SIS - CLADAG

CLADAG 2015

10° Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula,
October 8-10, 2015

BOOK OF ABSTRACTS

Editors:
Francesco Mola, Claudio Conversano
CUEC Editrice, Cagliari

CUEC
editrice

ISBN: 978 88 8467 749 9

Univeristà degli Studi
di Cagliari

Fondazione
Banco di Sardegna

CUEC EDITRICE
by Sardegna Novamedia Soc. Coop.

Via Basilicata n. 57/59
09127 Cagliari,
ITALY
Tel. & Fax +39 070 271573
www.cuec.eu
info@cuec.eu

ISBN: 978-88-8467-749-9

Preface

CLADAG 2015, the 10th Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society (SIS), will be held in Santa Margherita di Pula, Cagliari, Italy, from October 8th to October 10th 2015. The local organizer is the Department of Business and Economics of the University of Cagliari.

CLADAG 2015 will take place under the auspices of the International Federation of Classification Societies (IFCS) and of the Italian Statistical Society (SIS). It promotes advanced methodological research in multivariate statistics with a special vocation in Data Analysis and Classification. CLADAG supports the interchange of ideas in these fields of research, including the dissemination of concepts, numerical methods, algorithms, computational and applied results. It will also benefit of the support of Fondazione Banco di Sardegna.

CLADAG is a member of the International Federation of Classification Societies (IFCS). Among its activities, CLADAG organizes a biennial scientific meeting, schools related to classification and data analysis, publishes a newsletter, and cooperates with other member societies of the IFCS to the organization of their conferences.

The scientific program comprises three Keynote Lectures, an Invited Session, 10 Specialized Sessions, 15 Solicited Sessions and 15 Contributed Sessions. All the Specialized and Solicited Sessions have been promoted by the members of the Scientific Program Committee. The organizers wish to thank them for their cooperation in contributing to the success of CLADAG 2015.

The Book of Abstracts contains short papers of all the presentations scheduled in the conference program. It is organized according to type of session/lecture: Keynote Lectures, Specialized Sessions, Solicited Sessions and Contributed Sessions.

The editors would like to express their gratitude to the Rector of the University of Cagliari, the Director of the Department of Business and Economics and to all the statisticians working in the Department of Business and Economics for their enthusiasm in supporting the organization of this event from the very beginning, as well as to all people who worked hard to make it a success. Special thanks go to Dr. Massimo Cannas, Dr. Luca Frigau and Dr. Farideh Tavazoee for their editorial support

Last but not least, we thank all authors and participants, without whom the conference would not have been possible.

Cagliari, October 8 2015.

Francesco Mola,
Claudio Conversano

Conference Themes

The 10th Meeting is orientated towards all topics related to data analysis, classification, multivariate and computational statistics. Submission of papers addressing these topics in both methodological and practical perspective has been encouraged by the members of the Scientific Program Committee.

The list of topics includes, but is not limited to, the following:

A Classification Theory
Bayesian Classification Biplots Clustering models Consensus of Classifications Correspondence Analysis Discrimination and Classification Factor Analysis and Dimension Reduction Methods Fuzzy Methods Genetic Algorithms Hierarchical Classification Multidimensional Scaling Multiway Scaling Multiway Methods Neural Networks for Classification Non Hierarchical Classification Similarities and Dissimilarities Software algorithms for classification Unfolding and Related Scaling Methods

B Data Analysis
Bayesian data Analysis Big data analysis- Categorical Data Analysis Covariance Structure Analysis Data Mining Data Science Data Visualization Decision Trees Functional data analysis Mixture and Latent Class Models Multilevel data Analysis Non Linear Data Analysis Nonparametric and Semiparametric Regression Partial Least Squares Pattern recognition Robustness and Data Diagnostics Social networks- Software algorithms for multivariate analysis Spatial Data Analysis Symbolic Data Analysis.

Committees

Scientific Program Committee

Chair: Paolo Giudici (*University of Pavia*)

Members
Giuseppe Bove (*University of Roma Tre*)
Daniela Calo (*University of Bologna*)
Agostino Di Ciaccio (*University of Roma La Sapienza*)
Vincenzo Esposito Vinzi (*ESSEC, France*)
Francesca Greselin (*University of Milano Bicocca*)
Francesco Mola (*University of Cagliari*)
Francesco Palumbo (*University of Naples Federico II*)
Carla Rampichini (*University of Firenze*)
Giancarlo Ragozini (*University of Naples Federico II*)
Fabrizio Ruggeri (*CNR IMATI, Milan*)
Silvia Salini (*University of Milan*)
Adalbert F.X. Wilhelm (*Jacobs University Bremen, Germany*)

Local Organizing Committee

Chair: Francesco Mola (*University of Cagliari, Italy*).

Members: Stefano Cabras, Massimo Cannas, Claudio Conversano, Luca Frigau, Monica Musio, Mariano Porcu, Luisa Salaris, Isabella Sulis, Nicola Tedesco.

PARTICIPATING ORGANIZATIONS

International Federation of Classifican Societies (IFCS)

Società Italiana di Statistica (SIS)

SIS - CLADAG (Classification and Data Analysis Group of the Italian Statistical Society)

Università degli Studi di Cagliari

Fondazione Banco di Sardegna

Table of Contents

Causal Inference with Complex Data Structures
Organizer and Chair: Alessandra Mattei

Clustering in Time Series
Organizer and Chair: Michele La Rocca

Multiway Analysis
Organizer and Chair: Giuseppe Bove

Big Data Analysis
Organizer and Chair: Donato Malerba

Advances in Ordinal and Preference Data
Organizer and Chair: Antonio D'Ambrosio

Case studies in data science from Ligurian companies
Organizer and Chair: Delio Panaro

DIAGNOSTIC TOOLS FOR GAMLSS FITTED OBJECTS

Andrea Marletta[1], Mariangela Sciandra[1]

[1] Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo, (e-mail: andrea.marletta@unipa.it, mariangela.sciandra@unipa.it)

ABSTRACT: In the last years GAMLSS models were applied in many research fields representing a good solution to analyze data with huge variabilty. In this paper we propose a new approach to diagnostics in GAMLSS as an alternative to classical worm plot. An application will be shown where the class of GAMLSS is applied in order to detect the presence of liver fibrosis as a function of patients risk factors.

KEYWORDS: GAMLSS, liver fibrosis, mixture, worm plot, residuals analysis.

1 Introduction

We discuss some diagnostic tools for Generalized Additive Models for Location Scale and Shape (GAMLSS) in order to be able to identify possible departures from the model assumptions. Studying the adequacy of a GAMLSS model is not so obvious and little has been done in literature because of the several simultaneous assumptions each model includes about the different parameter involved in the model. So, for example, GAMLSS could show inaccuracies in the assumed linear predictors, one for each specified parameter or inadequacies related to overdispersion and misspecification in link functions. Moreover, due to the wide flexibility GAMLSS offer another common problem is related to misspecification of the family of conditional distribution. In this work we want to emphasize the problem of overdispersion, the most common form of unexpected variation. It occurs when the data exhibit variability exceeding that prescribed by the assumed distribution. As Fig. 1 shows, in the *liver fibrosis* example, data seem to be overdispersed. There is an increase of the variance with the mean ((a)) and a non linear relationship between means and the log transformed variances, as the two fitted curves show ((b)). In order to detect the presence of a variance-mean relationship, in Section 3 we propose the use of a mixture model using GAMLSS family of distributions.

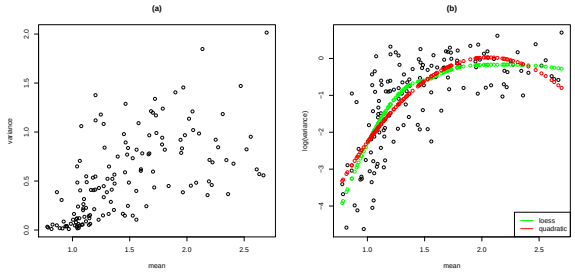

Figure 1. *Mean-variance (a) and mean-log(variance) (b) relationships in liver fibrosis data*

2 The GAMLSS models

General Additive Models for Location Scale and Shape were introduced firstly by Rigby and Stasinopoulos (2001) as a way of overcoming some of the limitations associated with Generalized Linear Models and Generalized Additive Models. They represent a flexible class of models for several reasons. Firstly, the distribution of the response variable can be selected by a very wide range of distributions including highly skewed and kurtotic continuous and discrete distributions. Moreover, once the response distribution has been fixed, they allow to model all the parameters of the chosen distribution using parametric and/or non parametric smooth functions of the explanatory variables. So, assuming the response variable Y to follow a four parameters distribution $Y \sim D(\theta)$ with $\theta = (\mu, \sigma, \nu, \tau)$, where μ and σ usually are location and scale parameters while ν and τ shape parameters, the formulation of GAMLSS given by Rigby and Stasinopoulos (2005) is

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k \beta_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \qquad k = 1,2,3,4 \tag{1}$$

where $g_k(.)$ are known monotonic link functions relating in a parametric way the distribution parameters to the explanatory variables $\mathbf{X}_k$ and h_{jk} represent the non-parametric additive terms. The vector of parameters β_k and the non parametric terms can be estimated following several approaches as described in Rigby and Stasinopoulos (2005).

3 Diagnostic tools for GAMLSS models

As in a classical model regression framework, once a model is fitted, next step deals with the problem of model selection. In a GAMLSS setting, model selection is usually performed by comparing various competing models in which different combinations of the components change; then, the overall adequacy of the selected model is assessed through the analysis of the *randomized quantile residuals* (Dunn & Smyth, 1996). They are defined as

$$r_i^q = \Phi^{-1}(u_i), \qquad i = 1, \ldots, n$$

where $\Phi(\cdot)$ represents the standard normal distribution function, and u_i is an uniform random variable on a specific region of the linear predictor. Strange pattern in the plot of these residuals against the predictors could suggest misspecified link functions. In order to identify regions of explanatory variables within which the models do not show an adequate fit an useful tool is given by the *worm plots* of residuals introduced by Van Buuren *et al.* (2001). The tool consists of a number of detrended Q-Q plots splitted according to some predictors. A model that fits the data well is characterized by "flat worms".

This paper proposes the use of a mixture approach for GAMLSS when standard diagnostic tools show overdispersed data. The central idea is: if the worm plot shows for example M-shape pattern, it could suggest bimodality in statistical units. Then, GAMLSS mixture model could be used to identify the underlying distributions.

4 A real dataset example

Liver fibrosis is one of the ten most frequent causes of death in the world and consists in a massive presence of connective tissue. It can be classified in 5 stages through the Metavir scoring system from a normal ($F0$) to a cirrhotic ($F4$) liver. In medicine, liver biopsy represents the gold standard test for staging liver disease. An alternative diagnostic technique is represented by the Acoustic Radiation Force Impulse (ARFI). ARFI measures the liver stiffness through mechanical excitation of tissue using acoustic pulses producing shear waves propoagation. The ARFI principle is that the stiffer the tissue, the faster be shear waves propagate. The dataset used in this example contains data about ARFI measurements collected in 2013 for 141 patients. To each elastography are associated a different number of measurements so the dataset shows a two-step hierarchical structure: a level for the exam and a second level for the measurements done during the same exam. The response variable is liver stiffness

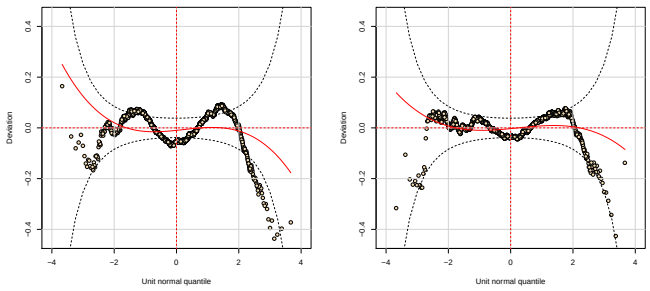

Figure 2. *Worm plots for two fitted GAMLSS objects*

(measured as wave speed in m/s) while explanatory variables are divided into two groups: patient specific explanatory variables (sex, age, size and weight) and explanatory variable of exam (depth, liver segment, patient position). In figure 2 two worm plots are shown: on the left a simple GAMLSS with a four parameters BCPE distribution shows a clear M-shape pattern; on the right we use a mixture to obtain a more flat worm and more points between boundaries. A possible hypothesis is that two groups could represent two classes of subjects: healthy and cirrhotic patients. To find further evidence about these conclusions, we will try to simulate data from different scenarios, for example by considering several degrees of overdispersion. Moreover, simulations will be also used to study the inferential properties of the proposed approach.

References

DUNN, P.K., & SMYTH, G.K. 1996. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 236-244.

RIGBY, R.A., & STASINOPOULOS, D. M. 2005. Generalized additive models for location, scale and shape). *Applied Statistics*, 54(3), 507-544.

RIGBY, R.A., & STASINOPOULOS, D. M. 2009. A flexible regression approach using GAMLSS in R.

VAN BUUREN, S., & FREDRIKS. 2001. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, 1259-1277.