distributions we have received and to apply the trimmed *k*-means ideas to the full set. Thus, we do not consider *k*-uples of probabilities but a set composed by *k* x *m* elements. Since our objects are probability distributions we need to handle a suitable metric between probabilities. Our selection has been the Wasserstein distance, with the associated barycenter as the natural candidate to replace the usual mean.

The talk will be (almost) self-contained and will include a numerical example which shows how, in some cases, it can be preferable to take several very small sub-samples from a large data set and apply the proposed procedure than to analyze the full sample at once. Some computational problems associated to the Wasserstein-barycenter will also be discussed. Finally, the method will be applied to a real data set.

Acknowledgements: The results in which the talk will be based have been obtained in collaboration with Profs. P. Alvarez-Esteban, E. del Barrio, C. Matrán and A. Mayo-Iscar, all of them at Universidad de Valladolid, Spain. The author has been partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant MTM2014-56235-C2-2-P.

10.00-11:00 Invited Talks and proposal for the next workshop (J.L. Vicente Villardón)

Two-way mixture distance-association model

José Fernando Vera and **Eva Boj**

Abstract: In distance-based regression analysis, the vector of continuous responses is projected in a Euclidean space given by multidimensional scaling (MDS). The MDS configuration is obtained by considering a dissimilarity matrix, typically of Euclidean distances, measured between the observed elements in the predictor space. One of the main problems in distance-based regression analysis for mixed variables is that of determining the number of dimensions or of latent predictor variables, in particular when a large dissimilarity data set arise to the observations in the predictor space. To have Euclidean distances, the observed proximities are translated by means of the well-known procedure of the additive constant. Nevertheless, this methodology usually increases the number of dimensions in MDS and therefore the number of latent predictor variables, in particular when the sampling size is larger with respect to the original number of predictor variables. To reduce the number of elements to be represented using dissimilarities, the use of cluster analysis in conjunction with MDS is an advisable procedure. The main aim in this methodology consists on the classification of the objects into clusters while simultaneously the cluster centres are represented in a low dimensional space. To determine a reduced number of latent predictor variables a cluster distance-based regression procedure is proposed. In this methodology, not the observed sample elements by itself but the coordinates of cluster centres are employed to determine a reduced space of latent predictor variables using a cluster-MDS procedure. Thus, given a dissimilarity matrix obtained from the original data set, a combination of a k-means procedure for dissimilarities and MDS is employed to determine a classification of the observed elements and to determine a reduced latent predictor space. The performance of the proposed procedure is illustrated with the analysis of real data sets.

Acknowledgements: Multivariate Statistics and Classification Group of the Spanish Society of Statistics and Operations Research; ECO2013-48413-R and MTM2014-56535-R of the Spanish Ministry of Economy and Competitiveness Spanish (co-financed by FEDER) and 2017SGR152 of the Generalitat de Catalunya, AGAUR.

Latent Block Distance-Association Model

José Fernando Vera

Abstract: Usually log-lineal models with a large number of parameter to be estimated results in tables with large number of cells. Several models with fewer parameters for the association, which also facilitate the interpretation, have been proposed for non-sparse data as the distance association (DA) model. For tables involving profiles, the DA model can be estimated but the given results may be difficult to interpret because the presence of a large amount of modalities and/or zeros. Collapsing rows is an advisable procedure for a profile by response table, when the response variable has a moderate number of modalities, and in particular for sparse data sets. Nevertheless as in the DA model, this procedure may still fail in the representation of associations for tables also having a large number of modalities in the response variable, and in particular for sparse tables as a profile by profile sparse contingency table. In this work, a latent block distance association model (LBDA) is formulated that aims the simultaneous partitioning of the rows and the columns of a contingency table, while the between blocks association is represented in a low dimensional space in terms of Euclidean distances. In the LBDA model, odds are defined in terms of the block-related main effects and of the distances, while odds ratio are defined only in terms of the squared distances.

Acknowledgements: This work has been partially supported by grants ECO2013-48413-R of the Spanish Ministry of Economy and Competitiveness (co-financed by FEDER).

Logistic biplots for binary data revisited

José Luis Vicente Villardón

Abstract: Classical Biplot methods allow for the simultaneous representation of individuals and continuous variables in a given data matrix. When variables are binary a classical linear biplot representation is not suitable. Some time ago we proposed a linear biplot representation based on logistic response models for binary data. The coordinates of individuals and variables are computed to have logistic responses along the biplot dimensions. The method is related to logistic regression in the same way that Classical Biplot Analysis (CBA) is related to linear regression. Thus we named the method “Logistic Biplot” (LB). In the same way as Linear Biplots are related to Principal Components Analysis, Logistic Biplots are related to Latent Trait Analysis or Item Response Theory. For the estimation of the parameters we have used Alternated Generalized Least Squares, Marginal Maximum Likelihood and Principal Coordinates Analysis followed by Logistic Regressions. The first two are more adequate when the number of individuals is higher than the number of variables and the last when the number of variables is high but, none of them works properly for big data matrices. In this work we revisit the geometry of those kinds of biplots and study new algorithms for the estimation of the biplot markers based on gradient descent and different optimization procedures to adapt the methods to bigger binary matrices. An R package is presented and some applications to diverse fields are revised.

Proposal for the next workshop:

AMyC2018

IV INTERNATIONAL WORKSHOP ON PROXIMITY DATA, MULTIVARIATE ANALYSIS AND CLASSIFICATION. UNIVERSITY OF SALAMANCA, SPAIN, OCTOBER 2018.

Organizing committee:

Eva Boj (UB), José Fernando Vera (UGR) and José Luis Vicente (USAL)

You are all invited to the IV Workshop AMyC in Salamanca to celebrate the 800 years of the University!

11:00-11:30 Coffee break

11:30-12:30 Session 4 (Chair: E. del Barrio Tellado)

Meta-analysis of clustering procedures based on *k*-barycenters in the Wasserstein space

Eustasio del Barrio, **Hristo Inouzhe** and Carlos Matrán.

Abstract: Cluster analysis addresses the detection of data grouping in data sets. Within this, too vague, description, model-based clustering aims to find particularly shaped groupings -clusters- according to specified distributions. In this setting, the clusters provided by the method are described by probability (often Gaussian) distributions that can be considered as elements of an abstract space. Particular interest has been deserved by the L2 Wasserstein distance, leading to a rich set-up for developing statistical concepts in a parallel way to those known on Euclidean spaces. This is the case of the k-barycenters, the abstract version of *k*-means, by large the widest used method in clustering problems, recently introduced in the Wasserstein space even in a robust version. We focus on the application of the (trimmed) Wasserstein *k*-barycenters to some of the fundamental problems present in cluster analysis. This includes parallelization or stabilization of procedures and even improvement of initial solutions for the algorithms involved in the methods, but we will also pay special attention to the meta-analysis tools arising from this robust aggregation procedure: Stability (or coherence) criteria, applied to the provided aggregation, will give descriptive signs on the number of clusters or on the adequacy of the clustering procedure. We present illustrative examples of the previously mentioned concepts.

Unsupervised classification of functional data: an application to classification with air navigation data

Paula Gordaliza Pastor and Pedro C. Álvarez-Esteban

Abstract: Technological advances in recent years have made it possible to incorporate measuring devices that are increasingly faster and more accurate in all areas of human activity and its surrounding environment. These devices continuously monitor a multitude of processes, generating large bases of functional data that are modelled as realizations of a stochastic process $X = \{X(t): t \in T\}$ taking values in a space of functions defined over a set T, usually a time interval. In particular, the trajectories described by moving objects have gained special interest in the last years, being one of the main objectives of their study grouping those with similar behavior to discover common patterns of movement in the objects that generate them and, in turn, distinguish others different or unwanted. In this project, we review the main methods of Functional Cluster Analysis existing today, to focus on those that best adapt to trajectory data. Finally, the classification of trajectories flown by different types of aircraft is presented as an application.

Robust functional clustering via trimming and constraints

Luis Angel García-Escudero, Diego Rivera-García, Joaquín Ortega and Agustín Mayo-Iscar

Abstract: Several approaches for functional clustering have been proposed. Unfortunately, they are not specifically aimed at dealing with the disturbing presence of outlying or anomalous curves. Therefore, a small fraction of anomalous curves may be extremely harmful when clustering the curves in our data set. For instance, main clusters can be artificially joined together and clusters made of few outlying curves can be detected too. Taking into account this problem, a robust model-based clustering methodology is proposed that relies on a "small-ball pseudo-density" for functional data (Delaigle and Hall, 2010) and the use of a data-driven (impartial) trimming approach. Moreover, constraints on the scatter parameters, within and between clusters, are also considered in order to avoid the detection of non-interesting spurious clusters. A computationally feasible algorithm is available for its implementation. The procedure is very flexible but it needs the choice of several tuning parameters that can be chosen throughout the careful monitoring of available graphical tools.

References:

Delaigle, A. and P. Hall (2010). Defining probability density for a distribution of random functions. Ann. Statist. 38 (2), 1171-1193.

12:30-13:30 AMyC Group Meeting

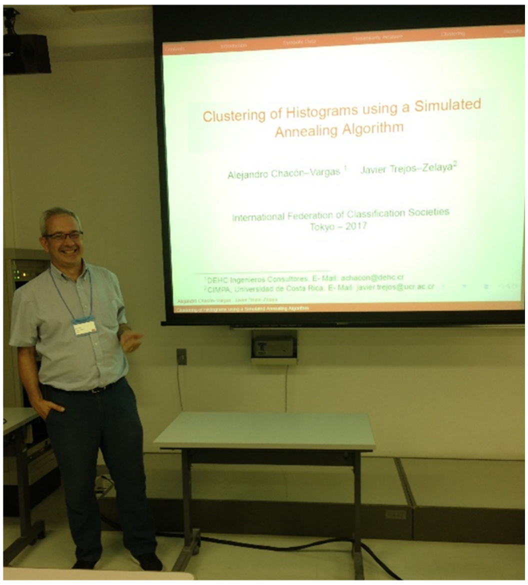
13:30-16:00 AMyC Group Meeting/Closing/Lunch

Desde la coordinación del Grupo AMyC queremos agradecer el esfuerzo realizado a todos los participantes del congreso de la International Federation of Classification Societies, IFCS2017, celebrado en Tokai University en Tokyo (Japón) en agostos de 2017, por hacer posible dos sesiones invitadas de nivel en esta edición de 2017 (http://ifcs.boku.ac.at/_conference/index.php/ifcs2017).

Algunas fotos:

Robust clustering for functional data based on trimming and constraints
Luis Angel García-Escudero[1]
joint work with
D. Rivera García[2], A. Mayo-Iscar [1] y J. Ortega [2]
[1]Universidad de Valladolid, Valladolid, España.
[2]Centro de Investigación en Matemáticas, Guanajuato, M
IFCS'17, Tokio, August 9th, 2017

Introduction
Definition of a Biplot (for continuous data)
Logistic Biplots
Fitting a Logistic Biplot
Example
(External) Logistic Biplots for Nominal and Ordinal Data - IFCS 2017 - Tokyo
José Luis Vicente Villardón
Universidad de Salamanca - Spain
800 AÑOS
1218-2018



Clustering of Histograms using a Simulated Annealing Algorithm

Alejandro Chacón–Vargas [1] Javier Trejos–Zelaya[2]

International Federation of Classification Societies
Tokyo – 2017

[1] DEHC Ingenieros Consultores. E- Mail: achacon@dehc.cr
[2] CIMPA, Universidad de Costa Rica. E- Mail: javier.trejos@ucr.ac.cr

Model-based clustering of count data based on the logistic-normal-multinomial distribution

M. Comas-Cufí[1]
marc.comas@udg.edu

J.A. Martín-Fernández[1]
josepantoni.martin@udg.edu

G. Mateu-Figueras[1]
gloria.mateu@udg.edu

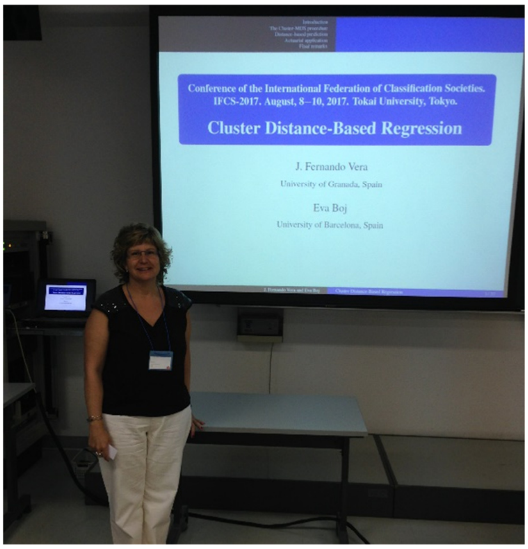
J. Palarea-Albaladejo[2]
javier.palarea@bioss.ac.uk

[1] Dpt. d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, Girona

Universitat de Girona
Departament d'Informàtica,
Matemàtica Aplicada i Estadística

[2] Biomathematics and Statistics Scotland, Edinburgh

BioSS

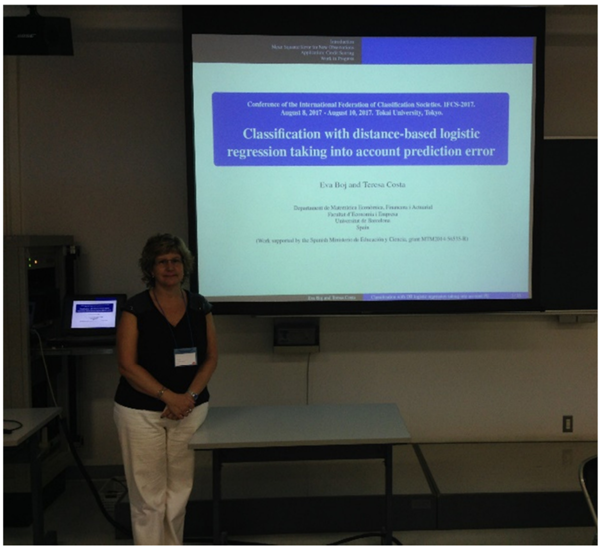


Conference of the International Federation of Classification Societies. IFCS-2017. August, 8–10, 2017. Tokai University, Tokyo.

Cluster Distance-Based Regression

J. Fernando Vera
University of Granada, Spain

Eva Boj
University of Barcelona, Spain



Conference of the International Federation of Classification Societies. IFCS-2017. August 8, 2017 - August 10, 2017. Tokai University, Tokyo.

Classification with distance-based logistic regression taking into account prediction error

Eva Boj and Teresa Costa

List of Contributors:

Alvarez Esteban, Pedro Cesar

Benítez Peña, Sandra

Boj del Val, Eva

Blanquero, Rafael

Carrizosa, Emilio

Castilla González, Elena M.

Comas-Cufí, Marc

Conde del Rio, David

Costa Cor, Mª Teresa

Cuesta Albertos, Juan A.

del Barrio Tellado, Eustasio

Fernández Temprano, Miguel A.

García Escudero, Luis Ángel

García Lapresta, José Luis

Ghosh, Abhik

González del Pozo, Raquel

Gordaliza Pastor, Paula

Gordaliza Ramos, Alfonso

Greselin, Francesca

Guerrero Lozano, Vanesa

Hennig, Christian

Ingrassia, Salvatore

Inouzhe Valdés, Hristo

Martin, Nirian

Martín-Fernández, Josep Antoni

Mateu-Figueras, María Gloria

Matrán Bea, Carlos

Mayo Iscar, Agustín

Ortega, Joaquín

Palarea Albaladejo, Javier

Pardo, Leandro

Ramírez Cobo, Pepa

Rivera García, Diego

Pérez Román, David

Romero Morales, Dolores

Rueda Sabater, Cristina

Salvador González, Bonifacio

Sauerbrei, Willi

Sillero Denamiel, M. Remedios

Vera, José Fernando

Vicente Villardon, José Luis

List of Participants:

Agulló Antolín, Marina	Universidad de Valladolid
Alvarez Esteban, Pedro Cesar	Universidad de Valladolid
Benítez Peña, Sandra	Universidad de Sevilla
Boj del Val, Eva	Universidad de Barcelona
Castilla González, Elena M.	Universidad Complutense de Madrid
Comas-Cufí, Marc	Universidad de Girona
Conde del Rio, David	Universidad de Valladolid
Cuesta Albertos, Juan A.	Universidad de Cantabria
del Barrio Tellado, Eustasio	Universidad de Valladolid
Fernández Temprano, Miguel A.	Universidad de Valladolid
García Escudero, Luis Ángel	Universidad de Valladolid
García Lapresta, José Luis	Universidad de Valladolid
González del Pozo, Raquel	Universidad de Valladolid
Gordaliza Pastor, Paula	Université Toulouse III - Paul Sabatier
Gordaliza Ramos, Alfonso	Universidad de Valladolid
Greselin, Francesca	Universitá degli Studi di Milano - Bicocca
Guerrero Lozano, Vanesa	Universidad Carlos III
Hennig, Christian	University College of London
Inouzhe Valdés, Hristo	Universidad de Valladolid
Larriba González, Yolanda	Universidad de Valladolid
Mata Crespo, Raquel	Universidad de Valladolid

Matrán Bea, Carlos	Universidad de Valladolid
Mayo Iscar, Agustín	Universidad de Valladolid
Pérez Román, David	Universidad de Valladolid
Poveda Marina, José Luis	Universidad de Salamanca
Rueda Sabater, Cristina	Universidad de Valladolid
Salvador González, Bonifacio	Universidad de Valladolid
Sillero Denamiel, M. Remedios	Universidad de Sevilla
Trandafir, Camelia	Universidad de Navarra
Vera, José Fernando	Universidad de Granada
Vicente Villardon, José Luis	Universidad de Salamanca

III International Workshop on Proximity Data,
Multivariate Analysis and Classification
S E I O

III International Workshop on Proximity Data, Multivariate Analysis and Classification

http://www.eio.uva.es/wamyc/

Universidad de Valladolid