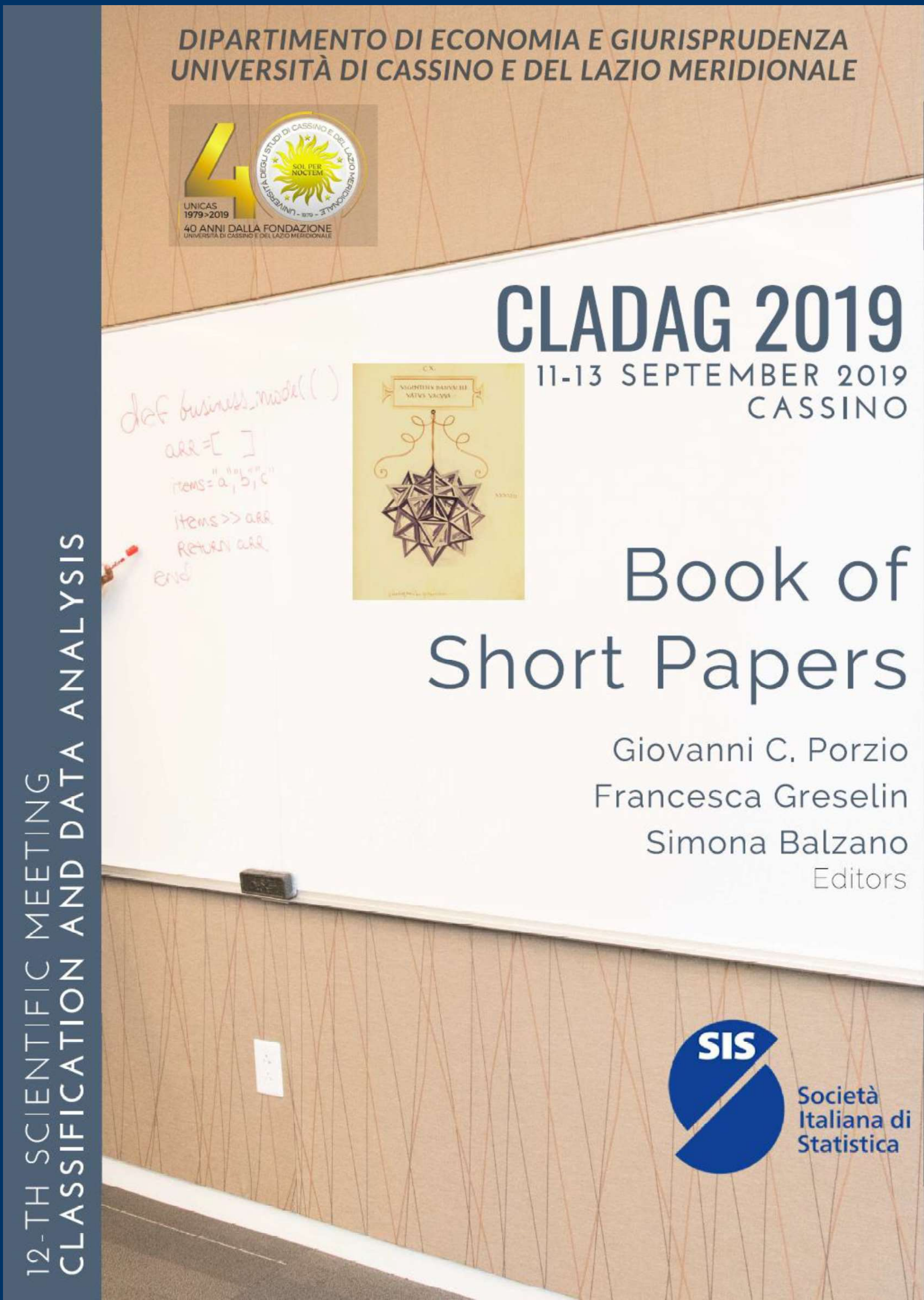
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA

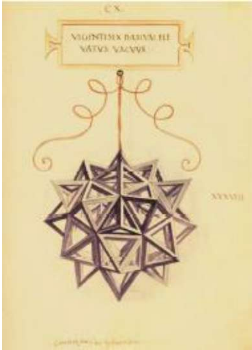


EUC
EDIZIONI UNIVERSITA' DI CASSINO

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

COMPOSITE LIKELIHOOD INFERENCE FOR SIMULTANEOUS CLUSTERING AND DIMENSIONALITY REDUCTION OF MIXED-TYPE LONGITUDINAL DATA

Antonello Maruotti[1, 2], Monia Ranalli[3] and Roberto Rocci[3, 4]

[1] Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne, Libera Università Maria Ss. Assunta, (e-mail: a.maruotti@lumsa.it)

[2] Department of Mathematics, University of Bergen,

[3] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, (e-mail: monia.ranalli@uniroma1.it)

[4] Dipartimento di Economia e Finanza, Università di Tor Vergata,

ABSTRACT: We introduce a multivariate hidden Markov model (HMM) for mixed-type (continuous and ordinal) variables. As some of the considered variables may not contribute to the clustering structure, we built a hidden Markov-based model such that we are able to recognize discriminative and noise dimensions. The variables are considered to be linear combinations of two independent sets of latent factors where one contains the information about the cluster structure, following an HMM, and the other one contains noise dimensions distributed as a multivariate normal (and it does not change over time). The resulting model is parsimonious, but its computational burden may be cumbersome. To overcome any computational issue, a composite likelihood approach is introduced to estimate model parameters.

KEYWORDS: mixed-type data, data reduction, HMM, composite likelihood, EM algorithm.

1 Introduction

In this work we focus our attention on longitudinal multivariate-mixed type data (continuous and ordinal variables). This means there are three major dependency structures: correlation between multivariate variables, temporal dependence and heterogeneity. Furthermore, to be realistic, we assume the presence of dimensions (named noise) that are uninformative for capturing the heterogeneity over time and could obscure the true data structure. To simplify, the aim of the proposal is to recover the cluster structure underlying the data that varies over time through some discriminative factors. Following the the Underlying Response Variable (URV) (see e.g. Jöreskog, 1990, Lee *et al.* ,

1990) approach, both the continuous and the categorical ordinal variables follow a Gaussian mixture model (Mclachlan & Peel, 2000), where the ordinal variables are only partially observed through their ordinal counterparts. To take into account the temporal dependence, we assume that the Gaussian mixture changes over time according to the realizations of an homogeneous first order Markov chain. In other words we are assuming a partially observed hidden Markov model (HMM). This extends the mixture model for mixed-type data (Everitt, 1988; Ranalli & Rocci, b 2017) over time. As regards the presence of noise variables, in literature there are approaches based on a family of mixture models which fits the data into a common discriminative subspace (see e.g. Bouveyron & Brunet, 2012; Kumar & Andreou, 1998; Ranalli & Rocci, 2017). The key idea is to assume a common latent subspace to all latent states that is the most discriminative. This allows to project the data into a lower dimensional space preserving the clustering characteristics over time, leading to a better and more parsimonious visualization and interpretation of the underlying structure of the data. The model can be formulated as a HMM with a particular set of constraints on the latent state parameters. The parameter estimates is based on a composite likelihood approach (Lindsay, 1988). The material is organized as follows. In section 2, we present the model specification. In section 3, we outline the model parameter estimation. 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 $\mathbf{x}_t = [x_1, \ldots, x_O]'$ and $\mathbf{y}_t^{\bar{O}} = [y_{O+1}, \ldots, y_P]'$ be O ordinal and $\bar{O} = P - O$ continuous variables, respectively, with $t = 1, \ldots, T$. 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 URV approach, the ordinal variables $\mathbf{x}$ are considered as a categorization of a continuous multivariate latent variable $\mathbf{y}_t^O = [y_1, \ldots, y_O]'$. We assume that the temporal evolution of these data is driven by a multinomial process in discrete time $\boldsymbol{\xi}_{1:T} = (\boldsymbol{\xi}_t, t = 1, \ldots, T)$, where $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tK})$ is a multinomial random variable with K classes. We specifically assume that such process is distributed as a homogeneous Markov chain, whose distribution, say $p(\boldsymbol{\xi}_{1:T}; \boldsymbol{p})$, is known up to a vector of parameters $\mathbf{p}$ that includes the initial probabilities and the transition probabilities of the chain. Conditionally on the value assumed each time by the Markov chain, the distribution of the data at time t depends on the specific component parameters of a partially observed multivariate normal. Formally, let define K initial probabilities as $p_k = P(\xi_{1k} = 1)$ with $\sum_{k=1}^{K} p_k = 1$ and K^2 transition probabilities as $p_{hk} = P(\xi_{tk} = 1 \mid \xi_{(t-1)h} = 1)$ with $h, k = 1, \ldots, K$ and $\sum_{h=1}^{K} p_{hk} = 1$. It follows that the Markov chain process is $p(\boldsymbol{\xi}_{1:T},\mathbf{p}) = \prod_{k=1}^{K} p_k^{\xi_{1k}} \prod_{t=1}^{T}\prod_{h=1}^{K}\prod_{k=1}^{K} p_{hk}^{\xi_{(t-1)h}\xi_{tk}}$. According to the URV, the joint distribution of $\mathbf{x}$ and $\mathbf{y}^{\bar{O}}$ can be constructed as follows. The latent relationship between $\mathbf{x}$ and $\mathbf{y}^{O}$ is explained by the threshold model, $x_i = c_i \Leftrightarrow \gamma_{c_i-1}^{(i)} \leq y_i < \gamma_{c_i}^{(i)}$, with $c_i = 1,\ldots,C_i$ and where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \ldots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ are the thresholds defining the C_i categories collected in a set $\boldsymbol{\Gamma}$ whose elements are given by the vectors $\boldsymbol{\gamma}^{(i)}$. To accommodate both cluster structure and dependence within the groups, we assume that the distribution $\mathbf{y}_t = [\mathbf{y}_t^{O\prime},\mathbf{y}_t^{\bar{O}\prime}]'$ given a particular point in time, say t and conditioning on $\boldsymbol{\xi}_t$, follows a partially observed multivariate normal, $f(\mathbf{y}_{nt} \mid \boldsymbol{\xi}_t) = \prod_{k=1}^{K} \phi_P(\mathbf{y}_{nt} \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{\xi_{nkt}}$, where the ξ_{nkt} is a Bernoulli variable that assumes value 1 if the $n-$th observation is classified in state k at time t, $\phi_P(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ is the density of a P-variate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Let us set $\boldsymbol{\psi} = \{\mathbf{p},\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K,\boldsymbol{\Gamma}\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space. For a random i.i.d. sample of size N, $(\mathbf{x}_1,\mathbf{y}_1^{\bar{Q}}),\ldots,(\mathbf{x}_N,\mathbf{y}_N^{\bar{Q}})$, the log-likelihood is

$$\ell(\boldsymbol{\psi}) = \sum_{n=1}^{N} \log\left[\sum_{\boldsymbol{\xi}_{1:T}} p(\boldsymbol{\xi}_t,\mathbf{p})\phi_{\bar{O}}(\mathbf{y}_{nt}^{\bar{O}} \mid \boldsymbol{\xi}_t,\boldsymbol{\mu}_k^{\bar{O}},\boldsymbol{\Sigma}_k^{\bar{O}})\pi_{nt}\left(\boldsymbol{\mu}_{nt;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}},\boldsymbol{\Gamma},\boldsymbol{\xi}_t\right)\right], \quad (1)$$

where, with obvious notation

$$\pi_{nt}\left(\boldsymbol{\mu}_{n;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}},\boldsymbol{\Gamma},\boldsymbol{\xi}_t,\right) = \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_O-1}^{(O)}}^{\gamma_{c_O}^{(O)}} \phi_O(\mathbf{u}_{nt};\boldsymbol{\mu}_{nt;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}})d\mathbf{u}_{nt},$$

where $\pi_n\left(\boldsymbol{\mu}_{nt;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}},\boldsymbol{\gamma}\right)$ is the conditional joint probability of response pattern $\mathbf{x}_{nt} = (c_1^{(1)},\ldots,c_O^{(O)})$ given the cluster k and the continuous variables $\mathbf{y}_{nt}^{\bar{O}}$. In order to identify the discriminative dimensions, it is assumed that there is a set of P latent factors $\tilde{\mathbf{y}}_t$, formed of two independent subsets.

In the first one, there are Q (with $Q \leq P$) factors that have some clustering information distributed as a mixture of Gaussians with class conditional means and variances equal to $E(\tilde{\mathbf{y}}^{Q} \mid k) = \boldsymbol{\eta}_k$ and $\mathrm{Cov}(\tilde{\mathbf{y}}^{Q} \mid k) = \boldsymbol{\Omega}_k$, respectively. In the second set there are $\bar{Q} = P - Q$ noise factors defining the so-called noise dimensions, that are independent of $\tilde{\mathbf{y}}^{Q}$ and their distribution does not vary from one class to another: $E(\tilde{\mathbf{y}}^{\bar{Q}} \mid k) = \boldsymbol{\eta}_0$ and $\mathrm{Cov}(\tilde{\mathbf{y}}^{\bar{Q}} \mid k) = \boldsymbol{\Omega}_0$. The link between $\tilde{\mathbf{y}}$ and $\mathbf{y}$ is given by a non-singular $P \times P$ matrix $\mathbf{A}$, as $\mathbf{y} = \mathbf{A}\tilde{\mathbf{y}}$. The final step is to identify the variables that could be considered as noise. Intuitively y_p is a noise variable if it is well explained by $\tilde{\mathbf{y}}^{\bar{Q}}$. Exploiting the independence between $\tilde{\mathbf{y}}^{Q}$ and $\tilde{\mathbf{y}}^{\bar{Q}}$, it is possible to compute proportions of each variable's variance that can be explained by the noise factors, and by one's complement, the proportions of each variable's variance that can be explained by the discriminative factors at each time point.

3 Construction of surrogate functions

The corresponding complete-data log likelihood involves multidimensional integrals that makes the maximum likelihood estimation computationally demanding and infeasible. To overcome this, we adopt a composite likelihood approach (Lindsay, 1988) based on $O(O-1)/2$ marginal distributions each of them composed of two ordinal variables and $\bar{O}$ continuous variables. The parameter estimates are carried out through an EM-like algorithm along with Baum-Welch recursion, that works in the same manner as the standard EM for HMMs.

References

BOUVEYRON, C., & BRUNET, C. 2012. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, **71**, 52–78.

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

JÖRESKOG, K. G. 1990. New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, **24**(4), 387–404.

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

LEE, S.-Y., POON, W.-Y., & BENTLER, P.M. 1990. Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters*, **9**(1), 91–97.

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.

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

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

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

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.