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

Modeling ordinal data
Organizer and Chair: Maurizio Carpita

Functional data analysis for environmental data
Organizer and Chair: Tonio Di Battista

Advances in quantile regression
Organizer and Chair: Cristina Davino

Directional Data
Organizer and Chair: Giovanni C. Porzio

ROBUST CLUSTERING FOR HETEROGENOUS SKEW DATA

Luis Angel García-Escudero,[1] Francesca Greselin[2] and Agustin Mayo-Iscar[1]

[1] Department of Statistics and Operations Research and IMUVA, University of Valladolid (e-mail: lagarcia@eio.uva.es, agustinm@eio.uva.es)

[2] Department of Statistics and Quantitative Methods, Milano-Bicocca University (e-mail: francesca.greselin@unimib.it)

ABSTRACT: The existing robust methods for model-based classification and clustering deal with elliptically contoured components. Here we introduce robust estimation for mixtures of skew-normal, by the joint usage of trimming and constraints. The model allows to fit heterogeneous skew data with great flexibility.

KEYWORDS: Clustering, robust estimation, skew data.

1 Introduction

In recent years, empirical evidences of asymmetric departures from normality in some real subpopulations suggested the introduction of skewed components in the classical mixture model approach. Therefore, increasing attention has been devoted to mixtures of multivariate skew normal and skew t, as well as to mixtures of normal-inverse-Gaussian distributions, shifted asymmetric Laplace distributions, generalized hyperbolic distributions and hierarchical mixtures (mixtures of mixtures), to capture asymmetric shapes in components. Most of these parametric families of skew distributions are clearly related, as it has been described in a careful review by Lee & McLachlan (2013a). In this paper we address robust estimation of mixtures of skew normal distributions. Our interest for the Finite Mixtures of Canonical Fundamental Skew Normal (FM CFUSN) has been motivated by the aim of adapting to the Gaussian case the methodology developed for Finite Mixtures of Canonical Fundamental Skew t (FM CFUST) in Lee & McLachlan (2014), based on skew distributions originated in Arellano-Valle & Genton (2005). By mimicking the CFUST properties, CFUSN has the special appealing of including as particular cases the restricted and unrestricted Skew Normal, respectively rMSN and uMSN. The great flexibility of CFUSN is due to its parameters, which allow to separately govern location, scale, correlation, and skewness. We will show, herein, that the FM CFUSN offers a very tractable model for many situations of departures from symmetry, being analytically and computationally easier than the FM CFUST. Taking advantage of its robust estimation, the model becomes even more adaptable to real phenomena. Indeed, robustness is often required by many datasets available nowadays, where populations are noisy, have some non-symmetric features, and could contain outliers asymmetrically disposed around the data clusters.

The first step to achieve robustness and obtain good breakdown properties for the estimators, is based on a *trimming procedure* incorporated along the iterations of the EM algorithm. The key idea is that a small portion of observations, which are highly unlikely to occur under the current fitted model assumption, is discarded from contributing to the parameter estimates. The same methodology has been proven to be very effective when addressing robust clustering for Gaussian and t mixtures models (see Neykov *et al.* , 2007; Gallegos & Ritter, 2009; García-Escudero *et al.* , 2014). Furthermore - and this is the second step - we implement a *constrained ML estimation* for the component covariances, aiming at reducing spurious solutions and preventing singularities of the likelihood, along the lines of Ingrassia & Rocci (2007).

Monte Carlo experiments show that bias and MSE of the estimator in several cases of contaminated data are dramatically inflated, while they return to be comparable to results obtained on skew data without noise, when the combined effect of trimming and constrained estimation is applied. Further, it can be shown that the estimator resists the influence of all classes of contaminating observations, as far as the outliers appear in a proportion lower than the trimming level α. As a final remark, we want to stress that the joint usage of trimming and constrained estimation allow to set the underlying mathematical and statistical problems as well posed ones.

2 Methodology

We consider the ML estimation for a g-component mixture model in which a random sample $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ follows a mixture of CFUSNs. The probability density function can be written as

$$\mathbf{Y}_j \sim \sum_{i=1}^{g} \pi_i f(\mathbf{y}_j|\mu_i, \Sigma_i, \Delta_i), \quad \pi_i \geq 0, \quad \sum_{i=1}^{g} \pi_i = 1,$$

where $f(\cdot)$ denotes a CFUSN density, $\Theta = (\theta_1, \ldots, \theta_g)$ with $\theta_i = (\pi_i, \mu_i, \Sigma_i, \Delta_i)$ are the unknown parameters of component i, and π_i are the weights of the groups.

To define the location-scale variant of the CFUSN distribution, let $(\mathbf{Y}_0, \mathbf{Y}_1)$ be a $p+q$ multivariate normal r.v., such that

$$\begin{bmatrix}\mathbf{Y}_0\\ \mathbf{Y}_1\end{bmatrix} \sim \mathcal{N}_{q+p}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathbf{I}_q & 0\\ 0 & \Sigma\end{bmatrix}\right) \tag{1}$$

where Σ is a positive definite scale matrix and 0 is a vector of zeros with appropriate dimension. Then, given a $p \times q$ matrix Δ, we obtain the stochastic representation for $\mathbf{Y}$, i.e. $\mathbf{Y} = \mu + \Delta|\mathbf{Y}_0| + \mathbf{Y}_1$, that follows the CFUSN distribution with density given by

$$f(\mathbf{y};\mu,\Sigma,\Delta) = 2^q \phi_p(\mathbf{y};\mu,\Omega)\Phi_q\left(\Delta^T\Omega^{-1}(\mathbf{y}-\mu);0,\Lambda\right)$$

where $\Omega = \Sigma + \Delta\Delta^T$, $\Lambda = \mathbf{I}_p - \Delta^T\Omega^{-1}\Delta$, and as usual $\phi_p(\mathbf{y};\mu,\Sigma)$ denotes the p-dimensional density of the multivariate Gaussian with mean μ and scale Σ evaluated at $\mathbf{y}$, while $\Phi_q(\cdot)$ denotes the cumulative distribution function.

In the ML approach, we consider the following *trimmed* log-likelihood function (see Neykov *et al.* , 2007; Gallegos & Ritter, 2009; García-Escudero *et al.* , 2014)

$$\mathcal{L}_{trim} = \sum_{i=1}^{n} z(\mathbf{y}_i)\log\left[\sum_{g=1}^{G}\phi_p(\mathbf{y}_i;\mu_g,\Omega_g)\Phi_q(\Delta_g^T\Omega_g^{-1}(\mathbf{y}_i-\mu_g);0,\Lambda_g)\pi_g\right]. \tag{2}$$

By $z(\cdot)$ we denote a 0-1 trimming indicator function that indicates whether observation $\mathbf{y}_i$ is trimmed off: $z(\mathbf{y}_i)$=0, or not: $z(\mathbf{y}_i)$=1. A fixed fraction α of observations, whose contributions to the likelihood are lower than their α-quantile, can be unassigned by setting $\sum_{i=1}^{n} z(\mathbf{y}_i) = [n(1-\alpha)]$ at each E-step, hence they do not contribute to the parameter estimation.

Further, we want to deal with the unboundedness of the target function $\mathcal{L}_{trim}$ when no constraints are imposed on the scatter parameters. In this case, the defining problem is ill-posed because the log-likelihood tends to ∞ when either $\mu_g = \mathbf{y}_i$ and $|\Sigma_g| \to 0$. As a trivial consequence, the EM algorithm can be trapped into non-interesting local maximizers, called "spurious" solutions, and the result of the EM algorithm strongly depends on its initialization. For this reason, we set a constraint on the maximization in (2), concerning the eigenvalues $\{\lambda_l(\Sigma_g)\}_{l=1,\ldots,d}$ of the scatter matrices Σ_g, by imposing

$$\lambda_{l_1}(\Sigma_{g_1}) \leq c\,\lambda_{l_2}(\Sigma_{g_2}) \qquad \text{for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G.$$

as in Ingrassia & Rocci (2007); García-Escudero *et al.* (2008); Gallegos & Ritter (2009). It is important to remark that the parameters related with location and skewness are not affected by the constrained maximization applied to the covariance matrices.

As usual, among many runs of the EM algorithm, we select the parameters given by the best final likelihood. To complete a synthetic description of the algorithm, our proposal for the initialization of the EM is to take a small random subsample for each component and to draw, for each observation in it, a random variate $\mathbf{Y}_0$ from $\mathcal{N}_p(0, \mathbf{I}_p)$. By applying multivariate regression on the selected observations in each component, and using the corresponding vectors $|\mathbf{Y}_0|$, we get an estimation of the model parameters (i.e. μ_g, Σ_g and Δ_g) for the g-th component. Finally, initial estimations for the group weights are drawn from a multinomial random variable.

References

ARELLANO-VALLE, R.B., & GENTON, M.G. 2005. On fundamental skew distributions. *Journal of Multivariate Analysis*, **96**(1), 93–116.

FRITZ, H., GARCIA-ESCUDERO, L.A., & MAYO-ISCAR, A. 2012. tclust: An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software*, **47**(12).

GALLEGOS, M.T., & RITTER, G. 2009. Trimmed ML estimation of contaminated mixtures. *Sankhya (Ser. A)*, 164–220.

GARCÍA-ESCUDERO, L.A, GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 1324–1345.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., & MAYO-ISCAR, A. 2014. A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, **8**(1), 27–43.

INGRASSIA, S., & ROCCI, R. 2007. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis*, **51**, 5339–5351.

LEE, S., & MCLACHLAN, G. 2013a. Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications*, **22**(4), 427–454.

LEE, S., & MCLACHLAN, G. 2013b. On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, **7**(3), 241–266.

LEE, S., & MCLACHLAN, G. 2014. Finite Mixtures of Canonical Fundamental Skew t-Distributions. *arXiv preprint arXiv:1405.0685*.

NEYKOV, N., FILZMOSER, P., DIMOVA, R., & NEYTCHEV, P. 2007. Robust Fitting of Mixtures Using the Trimmed Likelihood Estimator. *Computational Statistics & Data Analysis*, **52**(1), 299–308.