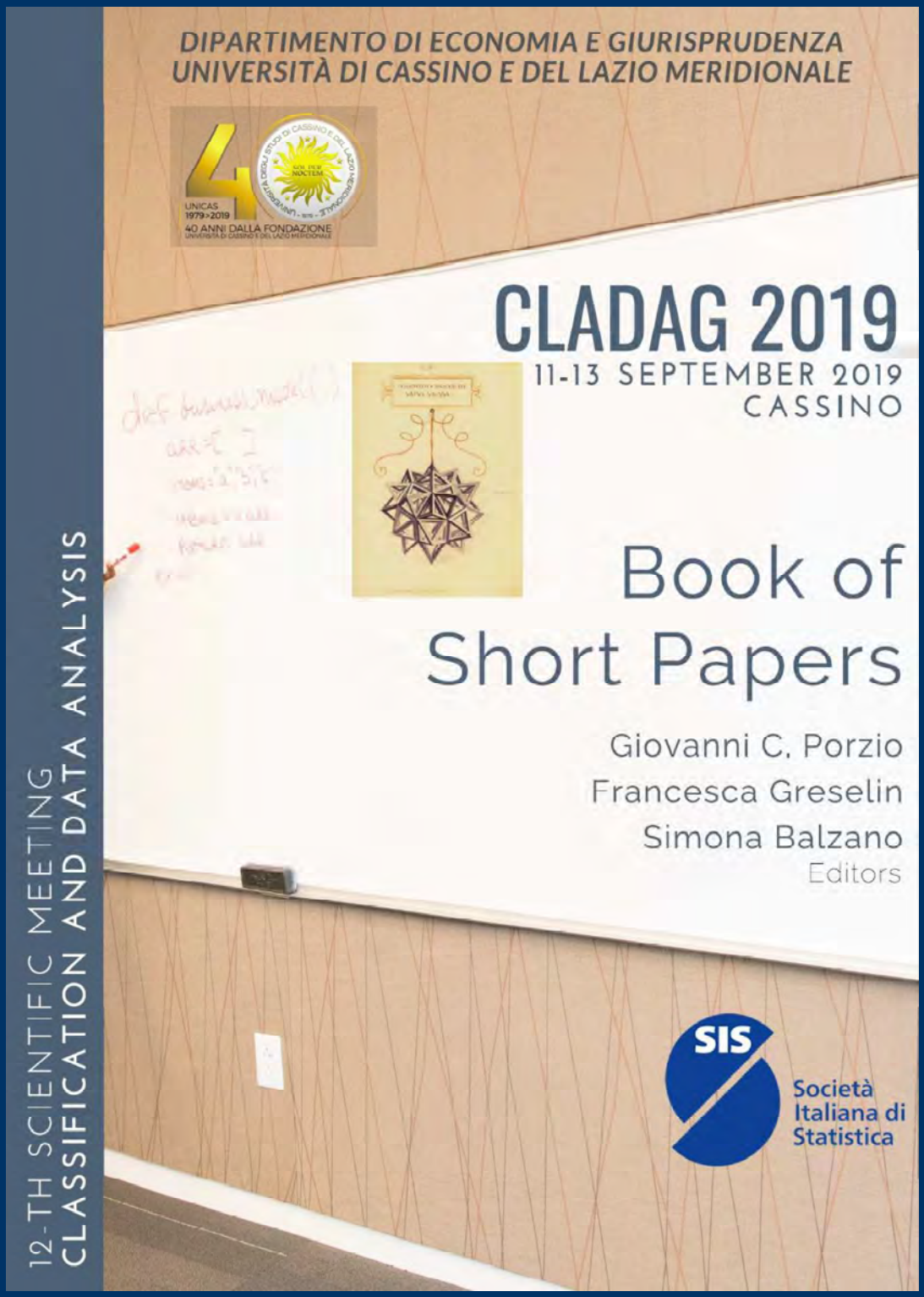
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA

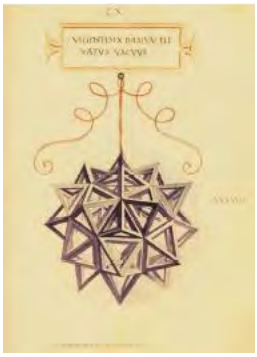


EDIZIONI UNIVERSITÀ DI CASSINO

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

PENALIZED VS CONSTRAINED MAXIMUM LIKELIHOOD APPROACHES FOR CLUSTERWISE LINEAR REGRESSION MODELING

Roberto Di Mari[1], Stefano Antonio Gattone[2] and Roberto Rocci[3, 4]

[1] Department of Economics and Business, University of Catania,
(e-mail: roberto.dimari@unict.it)

[2] Department of Philosophical and Social Sciences, Economics and Quantitative Methods, University G. d'Annunzio, Chieti-Pescara, (e-mail: gattone@unich.it)

[3] Department of Statistical Sciences, University of Rome La Sapienza,

[4] Department of Economics and Finance, University of Rome Tor Vergata,
(e-mail: roberto.rocci@uniroma2.it)

ABSTRACT: Several approaches exist to avoid singular and spurious solutions in maximum likelihood (ML) estimation of clusterwise linear regression models. We propose to solve the degeneracy problem by using a penalized approach: this is done by adding a penalty term to the log-likelihood function which increasingly penalizes smaller values of the scale parameters and the tuning of the penalty term is done based on the data. Another traditional solution to degeneracy consists in imposing constraints on the variances of the regression error terms (constrained approach). We will compare the penalized approach to the constrained approach in a broad simulation study and an empirical application, providing practical guidelines on which approach to use under different circumstances.

KEYWORDS: clusterwise linear regression, penalized likelihood, scale constraints.

1 Introduction

Let $y_1, \ldots, y_n$ be a sample of independent observations drawn from the response random variable Y_i, each observed alongside with a vector of J explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let us assume $Y_i|\mathbf{x}_i$ to be distributed as a finite mixture of linear regression models, that is

$$f(y_i|\mathbf{x}_i;\boldsymbol{\psi}) = \sum_{g=1}^{G} p_g \phi_g(y_i|\mathbf{x}_i, \sigma_g^2, \boldsymbol{\beta}_g) = \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\right], \tag{1}$$

where G is the number of clusters and p_g, $\boldsymbol{\beta}_g$, and σ_g^2 are the mixing proportion, the vector of $J+1$ regression coefficients that includes an intercept, and the variance term for the g-th cluster. The set of all model parameters is given by $\boldsymbol{\psi} = \{(p_1,\ldots,p_G;\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_G;\sigma_1^2,\ldots,\sigma_G^2) \in \mathbb{R}^{(G-1)+(J+1)G+G} : p_1+\cdots+p_G = 1, p_g > 0, \sigma_g^2 > 0, \text{ for } g = 1,\ldots,G\}$.

The likelihood function can be specified as

$$\mathcal{L}(\boldsymbol{\psi}) = \prod_{i=1}^{n}\left\{\sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi{\sigma_g}^2}} \exp\left[-\frac{(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\right]\right\}, \tag{2}$$

which we maximize to estimate $\boldsymbol{\psi}$ either by means of direct maximization or with the perhaps more popular EM algorithm (Dempster *et al.*, 1977). However, there is a well-known complication in ML estimation of this class of models: the likelihood function of mixtures of (conditional) normals with cluster-specific variances is unbounded (Kiefer & Wolfowitz, 1956; Day, 1969).

A traditional solution to the problem of unboundedness is based on the seminal work of Hathaway (1985) which, in order to have the likelihood function of univariate mixtures of normals bounded, suggested to impose a lower bound to the ratios of the scale parameters in the maximization step. The method is equivariant under linear affine transformations of the data. That is, if the data are linearly transformed, the estimated posterior probabilities do not change and the clustering remains unaltered. Recently, in the multivariate case, Rocci *et al.* (2018) incorporated constraints on the eigenvalues of the component covariances of Gaussian mixtures that are tuned on the data based on a cross–validation strategy. These constraints are built upon Ingrassia (2004)'s reformulation and are an equivariant sufficient condition for Hathaway's constraints. Estimation is done in a familiar ML environment Ingrassia & Rocci (2007), with data–driven selection of the scale balance. Di Mari *et al.* (2017) adapted Rocci *et al.* (2018)'s method to clusterwise linear regression, further investigating its properties.

Another possible approach for handling unboundedness is to modify the log-likelihood function by adding a penalty term, in which smaller values of the scale parameters are increasingly penalized. Representative examples can be found in Chen & Tan (2009); Chen *et al.* (2008); Ciuperca *et al.* (2003).

In this work we review the constrained approach of Di Mari *et al.* (2017) and develop a data-driven equivariant penalized approach for ML estimation. Next, we sketch the bulk of the methodologies.

2 The methodology

2.1 The constrained approach

Di Mari *et al.* (2017) proposed relative constraints on the group conditional variances σ_g^2 of the kind

$$\sqrt{c} \leq \frac{\sigma_g^2}{\bar{\sigma}^2} \leq \frac{1}{\sqrt{c}}, \tag{3}$$

or equivalently

$$\bar{\sigma}^2 \sqrt{c} \leq \sigma_g^2 \leq \bar{\sigma}^2 \frac{1}{\sqrt{c}}. \tag{4}$$

The above constraints are equivariant and have the effect of shrinking the variances to a suitably chosen $\bar{\sigma}^2$, the *target* variance term, and the level of shrinkage is given by the value of c. This constraints are easily implementable within the EM algorithm (Ingrassia, 2004; Ingrassia & Rocci, 2007), which is fully available in closed-form, and the selection of c is based on the data.

2.2 The penalized approach

An alternative to the constrained estimator is the penalized approach, in which a penalty $s_n(\sigma_1^2, \ldots, \sigma_G^2)$ is put on the component variances and it is added to the log-likelihood. Under certain conditions on the penalty function, the penalized estimator is know to be consistent (Chen & Tan, 2009). A function s_n that satisfies these conditions is

$$s_n(\sigma_1^2, \ldots, \sigma_G^2) = -\lambda \sum_{g=1}^{G} \left(\frac{\bar{\sigma}^2}{\sigma_g^2} + \log(\sigma_g^2) \right), \tag{5}$$

where $\bar{\sigma}^2$, the *target* variance, can be seen as our *prior* information on the scale structure and λ is the penalizing constant that is selected based on the data. Thus, the penalized log-likelihood can be written as

$$p\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + s_n(\sigma_1^2, \ldots, \sigma_G^2) \tag{6}$$

and the set of unknown parameters is found by ML with computation done by means of an EM algorithm that is available in closed-form. As well as with the constrained approach, the penalized approach is equivariant with respect to linear transformation in the response.

References

CHEN, J., & TAN, X. 2009. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, **100**(7), 1367–1383.

CHEN, J., TAN, X., & ZHANG, R. 2008. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 443–465.

CIUPERCA, G., RIDOLFI, A., & IDIER, J. 2003. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30**(1), 45–59.

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

DEMPSTER, A.P., LAIRD, N.M., & RUBIN, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

DI MARI, R., ROCCI, R., & GATTONE, S.A. 2017. Clusterwise linear regression modeling with soft scale constraints. *International Journal of Approximate Reasoning*, **91**, 160–178.

HATHAWAY, R.J. 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.

INGRASSIA, S. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, **13**(2), 151–166.

INGRASSIA, S., & ROCCI, R. 2007. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis*, **51**(11), 5339–5351.

KIEFER, J., & WOLFOWITZ, J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

ROCCI, R., GATTONE, S.A., & DI MARI, R. 2018. A data driven equivariant approach to constrained Gaussian mixture modeling. *Advances in Data Analysis and Classification*, **12**(2), 235–260.

CLADAG 2019
Cassino (ITALY)
11-13 September, 2019

Società Italiana di Statistica

UNICAS 1979>2019
40 YEARS FROM FOUNDATION
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE

The CLAssification and Data Analysis Group of the Italian Statistical Society (SIS) 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.

CLADAG is a member of the International Federation of Classification Societies (IFCS).

Among its activities, CLADAG organizes a biennial international 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.

Founded in 1985, the IFCS is a federation of national, regional, and linguistically-based classification societies. It is a non-profit, nonpolitical scientific organization, whose aims are to further classification research.