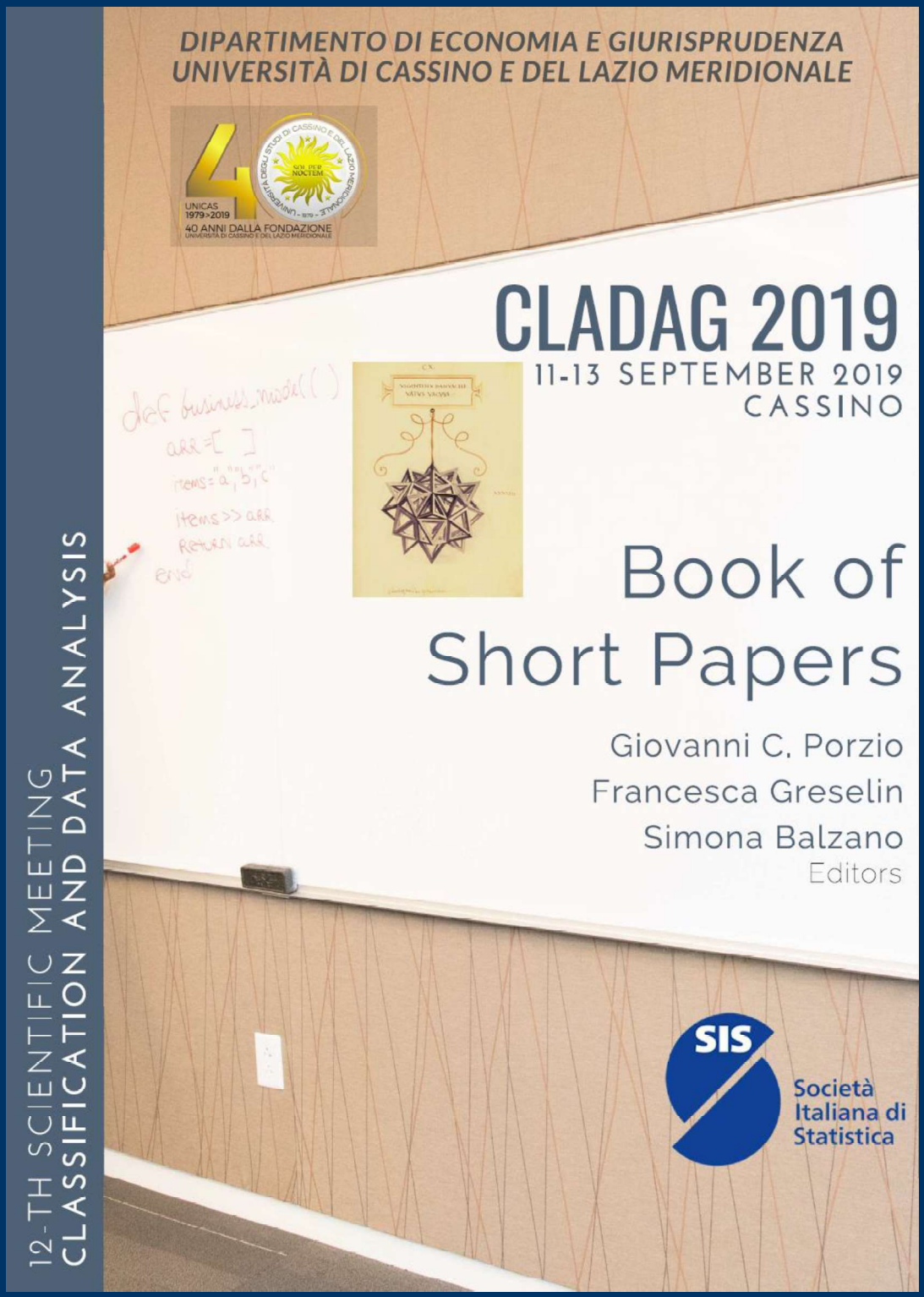
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

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

AN INDSCAL BASED MIXTURE MODEL TO CLUSTER MIXED-TYPE OF DATA

Roberto Rocci[1, 2] and Monia Ranalli[1]

[1] Department of Statistical Sciences, Sapienza University of Rome,
(e-mail: monia.ranalli@uniroma1.it)

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

ABSTRACT: A new parsimonious model to cluster mixed-type of data is presented. Continuous and ordinal data are modeled by a mixture of Gaussians partially observed. To be parsimonious, it is used a reparameterization of the covariance matrices of the multivariate Gaussians. This permits to control for the number of parameters and simplifies the interpretation of the results.

KEYWORDS: mixture models, mixed-type data, EM algorithm, parsimonious modelling.

1 Introduction

To cluster mixed-type data, i.e. ordinal and continuous variables (Everitt, 1988 and Ranalli & Rocci, 2017a), ordinal variables are assumed to be a discretization of some latent continuous variables jointly distributed with the continuous ones as a Gaussian mixture model (Mclachlan & Peel, 2000). However, a large number of parameters have to be estimated, especially when covariance matrices change over components. Several authors have proposed parsimonious reparametrizations, mainly for continuous data. For example, some constrain the eigenvalues and/or the eigenvectors of the covariance matrices to be the same across the groups (Banfield, 1993), while others reduce the number of parameters by using a factor analysis model for each covariance matrix (McLachlan *et al.* , 2003). In the same context, we find proposals where mixtures of factor analyzers are used to obtain different parsimonious models (McNicholas & Murphy, 2008). A different approach has been developed for continuous and ordinal data (see by Kumar & Andreou, 1998 and Ranalli & Rocci, 2017b, respectively). They assume that there exist two within uncorrelated sets of factors that generate the variables as linear combinations, whose distributions have group specific parameters only for the first set. In this way the reduction is not only in the number of parameters but also in the dimensionality of the

data. In fact, the component variables of the second set, without class specific parameters, can be considered noise dimensions. In this framework, we propose a new parsimonious reparameterization based on the assumption that the variables are linear combinations of within uncorrelated latent variables where only some of them are characterized by class specific parameters. The material is organized as follows. In section 2, we present the model specification. In section 3, we outline the model parameter estimation. Finally some remarks and considerations are discussed in section 4. The EM-like algorithm and an example of application on real data showing the effectiveness of the proposal will be presented elsewhere for lack of space.

2 Model specification

Let $\mathbf{x} = [x_1, \ldots, x_O]'$ and $\mathbf{y}^{\bar{O}} = [y_{O+1}, \ldots, y_P]'$ be O ordinal and $\bar{O} = P - O$ continuous variables. The associated categories for each ordinal variable are denoted by $c_i = 1, 2, \ldots, C_i$ with $i = 1, 2, \ldots, O$. Following the Underlying Response Variable approach (Muthén, 1984), the ordinal variables $\mathbf{x}$ are considered as a categorization of a continuous multivariate latent variable $\mathbf{y}^O = [y_1, \ldots, y_O]'$. The latent relationship between $\mathbf{x}$ and $\mathbf{y}^O$ is explained by the threshold model, $x_i = c_i \Leftrightarrow \gamma^{(i)}_{c_i - 1} \leq y_i < \gamma^{(i)}_{c_i}$, where $-\infty = \gamma^{(i)}_0 < \gamma^{(i)}_1 < \ldots < \gamma^{(i)}_{C_i - 1} < \gamma^{(i)}_{C_i} = +\infty$ are the thresholds defining the C_i categories collected in a set $\boldsymbol{\Gamma}$ whose elements are given by the vectors $\boldsymbol{\gamma}^{(i)}$. We assume that $\mathbf{y} = [\mathbf{y}^{O\prime}, \mathbf{y}^{\bar{O}\prime}]'$ follows a heteroscedastic Gaussian mixture, $f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi_p \left(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$, where the p_g's are the mixing weights and $\phi_p \left(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$ is the density of a P-variate normal distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. All the parameters are contained in $\boldsymbol{\psi}$.
For a random i.i.d. sample of size N, $(\mathbf{x}_1, \mathbf{y}^{\bar{O}}_1), \ldots, (\mathbf{x}_N, \mathbf{y}^{\bar{O}}_N)$, the log-likelihood is

$$\ell(\boldsymbol{\psi}) \quad = \quad \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g \phi_{\bar{O}}(\mathbf{y}^{\bar{O}}_n; \boldsymbol{\mu}^{\bar{O}}_g, \boldsymbol{\Sigma}^{\bar{O}}_g) \pi_n \left(\boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \boldsymbol{\Sigma}^{O|\bar{O}}_g, \boldsymbol{\Gamma} \right) \right], \tag{1}$$

where, with obvious notation $\pi_n \left(\boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \boldsymbol{\Sigma}^{O|\bar{O}}_g, \boldsymbol{\Gamma} \right) = \int_{\gamma^{(1)}_{c_1 - 1}}^{\gamma^{(1)}_{c_1}} \cdots \int_{\gamma^{(O)}_{c_O - 1}}^{\gamma^{(O)}_{c_O}} \phi_O(\mathbf{u}; \boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \boldsymbol{\Sigma}^{O|\bar{O}}_g) d\mathbf{u}$

where, $\pi_n \left(\boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \boldsymbol{\Sigma}^{O|\bar{O}}_g, \boldsymbol{\gamma} \right)$ is the conditional joint probability of response pattern $\mathbf{x}_n = (c^{(1)}_1, \ldots, c^{(O)}_O)$ given the cluster g and the continuous variables $\mathbf{y}^{\bar{O}}_n$. Finally p_g is the probability of belonging to group g subject to $p_g > 0$ and $\sum_{g=1}^{G} p_g = 1$. We assume that in each class the P variables are linear combinations of the same P latent factors, which are uncorrelated and change, from one

cluster to another, only in the means and variances. In formulas, if observation n comes from the subpopulation g ($g = 1, \ldots, G$), then the following model holds

$$\mathbf{y}_n = \mathbf{B}(\boldsymbol{\eta}_g + \mathbf{L}_g^{1/2}\mathbf{f}_n) \tag{2}$$

where $\mathbf{B}$ is a full rank ($P \times P$) matrix of component loadings, $\mathbf{f}_n$ is a random vector of P latent variables normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_P$ and $\boldsymbol{\eta}_g$ and $\mathbf{L}_g$ are a column vector and a positive definite diagonal matrix, respectively. This model implies that in component g-th the P observed variables are linear combination of P latent factors having $\boldsymbol{\eta}_g$ and $\mathbf{L}_g$ as mean vector and covariance matrix, respectively. The density of $\mathbf{y}_n$, given that observation n comes from the g-th subpopulation, is multivariate normal with mean $\boldsymbol{\mu}_g = \mathbf{B}\boldsymbol{\eta}_g$ and covariance matrix $\boldsymbol{\Sigma}_g = \mathbf{B}\mathbf{L}_g\mathbf{B}'$, obtaining a reparameterization of the covariance matrices well-known in the multidimensional scaling literature under the name INDSCAL (Carroll & Chang, 1970).

3 Model Estimation

To overcome the presence of multidimensional integrals, here, the full log-likelihood is replaced by a composite likelihood (Lindsay, 1988) formed of $O(O-1)/2$ marginal distributions each of them composed of two ordinal variables and the $\bar{O}$ continuous variables. This leads to the following surrogate function

$$c\ell(\boldsymbol{\psi}) = \sum_{n=1}^{N}\sum_{i=1}^{O-1}\sum_{j=i+1}^{O}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j}\delta_{nc_ic_j}^{(ij)}\log\left[\sum_{g=1}^{G}p_g\pi_{c_ic_j}^{(ij|\bar{O})}(\boldsymbol{\mu}_g^{(ij|\bar{O})},\boldsymbol{\Sigma}_g^{(ij|\bar{O})},\boldsymbol{\Gamma}^{(ij)})\phi_{\bar{O}}(\mathbf{y}_n^{\bar{O}};\boldsymbol{\mu}_g^{\bar{O}},\boldsymbol{\Sigma}_g^{\bar{O}\bar{O}})\right],$$

where $\delta_{nc_ic_j}^{(ij)}$ is a dummy variable assuming 1 if the n-th observation presents the combination of categories c_i and c_j for variables x_i and x_j respectively, 0 otherwise; $\pi_{c_ic_j}^{(ij|\bar{O})}(\boldsymbol{\mu}_g^{(ij|\bar{O})},\boldsymbol{\Sigma}_g^{(ij|\bar{O})},\boldsymbol{\Gamma}^{(ij)})$ is the conditional probability of variables x_j and x_i of being in category c_i and c_j respectively, given all the continuous variables $\mathbf{y}^{\bar{Q}}$. The parameter estimates are carried out through an EM-like algorithm, that works in the same manner as the standard EM on a composite complete log-likelihood.

4 Discussion

In this work we propose a parsimonious version of mixture of Gaussian partially observed based on a specific decomposition of the covariance matrices. This is always true when $G = 2$ but does not necessarily hold for G $\geq$ 3. In

this case, if needed, we can circumvent the lack of fit by relaxing some of the constraints, for example by requiring a block diagonal form rather than simply diagonal for the matrices $\mathbf{L}_g$, or by assuming that the number of factors is greater than P, i.e. the number of manifest variables (assuming $\mathbf{B}$ to be rectangular).

References

BANFIELD, J. D .AND RAFTERY, A. E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.

CARROLL, J. D., & CHANG, J.-J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, **35**(3), 283–319.

EVERITT, B.S. 1988. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, **6**(5), 305–309.

KUMAR, N., & ANDREOU, A.G. 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, **26**(4), 283 – 297.

LINDSAY, B. 1988. Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

MCLACHLAN, G., & PEEL, D. 2000. *Finite Mixture Models*. 1 edn. Wiley Series in Probability and Statistics. Wiley-Interscience.

MCLACHLAN, G. J., PEEL, D., & BEAN, R.W. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, **41**(3-4), 379–388.

MCNICHOLAS, P. D., & MURPHY, T. B. 2008. Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

MUTHÉN, B. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**(1), 115–132.

RANALLI, M., & ROCCI, R. 2017a. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, **110**(C), 87–102.

RANALLI, M., & ROCCI, R. 2017b. A Model-Based Approach to Simultaneous Clustering and Dimensional Reduction of Ordinal Data. *Psychometrika*.

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

Società Italiana di Statistica

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.

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

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.