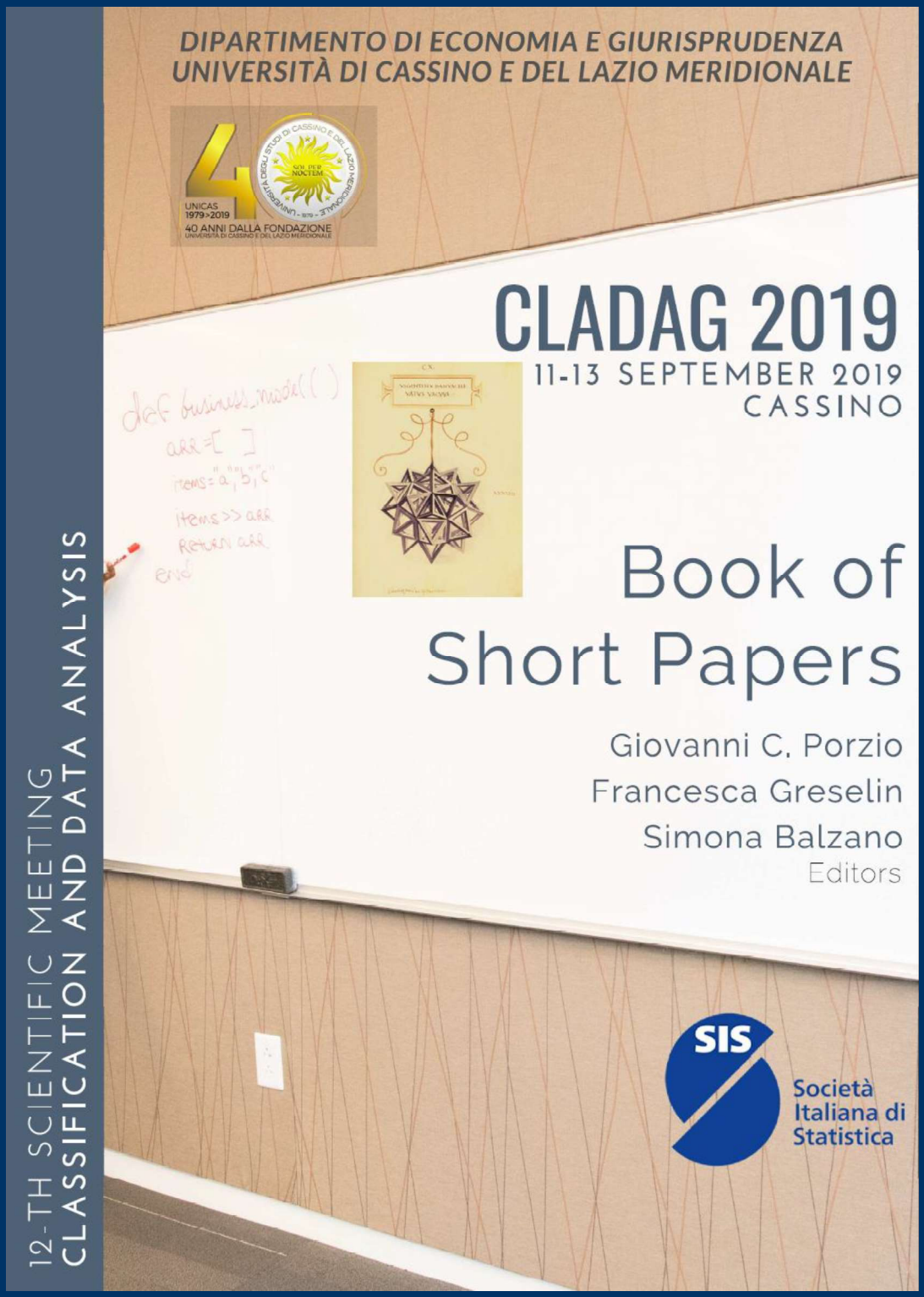
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA



EDIZIONI UNIVERSITA' DI CASSINO

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6

CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

Contents

Keynotes lectures

Invited and contributed sessions

Preface

This book collects the short papers presented at CLADAG 2019, the 12th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS).

The meeting has been organized by the Department of Economics and Law of the University of Cassino and Southern Lazio, under the auspices of the SIS and the International Federation of Classification Societies (IFCS). CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research.

Every two years, CLADAG organizes a scientific meeting, devoted to the presentation of theoretical and applied papers on classification and related methods of data analysis in the broad sense. This includes advanced methodological research in multivariate statistics, mathematical and statistical investigations, survey papers on the state of the art, real case studies, papers on numerical and algorithmic aspects, applications in special fields of interest, and the interface between classification and data science. The conference aims at encouraging the interchange of ideas in the above-mentioned fields of research, as well as the dissemination of new findings.

CLADAG conferences, initiated in 1997 in Pescara (Italy), were soon considered as an attractive information exchange market and became a most important meeting point for people interested in classification and data analysis. One reason was

certainly the fact that a selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

The Scientific Committee of CLADAG2019 conceived the Plenary and Invited Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2019 is particularly rich. All in all, it comprises 5 Keynote Lectures, 32 Invited Sessions promoted by the members of the Scientific Program Committee, 16 Contributed Sessions, a Round Table and a Data Competition. We thank all the session organizers for inviting renowned speakers, coming from 28 countries. We are greatly indebted to the referees, for the time spent in a careful review.

The editors would like to express their gratitude to the Rector of the University of Cassino and Southern Lazio and the Director of the Department of Economics and Law for having hosted the meeting. Special thanks are finally due to the members of the Local Organizing Committee and all the people who with their abnegation and enthusiasm have worked for CLADAG 2019.

Special thanks go to Alfiero Klain and Livia Iannucci for the editorial and administrative support.

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

Cassino, September 11, 2019

Giovanni C. Porzio
Francesca Greselin
Simona Balzano

ROBUST MODEL-BASED CLUSTERING WITH MILD AND GROSS OUTLIERS

Alessio Farcomeni[1] and Antonio Punzo[2]

[1] Department of Public Health and Infectious Diseases, Sapienza - University of Rome,
(e-mail: alessio.farcomeni@uniroma1.it)

[2] Department of Economics and Business, University of Catania,
(e-mail: antonio.punzo@unict.it)

ABSTRACT: We propose a model-based clustering procedure for mild and gross outliers. Our mixture model is based on heavy-tailed components (e.g., the contaminated normal distribution), but it is assumed to apply only to a subset of the data. Consequently, a proportion of observations is trimmed. We propose a penalized likelihood approach for estimation and selection of the proportions of mild and gross outliers, where the penalty parameter is fixed by formal optimality arguments. We conclude with an original real data example on the identification of the source from illicit drug shipments seized in Italy and Spain.

KEYWORDS: tclust, contaminated normal, penalized likelihood.

1 Introduction

In clustering based on the normal mixture model there are two main approaches to deal with contamination. One is based on the use of heavy-tailed or skewed component distributions. A recent example in this direction, preserving elliptical contours of clusters, are mixtures of contaminated normal (CN) distributions (Punzo & McNicholas, 2016). Component-wise methods are well suited to work with mild outliers (Ritter, 2015), and are sometimes labeled as weakly robust. A separate body of literature has instead worked with outliers in more general position, including gross outliers, and has usually proceeded by discarding or at least downweighting a proportion of the observations (Farcomeni & Greco, 2015). A good example is `tclust` (García-Escudero *et al.* , 2008), where a fixed proportion of observations is trimmed and the rest is assumed to follow a normal mixture model. These procedures have often formal robustness properties, e.g., positive breakdown point asymptotically.

In this work we merge the two approaches above by estimating a CN mixture after trimming a fixed proportion of gross outliers. Our model can be

seen from two different perspectives. On the one hand, clusters having a distribution with slightly heavy tails might be desired in order to assign as many observations to clusters as possible. In this case, it is indeed assumed that clean observations arise from, for example, a CN model. On the other hand, the trade off between mild and gross outliers is exploited in order to increase efficiency: some (mild) outliers are assigned to a cluster and contribute to centroid estimation, therefore decreasing the final mean squared error (MSE).

In this work we tackle also an additional open problem with trimming procedures, that of selecting the trimming proportion. Our proposal is based on a penalized likelihood approach, where the trimming proportion is in practice substituted by a penalty parameter. The advantage is that we can identify a heuristic but theoretically justified way of choosing an optimal penalty level, and therefore an optimal trimming proportion. Our fixed-penalty approach in some sense solves the issue of selecting the trimming proportion both for our model and the special case of trimmed normal mixture models (`tclust`). The methodology proposed in this paper has been implemented in `R` functions which can be downloaded from `https://github.com/afarcome/cntclust`.

2 Methodology

Let $x_1,\ldots,x_i,\ldots,x_n$ be a sample of n observations in d dimensions. Moreover, let $\alpha_0 \geq 0$ denote a trimming proportion of outliers which shall not be used to estimate model parameters. We assume data arise from the contaminated spurious outlier model

$$\prod_{i\in R}\sum_{j=1}^{k}\pi_j f_{\mathrm{CN}}(x_i;\mu_j,\Sigma_j,\alpha_j,\eta_j)\prod_{i\notin R}g_i(x_i), \tag{1}$$

where R denotes a set of non-trimmed observations of cardinality $\lfloor(1-\alpha_0)n\rfloor$ and g_i are pdfs generating the outliers in general position. Let $f_{\mathrm{N}}(\cdot;\mu,\Sigma)$ denote the probability density function (pdf) of a d-variate normal (N) distribution with mean vector μ and covariance matrix Σ. In (1), $f_{\mathrm{CN}}(x;\mu,\Sigma,\alpha,\eta) = (1-\alpha)f_{\mathrm{N}}(x;\mu,\Sigma)+\alpha f_{\mathrm{N}}(x;\mu,\eta\Sigma)$ denotes the pdf of a d-variate CN distribution with mean vector μ, scale matrix Σ, proportion of mild outliers $\alpha\in(0,1)$, and degree of contamination $\eta > 1$.

To estimate the parameters, we optimize the profile likelihood

$$\ell(\vartheta) = \sum_{j=1}^{k}\sum_{i\in R_j}\ell_i(\vartheta) = \sum_{j=1}^{k}\sum_{i\in R_j}\left[\ln\pi_j + \ln f_{\mathrm{CN}}\left(x_i;\mu_j,\Sigma_j,\alpha_j,\eta_j\right)\right], \tag{2}$$

where R_j denotes the set of observations assigned to the j-th cluster. To make maximization of (2) a well defined problem, we adopt the classical eigenvalue ratio constraint proposed by García-Escudero *et al.* , 2008.

Model (1) involves the difficult choice of $\alpha_0, \alpha_1, \ldots, \alpha_k$, where α_0 controls the proportion of gross outliers and α_j the proportion of mild outliers in the j-th cluster. We propose a LASSO-type penalized likelihood approach enforcing a sparse model selection in which some values in the set $(\alpha_0, \alpha_1, \ldots, \alpha_k)$ might be set to zero. A general form of penalized log-likelihood is given by

$$\ell(\vartheta) + P(\alpha_0, \alpha_1, \ldots, \alpha_k), \tag{3}$$

and we propose using $P(\alpha_0, \ldots, \alpha_k) = -\log(n) \sum_{j=0}^{k} \nu_j \alpha_j$. In order to reduce the number of penalty parameters, we set $\nu_0 = n\nu$ and $\nu_j = \nu$ for $j > 0$.

The choice of the penalty parameter ν has got direct consequences on the estimated trimming proportion α_0. If also $\alpha_1, \ldots, \alpha_k$ are included in the penalty, it also affects their estimates. Surprisingly enough, mapping the problem of selecting contaminating proportions to the scale of the likelihood gives an asymptotically "optimal" fixed value, $\nu = \sqrt{2d}$, which under certain assumptions guarantees that observations outside a chi-square type ellipse from a bulk of the data are trimmed.

Maximization of (2), and for fixed ν of (3), is carried out using a classification expectation-conditional maximization (CECM) algorithm, where eigenvalue ratio constraints are activated at the conditional maximization step is needed,

3 Example about clustering illicit drug shipments

We analyze data about $n = 151$ seizures of shipments of cocaine and heroin in Italy and Spain. They were sent to the forensic laboratories for checking the nature of the substance and quantifying the absolute and relative contents of each of several chemical compounds. In modern forensics it is believed that the contents of certain solvents might be useful for identifying the source, that is, clustering packages with respect to the illicit laboratory where the drug was processed. We verify this assumption by focusing on $d = 3$ compounds: hexane, acetone, and 2-propanol. We fix $k = 2$ and estimate a classical normal mixture model and a contaminated normal mixture model without trimming first. Then we use robust clustering methods: `tclust` and the contaminated-normal mixture model with trimming. In Table 1 we report, for values of the trimming level chosen using our penalized likelihood approach, the adjusted

Rand-index (ARI) showing the agreement between the class labels and the true underlying Italy/Spain location of seizure. With no or insufficient trimming one might conclude that there is no relationship between solvent contents and seizure location. On the other hand, after trimming the agreement becomes fairly high. As expected we note that the optimal trimming level using `tclust` is slightly larger than those using CNTCLUST0. While in our low sample size example this might not have strong consequences in terms of MSE, $\lceil 151(0.066-0.053)\rceil = 2$ seizures will not be attributed to a location using `tclust`, which can have forensic consequences.

Table 1. *Adjusted Rand-index (ARI) for location of drug seizure and clustering. In parentheses the trimming level. NM: normal mixture, CNM: contaminated normal mixture,* `tclust`*: trimmed NM, CNTCLUST0: trimmed CNM, CNTCLUST: penalized trimmed CNM with* $\nu = \sqrt{2d}$ *and fixed trimming level. The trimming level selected with our fixed-penalty approach is indicated with* $\hat{\alpha}_0$*.*

Method	ARI
NM	-0.076
CNM	-0.069
`tclust` ($\hat{\alpha}_0 = 0.066$)	0.657

Method	ARI
CNTCLUST0($\hat{\alpha}_0 = 0.053$)	0.660
CNTCLUST($\hat{\alpha}_0 = 0.053$)	0.658

References

FARCOMENI, A., & GRECO, L. 2015. *Robust Methods for Data Reduction*. Boca Raton, FL: CRC Press.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., MATRAN, C., & MAYO-ISCAR, A. 2008. A general trimming approach to robust cluster analysis. *Annals of Statistics*, **36**, 1324–1345.

PUNZO, A., & MCNICHOLAS, P. D. 2016. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

RITTER, G. 2015. *Robust Cluster Analysis and Variable Selection*. CRC Press.