Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018
Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors
Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press

fedOA

Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018

Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors

Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press

fedOA Press

ASMOD 2018 : Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference : Naples, 24-26 October 2018 / editors Stefania Capecchi, Francesca Di Iorio, Rosaria Simone. – Napoli : FedOAPress, 2018. – (Scuola di Scienze Umane e Sociali. Quaderni ; 11).

Accesso alla versione elettronica:
http://www.fedoabooks.unina.it

ISBN: 978-88-6887-042-3
DOI: 10.6093/978-88-6887-042-3
ISSN Collana: 2499-4774

Università degli Studi di Napoli Federico II
Centro di Ateneo per le Biblioteche "Roberto Pettorino"
Piazza Bellini 59-60
80138 Napoli, Italy
http://www.fedoapress.unina.it/

Published in Italy

Contents

Foreword ix

Invited papers

A. Agresti, *Simple ordinal model effect measures* 1

B. Francis, *Latent class models for multiple ordinal items* 3

B. Grün, G. Malsiner-Walli, *Bayesian latent class analysis with shrinkage priors: an application to the Hungarian heart disease data* 13

M. Kateri, *Modelling Ordinal Data: A ϕ-divergence based approach* 25

E. Ronchetti, *Robust statistical analysis of ordinal data* 27

G. Tutz, *Uncertainty, dispersion and response styles in ordinal regression* 33

Accepted papers

A. Barbiero, *Inducing a desired value of correlation between two point-scale variables* 45

A. Bonanomi, M.N. Ruscone, S.A. Osmetti, *Dissimilarity measure for ranking data via mixture of copulae* 53

G. Bove, E. Nuzzo, A. Serafini, *Measurement of interrater agreement for the assessment of language proficiency* 61

E. Brentari, M. Manisera, P. Zuccolotto, *Modelling perceived variety in a choice process with nonlinear CUB* 69

R. Colombi, S. Giordano, *A flexible distribution to handle response styles when modelling rating scale data* 77

M. Corduas, *Joint modelling of ordinal data: a copula-based method* 85

C. Davino, T. Naes, R. Romano, D. Vistocco, *Modeling preferences: beyond the average effects* 93

C. Davino, R. Simone, D. Vistocco, *Exploring synergy between CUB models and quantile regression: a comparative analysis through continuousized data* 101

E. di Bella, L. Leporatti, F. Maggino, L. Gandullia, *A Poset based indicator of gender equality at sub-national level* 109

M. Fattore, A. Arcagni, *Using mutual ranking probabilities for dimensionality reduction and ranking extraction in multidimensional systems of ordinal variables* 117

S. Golia, M. Carpita, *On classifiers to predict soccer match results* 125

L. Grilli, M.F. Marino, O. Paccagnella, C. Rampichini, *Multiple imputation and selection of ordinal level-2 predictors in multilevel models* 133

M. Iannario, A.C. Monti, P. Scalera, *Why the number of response categories in rating scales should be large* 139

A. Lorenzo-Arribas, M.J. Brewer, A.M. Overstall, *Solutions to issues with partial proportional odds models* 147

F. Pennoni, M. Nakai, *A latent variable model for a derived ordinal response accounting for sampling weights, missing values and covariates* 155

F. Pesarin, L. Salmaso, H. Huang, R. Arboretti, R. Ceccato, *Permutation tests for stochastic ordering with ordinal data* 163

A. Plaia, M. Sciandra, S. Buscemi, *Consensus measures for preference rankings with ties: an approach based on position weighted Kemeny distance* 171

M. Ranalli, R. Rocci, *Simultaneous clustering and dimensional reduction of mixed-type data* 179

V. Sansivieri, M. Matteucci, S. Mignani, *Bi-Factor MIRT observed-score equating under the NEAT design for tests with several content areas* 187

R. Simone, F. Di Iorio, C. Cappelli, *On the choice of splitting rules for model-based trees for ordinal responses* 195

F. Torti, S. Salini, M. Riani, *Robustness issues for categorical data* 203

M. Vives-Mestres, J. Antoni Martín-Fernández, S. Thió-Henestrosa, R.S. Kenett, *Applications and theoretical results of association rules and compositional data analysis: a contingency table perspective* 211

Foreword

This volume contains the peer-reviewed contributions presented at the 2nd International Conference on Advances in Statistical Modelling of Ordinal Data - ASMOD 2018 - held at the Department of Political Sciences of the University of Naples Federico II, (24-26 October 2018). The Conference brought together theoretical and applied statisticians to share the latest studies and developments in the field. In addition to the fundamental topic of latent structure analysis and modelling, the contributions in this volume cover a broad range of topics including measuring dissimilarity, clustering, robustness, CUB models, multivariate models, and permutation tests.

The Conference featured six distinguished keynote speakers: Alan Agresti (University of Florida, USA), Brian Francis (Lancaster University, UK), Bettina Gruen (Johannes Kepler University Linz, Austria), Maria Kateri (RWTH Aachen, Germany), Elvezio Ronchetti (University of Geneva, Switzerland), Gerhard Tutz (Ludwig-Maximilians University of Munich, Germany) who significantly contributed to making the Conference successful with their inspiring presentations.

Moreover, the Conference encompassed 22 contributions that were accepted as full papers for inclusion in this edited volume after a blind review process of two anonymous referees.

I would also like to take this opportunity to express my gratitude to the members of the Scientific Committee: Eugenio Brentari (University of Brescia), Anna Clara Monti (University of Sannio), Monica Pratesi (University of Florence), Roberto Rocci (University of Rome Tor Vergata), and Stefania Capecchi, Carmela Cappelli, Francesca Di Iorio, Maria Iannario, Rosaria Simone from the University of Naples Federico II for their helpful support. I am also very grateful to the members of the Organizing Committee: Stefania Capecchi, Carlo De Luca, Cinzia Della Monica, Giuliana Perretti, Maria Giovanna Porzio, Caterina Rinaldi, Filomenda Vilardi that contributed to the success of ASMOD 2018 and worked actively for its organization.

Finally, I wish to acknowledge the sponsorship of the Italian Statistical Society, the CLADAG (Classification and Data Analysis) Group, and the financial support of the Department of Political Sciences and the University of Naples Federico II.

Marcella Corduas
Chair of the Scientific Committee

INVITED PAPERS

Simple ordinal model effect measures

Alan Agresti[*]

Abstract: The survey effect measures for models for ordinal categorical data that can be simpler to interpret that the model parameters. For describing the effect of an explanatory variable while adjusting for other explanatory variables, we present probability-based measures, including a measure of relative size and partial effect measures based on an instantaneous rate of change. We also survey summary measures of predictive power that are analogs of R-squared and multiple correlation measures for continuous response variables. We suggest new measures of effect and of predictive power, illustrate the new and existing measures for an example, and provide R code for implementing them. The talk is based on recent papers with Claudia Tarantola and Maria Kateri.

Keywords: Average marginal effect, Ordinal probability comparison, R-squared.

[*] University of Florida, aa@stat.ufl.edu

Latent class approaches for modelling multiple ordinal items

Brian Francis[*]

Abstract: The modelling of the latent class structure of multiple Likert items is reviewd. The standard latent class approach is to model the absolute Likert ratings. Commonly, and ordinal latent class model is used where the logits of the profile probabilities for each item have an adjacent category formulation (DeSantis et al., 2008). an alternative developed in this paper is to model the relative orderings, using a mixture model of the *relative* differences between pairs of Likert items. This produces a paired comparison adjacent category log-linear model (Dittrich et al., 2007; Francis and Dittrich, 2017), with item estimates placed on a (0,1) "worth" scale for each latent class. The two approaches are compared using data on environmental risk from the International Social Survey Programme, and conclusions are presented.

Keywords: Multiple likert items, Ordinal latent class models, Paired comparisons.

1. Introduction

Collections of multiple Likert items in questionnaires are very common, and are usually used to measure underlying constructs. Scale from the Likert items can be built either through simply adding the item score or through using an IRT model such as a graded response model to build a score. This approach assumes that there is a single underlying construct to the items. The current paper, in contrast, takes a different view. It proposes that there is a latent class structure to the Likert items, with different classes having different patterns of high and low responses. In this approach, score building is not the aim; instead the aim is to understand the various patterns of responses that might exist in the population.

The standard latent class approach to multiple ordinal indicators essentially constructs a polytomous latent class model (Linzer and Lewis, 2011),

[*]University of Lancaster, UK, B.Francis@Lancaster.ac.uk

and constrains the latent class profile probabilities, imposing a linear score ordinal model on them (Magidson and Vermunt, 2004; DeSantis et al., 2008). This results in a latent class adjacent category ordinal model. The method however uses the *absolute* responses, and this has been criticised by some authors, as they state that each respondent has their own way of interpreting the Likert scale. Such interpretation may itself be culturally determined, or may depend on other covariates such as age, gender and so on. For example younger people and males may be more likely to express a firm opinion, using the end categories of a unipolar Likert scale, than older people and females. The alternative is to take a relative approach. While one method of doing this is to standardise the items for each respondent, subtracting the respondent mean. This is unsatisfactory as it ignores the categorical nature of the data. In this paper we instead develop a paired comparisons approach, which produces a worth scale for each latent class, ranking the items in order of preference. The paper compares the two methods and discusses the advantages and disadvantages of each method.

Some common notation is introduced which will be used to develop both models. The Likert items are assumed to be measured on the same response scale with identical labelling; it is assumed that there are H possible ordered response categories taking the values $1, \dots, H$ for each of the J Likert items indexed by j, and with N respondents indexed by i. $y_{ij}; \quad y_{ij} \in 1, 2, \dots, H$ is defined to be the (ordinal) response given by respondent i to item j. A set of H indicators for each item and respondent with the indicator z_{ijh} taking the value 1 if $y_{ij} = h$ and 0 otherwise.

2. The ordinal latent class model

We first introduce the ordinal latent class model, which models the absolute responses. Let y_{ij} be the ordinal response of respondent i to item j. It is assumed that there are K latent classes. The item response vector for respondent i is

$$\mathbf{y_i} = (y_{i1}, y_{i2}, \dots, y_{iJ}),$$

Then the ordinal latent class model is defined by:

$$\begin{aligned} P(\mathbf{y_i}) &= \sum_{k=1}^{K} \pi(k) P(\mathbf{y_i}|\mathbf{k}) \\ &= \sum_{k=1}^{K} \pi(k) \prod_{j} P(y_{ij}|k) \qquad \text{under conditional independence.} \end{aligned}$$

We write

$$P(y_{ij}|k) = \prod_{h=1}^{H} p_{jkh}^{z_{ijh}}$$

where p_{jkh} is the probability of observing the ordinal response h for indicator j given membership of latent class k - these are sometimes called the latent class profile probabilities.

Ordinality is imposed by using an adjacent categories ordinal model and we parameterise the model through regression parameters on the logit scale, which separates out the intercept parameter β_{jh} and the class specific parameters β_{jkh} for each item and response category.

$$\begin{aligned} \text{logit}(p_{jkh}) &= \beta_{jh} + \beta_{jkh} \\ &= \beta_{jh} + h\beta_{jk} \qquad \text{under a linear score model} \end{aligned}$$

The likelihood L is then given by

$$L = \prod_{i} \sum_{k=1}^{K} \pi(k) P(\mathbf{y_i}|\mathbf{k}).$$

Model fitting is usually carried out by using the EM algorithm - details are given in Francis et al. (2010) and Aitkin et al. (2014). Determination of the optimal number of classes is commonly achieved by choosing that model which minimises an information criterion, although a wide variety of other methods have been proposed. We have used the BIC in this paper.

3. The latent class ordinal paired comparison model

An alternative to the absolute latent class approach is to work on a *relative scale*. This perhaps is of greater interest. We take a paired comparison approach, using the difference in the ordinal likert responses. This allows the development of a "worth" scale between 0 and 1 with items placed on this scale. The sum of the item scores is defined to be 1. This section proceeds by developing the ordinal paired comparison model, and then extends that model by adding a mixture or latent class process to the model.

3.1. The ordinal paired comparison model

This model starts by constructing a set of paired comparisons - taking all possible pairs of items and comparing them in turn (Dittrich et al., 2007). For respondent i and for any two items $j = a$ and $j = b$, let

$$x_{i,(ab)} = \begin{cases} h & \text{if item } a \text{ preferred by } h \text{ steps to item } b & = y_{ia} - y_{ib} \\ 0 & \text{if Likert ratings are equal} & = 0 \\ -h & \text{if item } b \text{ preferred by } h \text{ steps to item } a & = y_{ia} - y_{ib} \end{cases}$$

The probability for a single PC response $x_{i,(ab)}$ is then defined by

$$p(x_{i,(ab)}) = \begin{cases} \mu_{ab} \left(\frac{\pi_a}{\pi_b}\right)^{x_{i,(ab)}} & : \text{ if } x_{i,(ab)} \neq 0 \\ \mu_{ab}\, c_{ab} & : \text{ if } x_{i,(ab)} = 0 \end{cases}$$

The πs represent the worths or importances of the items, c_{ab} represents the probability of no preference between items a and b and μ_{ab} is a normalising quantity for the comparison ab. Over all items, we now form a pattern vector $\boldsymbol{x}_i$ for observation i with $\boldsymbol{x}_i = (x_{i,(12)}, x_{i,(13)}, \ldots, x_{i,(ab)}, \ldots, x_{i,(J-1,J)})$ and count up the number of responses n_ℓ with that pattern. The *probability* for a certain pattern ℓ is

$$p_\ell = \triangle^* \prod_{a<b} p(x_{ab})$$

where $\triangle^*$ is a constant (the same for all patterns).

A log-linear model can now be constructed with observed counts n_ℓ. The *expected counts for a pattern* ℓ are defined as $\boxed{m_\ell = n\, p_\ell}$ where n is the total number of respondents defined by $n = n_1 + n_2 + \cdots + n_\ell + \cdots + n_L$ and where L is the number of all possible patterns.

Taking natural logs, the *log expected counts* are obtained by

$$\ln m_\ell = \alpha + \sum_{a<b} x_{ab}(\lambda_a - \lambda_b) + \mathbf{1}_{x_{ab}=0}\, \gamma_{ab}$$

For $x_{ab} = \quad h$ this is $\quad h(\lambda_a - \lambda_b)$, for $x_{ab} = -h$ this is $h(-\lambda_a + \lambda_b)$ and for $x_{ab} = \quad 0$ this is γ_{ab}. To show that this is an adjacent categories model, the log odds of a pair for any two adjacent categories on the ordinal scale can be examined - say h and $h+1$. Then, as $m_\ell = np_\ell$, we have

$$\begin{aligned} \ln\left(\frac{m_\ell(h)}{m_\ell(h+1)}\right) &= \ln(\mu_{ab}) + h(\lambda_a - \lambda_b) - \ln(\mu_{ab}) - (h+1)(\lambda_a - \lambda_b) \\ &= \lambda_a - \lambda_b \end{aligned}$$

which is true for any h as long as h or $h+1$ are not zero.

The worths π_j are calculated from the λ_j through the formula

$$\pi_j = \frac{\exp(2\lambda_j)}{\sum_{j=1}^{J} \exp(2\lambda_j)}.$$

3.2. Extending the model to incorporate latent classes

As before, we assume that there are K latent classes with different preference patterns (the lambdas). The likelihood L becomes:

$$\mathrm{L} = \prod_\ell \Big(\sum_{k=1}^{K} q_k\; n\; p_{\ell k}\Big) \quad \text{where} \sum_\ell p_{\ell k} = 1 \quad \forall\; k \;\text{ and } \sum_k q_k = 1.$$

$$\ln p_{\ell k} = \alpha + \sum_{a<b} x_{ab}(\lambda_{ak} - \lambda_{bk}) + \mathbf{1}_{x_{ab}=0}\, \gamma_{ab}$$

λ_j is replaced in the model byλ_{jk}, and we now have to additionally estimate the q_k. q_k is the probability of belonging to class k (the mass points or class sizes). Again, we use the EM algorithm to maximise the likelihood, and use the BIC to determine the number of classes. Typically, we need to use a range of starting values to ensure an optimal solution.

4. An Example

Six question items on the topic of environmental danger were taken from the 2000 sweep of the International Social Survey Programme , which focused on issues relating to the environment. As part of this survey, the respondents assessed the environmental danger of a number of different activities and items. The question is reproduced below; each question used the same response scale. The six Likert items are:

c air pollution caused by **c**ars (`CAR`)

t a rise in the world's **t**emperature (`TEMP`)

g modifying the **g**enes of certain crops (`GENE`)

i pollution caused by **i**ndustry (`IND`)

f pesticides and chemicals used in **f**arming (`FARM`)

w pollution of **w**ater (rivers, lakes, . . .) (`WATER`)

with the response scale for each of the items as follows:

In general, do you think *item* is

4. extremely dangerous for the environment
3. very dangerous
2. somewhat dangerous
1. not very dangerous
0. not dangerous at all for the environment

Both absolute and relative latent class models are fitted to this data. The standard ordinal latent class model (absolute) was fitted using Latent Gold 5.1 (Vermunt and Magidson, 2013), and the paired comparison ordinal latent class model (relative) was fitted using an extension to the `prefmod` package in R (Hatzinger and Maier, 2017). Both approaches used 20 different starting

Table 1. BIC values from fitting latent class models (a) the standard ordinal LC model and (b) the ordinal PC LC model

	(a) standard ordinal LC model absolute		(b) Ordinal PC LC model relative	
No. of classes K	BIC	no of parameters	BIC	no of parameters
1	24207.04	24		
2	22680.48	31	6823.11	26
3	22153.75	38	6359.56	32
4	22112.70	45	6204.76	38
5	22097.07	52	6303.71	44
6	22084.99	59		
7	22083.33	66		

values to ensure that the global maximum of the likelihood was reached. Table 1 shows the BIC values for both models, for a range of values of K. It can be seen that the standard latent class approach needs either six or seven classes (six classes is chosen here), whereas the paired comparison latent class model gives a minimum BIC for $K = 4$. The smaller number of classes found for the paired comparison approach is perhaps to be expected, as the standard approach needs to model both the absolute level of the Likert responses as well as the differences.

We examine the mean Likert rating for each of the items within each of the latent classes for the standard ordinal latent class model. In contrast, the worths provide the interpretation of the latent classes in the paired comparison LC model. Both plots are shown in Figure 1, which are oriented so that greater dangerousness (or greater danger worth) is towards the top of the plots.

It can be seen that for the standard ordinal latent class model, the first three classes - Class 1 (51%), Class 2 (24%) and Class 3 (12%) - all show little difference between the items, but differ according to their absolute level. The three remaining classes, in contrast, show considerable differences between the items. The paired comparison solution gives a similar story. The largest class shows little difference between the items, with the three remaining classes showing large differences in dangerousness between items. Al-

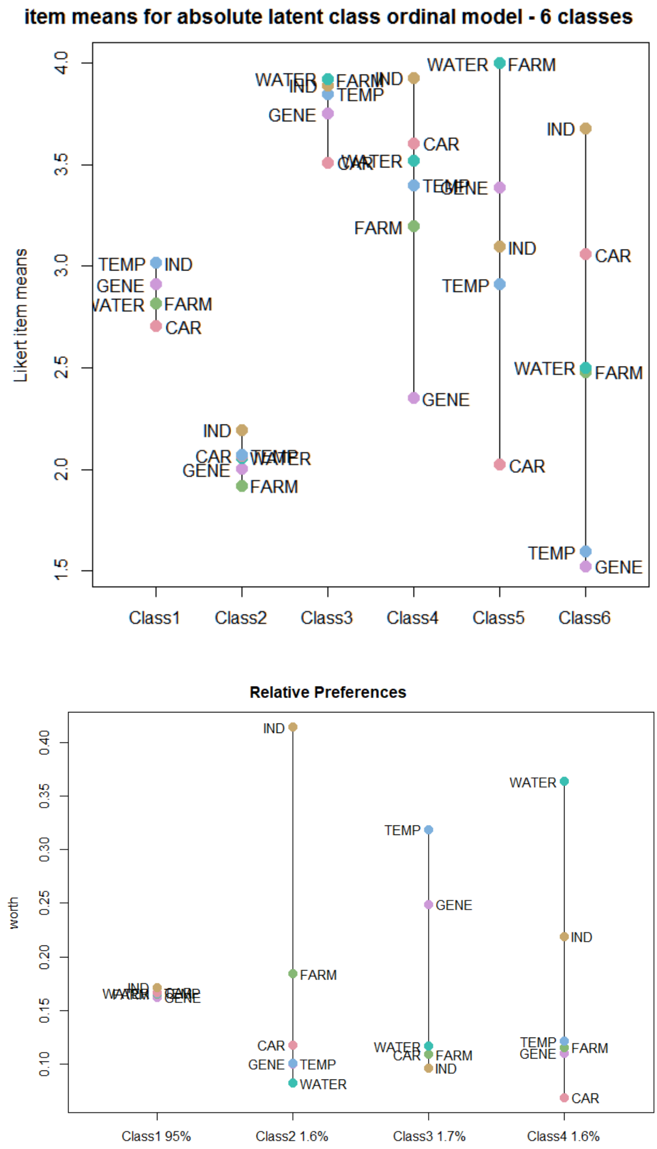


Figure 1. Item worths for (top) standard ordinal LC model and (bottom) ordinal paired comparison LC model

though the item rankings show some minor differences between the two methods, the results are similar.

5. Discussion and conclusions

This paper has demonstrated that the paired comparison ordinal model can be useful to understand the relative ordering of items in multiple Likert responses when the absolute level of the response is not of interest. The method leads to simpler models, which makes interpretation simpler. There are however some restrictions in using the model. The most important is that all Likert items must be measured on the same response scale. Differences between Likert items only make sense when this is true, and the paired comparison method relies on that. The PC method as currently implemented also assumes equidistance between the Likert categories, and further work is needed to relax this assumption.

Acknowledgements: I am grateful to Regina Dittrich and Reinhold Hatzinger for very helpful discussions on this topic.

References

Aitkin M., Vu D., Francis B. (2014) Statistical modelling of the group structure of social networks, *Social Networks*, 38, 74-87.

DeSantis S.M., Houseman E.A., Coull B.A., Stemmer-Rachamimov A., Betensky R.A. (2008) A penalized latent class model for ordinal data, *Biostatistics*, 9, 249-262.

Dittrich R., Francis B., Hatzinger R., Katzenbeisser W. (2007) A PairedASMOD 2018 Comparison Approach for the Analysis of Sets of Likert Scale Responses, *Statistical Modelling*, 7, 3-28.

Francis B., Dittrich R. (2017) Modelling multiple likert items through an adjacent categories ordinal paired comparison model. Presented at the 10th ERCIM comference on computational and Methodological Statistics, London; 16-18 December 2017.

Francis B., Dittrich R., Hatzinger R. (2010) Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do europeans get their scientific knowledge?, *The Annals of Applied Statistics*, 4, 2181-2202.

Hatzinger R., Maier M.J. (2017) prefmod: Utilities to Fit Paired Comparison Models for Preferences. R package version 0.8-34.

Linzer D., Lewis J. (2011) poLCA: An R Package for Polytomous Variable Latent Class Analysis, *Journal of Statistical Software*, 42.

Magidson J., Vermunt J. (2004) Latent class analysis. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Chapter 10, 175-198. Thousand Oaks, CA.: Sage Publications.

Vermunt J., Magidson J. (2013) Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc.

Bayesian latent class analysis with shrinkage priors: an application to the Hungarian heart disease data

Bettina Grün*, Gertraud Malsiner-Walli**

Abstract: Latent class analysis explains dependency structures in multivariate categorical data by assuming the presence of latent classes. We investigate the specification of suitable priors for the Bayesian latent class model to determine the number of classes and perform variable selection. Estimation is possible using standard tools implementing general purpose Markov chain Monte Carlo sampling techniques such as the software JAGS. However, class specific inference requires suitable post-processing in order to eliminate label switching. The proposed Bayesian specification and analysis method is applied to the Hungarian heart disease data set to determine the number of classes and identify relevant variables and results are compared to those obtained with the standard prior for the component specific parameters.

Keywords: Bayesian latent class analysis, Shrinkage prior, Variable selection.

1. Introduction

Latent class analysis (LCA) is a modeling approach for categorical data originally proposed by Lazarsfeld (1950). The observed association between the manifest categorical variables is assumed to be caused by latent classes. Conditional on class membership the categorical variables are assumed to be independent given the class specific variable distributions.

Issues in LCA are the selection of the number of classes and the identification of relevant variables. Within the frequentist framework using maximum likelihood estimation Dean and Raftery (2010) investigated the use of the BIC in combination with a headlong search algorithm to explore the model space to determine a suitable number of classes as well as subset of variables. They illustrate their approach using the Hungarian heart disease data set. Alternatively, White et al. (2016) use stochastic search methods to select the number of classes and relevant variables within the Bayesian framework.

*Johannes Kepler Universität Linz, bettina.gruen@jku.at

**Wirtschaftsuniversität Wien, gertraud.malsiner-walli@wu.ac.at

In this paper we investigate the use of sparse finite mixture models in combination with shrinkage priors. Malsiner-Walli et al. (2016) proposed the sparse finite mixture model with shrinkage priors on the means for the Gaussian finite mixture model. We extend this approach to the Bayesian latent class model. We also indicate how a general purpose Markov chain Monte Carlo (MCMC) sampler such as JAGS (Just Another Gibbs Sampler; Plummer 2003) can be used to obtain draws from the posterior and present suitable post-processing tools of the MCMC draws to eliminate label switching. This proposed model specification and analysis strategy is used to reanalyze the Hungarian heart disease data set.

2. Bayesian latent class model

Assume there are n observations $\boldsymbol{y}_i$, $i = 1, \dots, n$ given. Each observation $\boldsymbol{y}_i$ is a vector of length J, i.e., J variables are observed and each element y_{ij} contains values in $\{1, \dots, L_j\}$ implying that each variable j is a categorical variable with $L_j \geq 2$ different values.

The latent class model for observations $\boldsymbol{y}_i$, $i = 1, \dots .n$ is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{J} \prod_{l=1}^{L_j} \theta_{k,jl}^{\mathbb{1}(y_{ij}=l)} \right],$$

where $\boldsymbol{\pi} = (\pi_k)_{k=1,\dots,K}$, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_{k,jl})_{k=1,\dots,K;j=1,\dots,J;l=1,\dots,L_j}$, $\mathbb{1}()$ is the indicator function, and

$$\sum_{k=1}^{K} \pi_k = 1, \qquad \pi_k \geq 0, \forall k,$$

$$\sum_{l=1}^{L_j} \theta_{k,jl} = 1, \forall k, j, \qquad \theta_{k,jl} > 0, \forall k, j, l.$$

2.1. Prior specification

The parameter vector consists of $(\boldsymbol{\pi}, \boldsymbol{\Theta})$. In Bayesian finite mixture modeling one assumes in general that the component weights $\boldsymbol{\pi}$ and the component specific parameters $\mathbf{\Theta}$ are a-priori independent and that the component specific parameters are independently identically distributed (at least conditional on some hyperparameters). Furthermore conditionally conjugate priors are used to simplify MCMC sampling.

Component weights

For the component weights $\boldsymbol{\pi}$ a Dirichlet prior is assumed with a single parameter e_0:

$$\boldsymbol{\pi} \sim \text{Dirichlet}(e_0, \dots, e_0).$$

Rousseau and Mengersen (2011) show that e_0 is an influential parameter if an overfitted mixture model is estimated. Based on their results Malsiner-Walli et al. (2016) propose the sparse finite mixture model where an overfitting mixture with K, the number of components, much larger than the number of latent classes is fitted together with the specification of a very small and fixed value for e_0, e.g., $e_0 = 0.0001$. Under this prior setting the posterior of an overfitting mixture asymptotically concentrates on the region of the parameter space where superfluous components have negligible component weights instead of including duplicated components.

Standard prior for the component specific parameters

In Bayesian LCA one assumes that a-priori the parameters of the variables are independent within components. This implies that for each variable j and component k the component specific parameter vector $\boldsymbol{\theta}_{k,j\cdot}$ a-priori follows a Dirichlet distribution:

$$\boldsymbol{\theta}_{k,j\cdot} \sim \text{Dirichlet}(\boldsymbol{a}_j).$$

The value for $\boldsymbol{a}_j$ is selected to regularize the likelihood which in the case of an LCA model is often multi-modal, contains spurious modes and might have modes at the boundary of the parameter space.

Shrinkage prior for the component specific parameters

To shrink irrelevant variables towards a common Dirichlet parameter a hierarchical prior is specified on $\boldsymbol{a}_j$. For this purpose the Dirichlet parameter is re-parameterized into a mean and precision parameter plus a regularizing additive constant:

$$\begin{aligned} \boldsymbol{a}_j &= \boldsymbol{a}_{0,j} + \phi_j \boldsymbol{\mu}_j, & \boldsymbol{\mu}_j &\sim \text{Dirichlet}(\boldsymbol{m}_j),\ \forall j, \\ \phi_j &= \frac{1}{\lambda_j},\ \forall j, & \lambda_j &\sim \text{Gamma}(\nu_1, \nu_2),\ \forall j. \end{aligned}$$

Following Malsiner-Walli et al. (2016) we suggest to use $\nu_1 = \nu_2 = 0.5$. Furthermore we use uniform priors for $\boldsymbol{a}_{0,j}$ and $\boldsymbol{\mu}_j$, i.e., $\boldsymbol{a}_{0,j} = 1$ and $\boldsymbol{m}_j = \mathbf{1}$.

2.2. MCMC estimation

Estimation of the Bayesian latent class model consists of approximating the posterior distribution of $(\boldsymbol{\pi}, \boldsymbol{\Theta})$ using MCMC methods. Diebolt and Robert (1994) suggested to use data augmentation to facilitate MCMC estimation by adding the class memberships of the observations to the sampling scheme.

Standard prior for the component specific parameters

The sampling scheme is given by:

1. Draw the class memberships S_i for all observations $i = 1, \ldots, n$:

$$S_i \sim \text{Multinomial}(1, \boldsymbol{p}_i), \qquad p_{ik} \propto \pi_k \prod_{j=1}^{J} \prod_{l=1}^{L_j} \theta_{k,jl}^{\mathbb{1}(y_{ij}=l)}.$$

2. Conditional on $\boldsymbol{S} = (S_i)_{i=1,\ldots,n}$ draw $\boldsymbol{\pi}$ from a Dirichlet distribution:

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}(e_0 + n_1, \ldots, e_0 + n_K), \\ n_k &= \sum_{i=1}^{n} \mathbb{1}(S_i = k) \quad \forall k = 1, \ldots, K. \end{aligned}$$

3. Conditional on $\boldsymbol{S} = (S_i)_{i=1,\dots,n}$ draw $\boldsymbol{\theta}_{k,j\cdot}$ from a Dirichlet distribution:

$$\boldsymbol{\theta}_{k,j\cdot} \sim \text{Dirichlet}(a_{j1} + n_{k,j1}, \dots, a_{jL_j} + n_{k,jL_j}),$$
$$n_{k,jl} = \sum_{i=1}^{n} \mathbb{1}(S_i = k)\mathbb{1}(y_{ij} = l) \quad \forall k, j, l.$$

In each MCMC iteration the class memberships $\boldsymbol{S}$ induce a partition of the observations into K_+ classes, i.e., the number of non-empty components for this draw. In the overfitting mixture setting with K much larger than the number of classes and e_0 very small $K_+ \ll K$ and the posterior distribution of K_+ can be used to estimate the number of classes. Malsiner-Walli et al. (2016) proposed to use the mode as suitable point estimate.

Shrinkage prior for the component specific parameters

An additional sampling step is required to sample the hyperparameter values:

4. Conditional on $\boldsymbol{\Theta}$, sample $\boldsymbol{\mu}_j$ and λ_j for all j.

Model specification in BUGS and estimation using JAGS

The BUGS (Bayesian inference Using Gibbs Sampling; Lunn et al. 2009) model description language allows the specification of a Bayesian model based on a directed acyclic graph which contains the data as well as all parameters as nodes and where the edges are implied by the hierarchical specification of the Bayesian model.

For a Bayesian finite mixture models which is estimated using data augmentation the model specification not only includes the data $\boldsymbol{y}$ and the parameters $(\boldsymbol{\pi}, \boldsymbol{\Theta})$ but also the class memberships $\boldsymbol{S}$. The BUGS model specification for the model including the shrinkage prior is given in Figure 1. Note that for the standard prior the parameter `a[j, 1:L[j]]` is fixed and the four lines of code defining the relationships for `a`, `mu`, `phi` and `lambda` are dropped.

The model is estimated within R using package **rjags**. Only a list containing the data in an array `Y`, the dimensions `n`, `J`, `L` and the parameters needs

```
model {
  for (i in 1:n) {
    for (j in 1:J) {
      Y[i, 1:L[j], j] ~ dmulti(theta[S[i], j, 1:L[j]], 1)
    }
    S[i] ~ dcat(pi[1:K])
  }
  for (j in 1:J) {
    for (k in 1:K) {
      theta[k, j, 1:L[j]] ~ ddirch(a[j, 1:L[j]])
    }
    a[j, 1:L[j]] <- a0[1:L[j]] + phi[j] * mu[j, 1:L[j]]
    mu[j, 1:L[j]] ~ ddirch(m[1:L[j]])
    phi[j] <- 1 / lambda[j]
    lambda[j] ~ dgamma(nu1, nu2)
  }
  pi[1:K] ~ ddirch(e0[1:K]);
}
```

Figure 1. BUGS model specification for the sparse latent class model with shrinkage priors.

to be specified. Note that Y needs to be given as an array of dimension n $\times$ $\max(L_j)$ $\times$ J containing zeros and ones to indicate the observed values. n corresponds to the number of observations, J to the number of variables and L is a vector containing the number of categories for each variable. In addition the parameters specified are the number of components K and a vector e of length K containing e_0. Furthermore, for the standard prior a is a vector of ones of length $\max(L_j)$, whereas for the shrinkage prior, m and a0 are two vectors of ones of length $\max(L_j)$, and nu1 and nu2, the parameters of the Gamma prior on the shrinkage parameter λ, are both set equal to 0.5.

Then the model is defined using `jags.model()` and samples are drawn using `jags.samples()` while monitoring the parameters of interest using the argument `variable.names`.

For the presented results the call to `jags.model()` included an `inits` argument to set a specific random seed for reproducibility and an `n.adapt` argument to increase the number of iterations for adaptation to 5,000. Then `jags.samples` is called using 100,000 number of iterations with a thinning of 10.

2.3. Post-processing

The number of filled components K_+ are determined for each draw and an estimate $\hat{K}_+$ is obtained using the mode of the posterior distribution. If there is a distinct class structure in the data the MCMC sampler usually converges quickly to this number of classes and a clear mode can be identified (see Malsiner-Walli et al. 2016).

Conditional on the number of classes selected the draws are post-processed in the following way to obtain an identified model with suitable class specific parameter estimates as well as class assignments of the observations.

1. Discard all draws where $K_+ \neq \hat{K}_+$.

2. Discard all parameter draws $\boldsymbol{\theta}_{k,\cdot,\cdot}$ for empty components.

3. For each draw relabel the components to minimize the misclassification rate between the class assignments of this draw and the class assignments of the last draw.

Note that this is a very simple strategy to obtain an identified model which will only work if the data has a clear class structure. More elaborate approaches to deal with label switching have been proposed and might be required in more complicated settings to obtain good results (see Papastamoulis 2016).

3. Analyzing the Hungarian heart disease data

The Hungarian heart disease data consists of 284 patients on 5 categorical variables. For more details on the categorical variables with their levels see Table 1. Dean and Raftery (2010) analyzed this data set with LCA. They used maximum likelihood estimation in combination with the BIC to perform a joint approach for variable selection and determining the number of classes. They compared the classification results obtained with LCA to the known diagnosis of heart disease (angiographic disease status) available in the data set. The known diagnosis has two categories: "$< 50\%$" indicating less than 50%

Variable	Level	Class 1	Class 2
Chest pain type	Typical Angina	0.06 (0.02)	0.01 (0.01)
	Atypical Angina	0.57 (0.06)	0.07 (0.04)
	Non-anginal pain	0.26 (0.04)	0.08 (0.04)
	Asymptomatic	0.10 (0.07)	0.84 (0.06)
Exercise induced Angina	No	0.95 (0.03)	0.33 (0.11)
	Yes	0.05 (0.03)	0.67 (0.11)
Gender	Female	0.36 (0.04)	0.15 (0.04)
	Male	0.64 (0.04)	0.85 (0.04)
Resting Electrocardiographic results	Normal	0.81 (0.03)	0.77 (0.04)
	ST-T wave	0.15 (0.03)	0.21 (0.04)
	Estes' criteria	0.04 (0.02)	0.02 (0.01)
Fasting blood sugar $>$120 mg/dl	False	0.94 (0.02)	0.90 (0.03)
	True	0.06 (0.02)	0.10 (0.03)

Table 1. Posterior mean (and posterior standard deviations) of the class specific parameters for the identified 2-class sparse LCA model.

diameter narrowing and "> 50" indicating more than 50% diameter narrowing in any major vessel.

3.1. Sparse finite mixture model

An overfitting mixture model is estimated using $e_0 = 0.0001$ and $K = 10$. In addition a uniform prior is assumed for the class specific parameters, i.e., $a_{k,jl} = 1$. The posterior distribution of the number of non-empty components K_+ has a clear mode at 2 with 99.7% of the samples having 2 non-empty components. The remaining samples had 3 non-empty components (0.2%). Using the samples with 2 non-empty components to identify the model results in a posterior mean estimate for the component weight of the larger class π_1 of 0.579 with a posterior standard deviation of 0.075.

The class specific parameters for the categorical variables are given in Table 1. These results can be compared to those in Dean and Raftery (2010) who reported the maximum likelihood estimates for the parameters of a two-class latent class model. The posterior mean and the maximum likelihood estimates are similar. However, the Bayesian approach also provides uncer-

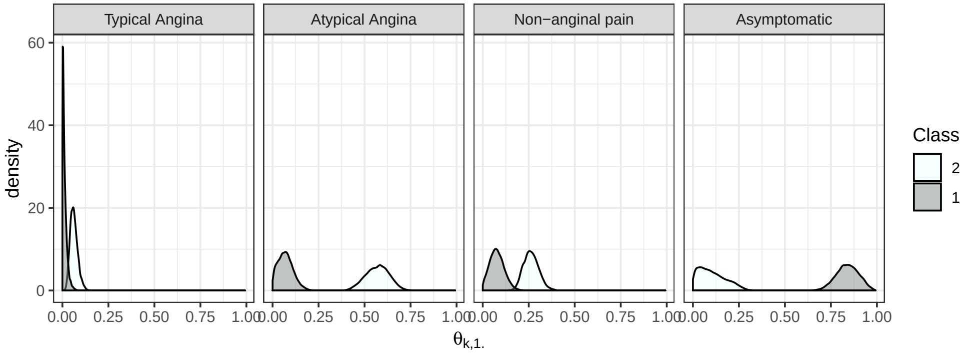


Figure 2. Posterior distribution of the class specific parameters for the variable "Chest pain type".

tainty estimates as given by the posterior standard deviations and the full posterior distributions which are visualized in Figure 2 for the variable "Chest pain type". In particular for parameter values which are estimated to be close to the boundary the posterior is non-normal and the full posterior allows to estimate suitable credible intervals for these parameters.

Observations can also be classified to the class they are most often assigned to during MCMC sampling after model identification. This partition is compared to the clinical partition contained in the data (see Table 3 on the left). The congruence between these two partitions is very high and results are similar to those reported in Dean and Raftery (2010).

3.2. Sparse finite mixture model with shrinkage prior

An overfitting mixture model is estimated using $e_0 = 0.0001$ and $K = 10$. In addition the shrinkage prior is imposed on the class specific parameters. The posterior distribution of the number of non-empty components K_+ has a clear mode at 2, with 99.9% of the samples having 2 non-empty components. The remaining samples had 3 non-empty components (0.2%). Using the samples with 2 non-empty components to identify the model results in a posterior mean estimate for the component weight of the larger class π_1 of 0.572 with a posterior standard deviation of 0.068. The class specific parameters for the variables are given in Table 2 and the congruence between the partitions in

Variable	Level	Class 1	Class 2
Chest pain type	Typical Angina	0.06 (0.02)	0.01 (0.01)
	Atypical Angina	0.57 (0.06)	0.07 (0.04)
	Non-anginal pain	0.26 (0.04)	0.08 (0.04)
	Asymptomatic	0.10 (0.07)	0.83 (0.06)
Exercise induced Angina	No	0.94 (0.03)	0.34 (0.10)
	Yes	0.06 (0.03)	0.66 (0.10)
Gender	Female	0.36 (0.04)	0.15 (0.04)
	Male	0.64 (0.04)	0.85 (0.04)
Resting Electrocardiographic results	Normal	0.81 (0.03)	0.79 (0.04)
	ST-T wave	0.16 (0.03)	0.19 (0.04)
	Estes' criteria	0.03 (0.01)	0.02 (0.01)
Fasting blood sugar >120 mg/dl	False	0.94 (0.02)	0.91 (0.03)
	True	0.06 (0.02)	0.09 (0.03)

Table 2. Posterior mean (and posterior standard deviations) of the class specific parameters for the identified 2-class sparse LCA model with shrinkage prior.

	Standard prior		Shrinkage prior	
	$<50\%$	$>50\%$	$<50\%$	$>50\%$
Class 1	139	15	135	14
Class 2	42	88	46	89

Table 3. Estimated versus clinical partition for the identified 2-component sparse LCA model with standard or shrinkage prior.

Table 3 on the right. Overall similar results are obtained for the two different component specific priors. However, using a shrinkage prior reduces the risk of overfitting heterogeneity and thus allows to obtain more precise estimates in case irrelevant variables are identified. Figure 3 shows the posterior distributions of the shrinkage parameters λ for each variable. Small values indicate that a variable is identified as not being relevant for distinguishing between the two classes and that similar parameter values are estimated for both classes. These results confirm those by Dean and Raftery (2010) who concluded that the variables "Resting Electrocardiographic results" and "Fasting blood sugar >120 mg/dl" are irrelevant.

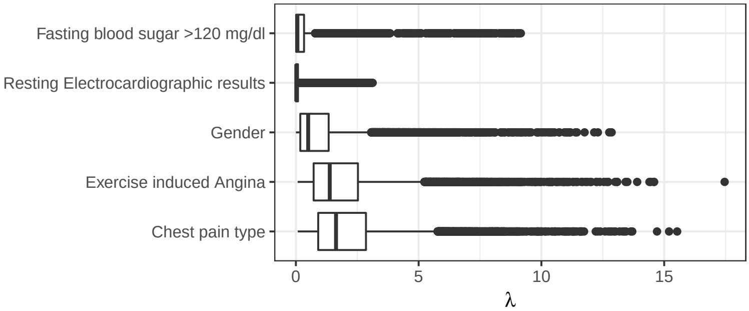


Figure 3. Box plot of the shrinkage parameter λ for each variable.

4. Conclusion

Suitable priors for Bayesian LCA are presented which regularize the likelihoods to avoid boundary solutions, induce sparse solutions with respect to the number of classes as well as shrinkage to perform implicit variable selection. Their application is demonstrated on the Hungarian heard disease data which was previously analyzed based on maximum likelihood estimation. This data set contains a clear structure with respect to the number of classes as well as the relevance of variables for clustering. Suitable priors for such a setting were proposed. Future research needs to investigate how these priors perform and need to be adapted in more challenging settings.

Acknowledgements: This research was funded by the Austrian Science Fund (FWF): P28740.

References

Dean N., Raftery A.E. (2010) Latent class analysis variable selection. *The Annals of the Institute of Statistical Mathematics*, 62, 11-35.

Diebolt J., Robert C.P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, 56, 363-375.

Lazarsfeld P. (1950) The logical and mathematical foundation of latent structure analysis. In *Measurement and Prediction, the American Soldier: Studies in Social Psychology in World War II*, IV, 362-412. Princeton University Press.

Lunn D., Spiegelhalter D., Thomas A., Best N. (2009) The BUGS Project: Evolution, Critique and Future Directions. *Statistics in Medicine*, 28, 3049-3067.

Malsiner-Walli G., Frühwirth-Schnatter S., Grün B. (2016) Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26, 303-324.

Papastamoulis P. (2016) label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software*, 69, 1-24.

Plummer M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Technische Universität Wien, Vienna, Austria.

Rousseau J., Mengersen K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society B*, 73, 689-710.

White A., Wyse J., Murphy T.B. (2016) Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. *Statistics and Computing*, 26, 511-527.

Modelling ordinal data: a ϕ-divergence based approach

Maria Kateri*

Abstract: The role of the ϕ-divergence (see Pardo, 2016) in constructing models for ordinal data will be discussed. In particular well-known models for contingency table analysis (see Kateri, 2014) and regression models for binary or ordinal responses (see Agresti, 2013) will be revisited and redefined through divergence measures like the Kullback-Leibler and the Pearson divergences. Since these divergences are members of the ϕ-divergence family, we shall proceed by embedding these models in generalized families of models derived by replacing the Kullback-Leibler or Pearson divergence through the ϕ-divergence. Properties of these model families will be considered and the role of the specific divergence measure used in describing the underlying dependence structure will be commented. More specific, ϕ-divergence based association models for two-way tables have been introduced in Kateri and Papaioannou (1995) and further discussed in Kateri (2018). Asymmetry models for square contingency tables are generalized by Kateri and Papaioannou (1997) while the quasi symmetry model for ordinal variables by Kateri and Agresti (2007). A ϕ-divergence based extension of the binary logistic regression model can be found in Kateri and Agresti (2010). Crucial quantities in developing and interpreting these models are the ϕ-scaled generalized odds ratios, based on the corresponding generalization of the odds ratio for 2×2 tables (see Espendiller and Kateri, 2016). The focus here will be on higher dimensional problems. Characteristic members of the presented ϕ-divergence based families of models, corresponding to the power divergence of Cressie and Read (1984), will be implemented on examples and discussed. The approach is maximum likelihood based. The maximum likelihood estimators (MLEs) of the models considered cannot be derived in closed-form expressions and have to be computed numerically. Finally, emphasis will be given in closed-form approximations to the MLEs that simplify the model fitting approach and can be valuable in model selection procedures in high-dimensional set-ups.

Keywords: Contingency tables, Generalized odds ratios, Logistic regression.

* Institute of Statistics, RWTH Aachen University, Germany, maria.kateri@rwth-aachen.de

References

Agresti A. (2013) *Categorical Data Analysis*, 3d ed. Wiley, Hoboken, NJ.

Cressie N., Read T.R.C. (1984) Multinomial Goodness-of-Fit Tests, *J. R. Statist. Soc. B*, 46, 440-464.

Espendiller M., Kateri M. (2016) A family of association measures for 2×2 contingency tables based on the ϕ-divergence, *Statistical Methodology*, 35, 45-61.

Kateri M. (2014) *Contingency Table Analysis: Methods and Implementation Using R*, Birkhäuser/Springer, New York.

Kateri M. (2018) ϕ-Divergence in Contingency Table Analysis, *Entropy*, 20, 1-12.

Kateri M., Agresti A. (2007) A class of ordinal quasi symmetry models for square contingency tables. *Stat. Probab. Letters*, 77, 598-603.

Kateri M., Agresti A. (2010) A generalized regression model for a binary response, *Stat. Probab. Letters*, 80, 89-95.

Kateri M., Papaioannou T. (1995) f-divergence association models, *Int. J. Math. & Stat. Sci.*, 3, 179-203.

Kateri M., Papaioannou T. (1997) Asymmetry models for contingency tables, *J. Amer. Statist. Assoc.*, 92, 1124-1131.

Pardo L. (2006) *Statistical Inference Based on Divergence Measures*, Chapman & Hall, New York.

Robust statistical analysis of ordinal data

Elvezio Ronchetti[*]

Abstract: This paper discusses the robustness issues of estimators and tests in the analysis of ordinal data based on ordinal response models. From a diagnostic point of view, we investigate the effects of outlying covariates and of specific deviations due to some respondents' behavior, on the reliability of maximum likelihood estimators and related test procedures. In particular we highlight the role of the link function in this context. Subsequently, we propose robust M-estimators as an alternative to maximum likelihood estimators. We show that M based inference outperform maximum likelihood inference, producing more reliable results in the presence of deviations from the underlying assumptions.

Keywords: Ordinal response models, Link functions, M-estimation.

1. Introduction

Ordinal data play an important role in applied research in many areas, such as medicine, psychology, sociology, political sciences, economics, marketing, and so on. They typically arise when items concerning opinions, preferences, judgements, evaluations, worries, etc., are expressed as ordered categories. The classical statistical approach to the analysis of ordered response models is based on the assumption that a (unobserved) latent variable drives the response and the model is then embedded within the Generalized Linear Model framework (McCullagh and Nelder (1989)); see standard books such as Agresti (2010) and Tutz (2012) among others. A different approach based on the so-called CUB models (Piccolo (2003); Iannario and Piccolo (2016)), parametrizes the probability of a given response as a mixture of a shifted Binomial and a discrete Uniform random variable. This approach does not require the specification of a model for the latent variable and describes directly the effect of the covariates on the feeling and the uncertainty underlying the respondents' choices.

[*]Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland, Elvezio.Ronchetti@unige.ch

In real situations, it has been recognized that respondents may deliberately or unconsciously choose a wrong category, as a consequence of a satisficing aptitude or a search for a *shelter* category (Iannario (2012)). This phenomenon, in addition to the occurrence of gross-errors or to erratic behavior by a few respondents, produces a contamination of the assumed model distribution, which can have an important impact on the resulting estimators and tests.

Robust statistics deals with deviations from the underlying assumptions and with their effects on the inferential procedures; see e.g. the books by Huber (1981, 2nd edition by Huber and Ronchetti, 2009), Hampel et al. (1986), Maronna et al. (2006). However, in spite of the huge body of literature in the past decades, the area of ordinal data has been somewhat neglected. A few exceptions are Hampel (1968), Victoria-Feser and Ronchetti (1997), Ruckstuhl and Welsh (2001), Moustaki and Victoria-Feser (2006), Croux et al. (2013), and Iannario et al. (2016).

Using ideas and tools of robust statistics, we propose robust M-estimators as an alternative to maximum likelihood estimators and we show that M based inference outperform maximum likelihood inference, producing more reliable results in the presence of deviations from the underlying assumptions.

2. *Maximum likelihood estimation*

In this paper we consider a rich class of ordinal response models based on a latent variable with covariates and different link functions. More specifically, let Y be an ordinal variable of interest which is linked to an underlying latent variable Y^* through the relationship

$$Y = j \qquad \Longleftrightarrow \qquad \alpha_{j-1} < Y^* \leq \alpha_j, \qquad j = 1, 2, \ldots, m, \tag{1}$$

where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = +\infty$ are the thresholds (cutpoints) of the continuous support of the latent variable, and m represents the given number of categories of Y.

The variable Y^*, in turn, depends on $p \geq 1$ covariates, so that for the i-th statistical unit we have the latent regression model

$$Y_i^* = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + \epsilon_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, 2, \ldots, n, \quad (2)$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ and ϵ_i is a random variable with distribution function $G(\epsilon)$ and density $g(\epsilon)$ respectively. Since Y^* is unobservable, a random sample is given by $(Y_i, \boldsymbol{X}_i)$, for $i = 1, 2, \ldots, n$.

Relationship (1) yields the following probability mass function for Y_i conditionally on $\boldsymbol{X}_i = \boldsymbol{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ip})'$

$$P(Y_i = j \mid \boldsymbol{x}_i) = P(\alpha_{j-1} < Y_i^* \leq \alpha_j) = G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}), \quad (3)$$

for $j = 1, 2, \ldots, m$. Common specifications of $G(\epsilon)$ are the Gaussian, the logistic and the extreme value distributions; see Agresti (2010) for an extensive review.

From (3) it is easy to write the likelihood function of the parameters and their score function. In particular, the k-th component of the score function of the regression parameter $\boldsymbol{\beta}$ is given by

$$\sum_{j=1}^{m} I[y_i = j] e_{ij}(\boldsymbol{\theta})\, x_{ik}, \quad (4)$$

where

$$e_{ij}(\boldsymbol{\theta}) = \frac{g(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - g(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})}{G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})}, \; i = 1, 2, \ldots, n\,, \; j = 1, 2, \ldots, m\,, \quad (5)$$

are the *generalized residuals* (Franses and Paap (2004), p. 123), $I[\cdot]$ is the indicator function, and $\boldsymbol{\theta} = ((\alpha_1, \ldots, \alpha_{m-1})', \boldsymbol{\beta}')'$.

By the standard theory of robust statistics, the influence function (Hampel, 1974) of the Maximum Likelhood Estimator (MLE) is proportional to its score function, which is unbounded in the covariate $\boldsymbol{x}$ and possibly in the generalized residuals depending on the distribution $G(\cdot)$. This leads to the conclusion that the MLE is locally non-robust. Notice that for the probit and the complementary log-log link, the generalized residuals are unbounded, whereas for the logistic link they are bounded. This implies that the logistic link provides robustness for the MLE at least with respect to deviations in the generalized residuals.

3. Robust alternatives

A class of robust alternatives to the MLE can be obtained by introducing weights in the ML score function (4). This defines M-estimators with a bounded influence function, which yields reliable estimates of the parameters, and can be used to derive robust testing procedures. A typical weight function is given by

$$w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{cases} 1, & \text{if } \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\| < c \\ \dfrac{c}{\sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\|}, & \text{if} \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\| \geq c\,. \end{cases} \tag{6}$$

A Mahalanobis distance $||\boldsymbol{x}_i|| = \left\{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_X)'\hat{\Sigma}_X^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_X)\right\}^{1/2}$ can be used for the norm of the covariates in (6), which needs however to be based on a robust multivariate estimator of location $\hat{\boldsymbol{\mu}}_X$ and scatter $\hat{\Sigma}_X$. Notice that these weights provide valuable diagnostic information on possible outliers and substructures in the data. The tuning constant c determines a trade-off between efficiency and robustness and can be computed by requiring a given efficiency (e.g. 95%) for the resulting estimator at the assumed model.

4. A numerical example: a shelter effect

We consider the following simple model. The response variable Y assumes 4 categories and depends on two qualitative variables W_i, for $i = 1, 2$. Each of them assumes three categories, coded by two dichotomous $0-1$ variables X_i^a

and X_i^b such that $X_i^a + X_i^b \leq 1$. The latent variable is $Y^* = 2.5X_1^a + 1.0X_1^b + 3.6X_2^a + 1.8X_2^b + \epsilon$, where $\epsilon \sim N(0,1)$ and the cutpoints are $\boldsymbol{\alpha} = (1.2, 2.8, 5)'$.

Now we consider the case when five Y_i, which originally take value 1, 2 or 3, are changed into 4. This kind of contamination occurs when the selected category (in this case "four") can be regarded as a *shelter* category: a choice that the respondents feel comfortable with, although it appears incoherent with their profiles in terms of covariates; see Iannario (2012), for a more extensive illustration of shelter choices.

The minimum and the maximum MSE-ratios of each parameter estimate between the MLE and M-estimators are shown in Table 1. Values of c between 1 and 1.5 seem thoroughly appropriate to achieve robust estimation according to the two criteria. If say $c = 1.25$ is taken, the gain in efficiency in the estimation of a single parameter varies between 37.6% and roughly 140%, which is a quite remarkable achievement obtained by M-estimators.

Table 1. Efficiency criteria when M-estimation is performed with the probit link and a shelter effect occurs.

c	1	1.25	1.5	1.75	2	2.5	3
Min(MSE-ratio)	1.354	1.376	1.373	1.342	1.299	1.201	1.124
Max(MSE-ratio)	2.556	2.398	2.227	2.029	1.819	1.485	1.253

Acknowledgements: This paper is a summary of Iannario M., Monti A. C., Piccolo D., Ronchetti E. (2017) Robust inference for ordinal response models, *Electronic Journal of Statistics*, 11, 3407-3445, where all the details can be found.

References

Agresti A. (2010) *Analysis of ordinal categorical data*, 2^{nd} ed., Wiley, New York.

Croux C., Haesbroeck G., Ruwet C. (2013) Robust estimation for ordinal regression, *Journal of Statistical Planning and Inference*, 143, 1486-1499.

Franses P.H., Paap R. (2004) *Quantitative models in marketing research*, Cambridge University Press, Cambridge.

Hampel F.R. (1968) Contribution to the theory of robust estimation, Ph.D Thesis, University of California, Berkeley.

Hampel F. R. (1974) The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 69, 383-393.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986) *Robust statistics: the approach based on influence functions*, Wiley, New York.

Huber P.J. (1981) *Robust statistics*, Wiley, New York.

Huber P.J., Ronchetti E. (2009) *Robust statistics*, 2^{nd} ed., Wiley, New York.

Iannario M. (2012) Modelling *shelter* choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, 21, 1-22.

Iannario M., Piccolo D. (2016) A comprehensive framework of regression models for ordinal data, *METRON*, 74, 233-252.

Iannario M., Monti A. C., Piccolo D. (2016) Robustness issues for CUB models, *TEST*, 25, 731-750.

Maronna R.A., Martin R.D., Yohai V. J. (2006) *Robust statistics: theory and methods*, Wiley, New York.

McCullagh P., Nelder, J. A. (1989) *Generalized linear models*, 2^{nd} edition, Chapman and Hall, London.

Moustaki I., Victoria-Feser M. P. (2006) Bounded-influence robust estimation in generalized linear latent variable models, *Journal of the American Statistical Association*, 101, 644-653.

Piccolo D. (2003) On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85-104.

Ruckstuhl A.F., Welsh A.H. (2001) Robust fitting of the binomial model, *The Annals of Statistics*, 29, 1117-1136.

Tutz G (2012) *Regression for categorical data*, Cambridge University Press, Cambridge.

Victoria-Feser M.P., Ronchetti E. (1997) Robust estimation for grouped data, *Journal of the American Statistical Association*, 92, 333-340.

Uncertainty, dispersion and response styles in ordinal regression

Gerhard Tutz*

Abstract: Alternative approaches to model uncertainty, dispersion and the tendency to middle or extreme categories in ordinal regression are considered. The focus is on repeated measurements when a person responds on several items. Then it is possible to account for individual response tendencies known as response styles. Extensions of the adjacent categories model are proposed that allow for the noncontingent response style, that is, a person answers randomly or nonpurposefully, and the extreme response style, which means a person has a tendency to prefer extreme or middle categories. Also mixture models for repeated measurements are discussed.

Keywords: Response styles, Uncertainty, Dispersion.

1. Introduction

Ordinal regression models aim at linking the choice of a response category to explanatory variables. In traditional models, for example in the most widely used proportional odds model, the explanatory variables determine primarily the location of the response on an underlying latent scale, which turns into the location on the range of observed categories, see, for example, Agresti (2009). However, alternative effects that determine the response on an ordinal scale might be present. One effect, which has been modeled in various versions of the CUB model (Iannario and Piccolo, 2016) is uncertainty. The manifest response in a category is determined not only by the preference towards a category but also by the the uncertainty of the respondent. CUB models refer to the underlying psychological mechanism that generates uncertainty. Alternative effects are the impact of explanatory variables on the dispersion of the categorical response and the preference for middle and extreme categories.

If a person answers to several items one has repeated measurements on a

*Ludwig-Maximilians-Universität München, tutz@stat.uni-muenchen.de

person. Then it is possible to include subject-specific effects that contain the individual tendency to respond as response styles, that means as a consistent pattern of responses that is independent of the content of the response. The presence of response styles may affect the response behaviour and, when neglected yield biases estimates, see, for example, Baumgartner and Steenkamp (2001), Van Vaerenberg and Thomas (2013).

In the following the modeling of uncertainty, dispersion and a tendency to middle and extreme categories in cross-sectional data is considered briefly. Then the modeling of response styles as individual traits is considered.

2. *Mixture Models in Ordinal Regression*

Natural candidates for the modelling of uncertainty in ordinal regression are mixture models. In particular the CUB model and its various extensions use this potential, see, for example, Piccolo (2003), Iannario and Piccolo (2016). A general mixture model with an indecision component has the form

$$P(R_i = r|\mathbf{x_i}) = \pi_i P_P(Y_i = r|\boldsymbol{x}_i) + (1-\pi_i)P_I(U_i = r), \tag{1}$$

where R_i represents the observed response and Y_i, U_i are the unobserved random variables taking values from $\{1, \ldots, k\}$. The variable Y_i represents the preference of a person for categories whereas U_i represents the indecision of a person. For the modelling of the preference, which is the deliberate choice, one can use any ordinal response model, for example, the proportional odds model, the adjacent categories model, or a shifted binomial model. The latter is used in the CUB model. For the mixture probability one typically uses a logit model $\operatorname{logit}(\pi_i) = \boldsymbol{x_i}^T\boldsymbol{\gamma}$.

The choice of the indecision component determines which form of indecision is specified. Classical CUB-type models assume the uniform distribution, $P_U(U_i = r) = 1/k$, which can be seen as the strongest form of *uncertainty*; a person chooses at random from the available categories. In a CUB model (binomial preference, uniform indecision) the Gini index, which measures deviation from the uniform distribution and therefore uncertainty, is monotonically increasing with the strength of the uncertainty $1 - \pi_i$.

Alternatively, one can specify indecision by using distributions with different shapes. If one uses use a beta-binomial distribution centered at the middle of the response categories (Tutz and Schneider, 2018) or a discretized beta distribution (Simone and Tutz, 2018) one models *dispersion* in the uncertainty component, which may be seen as a *response style*, more concrete a response style that allows for a tendency to middle or extreme categories. Alternative forms of response style distributions in the uncertainty component were used by Gottard et al (2016).

When using a centered distribution in the uncertainty component one allows for varying *dispersion*. For categorical data one might want to avoid the variance as a dispersion measure since it demands a higher scale level than the ordinal scale level. For $Y \in \{1, \dots, k\}$ a more appropriate measure is the sum $D = \sum_{j=2}^{k} var(Y_j) 4/(k-1)$, where $Y_j = 1$ if $Y \geq j$ and $Y_j = 1$ otherwise. One obtains $D = 0$ if Y has a one-point distribution and $D = 1$ if $p(Y = 1) = P(Y = k) = 0.5$. If indecision is modeled by a discretized beta distribution the dispersion of the indecision can vary between zero and one. Thus, for large indecision one obtains a wide range of dispersion for the response.

Dispersion can also be modeled in the preference part instead of the indecision part. The CUBE model (Iannario, 2014) uses the beta binomial distribution to model (over)dispersion in the preference part. Alternative approaches to model dispersion effects that are linked to explanatory variables are the location-scale model and the location-shift model. The *location-scale model* (McCullagh, 1980) uses the cumulative model in the extended form

$$P(Y \leq r) = F(\frac{\beta_{0r} + \boldsymbol{x}^T \boldsymbol{\beta}}{\boldsymbol{z}^T \boldsymbol{\alpha}}),$$

where $F(.)$ is a distribution function, typically the logistic function, and $\boldsymbol{z}$ is an additional vector of explanatory variables that determines dispersion. The model is also known as heterogeneous choice model or heteroscedastic logit model. The *location-shift version* of the cumulative model (Tutz and Berger, 2017) is

$$P(Y \leq r) = F(\beta_{0r} + \boldsymbol{x}^T \boldsymbol{\beta} + (r - k/2) \boldsymbol{z}^T \boldsymbol{\alpha}).$$

The term $(r - k/2)\boldsymbol{z}^T\boldsymbol{\alpha}$ shrinks or widens the thresholds of the cumulative model.

3. Responses on Several Items

If one has just one observation per person one may link uncertainty and a tendency to middle or extreme categories to explanatory variables but can not model them as individual traits. However, when a person responds on more than one item it is possible to model uncertainty and dispersion on the individual level. Then, uncertainty and dispersion may be seen as response styles that describe an individual's tendency to respond, which also can depend on explanatory variables. A widely used approach in item response modeling is the partial credit model, which is a member of the family of adjacent categories models.

Let $Y_{pi} \in \{0, 1, \ldots, k\}$, $p = 1, \ldots, P$, $i = 1, \ldots, I$ denote the ordinal response of person p on item i. The *partial credit model* (PCM) assumes for the probabilities

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)}\right) = \theta_p - \delta_{ir}, \quad r = 1, \ldots, k,$$

where θ_p represents the attitude or ability of a person and the δ_{ir} are item-specific thresholds on the latent scale. It turns into an ordinal regression model if the person parameter is determined by explanatory variables. If one replaces θ_p by $\theta_p + \boldsymbol{x}^T\boldsymbol{\beta}$ and considers the thresholds as item-specific intercepts one obtains the *adjacent categories model for repeated measurements*,

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)}\right) = \theta_p + \boldsymbol{x}_p{}^T\boldsymbol{\beta} - \delta_{ir}, \quad r = 1, \ldots, k,$$

with an subject-specific parameter θ_p, for which one assumes a normal distribution centered at zero. We will consider two extensions that include response styles as a continuous trait and will also discuss mixture models.

3.1. The Extreme Response Style Adjacent Categories Model

As in the location-shift model one may shift the thresholds of the model to obtain more concentration of the response in the middle or extreme categories. More concrete, one replaces the thresholds δ_{ir} by the term $\delta_{ir} - (m - r + 0.5)\alpha_p$, where $m = k/2$ and α_p is a subject-specific parameter that modifies the thresholds. The new thresholds are constructed such that intervals between thresholds are widened or narrowed by the subject-specific parameter α_p. In the case of five response categories one obtains the thresholds

$$\begin{array}{ccccccccc} 0 & | & 1 & | & 2 & | & 3 & | & 4 \\ & \delta_{i1} - 1.5\alpha_p & & \delta_{i2} - 0.5\alpha_p & & \delta_{i3} + 0.5\alpha_p & & \delta_{i4} + 1.5\alpha_p & \end{array}$$

It is seen that the difference between adjacent thresholds changes by the value α_p, which means a widening if α_p is positive and a narrowing if α_p is negative. The closed model has the form

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r-1)}\right) = (m-r+0.5)\alpha_p + \theta_p + \boldsymbol{x}_p{}^T\boldsymbol{\beta} - \delta_{ir}, \quad r = 1, \dots, k.$$

The role of the parameter α_p becomes obvious when considering extreme values:

> For $\alpha_p \to \infty$ one obtains that the probability mass is concentrated in the middle, that is, one has $P(Y_{pi} = m/2) = 1$, if k is even (odd number of categories), and $P(Y_{pi} = (k-1)/2) + P(Y_{pi} = (k+1)/2) = 1$ if k is odd (even number of categories).
>
> For $\alpha_p \to -\infty$ one obtains that the probability mass is concentrated in the extreme values, that is, $P(Y_{pi} = 0) + P(Y_{pi} = k) = 1$.

The parameter α_p represents the tendency to middle or extreme categories. Large values indicate a subject's tendency to choose middle categories while small values indicate a tendency to extreme categories. Often the latter tendency is referred to as extreme response style. The model considered here accounts for the tendency to middle categories *and* to extreme categories simultaneously. Nevertheless, for simplicity we refer to it as the *extreme response style adjacent categories model.* For earlier versions see also Tutz et al (2018).

3.2. The Uncertainty Adjacent Categories Models

An alternative is model is obtained by including a factor as in the location-scale model. Let again α_p denote a subject-specific parameter in the model

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = e^{\alpha_p}(\theta_p + \boldsymbol{x_p}^T\boldsymbol{\beta} - \delta_{ir}), \quad r = 1,\ldots,k, \quad (2)$$

In the case of ordered thresholds $(\delta_{ir} \leq \delta_{i,r+1})$ one obtains the following:

> For $\alpha_p \to \infty$ a person with $\theta_p + \boldsymbol{x_p}^T\boldsymbol{\beta} \in (\delta_{ir}, \delta_{i,r+1})$ has the probability $P(Y_{pi} = r) = 1$. One observes a distinct response, the person knows exactly which category he/she prefers. The property holds for all k if one defines in addition $\delta_{i0} = -\infty$, $\delta_{i,k+1} = \infty$.
>
> For $\alpha_p \to -\infty$ one obtains $P(Y_{pi} = r) = 1/(k+1))$ for all abilities/attitudes θ_p. The person has a discrete uniform distribution over the response categories, which means simple guessing.

The role of the parameter α_p differs from that in the extreme style model. Large values of α_p indicate a subject's tendency to a distinct response while small values indicate a tendency to random responding. The response style is also known as *noncontingent response style.* It is found if persons have a tendency to respond to items carelessly, randomly, or nonpurposefully. We use the general term uncertainty to characterize that tendency.

In both models one can let the person parameter α_p depend on explanatory variables by using $\alpha_p + \boldsymbol{z}^T\boldsymbol{\alpha}$ instead of α_p.

4. Response Styles and Mixture Models

An alternative approach to model response styles, which has been propagated in psychometrics, is to use finite mixture models. One assumes that the responses stem from a finite mixture

$$P((Y_{p1},\ldots,Y_{pI})) = \sum_{m=1}^{M} \pi_m P_m((Y_{p1},\ldots,Y_{pI})|\theta_p,\boldsymbol{\delta}_i^{(m)}).$$

It is assumed that the population is subdivided into M latent classes, where $P_m(.)$ denotes the model in the latent class m with parameters $\theta_p, \boldsymbol{\delta}_i^{(m)} = (\delta_{i1}^{(m)},\ldots,\delta_{ik}^{(m)})^T$, and π_m is the mixture probability (size) of the latent class m. The model is based on the mixture approach, which has some tradition in psychometrics, for an overview see Von Davier and Carstensen (2007).

The model is not without problems. Typically the same item response model, for example the partial credit model, is fitted within these classes, some components of the mixture may represent the substantive trait, some may represent response style behaviour, see, for example, Eid and Rauber (2000). However, the number of classes is unknown. One gets quite different models if one fits, for example, two or three classes since all the parameters change when allowing for one more class. Even if a number of classes is fixed it is frequently difficult to interpret what feature is represented by a class, it might be a response style or some other dimension that is involved when responding to items. The latter problem arises since the classes are not specified to represent specific traits.

It seems more appropriate to use mixtures of models that represent explicitly what one wants to identify. This approach is in the tradition of CUB models, which use different models in a mixture of two components, one for the deliberate choice and one for the uncertainty. In this spirit one can use a mixture of models that represent the preference and the explicit response style that is suspected to be present.

Let R_{pi} represent the observed response for person p and item i and Y_{pi}, U_{pi} be unobserved random variables taking values from $\{0,\ldots,k\}$. We propose the finite mixture model

$$P(R_{pi} = r|\theta_p, \pi_p, \boldsymbol{\delta}_i) = \pi_p P_M(Y_{pi} = r|\theta_p, \boldsymbol{\delta}_i) + (1 - \pi_p) P_R(U_{pi} = r).$$

The distribution of Y_{pi} is determined by $P_M(Y_{pi} = r|\boldsymbol{x}_i)$, with M standing for the model, in our case the partial credit model. The random variable U_{pi} represents the response style and can be specified in different ways.

As in CUB models one can use the uniform distribution, that is, $P_R(U_{pi} = r) = 1/k$. Then one models the noncontingent response style. An advantage is that the distribution is the same for all persons, therefore the tendency to using a response style is contained in the mixture probability π. If one uses for U_{pi} a discretized beta distribution as in Simone und Tutz (2018) one models the tendency to middle or extreme categories. Then an additional parameter for the person's tendency to respond is needed.

The strength of an individual's tendency to using a response style is contained in the mixture component, which is allowed to be subject-specific by assuming

$$\pi_p = \frac{\exp(\xi_0 + \xi_p)}{1 - \exp(\xi_0 + \xi_p)},$$

where ξ_p follows a normal distribution, $\mathrm{N}(0, \sigma^2)$. If $\sigma^2 = 0, \xi_0 \rightarrow -\infty$ one obtains the partial credit model as limiting case.

5. Concluding Remarks

It should be mentioned that the inclusion of response styles as continuous traits in the extreme response style and the uncertainty adjacent categories model are specifically designed for the adjacent categories model. If one tries to include similar terms in the cumulative model the effects have quite different meaning. As an example let us consider the proportional odds model for repeated measurements. In the model

$$\log\left(\frac{P(Y_{pi} \leq r)}{P(Y_{pi} > r)}\right) = e^{\alpha_p}(\theta_p - \delta_{ir}), \quad r = 1, \ldots, k,$$

the subject-specific factor e^{α_p} has an interpretation that differs from that in the corresponding adjacent categories model. For $\alpha_p \to \infty$ one obtains for a person with $\theta_p \in (\delta_{ir}, \delta_{i,r+1})$ the probability $P(Y_{pi} = r) = 1$, that means a person knows exactly what he/she wants. However, for $\alpha_p \to -\infty$ one obtains $P(Y_{pi} = 0) = P(Y_{pi} = k) = 0.5$, and therefore not the noncontingent response style. The strong probability in the extreme categories is more related to dispersion and the preference for extreme categories than to the noncontingent response style.

References

Agresti A. (2009) *Analysis of Ordinal Categorical Data, 2nd Edition*, Wiley, New York.

Baumgartner H., Steenkamp, J. (2001) Response styles in marketing research: A cross-national investigation, *Journal of Marketing Research*, 38, 143-156.

Eid M., Rauber M. (2000) Detecting measurement invariance in organizational surveys, *European Journal of Psychological Assessment*, 16, 20-30.

Gottard A., Iannario, M., Piccolo D. (2016) Varying uncertainty in CUB, *Advances in Data Analysis and Classification*, 10, 225-244.

Iannario M. (2014) Modelling uncertainty and overdispersion in ordinal data, *Communications in Statistics-Theory and Methods*, 43, 771–786.

Iannario M., Piccolo D. (2016) A comprehensive framework of regression models for ordinal data, Metron, 74, 233-252.

McCullagh P. (1980) Regression Model for Ordinal Data (with Discussion), *Journal of the Royal Statistical Society, Series B*, 42, 109-127.

Piccolo D. (2003) On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85-104.

Simone R., Tutz, G. (2018) Modelling Uncertainty and Response Styles in Ordinal Data *Statistica Neerlandica*, 72, 224-245.

Tutz G., Berger, M. (2017) Separating Location and Dispersion in Ordinal Regression Models. *Econometrics and Statistics*, 2, 131-148.

Tutz G., Schauberger G., Berger M. (2018) Response Styles in the Partial Credit Model. *Applied Psychological Measurement*, 42, 407-427.

Tutz G., Schneider M. (2017) Mixture Models for Ordinal Responses with a Flexible Uncertainty Component. *Technical Report 203, Department of Statistics LMU.*

Tutz G., Schneider M., Iannario M., Piccolo D. (2017) Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification*, 11, 281-305.

Van Vaerenbergh Y., Troy D (2013) Response styles in survey research: A literature review of antecedents, consequences, and remedies, *International Journal of Public Opinion Research*, 25, 195-217.

Von Davier M., Carstensen C. (2007) *Multivariate and mixture distribution Rasch models*, Springer, New York.

ACCEPTED PAPERS

Inducing a desired value of correlation between two point-scale variables

Alessandro Barbiero[*]

Abstract: Focusing on point-scale random variables, i.e., variables whose support space is given by the first m integers, we discuss how a desired value of Pearson's correlation can be induced between two assigned probability distributions, which are linked to a joint distribution via a copula function. After recalling how the value of the desired ρ is not free to vary within $[-1, +1]$, but is bounded to a narrower interval depending on the two marginal distributions, we devise a procedure to recover the same feasible value ρ for different dependence structures, focusing on one-parameter copulas encompassing the entire dependence spectrum.

Keywords: Attainable correlations, Copula, Ordinal variables.

1. Introduction

Datasets arising in the social sciences often contain ordinal variables. In particular, Likert scale items are those where, given a statement, the subject indicates strong agreement, agreement, neutrality, disagreement, or strong disagreement. A relevant example derives from questionnaires about customers' satisfaction. Satisfaction can be regarded as a multidimensional latent (i.e., unobservable) phenomenon, involving several aspects that can be usually measured using graded scales, such as "Very dissatisfied", "Dissatisfied", "Neither satisfied nor dissatisfied", "Satisfied" and "Very satisfied". Likert scales are often treated as interval scales, by scoring the ordered categories using the integers 1, 2, 3, . . . ; this amounts to assuming that the categories are evenly spaced. Though representing just an arbitrary assumption, it is quite a common and accepted practice as well as proceeding to further multivariate statistical analyses handling them as (correlated) univariate discrete variables.

[*] Department of Economics, Management and Quantitative Methods, University of Milan, alessandro.barbiero@unimi.it

Now, one may be interested in building and simulating a multivariate random vector whose univariate components are point-scale variables with assigned marginal distributions and whose pairwise correlations are chosen a priori as well. In the following we will limit our analysis to the bivariate case, which is by far easier to deal with, but whose results, with some caution, can be extended to the multivariate context. We consider two point scale random variables (r.v.s), X_1 and X_2, defined over the support spaces $\mathcal{X}_1 = \{1, 2, \ldots, m_1\}$ and $\mathcal{X}_2 = \{1, 2, \ldots, m_2\}$, respectively, with probability mass functions $p_1(i) = P(X_1 = i), i = 1, \ldots, m_1$, and $p_2(i) = P(X_2 = j), j = 1, \ldots, m_2$. We want to determine *some* bivariate probability mass function $p_{ij} = P(X_1 = i, X_2 = j), i = 1, \ldots, m_1; j = 1, \ldots, m_2$ such that its margins are p_1 and p_2 and the correlation ρ_{X_1,X_2} is equal to an assigned ρ. In order to give an answer to this question, we have first to recall two properties of Pearson's correlation, which applies to both the continuous and, to even a larger extent, the discrete case; this is the topic of Section 2. In Section 3, we briefly recall how to build copula-based bivariate discrete distributions. Section 4 is devoted to the description of the proposed procedure for inducing a desired value of correlation between two point-scale variables. Section 5 illustrates an application to CUB distributions.

2. Attainable correlations between two random variables

A first important but often neglected feature of Pearson's correlation is that given two marginal cumulative distribution functions (c.d.f.s) F_1 and F_2 and a correlation value $\rho \in [-1, +1]$, it is not always possible to construct a joint distribution F with margins F_1 and F_2, whose correlation is equal to the assigned ρ. We can state the following result, concerning "attainable correlations" (see McNeil et al. 2005, pp.204-205). Let (X_1, X_2) be a random vector marginal cdfs F_1 and F_2 and an unspecified joint cdf; assume also that Var$(X_1) > 0$ and Var$(X_2) > 0$. The following statements hold:

1. The attainable correlations form a closed interval $[\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} < 0 < \rho_{\max}$.

2. The minimum correlation $\rho = \rho_{\min}$ is attained if and only if X_1 and X_2

are countermonotonic. The maximum correlation $\rho = \rho_{\max}$ is attained if and only if X_1 and X_2 are comonotonic.

3. $\rho_{\min} = -1$ if and only if X_1 and $-X_2$ are of the same type, and $\rho_{\max} = 1$ if and only if X_1 and X_2 are of the same type.

For point-scale r.v.s X_1 and X_2, it is then clear that the maximum correlation is $+1$ if and only if they are identically distributed; whereas the minimum correlation can never be -1. The values $\rho_{\min}$ and $\rho_{\max}$ can be computed by building the cograduation and countergraduation tables (see, Ferrari and Barbiero, 2012, for an example of calculation).

A second fallacy of Pearson's correlation can be resumed as follows: Given two margins F_1 and F_2 and a feasible linear correlation ρ, the joint distribution F having margins F_1 and F_2 and correlation ρ is not unique. In other terms, the marginal distributions and pairwise correlations of a r.v. do not univocally determine its joint distribution. Even if this second fallacy may represent a limit from one side, on the other side represents a form of flexibility, since it means that given two point-scale r.v.s and a consistent value of ρ, there are different (possibly, infinite) ways to join them into a bivariate distribution with that value of correlation, as we will see in the next two sections.

3. Generating bivariate discrete distributions via copulas

How can we generate from a bivariate distribution respecting the assigned margins and correlation? Using copulas represent a straightforward solution. A d-dimensional copula is a joint c.d.f. in $[0,1]^d$ with standard uniform c.d.f.s U_j, $j = 1\ldots,d$:

$$C(u_1,\ldots,u_d) := P(U_1 \leq u_1,\ldots,U_d \leq u_d).$$

The importance of copulas in the study of multivariate c.d.f.s is summarized by the Sklar's theorem (see McNeil et al., 20005), whose version for $d = 2$ states that if F_1 and F_2 are the c.d.f.s of the point-scale r.v.s X_1 and X_2, the function

$$F(i,j) = C(F_1(i),F_2(j)), i = 1,\ldots,m_1; j = 1,\ldots,m_2 \tag{1}$$

defines a valid joint c.d.f. over $\mathcal{X}_1 \times \mathcal{X}_2$, whose margins are F_1 and F_2. The only requirement we have to impose is that the copula C is able to encompass the entire range of dependence, from perfect negative dependence ($\rho_{\min}$) to perfect positive dependence ($\rho_{\max}$). Among copulas enjoying this property, we recall the Gauss copula, the Frank copula, and the Plackett copula.

The Gauss copula

The d-variate Gauss copula is the copula that can be extracted from a d-variate normal vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ and is exactly the same as the copula of $\boldsymbol{X} \sim N_d(\mathbf{0}, P)$, where P is the correlation matrix of $\boldsymbol{Y}$. In two dimensions, it can be expressed, for $\rho \neq \pm 1$, as:

$$C^{Ga}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho_{Ga}^2}} e^{-\frac{s_1^2 - 2\rho_{Ga} s_1 s_2 + s_2^2}{2(1-\rho^2)}} \mathrm{d}s_1 \mathrm{d}s_2.$$

Independence, comonotonicity, and countermonotonicity copulas are special cases of the bivariate Gauss copula (for $\rho_{Ga} = 0$, $\rho_{Ga} = 1$, and $\rho_{Ga} = -1$, respectively).

The Frank copula

The one-parameter bivariate Frank copula is defined as

$$C^F(u_1, u_2; \theta) = -\frac{1}{\kappa} \ln \left[1 + \frac{(e^{-\kappa u_1} - 1)(e^{-\kappa u_2} - 1)}{e^{-\kappa} - 1} \right],$$

with $\kappa \neq 0$. For $\kappa \to 0$, we have that the Frank copula reduces to the independence copula; for $\kappa \to \infty$, it tends to the comonotonicity copula; for $\kappa \to -\infty$, it tends to countermonotonicity copula.

The Plackett copula

The one-parameter bivariate Plackett copula is defined as

$$C^P(u_1, u_2; \kappa) = \frac{1 + (\theta - 1)(u_1 + u_2) - \sqrt{[1 + (\theta - 1)(u_1 + u_2)]^2 - 4\theta(\theta - 1)u_1 u_2}}{2(\theta - 1)},$$

with $\theta > 0$. When $\theta = 1$, it reduces to the independence copula, whereas for $\theta \to 0$ it tends to the countermonotonicity copula and for $\theta \to \infty$ to the comonotonicity copula.

4. Inducing a desired value of correlation between two point-scale random variables

The bivariate p.m.f. corresponding to (1) can be computed as

$$p(i, j) = F(i, j) - F(i-1, j) - F(i, j-1) + F(i-1, j-1) \tag{2}$$

Computing the correlation coefficient for a bivariate point-scale variable (2) is very easy; since

$$\rho_{x_1 x_2} = (\mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2))(\text{Var}(X_1)\text{Var}(X_2))^{-1/2} \tag{3}$$

with $\mu_1 = \mathbb{E}(X_1) = \sum_{i=1}^{m_1} i p_1(i)$, $\text{Var}(X_1) = \sum_{i=1}^{m_1} (i - \mu_1)^2 p_1(i)$ (analogous results hold for X_2), and $\mathbb{E}(X_1 X_2) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} ijp(i, j)$.

Once the marginal distributions of X_1 and X_2 are assigned, their correlation coefficient ρ_{X_1,X_2} will depend only on the copula parameter $\theta \in [\theta_{\min}, \theta_{\max}]$; this relationship may be written in an analytical or numerical form, say $\rho_{X_1,X_2} = g(\theta)$. Since the function g is not usually analytically invertible, inducing a desired value of correlation ρ between two point-scale variables, falling in $[\rho_{\min}, \rho_{\max}]$, by setting an appropriate value of the θ, is a task that can be generally done only numerically, by finding the (unique) root of the equation $g(\theta) - \rho_{X_1,X_2} = 0$. If ρ_{X_1,X_2} is a monotone increasing function of the copula parameter, it can be implemented by resorting to the following iterative procedure (see Ferrari and Barbiero, 2012; Barbiero and Ferrari, 2015b):

1. Set $\theta^{(0)} = \theta^{\Pi}$ (with θ^{Π} being the value of θ for which the copula C reduces to the independence copula); $\rho^{(0)} = 0$.
2. Set $t = 1$ and $\theta = \theta^{(t)}$, with $\theta^{(t)}$ some value strictly greater (smaller) than $\theta^{(0)}$ if $\rho > (<) 0$
3. Compute $F(i, j; \theta^{(t)})$ using (1)

4. Compute $p(i, j; \theta^{(t)})$ using (2)

5. Compute $\rho^{(t)}$ using (3)

6. If $|\rho^{(t)} - \rho| < \epsilon$ stop; else
   set $t \leftarrow t + 1$,
   $\theta^{(t)} \leftarrow \min(\theta_{\max}, \theta^{(t-1)} + m(\rho - \rho^{(t-1)}))$ if $\rho > 0$, or
   $\theta^{(t)} \leftarrow \max(\theta_{\min}, \theta^{(t-1)} + m(\rho - \rho^{(t-1)}))$ if $\rho < 0$,
   with $m = \dfrac{\theta^{(t-1)} - \theta^{(t-2)}}{\rho^{(t-1)} - \rho^{(t-2)}}$; go back to 3.

The above heuristic algorithm makes sense if g is a monotone increasing function, which is often the case: for the Gauss, Frank, and Plackett copulas, the linear correlation is an increasing function of the dependence parameter θ, keeping fixed the two marginal distributions. The advantage of the proposed algorithm stands in the two following (connected) features: i) in the capacity of finding the appropriate value of θ without making use of any sample from the two marginal distributions, ii) in the possibility of controlling a priori the error ϵ (absolute difference between target and actual values of ρ_{X_1,X_2}); setting ϵ equal to 10^{-7} generally allows to recover θ in a few steps.

Existing procedures for solving the same problem are available in the literature, but do not enjoy the two features above mentioned. For example, the proposal by Demirtas (2006), requires the preliminary generation of a “huge” bivariate sample of binary data.

5. Application to CUB random variables

A CUB r.v. X is defined as the mixture of a shifted Binomial and a discrete Uniform distribution over the support $\{1, 2, \ldots, m\}$, for $m > 3$ (Piccolo, 2003). Its probability mass function is

$$P(X = i) = \pi \binom{m-1}{i-1} \xi^{m-j} (1 - \xi)^{j-1} + (1 - \pi)\frac{1}{m}$$

with (π, ξ) a parameter vector with the parametric space $(0, 1] \times [0, 1]$.

Corduas (2011) proposed using the Plackett distribution in order to construct a one parameter bivariate distribution from CUB margins; this proposal was later investigated by Andreis ad Ferrari (2012), also in a multivariate direction. Here, we reprise and extend these attempts of constructing a bivariate CUB r.v. Let suppose we want to build a bivariate model with margins $X_1 \sim \text{CUB}(m_1 = 5, \pi_1 = 0.4, \xi_1 = 0.8)$ and $X_2 \sim \text{CUB}(m_2 = 5, \pi_2 = 0.7, \xi_2 = 0.3)$; we can find the values of the attainable correlations using the function `corrcheck` in `GenOrd` (Barbiero and Ferrari, 2015a). It returns as minimum and maximum correlations the values $\rho_{\min} = -0.952003$ and $\rho_{\max} = 0.8640543$. We can then proceed and select a desired feasible value of correlation between the two CUB variates, say $\rho = 0.6$. We can then recover the values of ρ_{Ga} (for the Gauss copula), κ (for the Frank copula), and θ (for the Plackett copula), according to the iterative procedure illustrated in the previous section. Setting $\epsilon = 10^{-7}$, we obtain $\rho_{Ga} = 0.6898959$, $\kappa = 5.453455$, and $\theta = 11.30106$. The three joint p.m.f.s, sharing the same level of linear correlation, are reported in Table 1. It is easy to notice the differences among them. For example, the probability $P(X_1 = 2, X_2 = 3)$ takes the values 0.0922, 0.0948, and 0.1008, in the three joint distributions.

References

Andreis F., Ferrari P.A. (2013) On a copula model with CUB margins. *Quaderni di Statistica*, 15, 33-51.

Barbiero A., Ferrari P.A. (2015a) GenOrd: Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions, *R package version 1.4.0.*

Barbiero A., Ferrari P.A. (2015b) Simulation of correlated Poisson variables. *Applied Stochastic Models in Business and Industry*, 31, 669-680.

Corduas M. (2011) Modelling Correlated Bivariate Ordinal Data with CUB Marginals. *Quaderni di statistica*, 13, 109-119.

Demirtas H. (2006) A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, 76(11), 1017-1025.

Ferrari P.A., Barbiero A. (2012) Simulating ordinal data. *Multivariate Behavioral Research*, 47, 566-589.

McNeil A., Frey R., Embrechts P. (2005) *Quantitative risk management. Concepts, Techniques and Tools.* Princeton Series in Finance, Princeton.

Piccolo D. (2003) On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, 5, 85-104.

Table 1. Bivariate distribution with margins $X_1 \sim$ CUB$(m_1 = 5, \pi_1 = 0.4, \xi_1 = 0.8)$ and $X_2 \sim$ CUB$(m_2 = 5, \pi_2 = 0.7, \xi_2 = 0.3)$ and $\rho_{x_1 x_2} = 0.6$, obtained based on different copulas

(x_1, x_2)	1	2	3	4	5	total
1	0.0553	0.0711	0.0959	0.0551	0.0065	0.2838
2	0.0088	0.0317	0.0922	0.1178	0.0333	0.2838
3	0.0013	0.0077	0.0377	0.0869	0.0479	0.1814
4	0.0002	0.0020	0.0150	0.0566	0.0565	0.1302
5	0.0000	0.0004	0.0045	0.0319	0.0838	0.1206
total	0.0657	0.1129	0.2452	0.3481	0.2281	1

(a) Gauss copula

(x_1, x_2)	1	2	3	4	5	total
1	0.0498	0.0744	0.1042	0.0483	0.0071	0.2838
2	0.0126	0.0297	0.0948	0.1167	0.0300	0.2838
3	0.0022	0.0060	0.0301	0.0916	0.0515	0.1814
4	0.0007	0.0019	0.0108	0.0548	0.0621	0.1302
5	0.0003	0.0009	0.0053	0.0366	0.0775	0.1206
total	0.0657	0.1129	0.2452	0.3481	0.2281	1

(b) Frank copula

(x_1, x_2)	1	2	3	4	5	total
1	0.0518	0.0775	0.1001	0.0439	0.0105	0.2838
2	0.0093	0.0251	0.1008	0.1221	0.0266	0.2838
3	0.0025	0.0060	0.0276	0.1004	0.0450	0.1814
4	0.0012	0.0026	0.0105	0.0532	0.0627	0.1302
5	0.0008	0.0018	0.0062	0.0285	0.0833	0.1206
total	0.0657	0.1129	0.2452	0.3481	0.2281	1

(c) Plackett copula

Dissimilarity measure for ranking data via mixture of copulae

Andrea Bonanomi*, Marta Nai Ruscone**, Silvia Angela Osmetti***

Abstract: We propose a new dissimilarity measure for ranking data by using a mixture of copula functions. This measure evaluates the dissimilarity between subjects expressing their preferences by rankings in order to classify them by a hierarchical cluster analysis. The proposed measure is based on the Spearman's grade correlation coefficient on a transformation, operated by the copula, of the rank denoting the level of the importance assigned by subjects in the classification process. The mixtures of copulae are a flexible way to model different types of dependence structures in the data and to consider different situations in the classification process. The advantage by using mixtures of copulae with lower and upper tail dependence is that we can emphasize the agreement on extreme ranks, when extreme ranks are considered more important. An example on simulated data illustrates our proposal.

Keywords: Ranking data, Mixture of copulae, Distance measure.

1. Introduction

Cluster analysis of ranking data aims at the identification of groups of subjects with a homogenous, common, preference behavior. Ranking data occur when a number of subjects are asked to rank a list of objects according to their personal preference order. Cluster analysis input is a distance matrix, whose elements measure the distances between rankings of two subjects. The choice of the distance dramatically affects the final result. The issue when dealing with ordinal data lies in computing an appropriate distance matrix. Several distance measures have been proposed for ranking data (Alvo and Yu, 2014). The most important are referred to Kendall's τ, Spearman's ρ and Cayley distances (Critcholw et *al.*, 1991; Mallows, 1957; Spearman, 1904). When the aim is to emphasize top ranks, weighted distances for ranking data should be used (Tarsitano, 2005). In this context, Bonanomi et *al* (2017) propose a

*Università Cattolica del Sacro Cuore, Milano, andrea.bonanomi@unicatt.it

**LIUC Università Cattaneo, mnairuscone@liuc.it

***Università Cattolica del Sacro Cuore, Milano, silvia.osmetti@unicatt.it

distance measure for ranking data based on copula function with (lower) tail dependence for emphasize the agreement on top ranks, when the top ranks are considered more important than the lower ones.

In this work we propose a generalization of the distance using a mixture of copulae. In this way we have a more flexible instrument to model different types of data dependence structures and to consider different situations in the classification process. For example, by using mixture of copulae with lower tail dependence, we emphasize top ranks or by using mixture of copulae with upper tail dependence, we emphasize low rank. A mixture of copulae with both lower and upper tail dependence permits to assign more weight to both extreme ranks.

An example on simulated data illustrates our proposal.

2. *Our proposal of a dissimilarity measure*

Bivariate copula is a function that captures the dependence structure in a bivariate joint bivariate distribution function. Bivariate copula is, in fact, a class of bivariate distributions, whose marginals are uniform on the unit interval. It describes the dependence structure existing across pairwise marginal random variables (rv).

Sklar's theorem (see Nelsen, 2013) shows that every bivariate/multivariate distribution can be written via a copula representation. Let (Y_1, Y_2) be a bivariate rv with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cumulative distribution function (cdf) $F_{Y_1,Y_2}(y_1, y_2; \theta)$, then there always exists a copula function $C(\cdot, \cdot; \theta)$ with $C : I^2 \to I$ such that

$$F_{Y_1,Y_2}(y_1, y_2; \theta) = C\big(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta\big), \quad y_1, y_2 \in \mathbb{R}. \tag{1}$$

If the marginal cdfs are continuous then the copula $C(\cdot, \cdot; \theta)$ is unique. Moreover, if $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous the copula can be found by the inverse of (1):

$$C(u, v) = F_{Y_1,Y_2}(F_{Y_1}^{-1}(u), F_{Y_2}^{-1}(v); \theta) \tag{2}$$

with $u = F_{Y_1}(y_1)$ and $v = F_{Y_2}(y_2)$.

We consider a new family of copulae defining via finite mixtures (Nelsen, 2013). The idea is to create a new very flexible copula by combining two copulae, as follow:

$$C_M(u,v) = \alpha C_1(u,v;\theta_1) + (1-\alpha)C_2(u,v;\theta_2) \tag{3}$$

where $\alpha \in [0,1]$ is the weight of the mixture and C_1 and C_2 are two different copulae, with parameters θ_1 and θ_2, respectively. The two components of the mixture could not be from the same copula family.

We propose to use a mixture of copulae to define the distances between subjects in a hierarchical cluster analysis for ranking data.

We consider two subjects, a and b, expressing their preferences on k objects by rankings.

We consider the mixture of copulae C_M in equation (3) to describe the dependence structure of each pair of latent continuous variables (Y_a^*, Y_b^*) underlying the pair $(Y_a, Y_b) = \{i, j, p_{ij}\}$ for $i, j = 1, 2, .., k$. (Y_a, Y_b) is a bivariate ordinal variable where i and j represent the rank denoting the increasing or decreasing level of the importance assigned to the subjects on the k objects and p_{ij} is the joint frequency with values $1/k$, if the pair (i, j) is observed, and 0, otherwise.

Let F_1 and F_2 the cdfs of Y_a^* and Y_b^*, we assume that each pair (Y_a, Y_b) corresponds to the bivariate discrete random variable obtained by a discretisation of the continuous latent variable $(U = F(Y_a^*), V = F(Y_b^*))$ with support on $[0,1] \times [0,1]$, and cdf given by C_M.

Let $A_{ij} = [u_{i-1}, u_i] \times [v_{j-1}, v_j]$, $i, j = 1, 2, \ldots, k$, be the rectangles defining the discretisation. Let $p_{11}, \ldots, p_{kk}$ be the joint probabilities of the ordinal variables corresponding to the rectangles $A_{11}, \ldots, A_{kk}$.

Let $V_{C_M}(A_{11}), \ldots, V_{C_M}(A_{kk})$ be the volumes of the rectangles under the copula C_M, then there exists a unique element in the family of the mixture of copulae that satisfies the following relationship:

$$(V_{C_M}(A_{11}), \ldots, V_{C_M}(A_{ij}), \ldots, V_{C_M}(A_{kk})) = (p_{11}, \ldots, p_{ij}, \ldots p_{kk}). \tag{4}$$

Given the mixture of copulae C_M that satisfies the (4), we define the Spearman's grade correlation coefficients for the pair (Y_a, Y_b), with $a \neq b$, that performs well in measuring the agreement between two rankings:

$$\rho_S(C_M) = 12 \int_{I^2} [\alpha C_1(u, v; \theta_1) + (1 - \alpha) C_2(u, v; \theta_2)] dudv - 3 \qquad (5)$$

The Spearman's grade correlation coefficients of the convex combination of copulae corresponds to the convex combination of the individual Spearman's rho of the two copulae. Finally, the distance $d_{a,b}$ between the rankings of the subjects a and b is:

$$d^C_{a,b} = \sqrt{1 - \frac{\rho_s(C_M) + 1}{2}} \qquad (6)$$

We calculate the distances in (6) for each pair of n subjects. We propose to use the obtained $n \times n$ matrix as the dissimilarity matrix in a hierarchical cluster analysis.

By using (6) and the mixture of copulae in hierarchical cluster analysis, we can analyze different situations in the classification process.

For example, we consider three families of Archimedean copulae with different characteristics: Gaussian, Clayton, and Gumbel copula. The Gaussian copula is a symmetric copula that permits positive and negative correlation between the variables and does not allow the dependence in the tails. Instead, Clayton and Gumbel copulae are asymmetric. They permit only positive association and exhibit, respectively, strong left (lower) and right (upper) tail dependence.

By choosing only Gaussian copulae in the mixture, we assign the same weight to all ranks. It is possible to proof that, in a hierarchical cluster analysis, the use of Gaussian copula or classical Spearman rank correlation coefficient (Spearman approach) gives the same classification. By choosing a mixture of Gaussian and Clayton or Gumbel copulae, we can assign to the ranks different "weights" and emphasize the agreement in particular on the top or lower ranks. Therefore, by choosing a mixture of Clayton and Gumbel copulae, we emphasize the agreement only on extreme ranks.

3. An example on simulated data

In this section, we illustrate our proposal by an application to simulated data, analyzed in Bonanomi et *al.* (2017). The data consist on 10 rankings representing the judgements of 10 consumers about 6 aspects of a product, attributing "1" to the most important aspect and "6" to the least important one, reported in Table 1.

Table 1. Example: rankings of 6 products given by 10 consumers

Consumer	*Rankings*						*Consumer*	*Rankings*					
1	1	2	3	4	5	6	**6**	1	2	3	6	5	4
2	2	1	3	4	5	6	**7**	1	2	3	6	4	5
3	1	2	3	4	6	5	**8**	1	2	4	3	5	6
4	2	1	3	4	6	5	**9**	3	1	2	4	5	6
5	3	2	1	4	5	6	**10**	1	2	3	5	4	6

Our aim is to emphasize the extreme ranks (top ranks, lower ranks, or both simultaneously but with different emphasis as well), to develop a more flexible classification than the classical one obtained by Spearman rank correlation coefficient. To achieve this aim, we implement a hierarchical cluster analysis with a distance measure based on a mixture of Gumbel and Clayton copulae. This mixture allows positive associations between rankings and lower and upper tail dependence.

We performed the cluster analysis by using a complete linkage clustering method. We compare the dendrogram obtained by implementing a hierarchical cluster analysis based on the mixture of copulae with the one obtained by the Spearman rank correlation as similarity measure.

The Spearman approach assigns the same importance (weights) at every rank. The mixture of Gumbel and Clayton copulae with weight $\alpha = 0.5$ (equal weight for every copula) assigns to the ranks different weights emphasizing the agreement only on the extreme ranks.

Referring to Table 1, let consider the consumers **1**, **2**, **3** and **8**. If we address the issue of emphasizing top and lower ranks simultaneously, the preferences of consumers **1** and **8** are more similar than **1** and **2**. Moreover, consumer **1** and **8** would be both separated by consumer **3**.

In Figure 1 we compare the two dendrograms and we show the change of position of the subjects by using the two different approaches.

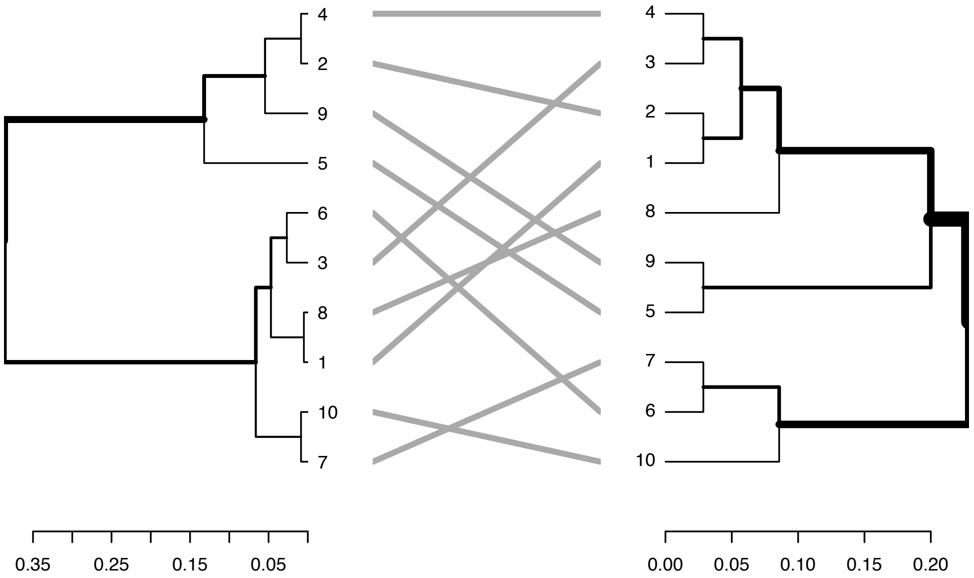


Figure 1. Comparison of dendrograms: Spearman grade correlation coefficient by mixture of copulae (Clayton and Gumbel copulae with weight $\alpha = 0.5$*) (on the left) and Spearman correlation coefficient (on the right)*

The consumers **1** and **8**, whose preferences differ only for the two central ranks, are grouped together at a very low height in the dendrogram obtained by using a mixture (left side of Figure 1), while they are grouped at a greater height in the dendrogram on the right side (Spearman approach).

The consumers **1** and **2**, whose preferences differ only for the two top ranks, are grouped together at a very low height in the dendrogram obtained by using the Spearman approach, while they are grouped at a greater height in the left side.

Moreover, a classification procedure that emphasizes both the top and the lower ranks approaches **1** and **8** and it separates consumer **3**.

In conclusion, the classical approach could be used when one wants to assign equal weights to all ranks in the definition of the distance between rankings. Spearman's grade correlation coefficient ρ_s using the mixture of Gumbel and Clayton copulae gives much more importance on top and lower ranks simultaneously, emphasizing the similarity of consumers with similar extreme ranks.

Acknowledgements: Authors would like to thank the Professor Giuseppe Boari for his useful suggestions.

References

Alvo A., Yu Philip L.H. (2014) *Statistical methods for ranking data*, Springer, New York.

Bonanomi A., Nai Ruscone M., Osmetti S.A. (2017) Defining subjects distance in hierarchical cluster analysis by copula approach, *Quality & Quantity*, 51, 849-872.

Critchlow D.E., Fligner M.A., Verducci, J.S. (1991) Probability models on rankings, *Journal of mathematical psychology*, 35, 294-318.

Mallows C.L. (1957) Non-null ranking models, *Biometrika*, 44, 114-130.

Nelsen R.B. (2013) *An Introduction to Copulas*, Springer, New York.

Spearman C. (1904) The proof and measurement of association between two things, *Am. J. Psychol*, 77-101.

Tarsitano A. (2005) Weighted rank correlation and hierarchical clustering, *Book of Short Paper*, CLADAG2005, Parma.

Measurement of interrater agreement for the assessment of language proficiency

Giuseppe Bove[*], Elena Nuzzo[**], Alessio Serafini[***]

Abstract: A proposal of a procedure to measure interrater absolute agreement for ordinal scales is provided capitalizing on the dispersion index for ordinal variables proposed by Giuseppe Leti. The procedure allows to avoid the problem of restriction of variance that sometimes affects traditional measures of interrater agreement in different fields of application. Rating data on a Likert scale regarding a study of assessment of language proficiency conducted at Roma Tre University are used for a comparison of the new procedure with some known measures of interrater absolute agreement.

Keywords: Interrater agreement, Ordinal scales, Language proficiency.

1. Introduction

Ordinal rating scales (e.g., Likert scales) are frequently developed to evaluate language proficiency in written or oral tasks. The levels of the rating scales have to be defined as clearly as possible, in order to allow their application by both expert and non-expert raters. Before their application, new rating scales are tested out by a group of raters, who assess the language proficiency of a corpus of argumentative (written or oral) texts produced by native and/or non-native writers. When each rater evaluates each writer, the raters provide comparable categorizations of the writers. The extent to which the raters categorizations coincide, the rating scale can be used with confidence without worrying about which raters produced those categorizations. So the main interest here is in analysing the extent that raters assign the same (or very similar) values on the rating scale (*absolute agreement*).

Many methods for measuring agreement among raters have been proposed and applied in many domains in the areas of psychology, education, sociol-

[*] Università Roma Tre, giuseppe.bove@uniroma3.it

[**] Università Roma Tre, elena.nuzzo@uniroma3.it

[***] La Sapienza Università di Roma, alessio.serafini@uniroma1.it

ogy, and medical research (reviews can be found, for instance, in Gwet 2014 and von Eye & Mun 2005). Those based on Pearson product-moment correlation are mainly considered for examining rating *consistency* (i.e., similarity of rank orders produced by the ratings). Other measures like those based on Cohen's Kappa coefficient and intraclass correlation coefficients seem more appropriate for the analysis of absolute agreement.

A problem that can be encountered when measuring interrater consistency or absolute agreement is that of *restriction of variance* (e.g., LeBreton *et al.* 2003), that consists in an attenuation of estimates of rating similarity caused by an artefact reduction of the between-subjects variance in ratings. This can happens in language studies when the same task is administered to native (L1) and non-native (L2) writers, and the analysis compares rater agreement in the two groups separately. Even in the presence of very good absolute agreement, traditional measures (e.g., Cohen's Kappa coefficient and intraclass correlations) can assume low values, especially for L1 group, because the range of ratings provided by the raters are concentrated in one or two very high levels of the scale (a range restriction that determines a between-writer variance restriction). In order to overcome this problem, measures for absolute agreement (or *consensus*) have been proposed (see LeBreton *et al.* 2003) that measure the within-writer variance of ratings (i.e., the between-rater variance) separately for each writer and summarize the results in a final average index (usually normalized in the interval $[0, 1]$). In this approach, the influence of the low level of the between-writer variance is removed by the separate analysis of the ratings of each writer. In the next two paragraphs of this contribution, after a more detailed presentation of the restriction of variance problem, we propose to measure interrater absolute agreement by using the dispersion index proposed by Leti (1983, pp. 290-297) for ordinal variables, in this way taking into consideration the ordinal level of the measurement scales.

2. Interrater absolute agreement and the restriction of variance problem

Measures of interrater absolute agreement like Cohen's *Kappa* (and extensions to take into account three or more raters) and intraclass correlations are usually applied when dealing with rating performed by ordinal scales. A first problem of these procedures is that they are not defined originally for ordinal scales and so have to be adapted. A second major problem is that of the restriction of variance mentioned in the introduction. A small fictitious example concerning scores on a six-point scale given by two raters to ten writers is provided in Table 1. The two raters provide very similar ratings and high absolute agreement, however concentrated in the two higher levels of the scale (level 5 and 6). Traditional interrater absolute agreement measures provide very low values for data in Table 1 (e.g., $Kappa = 0.17$; assuming a two-way random effects model we obtained $ICC(A, 1) = 0.18 -$ here the symbol for intraclass correlation is in accord with McGraw & Wong 1996). The reason for these results is that both indices are influenced by the range restriction of the scale: the $Kappa$ index is computed on the reduced 2×2 contingency table associated with only two levels of the scale; the between-writers variance in the numerator of the intraclass correlation $ICC(A, 1)$ is quite small as a consequence of the range restriction.

Table 1. Ratings on a six-point Likert scale

Writer	Rater 1	Rater 2
1	6	5
2	6	6
3	6	6
4	5	5
5	5	6
6	6	6
7	5	5
8	6	5
9	5	6
10	6	6

To circumvent the problem of low between-writers variance, one of the indices of interrater absolute agreement reviewed in LeBreton & Senter (2008) can be applied. As noted in the introduction, they are based on the idea to measure the within-writer variance of ratings (i.e., the between-rater variance) separately for each writer; in this way they should be insensitive to a lack of

Table 2. Values of the interrater absolute agreement indices

Writer	r_{WG}	CV(%)	d
1	0.91	9.09	0.2
2	1	0	0
3	1	0	0
4	1	0	0
5	0.91	9.09	0.2
6	1	0	0
7	1	0	0
8	0.91	9.09	0.2
9	0.91	9.09	0.2
10	1	0	0
Average	0.97	3.64	0.08

between-writer variance. One of the most popular indices in this group was proposed by James, Demaree and Wolf (1993); for a scale X the computation is given by

$$r_{WG} = 1 - \frac{s_x^2}{\sigma_E^2},$$

where s_x^2 is the observed between-rater variance of the ratings and σ_E^2 is the between-rater variance obtained from a theoretical null distribution representing a complete lack of agreement among raters. Raters in perfect agreement ($s_x^2 = 0$), provide a value $r_{WG} = 1$. In the applications, r_{WG} values greater than 0.7 (possibly 0.8) are considered associated with high level of interrater absolute agreement (see LeBreton & Senter 2008, p. 836 table 3). When the null distribution is assumed as uniform, the equation for the corresponding variance σ_{EU}^2 is

$$\sigma_E^2 = \sigma_{EU}^2 = \frac{A^2 - 1}{12},$$

where A refers to the total number of levels of the scale X. In the second column of Table 2 we reported the values obtained for r_{WG}, along with their average value in the last row of the table. All the values are greater than 0.9 and their average 0.97 testifies a substantial level of interrater agreement.

Very similar information could be obtained simply applying the *coefficient of variation* separately to the ratings of each writer (see the column labelled CV(%) in Table 2).

The index r_{WG} (as other indices reviewed in LeBreton & Senter 2008) allows to avoid the problem of variance restriction but as traditional measures of interrater agreement is defined only for interval data. Besides, depending on the choice of the null distribution, negative values could be obtained. For these reasons, in the next paragraph we propose a new procedure to measure absolute agreement for ordinal rating scales.

3. A descriptive measure of interrater agreement for ordinal scales

The dispersion of an ordinal categorical variable can be measured by the index proposed in Leti (1983, pp. 290-297), given in the following equation:

$$D^* = 2\sum_{k=1}^{K-1} F_k(1 - F_k),$$

where K is the number of categories of the variable and F_k is the cumulative proportion associated to category k. The index is nonnegative and it is easy to prove that $D^* = 0$ when all the observed categories are equal (absence of dispersion). The maximum value of the index (D^*_{max}) is obtained when all observations are concentrated in the two extreme categories of the variable (maximum dispersion), and it is

$$D^*_{max} = \frac{K-1}{2} \quad \text{for N even,}$$

$$D^*_{max} = \frac{K-1}{2}\left(1 - \frac{1}{\mathrm{N}^2}\right) \quad \text{for N odd,}$$

where N represents the total number of observations. For N moderately large, the maximum of the index can be assumed equal to $\frac{(K-1)}{2}$.

So it is possible to define a measure of dispersion normalized in the interval

$[0, 1]$ given by

$$d = \frac{D^*}{D^*_{max}}.$$

Two advantages of this proposal respect to measures of absolute agreement like r_{WG} are that d does not depend by the formulation of a null distribution for normalization and that can never be out of the range $[0, 1]$. The values of d computed separately for each writer in Table 1 (reported in the last column of Table 2), and the corresponding average value 0.08 confirm the substantial level of interrater agreement already revealed by r_{WG}. It is interesting to notice that D^* has properties of within and between dispersion decomposition analogous to the well-known variance decomposition (Grilli & Rampichini 2002).

4. An application to the assessment of language proficiency

To show and compare the performances of the indices of interrater absolute agreement considered in the previous sections on an empirical data set, we have analysed ratings obtained in a research conducted at Roma Tre University (see Nuzzo & Bove 2018, for a detailed description). The main aim of the study was to investigate the applicability of a six-point Likert scale for functional adequacy (an aspect of language proficiency) developed by Kuiken and Vedder (2017) to texts produced by native and non-native writers, and to different task types (narrative, instruction, and decision-making tasks). The scale comprises four subscales, corresponding to the four dimensions of functional adequacy identified by the authors of the scale: content, task requirements, comprehensibility, coherence and cohesion (the reader is referred to Kuiken and Vedder 2017 for a detailed presentation of scales and descriptors). 20 native speakers of Italian (L1) and 20 non-native speakers of Italian (L2) participated in the study as writers. All the texts produced by L1 and L2 writers (120 texts in total for the three tasks) were assessed by 7 native speakers of Italian on the Kuiken and Vedder's six-point Likert scale. The raters did not have any specific experience in judging written texts, and can therefore be categorized as being non-expert. For our purposes, we have selected ratings concerning only the narrative task and the subscale comprehensibility.

The results of the interrater agreement analysis for the subscale are summarized in Table 3, where the intraclass correlation $ICC(A,1)$ and the average values of r_{WG}, CV and d are shown for L1, L2 and total groups. The intraclass correlation $ICC(A,1)$ (that assumes a two-way random effects model) provides a low-moderate level of agreement for the total group of fourty students ($ICC(A,1) = 0.67$). The results for the average values of CV (12.16%) and d (0.22) seem in accord with $ICC(A,1)$, while the average value of r_{WG} (0,87, assuming as uniform the null distribution) highlights a higher level of agreement. When the analysis focuses separately on the two subgroups of L1 and L2 students, results of the L2 group show minor differences respect to the total group while those regarding the L1 group deserve particular attention. Interrater agreement measured by intraclass correlation is very low in the L1 group ($ICC(A,1) = 0.14$). Analysing the dispersion of the ratings given to this subgroup, it comes out that most of the raters used almost exclusively levels 5 and 6 of the scale. So, as in the fictitious example of Section 3, this range restriction caused the very low value of the intraclass correlation, despite the substantial agreement among the raters that scored all the L1 texts in the same high levels. This problem does not regard the results for the other three indices of Table 3 ($r_{WG} = 0.90$; $CV = 8.12\%$; $d = 0.17$) that show a very good level of absolute agreement.

Table 3. $ICC(A,1)$ and averages of r_{WG} , CV and d for the comprehensibility subscale in the L1, L2 and the Total groups

Group	N	$ICC(A,1)$	r_{WG}	CV(%)	d
L1	20	0.14	0.90	8.12	0.17
L2	20	0.63	0.84	16.20	0.28
Total	40	0.67	0.87	12.16	0.22

5. Conclusions

This contribution provides a new procedure to measure interrater absolute agreement for ordinal scales, capitalizing on the dispersion index for ordinal variables proposed by Leti (1983). Preliminary results obtained by applying the procedure to data regarding a study of assessment of language proficiency conducted at Roma Tre University seem encouraging. The procedure is not affected by the restriction of range of the ratings assigned by the raters, in accord with some other measures of interrater agreement for interval data, with the further advantage that it does not need the assumption of a null distribution to be computed. A number of issues needs to be investigated in future research, including: a study of the sampling properties of the proposed procedure; the possibility to define further measures of interrater absolute agreement, capitalizing on the dispersion decomposition provided in Grilli & Rampichini (2002); comparison with methods based on the latent variable approach (e.g., Raykov *et al.* 2012).

References

Grilli L., Rampichini C. (2002) Scomposizione della dispersione per variabili statistiche ordinali [Dispersion decomposition for ordinal variables], *Statistica*, 62, 111-116.

Gwet K.L. (2014) *Handbook of inter-rater reliability (4-th ed.)*, Advanced Analytics, LLC, Gaithersburg MD, USA.

James L. J., Demaree R. G., Wolf G. (1993) r_{wg}: An assessment of within-group interrater agreement, *Journal of Applied Psychology*, 78, 306-309.

Kuiken F., Vedder I. (2017) Functional adequacy in L2 writing. Towards a new rating scale *Language Testing*, 34, 321-336.

LeBreton J.M., Burgess J.R.D., Kaiser R.B., Atchley E.K., James L.R. (2003) The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar?, *Organizational Research Methods*, 6, 80-128.

Leti G. (1983) *Statistica descrittiva*, Il Mulino, Bologna.

McGraw K.O., Wong S.P. (1996) Forming inferences about some intraclass correlation coefficients, *Psychological Methods*, 1, 30-46.

Nuzzo E., Bove G. (2018) Assessing functional adequacy across tasks: A comparison of learners and native speakers' written texts, (*submitted for publication*).

Raykov T., Dimitrov D.M., von Eye A., Marcoulides G.A. (2012) Interrater agreement evaluation: a latent variable approach, *Educational and Psychological Measurement*, 73, 512-531.

von Eye A., Mun E.Y. (2005) *Analyzing rater agreement. Manifest variable methods*, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Modelling perceived variety in a choice process with nonlinear CUB

Eugenio Brentari*, Marica Manisera**, Paola Zuccolotto***

Abstract: In consumer research, marketing, public policy and other fields, individuals' choice depends on the number of possible alternatives. In addition, according to the literature, the choice satisfaction is influenced not only by the number of options but also by the perceived variety. The aim of the present study is to apply a novel statistical approach to model perceived variety, in order to better understand the perceptions of individuals about the variety of the possible choice options. We resort to the class of CUB (Combination of Uniform and Binomial random variables) models, in particular to the Nonlinear extension of CUB, in order to (*i*) provide a measure for perceived variety, (*ii*) add a measure of uncertainty, (*iii*) give insights on the state of mind of respondents toward the response scale. The application of the Nonlinear CUB to real data previously published shows interesting results.

Keywords: Consumer choice, Rating data, Transition probabilities.

1. Introduction

According to several studies in the field of human judgement and decision making, the number of options in a choice process influence choice satisfaction. Some authors support the existence of a phenomenon called "choice overload" or "overchoice", occurring when several - approximately equivalent - options are available and individuals find it difficult to make a decision and report lower choice satisfaction. Satisfaction with choices can be described as an inverted U-shaped function of the number of options: choice satisfaction increases as the number of options increases, but after a certain point starts to decrease. Actually, choice overload is a debated issue and the research has been more recently also focused on understanding why choice overload

*University of Brescia, eugenio.brentari@unibs.it

**University of Brescia, marica.manisera@unibs.it

***University of Brescia, paola.zuccolotto@unibs.it

occurs (Chernev et al., 2015). This topic is relevant in any field where individuals' choice depends on the number of possible alternatives: for example, in consumer research, when marketing managers must decide how many types to include in a product line, or when retailers must decide how many brands of the same product to place on shelves, or in public policy, when public policy agents must decide how many and which alternatives to offer to citizens, for example health plans (Szrek, 2017). Having more choices is advantageous initially and variety is positive. Nevertheless, too many choices increases complexity, that is often considered negative for choice because it can drive individuals to delay their decision or even be indecisive. From a psychological point of view, however, choice satisfaction is influenced not only by the number of options: instead, it is the perceived variety that plays a crucial role (Szrek, 2017). Unlike the number of options, perceived variety is not easy to measure: individuals perceive easily the number of options, while the actual variety, that is the true level of variety of an assortment (for example, assortment of a product category in a supermarket), may be difficult for individuals to perceive correctly. While the literature in marketing and consumer research fields is more interested in revealing the relationships among perceived costs, perceived benefits and choice satisfaction (both choice process satisfaction and choice outcome satisfaction), the aim of the present study is to apply a novel statistical approach to model perceived variety, in order to better understand the perceptions of individuals about the variety of the possible choice options. Like other individuals' perceptions, perceived variety is often investigated by means of questionnaires, composed of questions (items) with ordered response categories (ratings). The resulting rating data can be modelled by means of several statistical methods. Among them, a very interesting class of models, called CUB (Combination of Uniform and shifted Binomial), has been proposed by Piccolo (2003) and D'Elia and Piccolo (2005). In the CUB framework, the respondents' psychological decision process is interpreted as a combination of two latent components, called *feeling* and *uncertainty*, that express, respectively, the level of agreement with the item being evaluated and the human indecision surrounding any discrete choice. Thanks to a very productive research group headed by Domenico Piccolo and their fruitful collaborations with several other researchers, CUB models have been applied widely and further developed and extended (see references in Iannario and Piccolo, 2016). Among these developments, in this paper we focus on the Nonlinear CUB model (NLCUB; Manisera and Zuccolotto, 2014), introduced as a possible generalization of standard CUB based on the idea that the response categories can be "unequally spaced" in the respondents' perception.

In this paper, we apply the NLCUB model to the real data analysed in a previous study (Szrek, 2017) focused on number of options, perceived variety and choice satisfaction. The aim, as mentioned before, is to model the perceived variety in order to deepen its understanding, by (*i*) providing a measure for perceived variety, obtained by a model appropriately conceived for rating data, (*ii*) adding a measure of uncertainty, (*iii*) giving insights to the possibly unequal spacing among response categories in the respondents' mind. The paper is organised as follows. In Section 2, CUB and NLCUB models are briefly recalled, while Section 3 describes the results and gives some concluding remarks.

2. *CUB and Nonlinear CUB models*

CUB and NLCUB models assume that the response of each individual to a given item with a response scale of m ordered categories is the combination of a *feeling* attitude (agreement) towards the item and an intrinsic *uncertainty* component surrounding the discrete choice. Both models fit rating data by means of a mixture of two random variables (r.v.) V and U, aimed to model the feeling and the uncertainty component, respectively.

In CUB models, the distribution probability of the discrete r.v. *R* generating the observed ratings r ($r = 1, \dots, m$) is given by

$$Pr(R = r; \theta) = \pi Pr(V(m, \xi) = r) + (1 - \pi) Pr(U(m) = r)$$

with $\theta = (\pi, \xi)'$, $\pi \in (0, 1]$, $\xi \in [0, 1]$, and where, for a given m, $V(m, \xi)$ is the Shifted Binomial r.v., with trial parameter m and success probability $1 - \xi$ and U is a discrete Uniform r.v. defined over the support $\{1, \dots, m\}$; $1 - \xi$ and $1 - \pi$ are called *feeling* and *uncertainty* parameters, respectively.

In the NLCUB model, while the uncertainty is still modelled by a discrete Uniform, the r.v. modelling feeling is defined differently from CUB. Starting from a given value $T \geq m-1$ and a set of values $k_0, k_1, \cdots, k_m$, such that $0 = k_0 \leq k_1 \leq \cdots \leq k_m = T+1$, the discrete r.v. *R* generating the observed rating r has the probability distribution given by

$$\begin{aligned} Pr(R = r|k_0, k_1, \cdots, k_m; \theta) = \pi & \sum_{v=k_{r-1}+1}^{k_r} Pr(V(T+1, \xi) = v) \\ & + (1-\pi) Pr(U(m) = r). \end{aligned}$$

Usually, the values $g_s = k_s - k_{s-1}$ are used, because of their easy interpretation in terms of the decision process; each $g_s = k_s - k_{s-1}$ indicates the number of terms in the sum in the previous formula for $r = s$ (Manisera and Zuccolotto, 2016).

For a given g $= (g_1, \ldots, g_m)'$, the previous probability distribution can be written as

$$Pr(R = r|k_0, k_1, \cdots, k_m; \theta) = \pi \sum_{i=g_0+\cdots+g_{r-1}}^{g_0+\cdots+g_r-1} \binom{T}{i} (1-\xi)^i \xi^{T-i} + \frac{1-\pi}{m}$$

where $g_0 := 0$ and $T = g_1 + \ldots + g_m - 1$.

The standard CUB model is a special case of NLCUB with $T = m-1$ and $g_s = 1$ for all $s = 1, \ldots, m$. Manisera and Zuccolotto (2015a, 2017) investigated the conditions for NLCUB model identifiability and proposed the use of the EM algorithm for parameter estimation.

An interesting feature of the NLCUB is the possibility to define the so-called transition probabilities. It is worth recalling that the NLCUB model is derived as a special case of a general framework proposed to describe the Decision Process (DP) driving individuals to answer questions with ordered response categories (Manisera and Zuccolotto, 2014). Very briefly, according to this idea, two decision approaches coexist in the DP, the feeling and the uncertainty approach. The final rating expressed can derive from a feeling or an uncertainty approach, with given probabilities.

The feeling approach proceeds through T consecutive steps. At each step

t, the individual gives a provisional rating r_t, which updates the one given at previous steps, until step T is reached: the last rating r_T is the rating generated by the feeling approach. The transition probabilities are then defined as $\phi_t(s) = Pr(R_{t+1} = s+1|R_t = s)$, $s = 1, \ldots, m-1$ and describe the respondents' state of mind about the response scale used to express judgments in the feeling approach. The decision process has been defined to be linear or nonlinear according to whether the transition probabilities $\phi_t(s)$ are constant on non-constant for different t and s. A graphical representation of $\phi_t(s)$, called "transition plot", can be easily constructed (Manisera and Zuccolotto, 2014, 2016), with the response scale on the x-axis and the corresponding perceived ratings on the y-axis. The shape (linear or not) of the resulting piecewise linear curve indicates whether the DP is linear or nonlinear.

Starting from the transition probabilities, we can also define the expected number μ of one-rating-point increments during the feeling path and the unconditional probability of increasing one rating point in one step of the feeling path $\phi = \mu/T$. In the CUB models, the transition probabilities are constant over t, s and equal to $\phi_t(s) = 1 - \xi$, for all t, s, with $s = 1, \ldots, m-1$, $t = 1, \ldots, m-1$ and $\phi_0(1) := 1 - \xi$. In addition, we also have $\phi = 1 - \xi$ and $\mu = (m-1)(1-\xi)$. In other words, the CUB models imply a linear DP and the feeling parameter $1 - \xi$ indicates the probability of increasing one rating point in one step of the feeling path. On the contrary, NLCUB is able to model nonlinear decision processes, as it allows non-constant transition probabilities, given by

$$\phi_t(s) = (1-\xi)\frac{\binom{t}{k_s - 1}(1-\xi)^{k_s-1}\xi^{t-k_s+1}}{\sum\limits_{i=k_{s-1}}^{k_s-1}\binom{t}{i}(1-\xi)^i\xi^{t-i}} \qquad t < T, \quad s = 1, \ldots, m-1$$

with $k_{s-1} \leq t < T$, $\phi_0(1) := 0$ if $g_1 > 1$ and $\phi_0(1) := 1 - \xi$ if $g_1 = 1$; g_s play a fundamental role in the noncostantness of the transition probabilities. Linear DPs can also be modelled with NLCUB, because when $T = m - 1$ and $g_s = 1$ for all s, NLCUB collapses to the traditional CUB. In NLCUB, the expected number of one-rating-point increments during the feeling path is given by

$$\mu = \phi_0(1) + (1-\xi)\sum_{t=1}^{T-1}\sum_{s=1}^{m-1}\binom{t}{k_s-1}(1-\xi)^{k_s-1}\xi^{t-k_s+1}$$

and $1+\mu$ is the expected rating of the feeling approach, without the effect of the uncertainty approach. With NLCUB models, μ is used as feeling parameter in place of $1-\xi$, because it allows the comparison among NLCUB models having different g, while $1-\pi$ is still the uncertainty parameter.

3. Understanding choice perceived variety: a case study

Data refer to an experiment with prescription drug plans and are available as supplementary material of Szrek (2017) downloadable from the website of "Judgment and Decision Making", the journal of the Society for Judgment and Decision Making (SJDM) and the European Association for Decision Making (EADM). Data come from a survey involving 545 individuals, who have been randomized to a set of 2, 5, 10 or 16 drug plan options and asked to select one plan from the set shown to them. In addition, they were asked to rate their perceived variety, answering the question "Do you think that the selection should have included a greater variety of plans?", with responses on a 1-7 scale (1=I had too little variety; 4=I had the right amount of variety; 7=I had too much variety), besides some other information about outcome and process satisfaction, perceived benefits and costs, and individual characteristics (gender, age, education).

NLCUB was used to model the perceived variety, separately for individuals assigned to plans with 2, 5, 10 and 16 options. The goodness-of-fit indices are very good for all the four models: the dissimilarity index (a normalized index measuring the distance between the observed and the estimated frequencies) ranges from 0.02 (5 options) to 0.12 (10 options). Figure 1 represents the perceived variety in a unique representation (Manisera and Zuccolotto, 2015b): for each of the four groups of respondents, the corresponding NLCUB is represented by a very small and stylized transition plot positioned according to the estimated measure of uncertainty ($1-\hat{\pi}$, *x*-axis) and feeling ($\hat{\mu}$, *y*-axis) in the parameter space $[0,1) \times [0, m-1]$.

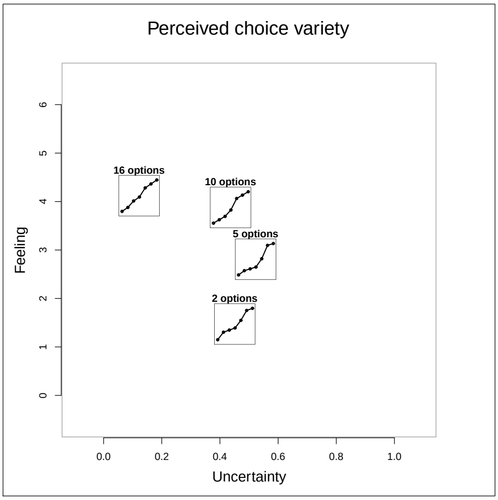


Figure 1. Transition plots for 2, 5, 10 or 16 drug plan options

The different locations of the four groups on the y-axis indicate that the perception of the amount of variety varies with the number of options: as the number of options increase, the amount of variety perceived increase. The estimated μ equals 1.59, 2.82, 3.81, 4.04 for the four groups with 2, 5, 10 and 16 options, respectively. As expected, respondent with few options perceived variety being less than they wanted, while, on the contrary, respondents with more options perceived variety being more than they wanted. We are able to add a measure of uncertainty to this result already highlighted in (Szrek, 2017). The uncertainty (x-axis in Figure 1) associated with respondents with 16 plans is very low (0.15), while the other three groups show a moderate level of uncertainty (0.45, 0.52, 0.44 for the groups with 2, 5 and 10 options, respectively). Another interesting insight we can add to this results is given by interpreting the transition plots (the non-stylized plots are not shown here due to space constraints). The degree of nonlinearity can be measured by a normalized index (Manisera and Zuccolotto, 2013) that equals 0.43, 0.44, 0.33, 0.26 in the plans with 2, 5, 10 and 16 options, respectively. The group facing only 2 or 5 options show similar convex transition plots, suggesting that respondents find it difficult moving from category 4 to 5 and from 5 to 6. Indeed, the majority of respondents of both groups (77% and 69%) gave a rating equal to or lower than 4. The main difference between the two groups is that the mode category is 1 for the group with 2 options and 4 for the group with 5 options. This is reflected in the transition plot, showing that individuals with only 2 options find it difficult moving also from category 1 to 2. The transition plot of individuals with 10 options shows a lower level of nonlinearity, with a higher difficulty of moving from category 4 to 5. The transition plot of individuals with 16 options, instead, is pretty linear, suggesting that it is almost equally difficult (or easy, since the estimated feeling $\hat{\mu}$ is 4.04) moving from one category to the next one. Further studies will be focused on the relationship between perceived variety (feeling but also uncertainty measures) on one hand and choice satisfaction (both in terms of process and outcome satisfaction) and perceived costs and benefits on the other hand. In addition, it could be interesting to examine the effect of individuals' characteristics on such relationship.

References

Chernev A., Böckenholt U., Goodman J. (2015) Choice overload: A conceptual review and meta-analysis, *Journal of Consumer Psychology*, 25, 333-358.

D'Elia A., Piccolo D. (2005) A mixture model for preference data analysis, *Comput Stat Data An*, 49, 917-934.

Iannario M., Piccolo D. (2016) A comprehensive framework of regression models for ordinal data, *Metron*, 74, 233-252.

Manisera M., Zuccolotto P. (2013) Nonlinear CUB models: some stylized facts, *QdS - J Methodol Appl Statist*, 1-2.

Manisera M., Zuccolotto P. (2014) Modeling rating data with Nonlinear CUB models, *Comput Stat Data An*, 78, 100-118.

Manisera M., Zuccolotto P. (2015a) On the identifiability of Nonlinear CUB models, *Journal of Multivariate Analysis*, 140, 302-316.

Manisera M., Zuccolotto P. (2015b) Visualizing multiple results from Nonlinear CUB models with R grid viewports, *Electronic Journal of Applied Statistical Analysis*, 8, 360-373.

Manisera M., Zuccolotto P. (2016) Treatment of 'don't know' responses in a mixture model for rating data, *Metron*, 74, 99-115.

Manisera M., Zuccolotto P. (2017) Estimation of Nonlinear CUB models via numerical optimization and EM algorithm, *Commun Stat-Simul Comput*, 46, 5723-5739.

Piccolo D. (2003) On the moments of a mixture of uniform and shifted binomial random variables, *Quad Stat*, 5, 85-104.

Szrek H. (2017) How the number of options and perceived variety influence choice satisfaction: An experiment with prescription drug plans, *Judgm Decis Mak*, 12, 42-59.

A flexible distribution to handle response styles when modelling rating scale data

Roberto Colombi[*], Sabrina Giordano[**]

Abstract: It is commonly known that the respondents to rating scale questions, when are not aware, can select their own response using only certain response categories regardless the item content. This behavior is described as response style. Thus, the observed response can be a real opinion or dictated by a response style behavior. Marginal models (HMMLU) for multivariate responses by Colombi et al., 2018, enables us to distinguish these two behaviors and allows to specify the distributions of uncertain responses. We extend the class of HMMLU models with a new family of discrete distributions whose two parameters allow the uncertain distributions to be U-shaped, bell-shaped, unimodal, symmetric, skewed or uniform, for capturing different response styles.

Keywords: Mixture models, Latent variables, Marginal models.

1. Introduction

Questionnaires with rating scale items are widely used in psychological, social or marketing surveys to measure opinions, interests, or attitudes. In such contexts, it is commonly observed that a respondent, when in doubt, may consistently use only a few of the given options irrespective of his/her opinion. Someone may skip the endpoints, others have tendency to mark the extremes or the middle category (extreme or midpoint response styles), others respond with agreement/disagreement (acquiescence) regardless of item content, optimists may overvalue their feelings and pessimists may underrate them (one side contraction). The term response style indicates this systematic tendency and it is extensively debated in the literature (e.g. Baumgartner and Steenkamp, 2001).

A family of marginal models (HMMLU) for multivariate responses has been introduced by Colombi et al., 2018, to take into account that an ob-

[*] University of Bergamo, colombi@unibg.it

[**] University of Calabria, sabrina.giordano@unical.it

served response can be the real respondent's attitude (aware response) or ensued from a response style. An HMMLU model enables us to specify the distribution of responses due to response styles (called uncertainty distribution) and distinguish it from that of responses dictated by awareness. Uniform and shifted Parabolic probability functions (Colombi et al., 2018) have been used as uncertainty distributions in the HMMLU model. Since the proposed distributions are symmetric or can cope with only one response style, the class of HMMLU models is, in this paper, enriched by a new family of distributions with two shape parameters. This gives the opportunity of choosing among several alternatives of uncertainty distributions (U-shaped, bell-shaped, unimodal, symmetric, skewed, uniform distributions) which capture different response styles. Covariates are also inserted to account for individual differences in response styles.

2. The family of shifted reshaped parabolic distributions

Remind that a probability function $p(i)$, $i = 1, 2, \ldots, m$, of a discrete variable with m levels can be specified by a set of local logits l_i, $i = 1, 2, \ldots, m-1$, as shown below

$$p(1) = \frac{1}{1 + \sum_{j=2}^{m} \exp\{\sum_{i=1}^{j-1} l_i\}},\ p(i) = \frac{\exp\{\sum_{j=1}^{i-1} l_j\}}{1 + \sum_{j=2}^{m} \exp\{\sum_{i=1}^{j-1} l_i\}}, i = 2, \ldots, m.$$

We derive a new family of distributions by a linear transformation of the local logits of the Parabolic random variable with probability function

$$p(i) = \frac{6(m+1-i)i}{(m+2)(m+1)m}, \qquad i = 1, 2, \ldots, m.$$

More precisely the Local Shifted Reshaped Parabolic (LSRP) distribution is specified by the local logits l_i given by the linear transformation

$$l_i = \phi_0 + \phi_1 \log \frac{p(i+1)}{p(i)}, \qquad i = 1, 2, \ldots, m-1.$$

The LSRP distribution family contains, as a special case, the Uniform distribution ($\phi_0 = \phi_1 = 0$), while for negative (positive) values of ϕ_1 it is U-

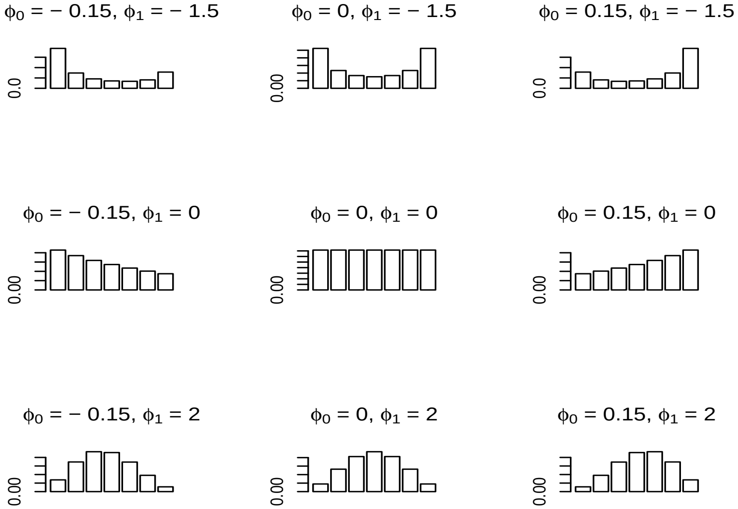


Figure 1. Local Shifted Reshaped Parabolic distributions with different shape parameters

shaped (bell-shaped).

Thus, parameter ϕ_1 rules the frequencies for extreme and middle points. Specifically, high (low) values of ϕ_1 correspond to distributions where indecision leads to focus on middle categories (extreme categories). This allows us to describe adequately extreme and midpoints response styles. Parameter ϕ_0 governs the skewness of the LSRP distributions. In fact, if $\phi_0 = 0$, these distributions are symmetric, right skewed for $\phi_0 > 0$ and left skewed otherwise. Moreover, for a given ϕ_1, LSRP distributions are monotonically ordered, according to the likelihood ratio stochastic ordering as function of ϕ_0. Positive (negative) values enable to capture the acquiescence response style of respondents who tend to endorse the agreement (disagreement) side of the rating scale. Figure 1 shows some examples. Following the same reasoning, distributions like LSRP can be obtained starting from other probability functions independent from unknown parameters or using different logits. The idea of deriving a new distribution by transforming linearly logits of a free parameter discrete probability function is convenient when marginal models are used to fit ordinal data. Alternative uncertainty distributions (e.g. Tutz and Schneider, 2017) could be proposed in the context of HMMLU models but with less advantages in terms of computational ease.

3. A mixture model with LSRP uncertainty distributions

We present the simple bivariate version of the HMMLU model, introduced by Colombi et al., 2018, in a more general extent.

Let R_1 and R_2 be two ordinal variables with support $\{1, 2, \ldots, m_1\}$ and $\{1, 2, \ldots, m_2\}$, respectively. We assume the existence of two binary latent variables, U_l, $l = 1, 2$, such that the respondent answers the l^{th} question according to his/her awareness when $U_l = 1$ or his/her response style when $U_l = 0$. We assume that each observable variable R_l depends only on its latent variable U_l, $l = 1, 2$, and that the observable responses R_1 and R_2 are independent when at least one of them is given under uncertainty. Therefore, the joint distribution of the observable variables is specified by the mixture

$$\begin{aligned} P(R_1 = r_1, R_2 = r_2) &= \pi_{00}\, g_1(r_1, \phi_{01}, \phi_{11})\, g_2(r_2, \phi_{02}, \phi_{12}) \\ &\quad + \pi_{01}\, g_1(r_1, \phi_{01}, \phi_{11})\, P(R_2 = r_2 \mid U_2 = 1) \\ &\quad + \pi_{10}\, P(R_1 = r_1 \mid U_1 = 1)\, g_2(r_2, \phi_{02}, \phi_{12}) \\ &\quad + \pi_{11}\, P(R_1 = r_1, R_2 = r_2 \mid U_1 = 1, U_2 = 1) \end{aligned} \tag{1}$$

for every $r_1 = 1, 2, \ldots, m_1$ and $r_2 = 1, 2, \ldots, m_2$, where $\pi_{ij} = P(U_1 = i, U_2 = j)$, $i = 0, 1$, $j = 0, 1$, are the joint probabilities of the latent variables. Specifically they are the probabilities that both the answers are given with awareness (π_{11}), both with uncertainty (π_{00}) or one with uncertainty and the other one with awareness (π_{01} and π_{10}). Moreover, $g_l(r_l, \phi_{0l}, \phi_{1l})$, $l = 1, 2$, denotes the distribution of responses under uncertainty, which here belong to the family of Shifted Reshaped Parabolic introduced in Section 2. Finally, $P(R_1 = r_1, R_2 = r_2 \mid U_1 = 1, U_2 = 1)$ is the joint distribution of the two aware responses and $P(R_l = r_l \mid U_l = 1)$ are the marginal ones, with $r_l = 1, 2, \ldots, m_l$, $l = 1, 2$. The probabilities π_{ij}, $i = 0, 1$, $j = 0, 1$, are parameterized through two marginal logits λ_l, $l = 1, 2$, measuring the probability of being uncertain on each specific item, plus a log odds ratio λ_{12}. When this parameter is positive, respondents tend to have the same behavior of uncertainty/awareness on the two items. A marginal parametrization is also adopted for the joint probabilities $P(R_1 = r_1, R_2 = r_2 \mid U_1 = 1, U_2 = 1)$. So that, to parameterize the probabilities $P(R_1 = r_1 \mid U_1 = 1)$, and the probabilities $P(R_2 = r_2 \mid U_2 = 1)$, we introduce the vectors $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ of $(m_1 - 1)$ and $(m_2 - 1)$ marginal local logits, respectively. These logits and the vector $\boldsymbol{\eta}_{12}$ of $(m_1 - 1)(m_2 - 1)$ local log odds ratios parameterize the joint distribution $P(R_1 = r_1, R_2 = r_2 \mid U_1 = 1, U_2 = 1)$. As the number of parameters is $m_1 m_2 + 7$ the mixture is not identifiable without further constraints. If a set of covariates accounts for respondents heterogeneity, identifiability can be assured by linear models for the logits λ_l, $l = 1, 2$ and the shape parameters ϕ_{11}, ϕ_{12}, parallel linear models for $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ and by the assumption that ϕ_{01}, ϕ_{02}, λ_{12} do not depend on covariates (see Colombi et al., 2018, for more details). Useful restrictions on $\boldsymbol{\eta}_{12}$ are the conditions of homogeneous and uniform association.

4. An example

We analyse the data from the module on health and care seeking of the European Social Survey (ESS2 2004). Respondents are asked to reply to questions on alternative forms of health care, such as $R_1 = Sex$ (Approve if healthy people use medicines to improve sex life), and $R_2 = Happy$ (Approve if healthy people use medicines to feel happier). The responses are given on a 5-points scale (1 = “strongly approve", 2 = “approve", 3 = “Neither approve nor disapprove", 4 = “disapprove", 5 = “strongly disapprove"). In addition, we consider two explanatory variables: *Gender* (0 = “Male", 1 = “Female") and *Country* (0 = “France" and 1 = “United Kindom").

We believe that the observed responses could be contaminated by some response styles. Thus, some HMMLU models are adapted to the data at hand in order to account for such behavior in the responses. We focus our attention also on detecting whether the shape of the uncertainty distributions varies according to the respondent’s characteristics.

With this aim, we fit HMMLU models specified under different hypotheses on $g_l(r_l, \phi_{0l}, \phi_{1l})$, $l = 1, 2$ in model (1). The parameters are now denoted as $\phi_{0l}^{GC}, \phi_{1l}^{GC}$, with $G, C = 0, 1$, $l = 1, 2$ in the strata identified by the combinations of *Gender* and *Country*. We consider the following uncertainty distributions, *U:* Uniform ($\phi_{0l}^{GC} = 0, \phi_{1l}^{GC} = 0$, $G, C = 0, 1$, $l = 1, 2$); *RP:* Reshaped Parabolic ($\phi_{0l}^{GC} = 0, \phi_{1l}^{GC} = \phi_{1l}$, $G, C = 0, 1$, $l = 1, 2$); *SRP:* Shifted Reshaped Parabolic ($\phi_{0l}^{GC} = \phi_{0l}, \phi_{1l}^{GC} = \phi_{1l}$, $G, C = 0, 1$, $G, C = 0, 1$, $l = 1, 2$); *HRP:* Heterogeneous Reshaped Parabolic ($\phi_{0l}^{GC} = 0, \phi_{1l}^{GC} = \beta_{0l} + \beta_l^G + \beta_l^C$, $G, C = 0, 1$, $l = 1, 2$); *HSRP:* Heterogeneous Shifted Reshaped Parabolic ($\phi_{0l}^{GC} = \phi_{0l}, \phi_{1l}^{GC} = \beta_{0l} + \beta_l^G + \beta_l^C$, $G, C = 0, 1$, $l = 1, 2$). These models are compared in Table 1, where it is evident that model $\mathcal{M}_5$ shows the best fit.

Table 1. Models Comparison

Model	*Unc. distr.*	*loglik*	*n.par.*	*Compared Models*	*LRT*	p-value
$\mathcal{M}_1$	*U*	-9364.435	20			
$\mathcal{M}_2$	*RP*	-9198.537	22	$\mathcal{M}_1$ vs $\mathcal{M}_2$	331.795	0.0000
$\mathcal{M}_3$	*SRP*	-9171.389	24	$\mathcal{M}_2$ vs $\mathcal{M}_3$	54.2966	0.0000
$\mathcal{M}_4$	*HRP*	-9172.123	26	$\mathcal{M}_2$ vs $\mathcal{M}_4$	52.8283	0.0000
$\mathcal{M}_5$	*HSRP*	-9155.784	28	$\mathcal{M}_4$ vs $\mathcal{M}_5$	32.6785	0.0000

In each model the probabilities to give aware answers or to have a tendency towards a response style vary according to *Gender* and *Country* since there is parallel additive effect of the covariates on the logits mentioned in Section 3. The association among aware responses is assumed uniform homogeneous.

According to model $\mathcal{M}_5$ respondents belonging to the four covariate groups behave differently when uncertain. Plots in Figure 2 illustrates the uncertainty distributions for the two items and the four strata. English respondents totally avoid the extremes and take a shelter in the neutral category when the question concerns the admissible use of pils to improve sex performances. In France, instead, for the same question people seem less elusive and concentrate their answers on the positive/intermediate side. As the matter refers happiness, unaware people, both in UK and France, distribute their preferences quite equally from “approve" to “disapprove". In UK, the distribution is a bit more picked on the middle category. Extremes are not totally excluded, especially in France. Men and women show a very similar uncertain behavior in answering both the questions.

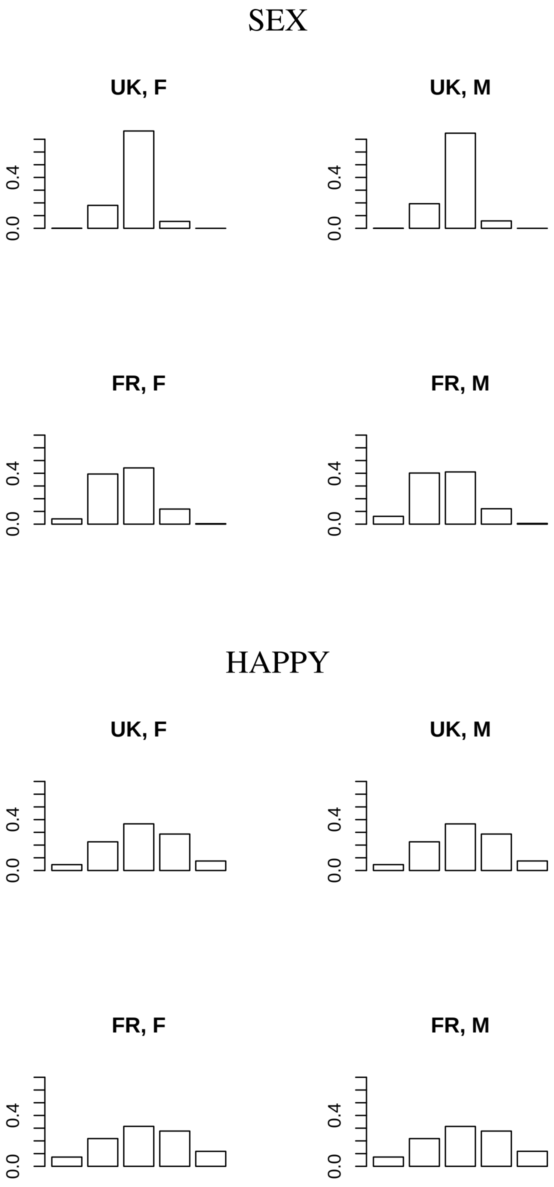


Figure 2. Uncertainty distributions of respondents belonging to the four Gender-Country groups for the two items: Sex and Happy

References

Baumgartner H., Steenkamp J.B. (2001) Response styles in marketing research: A cross-national investigation, *Journal of Marketing Research*, 38, 143-156.

Colombi R., Giordano S., Gottard A., Iannario M. (2018) Hierarchical marginal models with latent uncertainty, arxiv.org/pdf/1607.00882.

Tutz G., Schneider M. (2017) Mixture models for ordinal responses with a flexible uncertainty component, Technical Report Number 203, University of Munich.

Joint modelling of ordinal data: a copula-based method

Marcella Corduas[*]

Abstract: In this article we present an innovative technique to construct a multivariate distribution from margins described by CUB models. In particular, we use the Plackett distribution as a copula function, and we apply the discrete vine pair copula construction method to achieve a computational efficient solution. The proposed approach will be applied to model the importance of three key drivers of extra-virgin oil consumption in Italy.

Keywords: Copula distribution, CUB model, Plackett distribution.

1. Introduction

Likert scales are commonly used when interviewees are requested to rate the importance of certain factors in determining their choices. This type of data deserves special attention because the judgments may depend on covariates characterizing the raters which may be clustered into sub-groups exhibiting more homogeneous choices. In addition, although items are rated individually, the judgements about connected items may be correlated. Thus, the joint modelling of ratings could be useful for better understanding the preferences and choices of the interviewees. In this work, we present an innovative technique for modelling multivariate ordinal data. In particular, moving from the approach discussed by Corduas (2015), we propose to model each component of the multivariate random variable by a CUB model. This is a univariate mixture distribution defined by the convex combination of a discrete Uniform and a shifted Binomial distribution that has been widely investigated in recent years, and fruitfully applied to various fields (Piccolo, 2003). Then, we use the Plackett distribution as a copula to estimate a bivariate distribution from given margins. Finally, we derive the multivariate distribution by means of the discrete pair copula construction (PCC) algorithm. This is a computational efficient procedure based exclusively on the use of bivariate copulas. In such a way, the estimation procedure becomes feasible even when the number

[*] University of Naples Federico II, corduas@unina.it

of dimensions increases.

The plan of the article is the following. In Section 2 we introduce the model for the marginal distribution. Then, in Section 3 we discuss the problems related to the estimation of a joint distribution using a discrete copula. Finally, in Section 4 the proposed technique is applied to the study of key drivers of extra virgin olive oil consumption in Italy. Some final remarks conclude the article.

2. The model for margins

Evaluation data originated by rating a given item can be modeled by means of the CUB distribution (Piccolo, 2003):

$$Pr(X=x;\boldsymbol{\theta}_x)=\pi_x\left(\begin{array}{c} m-1 \\ x-1 \end{array}\right)(1-\xi_x)^{x-1}\xi_x^{m-x}+(1-\pi_x)\frac{1}{m}, \qquad x=1,2,...,m.$$

where $\boldsymbol{\theta}_x = (\pi,\xi)' \in \Omega(\boldsymbol{\theta}_x)$ and the parameter space $\Omega(\boldsymbol{\theta}_x) = \{(\pi,\xi) : 0 < \pi \leq 1, 0 \leq \xi \leq 1\}$ is the (left-open) unit square. We will refer to this probability mass distribution (*pmf*) as $X \sim CUB(\pi,\xi)$. The model mimics a simplified choice mechanism which is supposed to underly the moulding of the judgements when a rater is requested to express preferences, degree of satisfaction about a certain item or, generally speaking, the agreement with a given statement by means of a Likert scale (Iannario and Piccolo, 2016 and therein references). The condition $m > 3$ guarantees that the model is statistically identifiable (Iannario, 2010). The interest for CUB model relies on its flexibility in representing observed data by means of a parsimonious formulation and on the fact that the interpretation of the estimated parameters can be easily found. In particular, $(1-\pi_x)$ is a measure of the degree of *uncertainty* that affects the rater's judgements whereas $(1-\xi_x)$ describes the strength of attraction (*feeling*) that the rater feels towards the item under evaluation. Parameters may be related to explanatory variables characterizing raters by means of a logistic link function, but this aspect will not be considered in the present work. The model has found application to various fields of analysis, including linguistics, social analysis, economics and medicine, and has been extended in order to take the presence of a shelter effect, dominant preferences, random effects of the components, "don't know" answers into account (for a review, see Piccolo 2018).

Some recent contributions have investigated the problem of modelling multivariate ordinal data with CUB margins (Corduas, 2011, 2015; Andreis and Ferrari, 2013; Colombo and Giordano, 2016). In the same line, in this article, we propose to build a multivariate distribution with given margins using the discrete vine representation with the bivariate Plackett distribution as a possible pair copula.

3. The joint distribution

Firstly, we briefly recall the method introduced by Plackett (1965) for constructing a one parameter bivariate distribution from given margins. A bivariate Plackett random variable (X, Y) is characterised by the following joint cumulative distribution function (*cdf*):

$$C(F(x), G(y); \psi) = \frac{M(x,y) - [M^2(x,y) - 4\psi(\psi-1)F(x)G(y)]^{1/2}}{2(\psi-1)},$$

where $\psi \in (0, \infty)$. Here, $F(x)$ and $G(y)$ are the pre-defined marginal *cdfs*. Moreover, $M(x,y) = 1 + (F(x) + G(y))(\psi - 1)$. The parameter ψ is a measure of association between X and Y; specifically, $\psi = 1$ implies that X and Y are independent, whereas $\psi < 1$ and $\psi > 1$ refer to negative and positive association, respectively. Although Molenberghs and Lesaffre (1994) has introduced the multivariate Plackett's distribution, this has had limited applicability in practical situations. As matter of facts, its construction is in general rather demanding because the computation burden increases remarkably with the increase of the number of dimensions.

However, the bivariate Plackett's distribution may become the building block for constructing a multivariate discrete distribution where the scalar variables follow a CUB distribution. Specifically, let $\mathbf{Y} = (Y_1, ..., Y_k)'$ be a k-variate discrete random variable. The joint probability can be decomposed as:

$$Pr(Y_1 = y_1, ..., Y_k = y_k) = Pr(Y_1 = y_1 | Y_2 = y_2, ..., Y_k = y_k) \times \\ Pr(Y_2 = y_2 | Y_3 = y_3, ..., Y_k = y_k) \times ....Pr(Y_k = y_k).$$

Introducing a general notation, each term on the right hand side of the above formula is a conditional probability such as: $Pr(Y_j = y_j|\mathbf{V} = \mathbf{v})$ where $\mathbf{V}$ is a subset of random variables in $\mathbf{Y}$. Moreover, this can be written in terms of a copula. Specifically, following Panagiotelis et al. (2012):

$$Pr(Y_j = y_j|\mathbf{V} = \mathbf{v}) = \frac{Pr(Y_j = y_j, V_h = v_h|\mathbf{V}_{\backslash \mathbf{h}} = \mathbf{v}_{\backslash \mathbf{h}})}{Pr(V_h = v_h|\mathbf{V}_{\backslash \mathbf{h}} = \mathbf{v}_{\backslash \mathbf{h}})}$$

and

$$\begin{aligned} &Pr(Y_j = y_j, V_h = v_h|\mathbf{V}_{\backslash \mathbf{h}} = \mathbf{v}_{\backslash \mathbf{h}}) = \\ &\quad \sum_{i_j=0,1} \sum_{i_h=0,1} (-1)^{i_j+i_h} C_{Y_j,V_h|\mathbf{V}_{\backslash \mathbf{h}}}(F_{Y_j|\mathbf{V}_{\backslash \mathbf{h}}}(y_j - i_j|\mathbf{v}_{\backslash \mathbf{h}}), F_{V_h|\mathbf{V}_{\backslash \mathbf{h}}}(v_h - i_h|\mathbf{v}_{\backslash \mathbf{h}})) \end{aligned}$$

where $F_{Y_j|\mathbf{V}_{\backslash \mathbf{h}}}$ and $F_{V_h|\mathbf{V}_{\backslash \mathbf{h}}}$ are the conditional *cdfs* and C is a bivariate copula function. As an illustrative example, we consider the case of a three dimensional variable: (Y_1, Y_2, Y_3) where each scalar random variable Y_i takes values $y_i \in S(Y_i) = \{1, 2, ..., m\}$, m is given and the ordinal scale is such that 1 is associated to the worst judgement and m to the best one. In addition, in order to simplify the notation, whenever possible we drop the reference to the argument of the function. Given a sample of ordinal data, (y_{1r}, y_{2r}, y_{3r}), for $r = 1, 2, ...n$, the estimation algorithm is summarised as follows.

1. A CUB model is fitted to each sample (y_{hr}), for $r = 1, \dots n$, $h = 1, 2, 3$. This yields the marginal models: $CUB_1(\pi_1, \xi_1)$, $CUB_2(\pi_2, \xi_2)$, $CUB_3(\pi_3, \xi_3)$ and the corresponding *cdfs*: F_1, F_2, F_3;

2. Estimate the joint distributions using the Plackett bivariate copula C: $F_{12} = C(F_1, F_2; \psi_{12})$ and $F_{32} = C(F_3, F_2; \psi_{32})$. These yield the evaluation of the corresponding joint *pmf* P_{12} and P_{32};

3. Compute the conditional distributions: $F_{1|2}(y_1|y_2 = i; \psi_{1|2=i})$ and $F_{3|2}(y_3|y_2 = i; \psi_{3|2=i})$, $i = 1, \dots, m$;

4. Estimate the joint conditional distribution functions by means of the copula: $F_{13|2=i} = C(F_{1|2}(y_1|y_2 = i), F_{3|2}(y_3|y_2 = i); \psi_{13|2=i})$, and then, from those, the conditional *pmf* $P_{13|2}$ for $i = 1, \dots, m$;

5. Compute the multivariate *pmf*: $P_{123} = P_{13|2}P_2$.

Note that in each step at most a bivariate copula is needed. The estimation can be performed either by means of the IFM (Inference For the Margins) method (Joe, 1997), or by full maximum likelihood when k is rather small. We briefly illustrate the general framework of the IFM procedure, that can be easily particularized to the specific implementation required by the PCC algorithm. Consider a bivariate copula-based model:

$$G(x, y; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\delta}) = C(G_1(x; \boldsymbol{\alpha}_1), G_2(y; \boldsymbol{\alpha}_2); \boldsymbol{\delta})$$

where G_1 and G_2 are univariate *cdfs* with parameters $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and C is a copula with parameters $\boldsymbol{\delta}$. Given a sample of i.i.d. observations $\{(x_r, y_r),\ r = 1, ..., n\}$, in the first step, the marginal models are separately estimated by maximum likelihood to get $\widehat{\boldsymbol{\alpha}_1}$ and $\widehat{\boldsymbol{\alpha}}_2$. For instance, in the above PCC algorithm, the CUB parameter estimates needed in (1) can be obtained using the EM estimation algorithm illustrated by Piccolo (2006). In the second step of the IFM approach, the function $L(\boldsymbol{\delta}, \widehat{\alpha}_1, \widehat{\alpha}_2)$ is maximized over $\boldsymbol{\delta}$, being:

$$L(\boldsymbol{\delta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \sum_{r=1}^{n} \log(g(x_r, y_r; \boldsymbol{\delta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2))$$

the log-likelihood function for the joint distribution. Furthermore, the asymptotic covariance matrix can be derived from the Godambe information matrix, or by the jackknife method as suggested by Joe (1997).

Finally, going back to the fitted joint distribution, the adequacy is assessed by means of the pseudo-R^2:

$$R^2_{CU} = (1 - \exp(LR_{no}/n))/(1 - \exp(LR_{max}/n)) \tag{1}$$

where $LR_{no} = 2(L_M - L_0)$ and $LR_{max} = 2(L_{max} - L_0)$, being: L_M the maximized log-likelihood value of the considered model, L_0 is the value of the log-likelihood of the null model where independence among the scalar random variables is assumed, L_{max} is the log-likelihood value of the model with a perfect fit (Cragg and Uhler, 1970; Meinel, 2009).

4. An empirical application: extra-virgin olive oil data

As an illustration of the proposed model, we present a study about the perception of Italian consumers on extra virgin olive (EVO) oil quality and, specifically, about the importance of some product attributes on purchase decision. Various authors have investigated the use of the CUB distribution for modelling ratings about food products (see, for instance, Corduas et al., 2013; Capecchi et al., 2016; Iannario et al., 2012), but the joint modelling of consumer ratings is still unexplored. In this regards, we recall that Italy is one of the major producing and consuming countries of olive oil. However, the factors that affect olive oil purchasing behaviour are not clear because consumers are not accustomed to associate the organoleptic properties to quality signals. Numerous contributions have investigated the liking/disliking of consumers about EVO oil focusing on consumption driving factors, such as the perceived health benefits, the importance of the region of origin, the role of sensory cues (Dekhili *et al.*, 2011).

The sample under study consists of ratings given by 1000 Italian consumers belonging to the AC Nielsen panel (Corduas, 2015). Each interviewee was asked to rate the importance of three EVO oil attributes (colour, taste, Italian origin) in determining his/her purchase decision on a 7 point Likert scale (where 1 denoted "not important at all" and 7 "extremely important").

Table 1. Joint distribution of (Colour, Taste, Italian Origin) of EVO oil

	Colour	Taste	Italian Origin
π	0.873 (0.024)	0.469 (0.007)	0.781 (0.022)
ξ	0.308 (0.015)	0.353 (0.031)	0.068 (0.005)
(Colour,Taste)		(Italian Origin, Taste)	
$\psi_{12} = 3.860\,(0.412)$		$\psi_{32} = 1.909\,(0.208)$	
(Colour,Italian Origin$\mid$Taste)			
$\psi_{13\mid 2=1} = 3.881\,(2.135)$		$\psi_{13\mid 2=2} = 4.221\ (1.994)$	
$\psi_{13\mid 2=3} = 6.308\,(2.265)$		$\psi_{13\mid 2=4} = 2.714\,(0.690)$	
$\psi_{13\mid 2=5} = 2.836\,(0.858)$		$\psi_{13\mid 2=6} = 2.864\,(0.777)$	
$\psi_{13\mid 2=7} = 4.913\,(1.572)$			
$L_{max} = -4412.90$	$L_M = -4709.56$	$L_0 = -6034.39$	$R^2_{CU} = 0.96$

Note: standard errors in parentheses

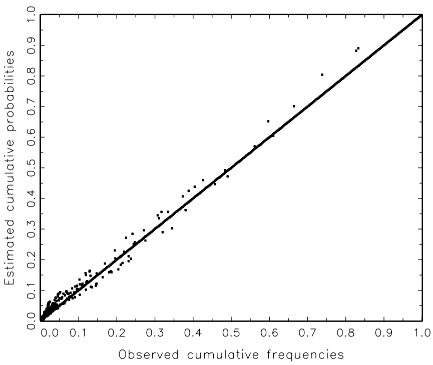


Figure 1. Observed cumulative frequencies vs estimated cumulative probabilities

Since there are only few pairs of copula in a three-dimensional model, it is possible to estimate all possible models (by changing the conditioning variable), and select the model that achieves the best fit. In Table 1, we report the results obtained by the IFM method with the jackknife for the estimation of standard errors. The judgements about *colour* and *taste* are positively correlated, and although in a lower extent, the same consideration applies to *taste* and *Italian origin of the olives*. This is probably due to the fact that consumers easily recognise attributes perceived through senses with respect to other features. The judgements about the three considered attributes are well fitted by the joint multiple distribution as shown by the high value of the fitting measure and the plot of the observed cumulative frequencies against the fitted cumulative probabilities (Figure 1).

In conclusion, the approach seems to achieve meaningful results useful for marketers that need to know the product features that are most closely related to actual purchase decisions in order to build effective marketing strategies.

References

Andreis F., Ferrari P. (2013) On a copula model with CUB margins, *QdS. Journal of Methodological and Applied Statistics*, 15, 33-51.

Capecchi S., Endrizzi I., Gasperi F., Piccolo D. (2016) A multi-product approach for detecting subjects' and objects' covariates in consumer preferences, *British Food Journal*, 118, 515-526.

Colombi R., Giordano S. (2016) A class of mixture models for multidimensional ordinal data, *Statistical Modelling*, 16, 322-340.

Corduas M. (2011) Modelling correlated bivariate ordinal data with CUB marginals, *Quaderni di Statistica*, 13, 109-119.

Corduas M. (2015) Analyzing bivariate ordinal data with CUB margins, *Statistical Modelling*, 15, 411-432.

Corduas M., Ievoli C., Cinquanta L. (2013) The importance of wine attributes for purchase decisions: a study of Italian consumers' perception, *Food Quality and Preference*, 28, 407-418.

Cragg J.G., Uhler R.S. (1970) The demand for automobiles, *Canadian Journal of Economics*, 3, 386-406.

Dekhili S., Sirieix L., Cohen E. (2011) How consumers choose olive oil: The importance of origin cues, *Food Quality and Preference*, 22, 757-762.

Iannario M. (2010) On the identifiability of a mixture model for ordinal data, *METRON*, LXVIII, 87-94.

Iannario M., Piccolo D. (2016) A comprehensive framework of regression models for ordinal data, *METRON*, 74, 233-252.

Iannario M., Manisera M., Piccolo D., Zuccolotto, P. (2012) Sensory analysis in the food industry as a tool for marketing decisions, *Advances in Data Analysis and Classification*, 6, 303-321

Joe H. (1997) *Multivariate Models and Dependence Concepts*, London: Chapman & Hall.

Meinel N. (2009) Comparison of performance measures for multivariate discrete models, *Advances in Statistical Analysis*, 93, 159-174.

Molenberghs G., Lesaffre E. (1994) Marginal modelling of correlated ordinal data using multivariate Plackett distribution, *Journal of the American Statistical Association*, 89, 633-644.

Panagiotelis A., Czado C., Joe H. (2012) Pair copula constructions for multivariate discrete data, *Journal of the American Statistical Association*, 107, 1063-1072.

Piccolo D. (2003) On the moments of a mixture of Uniform and shifted Binomial random variables, *Quaderni di Statistica*, 5, 85-104.

Piccolo D. (2006) Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33-78.

Piccolo D. (2018) A new paradigm for rating data models. Palermo: *Proceedings of 49th Scientific Meeting of the Italian Statistical Society*, 1-12.

Plackett R.L. (1965) A class of bivariate distributions, *Journal of the American Statistical Association*, 60, 516-522.

Modeling preferences: beyond the average effects

Cristina Davino[*], Tormod Naes[**], Rosaria Romano[***],
Domenico Vistocco[****]

Abstract: Preference mapping are a collection of multivariate statistical techniques widely used by marketing and R&D divisions to understand which sensory characteristics drive consumer acceptance of goods. These techniques provide a perceptual map of the products based on the so-called sensory dimensions, on which the liking values for each consumer are regressed. This study proposes an innovative preference mapping based on the quantile regression. Using the quantile regression instead of the classical least squares regression allows to explore the whole distribution of the consumer preference. This permits to obtain additional information both at the individual consumer level, analyzing how the preference varies with respect to the different quantiles and at the general level, highlighting on the preference map consumers with homogeneous behaviors with respect to the different quantiles.

Keywords: Preference mapping, Rating data, Quantile regression.

1. Preference mapping

Preference mapping is a collection of multivariate statistical techniques that aim to analyze consumer acceptance of food and beverages products (Meilgaard, Civille, Carr, 2007). There are two different types of these methods, namely *internal preference mapping* and *external preference mapping* (Meullenet, Xiong, Findlay, 2008). Internal preference mapping uses consumer acceptance ratings to determine a multidimensional representation of products and consumers in a common space. External preference mapping (PREFMAP) uses sensory descriptive attribute ratings to obtain a multidimensional representation of products, sensory characteristics and consumers in a common space. PREFMAP is crucial to the food and beverages indus-

[*] University Federico II of Naples, cristina.davino@unina.it

[**] Nofima AS, tormod.naes@nofima.no

[***] University of Calabria, rosaria.romano@unical.it

[****] University of Cassino e del Lazio Meridionale, vistocco@unicas.it

tries to understand which sensory characteristics drive consumer acceptance of goods. This information is used by marketing and R&D divisions to adapt existing products or create new products that meet consumers' expectations.

The most common PREFMAP method consists of a two step procedure that combines principal component analysis (PCA) and least squares regression (LSR) (Naes, Brockhoff, Tomic, 2010). In the first step a *perceptual map* of the products is obtained through a PCA of the product-by-attribute sensory matrix, and the principal components obtained from the analysis are called *key sensory dimensions* (Meilgaard, Civille, Carr, 2007). In the second step, a regression model is used to fit each consumer in the perceptual space. The main assumption is that the preference of each consumer depends linearly on the sensory attributes. Furthermore, as the method is grounded on LSR, it focuses on the average effects of sensory dimensions.

In some situations it is also useful to study the whole distribution of the liking. At this aim, quantile regression (QR) (Koenker, 2005; Davino, Furno, Vistocco, 2013) can be used to provide an estimate of conditional quantiles of the dependent variable instead of conditional mean. QR was recently used in consumer study for relating liking to consumer factors (Davino, Romano, Naes, 2015), and for handling consumer heterogeneity (Davino, Romano, Vistocco, 2018). The aim of this study is to extend the use of QR to the PREFMAP in order to provide additional information, not only about how the sensory dimensions link to consumer preference on average. The classical approach to PREFMAP based on the LSR does not allow to distinguish consumers able to discriminate preferences among products from consumers with uniform liking. The use of QR is advisable to highlight precise consumers, that is consumers with a strong difference in the liking pattern. Applying the classical approach would obscure this information, treating uniform consumers alike the precise ones.

The study is structured as follows: i) the classical approach to PREFMAP based on LSR is presented in Section (2); ii) a new approach based on QR is described in Section (3); iii) results of the proposed method on a case study concerning consumer liking of apple juice are shown in Section (4).

2. External preference mapping by least squares regression

Let X be the sensory matrix ($I \times K$), where the entry x_{ik} is the measured value of product i and sensory attribute k ($i = 1, \ldots, I; k = 1, \ldots, K$). The PCA model to develop a perceptual map based on the sensory characteristics can be written as

$$X = TP^T + E \tag{1}$$

where T is the matrix ($I \times A$) of the *principal component scores* that are linear combinations of the original data X, and P is the matrix ($K \times A$) of the *loading values* that define the contribution of each of the original variables in the computation of the principal components. The matrix E represents random noise and A is the number of components included in the model.

Let Y be the liking matrix ($I \times J$), where the entry y_{ij} is the measured value of product i and consumer j ($j = 1, \ldots, J$). The liking values for each consumers are regressed onto the first *sensory dimensions*, generally the first two PC's (i.e., $A = 2$):

$$y_{ij} = \beta_{j1} t_{i1} + \beta_{j2} t_{i2} + \epsilon_{ij} \tag{2}$$

The final results of this two step procedure provide a perceptual map and a loadings plot. In the first, products are located on the basis of the sensory characteristics, while in the second consumers' preferences are visualized and the direction for their preferences are identified.

3. External preference mapping by quantile regression

QR can provide complementary information to the classical PREFMAP. At this aim, it is introduced in the second step of the previously described procedure, when liking for each consumer is related to the first sensory dimensions. As classical linear regression provides the estimation of the conditional mean of a response variable distribution as a function of a set of predictors, QR provides the estimation of the conditional quantiles of a response variable distribution as a function of a set of predictors. It results that Equation (2) can be generalized to the QR framework as:

$$y_{ij}(\theta) = \beta_{j1}(\theta)t_{i1} + \beta_{j2}(\theta)t_{i2} + \epsilon_{ij} \tag{3}$$

where $(0 < \theta < 1)$. The interpretation of the QR coefficients is analogous to LSR coefficients: they measure the rate of change of the θth quantile of the dependent variable distribution per unit change in the value of a given predictors, holding the others constant. It is potentially possible to estimate an infinite number of regression lines, but in practice a finite number is numerically distinct, which is known as the quantile process. In practice, it is quite common that each researcher defines the quantiles of interest which, in most cases, are the three quartiles. For each quantile of interest, a regression line is estimated and, consequently, a set of coefficients and a fitted response vector can be obtained.

With respect to each consumer, the introduction of QR in PREFMAP provides a set of coefficients for each quantile of interest. This information allows to measure what is the impact of a change in the sensory dimensions on the liking for the most and least preferred products. Note that consumers showing large differences between coefficients at two extremes quantiles correspond to consumers with a precise liking pattern. With respect to the whole panel of consumers, QR allow to obtain a consumer loading plot that visualizes groups of consumers who are similarly affected by a given change on the sensory dimensions. In the case study section, it is also suggested a conjoint representation able to simultaneously represent results related to two opposite quantiles (e.g $\theta = 0.25$ and $\theta = 0.75$).

4. Case study

The data used for this study have been obtained from the article by Rdbotten *et al.* (2009). Apple juice samples were selected according to an experimental design (a 2*3 factorial design) with two levels of acid concentration (H=high, L=low) and three levels of sugar concentration (H=high, M=medium, L=low). The 6 samples were tested by 125 consumers using the 9-point hedonic scale (Peryam and Pilgrim, 1957). Descriptive sensory analysis was also carried out, and details of the procedure are given in (Rdbotten *et al.*, 2009). Results from classical PREFMAP are given in Figure (1). A joint interpretation of the

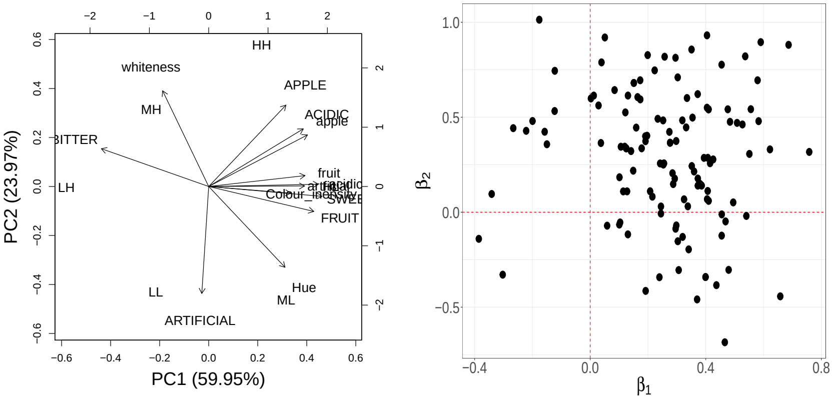


Figure 1. External preference mapping results

two plots shows that almost all consumers prefer sweet products, but some of them prefer products with high acid content, while the others prefer a low acid content. Note that products were evaluated both for flavor (descriptors labelled in capital letters) and smell (descriptors labelled in lower case).

4.1. Exploiting QR for single consumers

Consider estimating a QR model for each individual consumer and for a set of quantiles of interest, that is the three quartiles (θ = [0.25, 0.5, 0.75]). Three sets of coefficients are then estimated for each consumer. Figure 2 shows QR coefficients for two consumers, namely C27 (left-hand plot) and C49 (right-hand plot). For each plot, each panel represents a single regression coefficient (for sake of brevity, the the intercept is not shown). The horizontal axis displays the different quantiles, while the effect of each regressor holding the other constant is represented on the vertical axis. Standard errors and confidence intervals can also be added to the graph. For consumer C27 the two PCs have a different impact on the liking of C27: the coefficients β_1 are always higher than coefficients β_2 that are even negative. As discussed in Section (3), a regression coefficient at a specific quantile provides information of the effect of predictors on the selected conditional quantile of the liking distribution. For instance, C27 shows a $\beta(0.25)$ equal to 0.5 while the $\beta(0.75)$ is larger. This means that the effect of the predictors on the conditioned upper part of the liking distribution is stronger: increasing the level of sweetness (positive verse of PC1) increases more the preference for the most preferred products than for the less liked ones. Considering the lines related to the β_1 coefficients, it results that, modifying the sensory attributes explaining the first PC, has always positive impact on the liking of C27 but, moving from lower to higher quantiles, the effect of PC1 on liking increases showing that the most preferred products could take more advantage of an increment of the sensory attributes correlated to PC1. The opposite holds for β_2. Figure 2 (right-hand side) shows results for another consumer (C49). Here, the sensory dimensions not only have a different size compared to the different quantiles, but also a different sign. An increase in the level of sweetness (PC1) would increase the preference for the less preferred products ($\theta < 0.50$), while it would reduce the preference for the most preferred ones ($\theta = 0.75$).

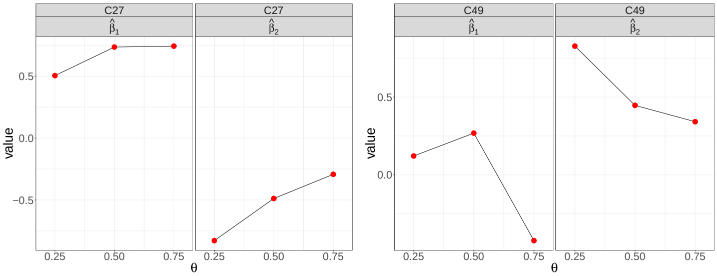


Figure 2. QR coefficients for single consumers

4.2. Exploiting QR for the whole panel

One of the main strengths of preference mapping is to suggest possible drivers to increase the liking, taking into account that whatever action will have different impacts on different groups of consumers.

Exploiting the proposed quantile approach, a further representation is proposed to simultaneously provide QR results at the two extreme considered quantiles. Figure 4 represents a group of 11 consumers. They have been selected as sample consumers because they shows different behaviors. Each

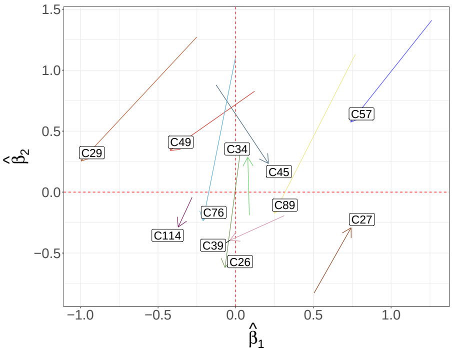


Figure 3. QR loading plot considering two extreme quantiles

consumer is represented according to the β_1 and β_2 coefficients estimated at the two quantiles. The two points representing each consumers are linked by an arrow depicted in the direction from $\theta = 0.25$ to $\theta = 0.75$. Considering consumer number 57 (from now C57), it is possible to appreciate that both coefficients related to PC1 and PC2 are positive but any action on the variables correlated to them will have a higher impact on the liking of the less preferred products. Arrows crossing two quadrants represent consumers with non-concordant signs at $\theta = 0.25$ and $\theta = 0.75$. It is the case of consumer C49, previously discussed. It is worth to note that consumers able to discriminate preferences among products are represented by a longer arrow, than consumers with uniform liking.

Finally, a plot that combines the results of the LSR and QR approach to PREFMAP is shown in the Figure (4). The different symbols correspond to all the different possible directions for the arrows in the previous plot. Specifically, the symbols corresponding to two equal numbers indicate consumers who are located in the same quadrant since coefficients of the two quantiles with respect to the two components have the same signs. For instance, consumer C27 included in the (4,4) group, has β_1 coefficients both positive for the two quantiles, and β_2 both negative. While consumer C49, included in the (1,2) group, has coefficients at $\theta = 0.25$ in the first quadrant and coefficients at $\theta = 0.75$ in the second quadrant. The information provided by this plot is very important because it allows to visualize the variability of preferences

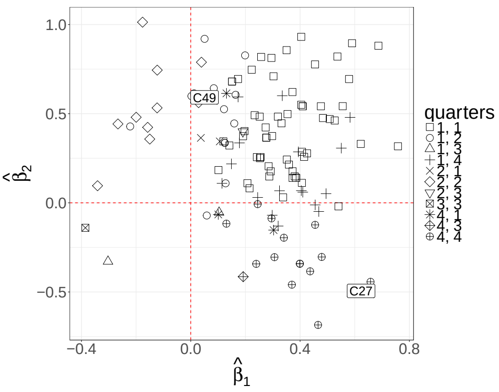


Figure 4. Loadings plot combining LSR and QR approach

with respect to preference directions. As an example, if we consider the direction of maximum preference, i.e. consumers in the first quadrant, we note that not all of them have coefficients consistent with the different quantiles. Consumers labeled with a cross show discrepancies. For a detailed analysis of these discrepancies we must then consider the arrows plot in Figure (4).

References

Davino C., Romano R., Naes T. (2015) The use of quantile regression in consumer studies, *Food quality and preference*, 40, 230-239.

Davino C., Furno M., Vistocco D. (2013) *Quantile regression: theory and applications*, John Wiley & Sons Ltd, United Kingdom.

Davino C., Romano R., Vistocco, D. (2018) Modelling drivers of consumer liking handling consumer and product effects, (forthcoming).

Koenker R. (2005) *Quantile regression*, Econometric Society Monograph, 38, Cambridge University Press, New York.

Meilgaard M.C., Carr B.T., Civille G.V. (1999) *Sensory evaluation techniques*, CRC press, Boca Raton.

Meullenet J.F., Xiong R., Findlay C.J. (2008) *Multivariate and probabilistic analyses of sensory science problems*, John Wiley & Sons Ltd, United Kingdom.

Naes T., Brockhoff P.B., Tomic O. (2010) *Statistics for Sensory and Consumer Science*, John Wiley & Sons Ltd, United Kingdom.

Peryam D.R., Pilgrim F.J. (1957) Hedonic scale method of measuring food preferences, *Food technology*, 11, 9-14.

Rdbotten M., Martinsen B.K., Borge G.I., Mortvedt H.S., Knutsen S.H., Lea P., Naes T. (2009) A cross-cultural study of preference for apple juice with different sugar and acid contents, *Food quality and preference*, 20, 277-284.

Exploring synergy between CUB models and quantile regression: a comparative analysis through continuousized data

Cristina Davino[*], Rosaria Simone[**], Domenico Vistocco[***]

Abstract: The paper investigates a parallel between CUB models and quantile regression through an illustrative case study on rating data. While CUB models have been proposed for modeling ordinal variables, quantile regression is mostly convenient for quantitative responses. The goal is to advance a comprehensive approach in which discrete ordinal outcomes on one hand and their continuousized version on the other coexist so to take advantage of two modern modeling frameworks.

Keywords: CUB models, Quantile Regression, Continuousized data.

1. Introduction and Motivation

The generalization of empirical findings from average is one of the factors that generates the common sense of diffidence about Statistics in the layman. It is efficiently described in the flaw of averages: “plan based on the assumptions that average conditions will occur are usually wrong” (Savage, 2002). The focus on the mean is a widespread approach even among insiders, since most applied Statistics is related to the estimation of average effects. The sentence that introduces regression in the book of Mosteller and Tukey (1977) is a clear invitation for insiders to go beyond the mean: “Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions”. The if and how the insiders have welcomed (and will welcome) this invitation can help to dissipate layman’s mistrust in statistical tools. The flaw of averages is becoming increasingly important in recent times because of the huge data dimension and of the complexity of the relationships among the data itself (Aguilar, 2018).

[*] University of Naples Federico II, cristina.davino@unina.it

[**] University of Naples Federico II, rosaria.simone@unina.it

[***] University of Cassino and Southern Lazio, vistocco@unicas.it

In this framework, Quantile Regression (QR), which was introduced as far back as 1978 (Koenker and Basset 1978), can be revitalised and regarded as one of the most modern and challenging methods in the era of big data. QR is based on the estimation of a set of conditional quantiles of a response variable as a function of a set of covariates. The method allows to verify if the effect played by the regressors varies on the low, middle and upper parts of the dependent variable thus suggesting different interpretation paths and revealing a scale and/or shape effect (Davino et al. 2014). If on one hand QR can be considered complementary to classical ordinary least squared regression (OLS), on the other hand the method represents a proper and suitable option when the homoschedastic assumption of the classical regression model cannot be satisfied, if the dependent variable has a skewed distribution or in presence of outliers. Nevertheless QR and the implied interpretation are not always suitable for ordinal data analysis, especially in cases where the response is on a finite discrete support and with a low number of possible answers. This is very common in survey data where the number of categories typically ranges from 5 up to 10 and thus a straightforward quantile modeling can raise some issues being not always greatly informative. In this respect, a manyfold perspective can be adopted with CUB models (D'Elia and Piccolo, 2005). The main feature of this class of models is the parsimonious yet flexible specification of both perceptual and decisional aspects of the rating process as a mixture of *feeling* and *uncertainty* directly on the measurement scale (Piccolo *et al.*, 2018). Thus, both QR and CUB models are appealing statistical frameworks for the analysis of evaluation–type data, for continuous and ordinal responses respectively. This contribution aims to investigate the connection between the two approaches. A combined analysis of CUB models and QR can be pursued if continuous variables are collected and then discretized, or conversely if genuine ordinal outcomes are *continuosized.* We opt here for the latter strategy, exploiting a solution proposed by Tamhane et al. (2002). In particular, let R be an ordinal variable collected on a rating scale coded with integers $1, \dots, m$ ($m > 3$). If n_j is the observed cell count for $R = j$, then continousized data in $[0, 1]$ can be obtained by uniformly spreading such observation in the interval $(\frac{j-1}{m}, \frac{j}{m}]$ to be then rescaled in the interval $[1, m]$. The approach can be easily adapted in case categories are not equally spaced. In the following a brief introduction of the two methods, CUB and QR, along with remarks on their possible integrated use will be provided through an illustrative case study on rating data.

2. *CUB and QR in a nutshell*

For quantitative variables measuring latent traits like happiness, social behaviors, self-evaluations, and so on, it is often preferable to pursue a discretization to summarize the phenomenon into ordered categories. Since in these cases it is of primary importance to understand the psychological mechanism driving the response process, the framework of CUB models offers advantageous interpretation of results by allowing a combined modeling of perceptual and decisional aspects of the choice. The rationale of this class of models is that each respondent has a propensity to provide a deliberate answer which is unavoidably mixed with the indeterminacy produced by the discretization of the latent trait. As an extreme circumstance, such indeterminacy collapses to a random choice. Thus, if R_i is the rating response given by the i–th subject and collected on a rating scale coded with integers $1, \dots, m$ ($m > 3$), then a two-component mixture is specified between a shifted binomial and a discrete uniform distribution:

$$Pr\big(R_i = r \mid \pi_i, \xi_i\big) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} \, + \, (1-\pi_i)\frac{1}{m},$$

$$logit(\pi_i) = \beta^{'} y_i, \qquad logit(\xi_i) = \gamma^{'} w_i,$$

where y_i, w_i are subjects' covariates specified to identify response profiles. The parameter ξ_i is referred to as the *feeling* parameter since $1 - \xi_i$ measures the preference of a category over the lower ones in a sequence of pairwise comparisons among categories. The mixing weight π_i, instead, is called the *uncertainty* parameter since $1 - \pi_i$ measures the overall uncertainty of the respondent's assessment: then, in particular, the larger it is, the higher the overall heterogeneity in the response distribution. ML estimation of parameter can be implemented by running the EM algorithm, and significance of variables' effects can be checked according to Wald test (Piccolo, 2006). Quantile regression has been instead proposed to model the whole conditional distribution of a response y given a set of p covariates $\mathbf{X}$, data observed on n units. Although models to deal with binary, nominal and categorical responses recently appeared in literature, QR is mostly used in case of numerical responses. In this paper we restrict our consideration to the case of linear effects. In such a case, QR estimates separate linear models for different quantiles $\theta \in [0, 1]$:

$$y_i(\theta) = \beta_0(\theta) + \mathbf{x}_i^\top \beta(\theta) + \epsilon_i, \tag{1}$$

such that $P(\epsilon_{i\theta} \leq 0) = \theta$ and $i = 1, \ldots, n$. The separate models are interpretable in terms of regression models for the quantiles of the response. The conditional distribution of the response can be estimated using a dense set of conditional quantiles. QR is distribution free since it does not pose any parametric assumption for the error (and hence response) distribution. The coefficients are commonly estimated through a variant of the simplex algorithm, while interior–point methods are especially suitable to deal with large scale problems (Koenker, 2005). Alternative estimators have been recently proposed exploiting the asymmetric Laplace distribution as a convenient model for the error distribution, thus allowing to embed QR in the likelihood framework and to extend it in a bayesian approach (Furno, Vistocco, 2018). As regards inference, QR estimators are asymptotically normal distributed with different forms of the covariance matrix depending on the model assumptions; resampling methods being a valid and widespread option.

3. The case study on relational goods and leisure time

The combined analysis between CUB models and QR will be discussed on the basis of a self-evaluation of the family making ends meet collected during a survey at University of Naples Federico II in December 2014. The purpose of the survey was to carry out an observational study on relational goods and activities for leisure time. Questionnaires were filled by students who were in turn asked to administer it also to acquaintances of theirs, according to a snowball sampling scheme. Every participant was asked to evaluate family end meet (from now, *EndsMeet*) on a 10-point Likert scale, ranging from 1 = 'never, at all' to 10 = 'always, a lot'. In the end, a sample of $n = 2181$ observations is considered. The goal is investigating the effects that the following covariates have on *EndsMeet*: *Child* and *Residence*, two dichotomous factors respectively with level 1 if there is any child aged less than 12 in the family and if the respondent lives in Naples or in its province. The solution proposed by Tamhane et al. (2002) has been used to transform the ordinal variable *EndsMeet* into continuousized data, so to use it as response variable in the QR model. Figure 2 (left-hand side) shows the observed frequency distribution of *EndsMeet*, with the kernel density of the corresponding continuousized data superimposed. The distribution of continuousized *EndsMeet* in the categories of the two covariates is shown in the right-hand part of Figure 2. The distribution of the response variable appears asymmetric in the group of families with at least one child less than 12 years old. It is worth of mention that just 20% of the interviewed belongs to this category and that almost 74% lives in Naples or in its province. The complete dataset with detailed description of all the collected variables is loaded within the R package CUB (Iannario *et al.*, 2018), which has been used for CUB models, tests and validation; for quantile regression, the R package quantreg (Koenker, 2018) has been used.

3.1. A parallel between CUB and QR results

The simplest QR model with a dichotomous regressor can help in testing the synergy between QR and CUB. Table 1 (first block of rows) reports the OLS and QR coefficients at five chosen quantiles, θ=[0.1, 0.25, 0.5, 0.75, 0.9] in a model with only *Child* as regressor. Either the OLS and the QR coefficients are significant with p-values less than 0.001 (standard errors have been estimated through resampling methods). but QR integrates results provided by classical regression. For example, having children less than 12 years old negatively impacts on the capability to get end of the month but this effect is higher on the lowest part of the distribution (at the 10% percentile is almost twice the average effect) and it becomes negligible and not significant on the

Table 1. OLS and QR estimated parameters for the two considered models

		OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
Child	$\hat{\beta}_0$	6.33	3.08	5.01	6.57	7.96	9.11
	$\hat{\beta}_1$	-0.45	-0.88	-0.68	-0.52	-0.27	-0.10
Residence	$\hat{\beta}_0$	0.60	0.22	0.45	0.64	0.79	0.91
	$\hat{\beta}_1$	-0.03	-0.02	-0.02	-0.04	-0.04	-0.03

highest part of the distribution (estimating a much more dense of quantiles, it results that the lowest slope is equal to -1.14 and the highest to 0.006). Thus, there is evidence for heterogeneity of effects of the regressor along the measurements scale. This claim is fully supported by inspecting CUB regression fit to the ordinal data ($BIC = 9689.38$):

$$logit(1-\hat{\pi}_i) = \underset{(0.099)}{0.100} + \underset{(0.256)}{0.687}\, Child_i, \quad logit(1-\hat{\xi}_i) = \underset{(0.040)}{0.694} - \underset{(0.114)}{0.255}\, Child_i.$$

As a result, responses are more heterogeneous in case there is a child aged less than 12 years in the family (uncertainty importance in the sense of weight for the uniform distribution increases from $1 - \hat{\pi}_0 = 0.529$ to $1 - \hat{\pi}_1 = 0.697$ when switching from $Child = 0$ to $Child = 1$, whereas perceived easiness in making ends meet (as measured by $1 - \hat{\xi}_i$) decreases from 0.668 to 0.617.

A further investigation of the synergy deriving from a conjoint use of QR and CUB is realised using the second regressor, *Residence*, which is dichotomous too but with a different impact on the response variable. Indeed, it affects only the location of the distribution being statistically significant only for the feeling component (as evident also from the right panel of Figure 3):

$$1 - \hat{\pi} = \underset{(0.022)}{0.557}, \quad logit(1 - \hat{\xi}_i) = \underset{(0.084)}{0.873} - \underset{(0.094)}{0.281}\, Residence_i$$

Specifically, being resident in the metropolitan area of Naples decreases (perceived) easiness in making ends meet. The constant uncertainty level given *Residence* gains insight when looking at QR results on the continuousized response: the impact of living in Naples or in its province is negative but almost constant along the distribution (see second block of rows in Table 1). More-

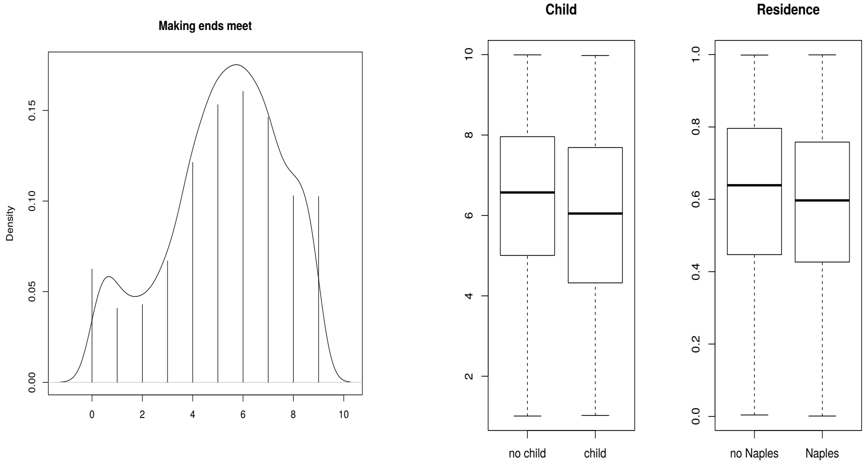


Figure 1. Rating data and continuousized version for the rating: Do you easily make ends meet? (left). Boxplot for the continuousized version given Child and Residence (right))

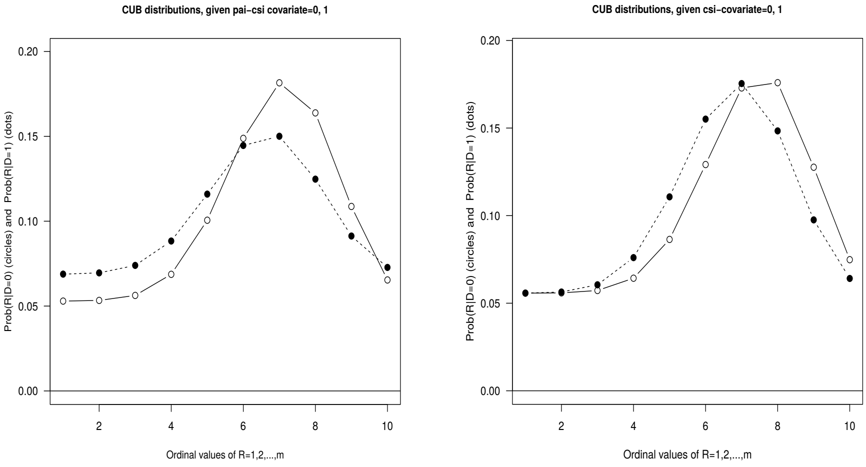


Figure 2. Conditional CUB distributions given $Child$ (left-hand side) and $Residence$(right-hand side))

over, this effect is related to very high standard errors in the lowest part of the distribution.

4. Conclusions and future research

The paper has advanced a comparative application of quantile regression methods for quantitative responses and CUB models for rating data: continuousized data allows to switch from one setting to the other with the goal of understanding mutual advantages, analogies and differences of the two approaches. Particular emphasis has been given to the interpretation of uncertainty and heterogeneity of regressors' effects. In this vein, future research developments can be outlined by simulation studies and more challenging empirical evidence.

References

Aguilar S. J. (2018) Learning Analytics: at the Nexus of Big Data, Digital Innovation, and Social Justice in Education, *TechTrends*, 62, 37-45.

D'Elia A., Piccolo D. (2005) A mixture model for preference data analysis, *Computational Statistics and Data Analysis*, 49, 917-934.

Davino C., Furno M., Vistocco D. (2014) *Quantile Regression: Theory and Applications*, John Wiley & Sons.

Furno M., Vistocco D. (2018) *Quantile Regression: Estimation and Simulation*, John Wiley & Sons.

Iannario M., Piccolo D., Simone R. (2018) CUB: A Class of Mixture Models for Ordinal Data, R package version 1.1.2. https://CRAN.R-project.org/package=CUB.

Koenker R.W., Basset G. (1978) Regression quantiles, *Econometrica*, 46, 33-50.

Koenker, R. (2005) *Quantile Regression*, Econometric Society Monographs, Cambridge: Cambridge University Press.

Koenker R. (2018) quantreg: Quantile Regression. R package version 5.35. https://CRAN.R-project.org/package=quantreg

Mosteller F., Tukey J. (1977) *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison–Wesley.

Piccolo D. (2006) Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33-78.

Piccolo D., Simone R., Iannario M. (2018) Cumulative and CUB models for rating data: a comparative analysis, *International Statistical Review*, forthcoming.

Savage, S. (2002) The flaw of averages, *Harvard Business Review*, 80, 20-22.

Tamhane A., Ankemanman B., Yang Y. (2002) The Beta distribution as a latent response model for ordinal data (I): Estimation of location and dispersion parameters, *Journal of Statistical Computation and Simulation*, 72, 473 - 494.

A Poset based indicator of gender equality at sub-national level

Enrico di Bella[*], Lucia Leporatti[**], Filomena Maggino[***],
Luca Gandullia[****]

Abstract: Gender equality represents a central issue in the socio-economic background of our society and, consequently, its study is gaining increasing attention in the international debate. During the last 20 years, the international literature proposed a number of indicators that aim at measuring gender equality but there are few experiences of gender equality indicators at sub-national level. The aim of this work is twofold: on one hand we propose a regional decomposition of the European Institute for Gender Equality index (R-GEI); on the other we compare the synthetic indicator obtained following the EIGE methodology with a poset based synthetic indicator (POR-GEI). The new R-GEI is obtained reproducing the EIGE methodology, and it is compared to the POR-GEI index that exploits poset theory for aggregating indicators.

Keywords: EIGE, Gender equality, Poset.

1. Introduction

The interest for gender inequality in several areas of life is increasingly at the center of the international debate given its strong socio-economic implications. Much effort has been devoted to the identification of appropriate ways to measure gender inequality and, with this aim, over the years various composite indicators have been proposed internationally. The three most relevant solutions fit for the purpose can be identified in the Global Gender Gap Index of the World Economic Forum (WEF, 2017), in the Gender Development Index (GDI) proposed by the United Nations (UNDP, 2017) and in the Gender Equality Index (GEI) of the European Institute for Gender Equality (EIGE, 2017a, b). Although having some common features, the three indica-

[*] University of Genova, edibella@unige.it

[**] University of Genova, lucia.leporatti@unige.it

[***] University of Roma "La Sapienza", filomena.maggino@uniroma1.it

[****] University of Genova, luca.gandullia@unige.it

tors differ from a number of issues. First, there is not a full overlapping in the variables selected to describe the domains and in the domains themselves since this choice inevitably depends on the actual availability of data. In addition, the indicators differ substantially for weighting and aggregation procedures. Despite the undeniable effort devoted to the identification of proper indicators, a relatively unexplored topic concerns the evaluation of gender inequality at a sub-national level. This perspective may be particularly interesting in a country like Italy characterized by persistent regional disparities in terms of economic development, population structure and size. According to last available GEI data (year 2015), Italy ranks 14^{th} among the EU28 countries for gender equality, with a GEI value equal to 62.1 compared to a EU28 average of 66.2. Although some studies have been proposed to analyze gender inequality in Italy at the sub-national level, these are mostly referred to specific areas of the country and there are no systematic attempts to measure the phenomenon under a comprehensive perspective. One of the reason for this lack, is ascribable to the paucity of publicly available indicators at a NUTS2 level. This study is a preliminary work aimed at proposing a method to breakdown at sub-national level the Italian GEI index computing regional indices as faithful as possible to that produced by EIGE. The decision to select EIGE indicator among the pool of available ones is explained by its completeness (compared to the other two mentioned above) in terms of domains and variables. In addition most of the variables used in GEI are based on European surveys (mainly EU-SILC and Labor Force Survey) that guarantee a territorial detail representative at a NUTS-2 level. For variables that can not be used at NUTS-2 level due to a low representatives of survey data or to their weak meaning at a regional level (think for example to the number of women in Parliament), alternatives are proposed trying to be as much coherent as possible with the original meaning of replaced variables. The resulting regional gender equality (R-GEI) for Italy is then analyzed in a partially ordered sets (posets) approach (e.g.: Fattore, 2018) to introduce an additional method to define a composite indicator of gender equality at regional level.

2. *Data and methods*

In order to build the two alternative regional gender equality indicators we followed three main steps:

1. **assessment of the original GEI variables** in terms of data sources and relevance at a regional level. This step has been structured in two activities: a) computation of single indicators at a NUTS2 level using GEI survey microdata (when representative), alternative surveys (when original surveys are not representative) or other official databases; b) substitution of meaningless variables with others more consistent with the regional perspective;

2. **R-GEI computation**: use of the GEI methodology (third edition) to build the R-GEI;

3. **POR-GEI computation**: use of the poset methodology (third edition) to build the POR-GEI.

The next sub-sections will briefly describe each of the three steps.

2.1. Indicators for the regional gender equality index at sub-national level.

EIGE's GEI is based on 6 core domains (Work, Money, Knowledge, Health, Time and Power), 14 sub-domains and 31 variables and it ranges from 1 (i.e. total inequality) to 100 (i.e. full equality). The variables employed for the computation of GEI come from different data sources. For what it concerns survey datasets, the four sources used are: Labor Force Survey (LFS); European Union Statistics on Income and Living Conditions (EU-SILC); European Health Interview Survey (EHIS) and European Working Conditions Surveys (EWCS). The first three surveys are provided by Eurostat and their sample size and scheme guarantee the representativeness at a NUTS2 level. The LFS is mainly used to fill in the domains of Work and Knowledge. The EU-SILC survey is instead used in the fields of Money and Health while EHIS is used to assess Health. The EWCS survey is instead developed by Eurofound and it is mainly used to address the issue of Time. However, the number of observations sampled for EWCS is not sufficient to obtain reliable regional estimates (1,402 observations for all Italy). Therefore, most of the regional variables connected to this domain will be based on the ISTAT survey "Aspects of daily life". Another domain that requires a significant revision is that of Power: indeed, under a regional perspective, most of the original indicators lose their relevance. As a consequence, the variables connected to Political Power have been replaced by more local/regional measures such as the share of women in regional boards, municipality and regional assessors, city majors (source: Italian Ministry of Interior). Also the variables connected to Economic Power has been revised due to the unavailability of original EIGE data at a NUTS2 level (source of new data: INPS). No alternatives have instead been found for the sub-domain of Social Power. Given these premises, after the adjustments, 10 out of the 31 original variables are exactly based on GEI definition and data, 15 are instead based on a definition as close as possible to that adopted in GEI but using data representative at a regional level.

2.2. The regional level Gender Equality Index (R-GEI).

After selecting the relevant indicators, we have followed the EIGE methodology to build up the regionalised composite indicator. In particular, as suggested by EIGE (2017b) we started with the computation of gender gaps $(\Upsilon_{(x_{it})})$ that measures the gaps between men and women for each variable X and for the i-th region and the t time period as:

$$\Upsilon_{(x_{it})} = \left| \frac{\bar{X}_{it}^{W}}{\tilde{X}_{it}^{a}} - 1 \right| \tag{1}$$

where $\bar{X}_{it}^{W}$ represents the value of the X variable for women and $\tilde{X}_{it}^{a}$ is the average unweighted values of men and women. In order to interpret the measure as gender equality and not as gender gaps, the second step consists in taking the complementary of the gaps $1 - (\Upsilon_{(x_{it})})$.

In addition, a correction coefficient $(\alpha_{(x_{it})})$ is applied to take into account both gender gaps and the overall level of achievement. The correction procedure is developed by comparing, for each variable, the performance of each region with the best performance recorded:

$$\alpha_{(x_{it})} = \left(\frac{\tilde{X}_{it}^{T}}{max(\tilde{X}_{it}^{T})} \right)^{1/2} \tag{2}$$

As a consequence, the higher the gender gaps and the higher the distance form the best performing region, the lower would be the final value for the R-GEI. The resulting indicators, for each variable and region are then computed as follows:

$$\Gamma_{(x_{it})} = 1 + \left[\alpha_{(x_{it})} \cdot \left(1 - \Upsilon_{(x_{it})} \right) \right] \cdot 99 \tag{3}$$

The last part of the procedure consists in the weighting and aggregation of the variables, sub-domains and domains in one composite indicator. The aggregative procedure suggested by GEI operates through various steps (see EIGE, 2017b, for details), the last one of which is the aggregation of domain specific indicators in one composite indicator using a weighted geometric mean with a weighting scheme defined by experts (Work = 0.19; Money = 0.15; Knowledge = 0.22; Time = 0.15; Power = 0.19; Health = 0.10).

2.3. The poset based regional level Gender Equality Index (POR-GEI).

The poset approach to build synthetic indicators keeps all the information inherent to each indicator separate from the other resulting in synthetic indicators that are not aggregative in nature (Fattore, 2018; di Bella, 2018). The units of analysis of posets are not the elementary indicators but the vectors of values of each indicator for any single statistical unit, called *profiles*. Poset theory focuses on the concept of comparability or incomparability between couples of profiles. Two profiles are comparable when the values of the indicators of one profile are not lower (or, vice versa, not bigger) than the corresponding values of the other profile; on the contrary, if at least one is bigger (or lower), then the two profiles are claimed to be incomparable. The visual representation of a (finite) poset is generally made by means of the diamond scalogram or Hasse diagram, a graph in which profiles are graphed vertically in levels that define their relative position in the ordering; those profiles which cannot be directly compared are positioned on the same level of the graph and they are not directly connected. The Hasse diagram is not only a graphical output for poset analysis but also, through its linear extensions, a tool for the definition of poset-based non aggregative composite indicators. Analysing (via computational methods) all the possible linear extensions of a Hasse diagram makes it possible to identify the position of each statistical unit (in our case each Italian region) in each linear extension deriving a metric that is called “average height" (Bruggemann and Annoni, 2014). Such a metric is used as a synthetic indicator for the system of indicators that is under study and can be used for a ranking of statistical units. In this work we processed the data according to the following steps (see Figure 1):

1. we clustered each variable using a Duda-Hart stopping rule (Duda *et al.*, 2001) in order to discretise the data and to reduce random incompatibilities;

2. we derived for each domain of gender equality a Hasse diagram and we computed for each of them the “extended average height" a domain specific synthetic indicator (SI);

3. we clustered the domain specific synthetic indicators using again the Duda-Hart stopping rule;

4. we derived the global Hasse diagram and we used the average heights of its elements (i.e. the Italian regions) as POR-GEI values.

Steps 1 and 3 were run in STATA by StataCorp LLC whilst steps 2 and 4 were done using the PyHasse online tool (LPOM package) by prof. Reiner Bruggemann freely available at: https://www.pyhasse.org/.

3. Results and conclusions

Table 1 provides a comparison of the rankings of the Italian regions using R-GEI and POR-GEI. Spearman’s rank correlation coefficient between the two rankings is 0.947 that is a pretty high value but, as it can be easily seen, the two rankings although similar have some important differences. There

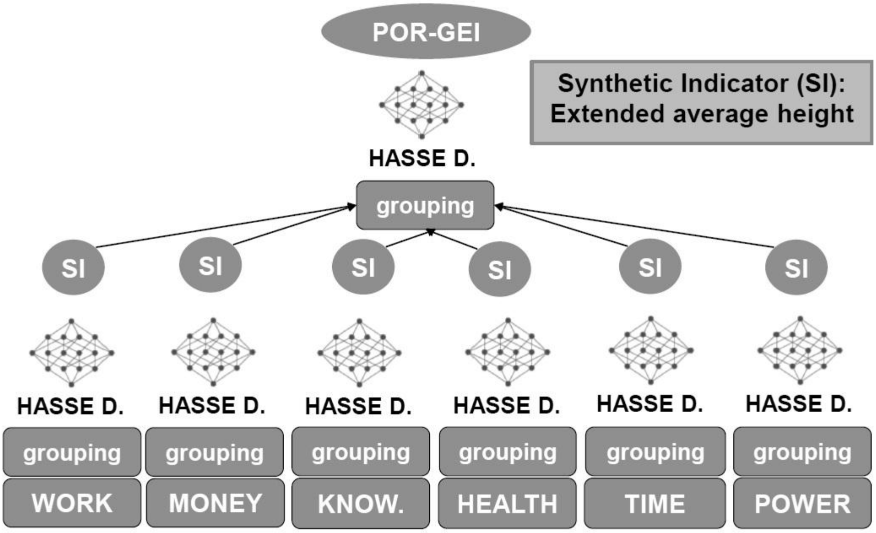


Figure 1. The construction of the POR-GEI synthetic indicator.

Table 1. Rankings of Italian regions according to R-GEI and POR-GEI.

Region	R-GEI	POR-GEI		Region	R-GEI	POR-GEI
Lombardia	1	2		Umbria	11	8
Emilia-Romagna	2	3		Sardegna	12	10
Toscana	3	1		Valle d'Aosta	13	13
Piemonte	4	7		Abruzzo	14	14
Friuli V.G.	5	9		Molise	15	15
Trentino A.A.	6	5		Puglia	16	18
Lazio	7	5		Basilicata	17	17
Veneto	8	4		Calabria	18	20
Liguria	9	11		Campania	19	19
Marche	10	12		Sicilia	20	16

are at least two main results that we want to point out in this preliminary work. First, we think that it is possible to find variable at sub-national level that are consistent with EIGE's framework. Second, it is possible to develop a synthetic indicator for gender equality without any subjective choice (e.g. weighting of domains). Further work will explore the advantages of POR-GEI against R-GEI in particular analysing the actual multi-dimensionality of the battery of variable used in this study and the advantages of a non compensative approach against traditional aggregative (and compensative) synthetic indicators.

References

Bruggemann R., Annoni P. (2014) Average Heights in Partially Ordered Sets, *MATCH*, 71, 177-142.

di Bella E. (2018) Partial Order Scalogram Analysis with base coordinates (POSAC) *Wiley StatsRef: Statistics Reference Online*, code: stat08113.

Duda R.O., Hart P.E. and Stork D.G. (2001) *Pattern Classification* 2nd ed. New York: Wiley.

EIGE (2017a) *Gender Equality Index 2017: Measuring gender equality in the European Union 2005-2015*, European Institute for Gender Equality Report.

EIGE (2017b) *Gender Equality Index 2017: Gender Equality Index 2017: Methodological Report*, European Institute for Gender Equality Report.

Fattore M. (2018) Partially Ordered Sets *Wiley StatsRef: Statistics Reference Online*, code: stat08101.

UNDP (2017) *2016 HDR Report*, UN-Devolpment Program.

WEF (2017) *The Global Gender Gap Report 2017*, World Economic Forum Report.

Using mutual ranking probabilities for dimensionality reduction and ranking extraction in multidimensional systems of ordinal variables

Marco Fattore*, Alberto Arcagni**

Abstract: In this paper, we address the extraction of rankings from multi-indicator systems, as a problem of approximation between the so-called “mutual ranking probability” matrices, associated to the partial order relations derived from the data. After providing a theoretical treatment of the topic, we propose a practical algorithm for ranking extraction and show it in action on a real example, pertaining to regional competitiveness.

Keywords: Multi-indicator system, Partially ordered set, Ranking.

1. Introduction

Ranking is one of the most typical goals of statistical evaluation studies, particularly in socio-economics. The starting point for ranking construction is usually a *multi-indicator system* (MIS), i.e. a collection of attributes related to some concept of interest, against to which statistical units are scored. The score vector of each unit (i.e. its *score profile*) is then collapsed into a single number, by means of some composite indicators, and a ranking is finally built. This approach, however, breaks down when the statistical variables comprised in the MIS are of an ordinal kind and cannot be aggregated. As a matter of fact, there is no well-founded theory of ranking construction for ordinal multi-indicator systems yet, although some proposals can be found in literature (see Bruggemann, R. & Patil, G. P. 2011, Fattore M. 2017). In this short paper, we address the ranking problem on ordinal MISes as a problem of *dimensionality reduction* of the ordinal relations associated to the indicator systems and develop a practical ranking algorithm which does not involve any variable aggregation. The key to this result is the use of *partial order theory*,

* University of Milano - Bicocca, marco.fattore@unimib.it

**University of Milano - Bicocca, alberto.arcagni@unimib.it

which is the most natural formal framework for the description of multidimensional ordinal data and allows to turn "optimal" ranking extraction, into a problem of matrix approximation. The paper is organized as follows. Section 2 sets the formal stage and introduces some essential concepts of partial order theory; Section 3 develops the approximation approach to ranking extraction, provides a heuristic algorithm to perform it in practice and works out a simple, but real, example; Section 4 concludes.

2. *The formal setting*

Given a (here, finite) set X, a *partial order relation* $\trianglelefteq$ on it is a binary relation which is (Davey B. A., Priestley B. H. 2002, Schröder, B. 2002) *reflexive* ($x \trianglelefteq x$, for all $x \in X$), *anti-symmetric* ($x_1 \trianglelefteq x_2$ and $x_2 \trianglelefteq x_1$ implies $x_1 = x_2$) and *transitive* ($x_1 \trianglelefteq x_2$ and $x_2 \trianglelefteq x_3$ implies $x_1 \trianglelefteq x_3$). A set X endowed with a partial order relation is called a *partially ordered set*, or a *poset* for short. Posets are the natural data structures associated to multi-indicator systems of ordinal variables. To realize why, suppose to score n objects against k ordinal attributes, getting a score *profile* (i.e. a sequence of k scores) for each unit. Dealing with ordinal scales, the only operation that can be legitimately performed on the set of profiles associated to the MIS is just to *multidimensionally compare* them. If the scores[1] of profile x_1 are all not higher than those of profile x_2, and at least one is strictly lower, then the two profiles are *comparable* and, in this case, profile x_2 *dominates* profile x_1 (we write $x_1 \triangleleft x_2$ to mean $x_1 \trianglelefteq x_2$ and $x_1 \neq x_2$). But if the two profiles have so-called *conflicting scores* (i.e. profile x_1 has some scores higher than profile x_2 and some scores lower), then the two profiles are *incomparable* (in formulas, $x_1 || x_2$). As a consequence, the set of profiles can be ordered only partially and thus it is naturally structured as a poset[2]. In principle, one could introduce additional criteria to order those profiles which are incomparable in the poset, so defining new posets which, technically speaking, *extend* (i.e. add comparabilities to) the former. Among such *extensions*, some have no incomparabilities;

[1] We are assuming that all the variables are oriented in the same way.

[2] Technically speaking, this requires units' profiles to be different, so as to fulfill the anti-symmetry property; if two units share the same profile, they must be clustered together.

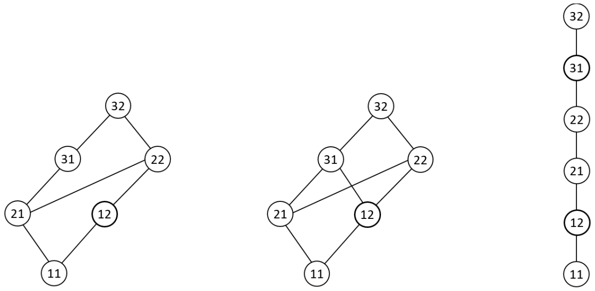


Figure 1. A poset (on the left) and two of its extensions (on the right, a linear extension). Posets are depicted as Hasse diagrams, i.e. as graphs where $x_1 \trianglelefteq x_2$ *if and only if there is a downward sequences of edges, linking the corresponding nodes.*

these are called *linear extensions* and, in practice, correspond to rankings of the input objects. Figure 1 clarifies pictorially the above discussion. Extracting a ranking out of a MIS, i.e. out of the poset associated to it, is thus the same as picking up a suitable element out of the set of its linear extensions. The key point is then determining the criterion to select such an element and turning it into a practical ranking algorithm. We address both issues in the next section.

3. Ranking extraction from ordinal multi-indicator systems

Any finite poset can be reconstructed from its set of linear extensions; in fact, it can be proved that the set of comparabilities of a poset coincides with the set of comparabilities *common* to its linear extensions (Schröder, B. 2002). In shorter terms, any finite poset is the *intersection* of its linear extensions. In general, however, the same poset can be reconstructed as the intersection of smaller subsets of linear extensions and the smallest cardinality of such sets is called the *dimension* of the poset (Schröder, B. 2002). Clearly, a linear order is a poset of dimension 1 and any poset which is not linear has dimension strictly greater than 1. So, ranking the objects scored in the input MIS is equivalent to reduce the dimensionality of the associated poset to 1. Any dimensionality reduction process involves some information loss and our aim is to pick up the ranking which minimizes it. We must thus introduce a cost function, driving the extraction process, and this leads to discussing how posets can be algebraically *represented* in matrix terms. Finite posets can be algebraically represented in two main ways, namely by using the *cover* matrix G or the *incidence* matrix Z. Let Π be the poset associated to the MIS and let $x_i, x_j \in \Pi$; we say that x_j *covers* x_i ($x_i \prec x_j$), if $x_i \lhd x_j$ and there is no other element $x_h \in \Pi$ such that $x_i \lhd x_h \lhd x_j$. The covering relation is naturally described by the cover matrix $G_{n\times n}$ (n being the number of elements in the poset), whose entries are defined as $G_{ij} = 1$, if $x_i \prec x_j$, and $G_{ij} = 0$ otherwise. The covering relation $\prec$ determines, by transitivity, the partial order relation $\unlhd$ (in fact, $\unlhd$ is said to be the *transitive closure* of $\prec$), so that G can be considered as a representation of Π itself. On the other hand, Π can be represented by simply listing the pairs (x_i, x_j) of elements, such that $x_i \unlhd x_j$; this is the same as defining the incidence matrix $Z_{n\times n}$, where $Z_{ij} = 1$ if and only if $x_i \unlhd x_j$ and $Z_{ij} = 0$ otherwise. Given such matrix representations, one is tempted to assess the information loss implied by the construction of a ranking ℓ from the input poset Π, by measuring the distance between the corresponding matrices G_ℓ and G_Π, or Z_ℓ and Z_Π, using some metric, e.g. the L^1 distance[3]. This, however, is ineffective, and we now show why. Consider Figure 2 where a small poset and its three linear extensions are depicted. Since the red dot is incomparable with both the other elements, any of the linear extensions can be legitimately selected as the candidate final ranking (which, in this case, is not unique). The G matrices of the poset and its linear extensions are listed below:

$$G_\Pi = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \quad G_{\ell_1} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \quad G_{\ell_2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}; \quad G_{\ell_3} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The L^1 distances between the G matrix of the poset and the G matrices of the three linear extensions are different (namely, the distance equals 3 for linear extension 2 and equals 1 for both linear extensions 1 and 3). This

[3] Various metrics are available for matrix comparisons; for simplicity's sake, here we consider L^1 distance, which is often used in this context (De Baets B., De Meyer H. 2003). In any case, the results worked out in the following does not depend, in their essence, on the metric chosen.

shows that using cover matrices introduces a bias, not justfiable in covering terms, in ranking selection, ruling out candidates that could be legitimately chosen. On the other hand, the uselessness of incidence matrices for ranking selection is made evident by observing that, as it can be easily checked, the distances between *any* Z_ℓ matrix of a linear extension and the Z_Π matrix of the input poset are all the same, so that L^1 distances between incidence matrices cannot discriminate among different candidate rankings. To solve the ranking problem, we thus need a different way to represent finite partial orders. This requires introducing the concept of *mutual ranking probability* between poset objects. Consider a pair of incomparable elements x_i and x_j of Π; in some linear extensions, x_i is ordered below x_j while in some others the order is reversed (if not, the two elements would be comparable in Π). Let ω_{ij} be the number of linear extensions where $x_i \lhd x_j$ and let ω be the cardinality of the set $\Omega(\Pi)$ of linear extensions of Π, then $p_{ij} = \omega_{ij}/\omega$, i.e. the share of linear extensions where x_i is ordered below x_j, is called *mutual ranking probability* (MRP) between x_i and x_j. Clearly, $p_{ij} = 1 - p_{ji}$ and $p_{ij} = 1$ if and only if $x_i \unlhd x_j$ in Π. Arranging MRPs into a $n \times n$ matrix P_Π, we get a new matrix representation of the poset, which proves much more useful for ranking extraction, than the cover or the incidence ones. In fact, even if two objects are incomparable in Π, it may well be that one of the two dominates the other in most of the linear extensions, so as to "almost dominate" the latter, also in the input poset. As a simple example, the MRP matrix for the poset depicted on the left side of Figure 2 is reported hereafter:

$$P_\Pi = \begin{bmatrix} 1 & 0 & 1/3 \\ 1 & 1 & 2/3 \\ 2/3 & 1/3 & 1 \end{bmatrix}.$$

These "dominance degrees" are completely overlooked in the G_Π and Z_Π matrices, while they are accounted for in P_Π, which is not binary[4] and provides a richer and more explicit information on the pairwise dominance structure of the poset. We thus search for the best ranking of the n objects, by

[4] The MRP matrix is binary if and only if it represents a linear order; in such a case, it coincides with the incidence matrix, since in a linear order there are no incomparabilities.

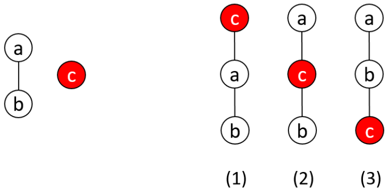


Figure 2. A poset on three elements (left) and its three linear extensions (right).

searching for the linear extension ℓ^* of Π which minimizes $\|P_\Pi - P_{\ell^*}\|_1$, in the set of linear extensions of Π. Notice that, in general (e.g. when the input poset has internal simmetries) the best approximating linear order is not unique.

3.1. A heuristic algorithm

To extract "the" (or "a") best approximating linear order, we must minimize $\|P_\Pi - P_\ell\|_1$ over all of the $n \times n$ binary matrices P_ℓ representing linear extensions of Π. In general, this is a hard computational task, which cannot be accomplished simply by listing all the linear extensions of Π and computing the cost function, since the cardinality of $\Omega(\Pi)$, for real posets, is extremely huge. We thus provide a greedy algorithm which reduces the computational burden, while finding out a reasonable candidate for "the" best approximating linear extension. The algorithm is described as follows: (i) **input matrix computation**: from the input poset Π, compute the cover matrix G_Π and the MRP matrix P_Π; (ii) **initialization**: initialize a cover matrix G' to G_Π; (iii) **updating G'**: search for the highest entry of P_Π, less than 1, and turn to 1 the corresponding entry of G'; (iv) **computing Z'**: from G', compute the transitive closure Z', which represents a partial order, where the comparabilities implied by the updating of G' have been added; (v) **repetition**: repeat steps 3-4, until Z' represents a linear order ℓ^*.

As mentioned, the goodness of approximation of the selected linear extension to the input poset can be computed by considering the L_1 distance between the input MRP matrix P_Π and the final MRP matrix P_{ℓ^*}; a practical goodness of fit index is thus:

$$GoF = 1 - \frac{\|P_{\Pi} - P_{\ell^*}\|_1}{\|P_{\Pi}\|_1}.$$

3.2. An application

We consider data on the competitiveness level of Belgium regions, published by the Joint Research (Center Annoni P., Dijkstra L. 2013). In particular, we consider the ranks of each country on three main competitiveness pillars, named *Basic*, *Efficiency* and *Innovation*, respectively (see Table 1). Three regions have the same profiles (Bruxelles-Capitale, Vlaams-Brabant, Brabant Wallon), so we clustered them into a macro-region M. The corresponding partial order is depicted in Figure 3 (left), while the extracted ranking is shown in the middle of the same picture. The algorithm requires 16 iterations (see Figure 3, right panel) to get to the final ranking, which has a GoF equal to 0.774.

Table 1. Belgium regional ranks on the three competitiveness pillars (M = Bruxelles-Capitale + Vlaams-Brabant + Brabant Wallon).

Code	Region	Basic	Efficiency	Innovation
M	/	3	2	1
BE21	Antwerpen	1	5	4
BE22	Limburg	2	7	6
BE23	Oost-Vlaanderen	7	1	5
BE25	West-Vlaanderen	8	6	9
BE32	Heinaut	9	11	10
BE33	Liege	6	9	8
BE34	Luxembourg	11	10	11
BE35	Namur	10	8	7

4. Conclusion and further research

In this paper, we have proposed a new approach to ranking construction on ordinal multi-indicator systems, as a dimensionality reduction process on

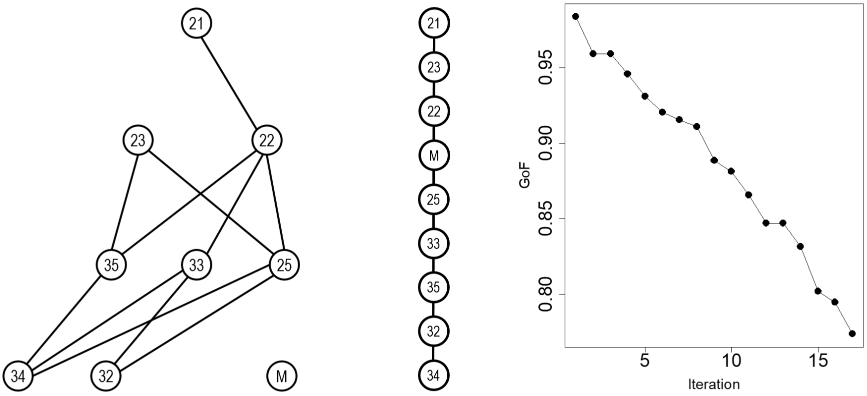


Figure 3. Belgium competitiveness poset (left), the final ranking (middle) and the GoFs of the extensions, as iterations proceed.

the associated posets. The methodology is very general and can be applied to any kind of finite poset, not necessarily derived from systems of indicators. Future research will be mainly devoted to the study of the link between the optimal dimensionality reduction process proposed in the paper and other ranking extraction procedures available in literature and to possible logic and computational improvements of the algorithm itself.

References

Annoni P., Dijkstra L. (2013) *EU Regional Competitiveness Index - RCI 2013*, JRC Scientific and Policy Reports.

Bruggemann R., Patil G.P. (2011) *Ranking and Prioritization for Multi-indicator Systems*. New York: Springer-Verlag.

Bubley R., Dyer M. (1999) Faster random generation of linear extensions, *Discrete Mathematics*, 201, 81-88.

Davey B.A., Priestley B.H. (2002) *Introduction to Lattices and Order*. Cambridge: CUP.

De Baets B., De Meyer H. (2003) Transitive approximation of fuzzy relations by alternating closures and openings, *Soft Computing*, 7, 210-219.

Fattore M. (2017) Synthesis of indicators: the non-aggregative approach, In: F. Maggino (ed.), *Complexity in societies: From Indicators Construction to their Synthesis*, Springer.

Schröder B. (2002) *Ordered set. An introduction*. Basel: Birkhäuser.

On classifiers to predict soccer match results

Silvia Golia[*], Maurizio Carpita[**]

Abstract: Many statistical models are widely used to predict the result of a soccer match; the standard predictive criterium of classification is the the majority rule, which corresponds to the mode in a polytomous case. In this study, other predictive criteria are proposed and compared with the modal one. The predictive performances are evaluated considering a set of indicators built from the resulting 3×3 confusion matrix. The data used come from the Kaggle European Soccer Database and refer to the seasons from 2009/2010 to 2015/2016 of the Italian League Serie A.

Keywords: Polytomous classifier, Bayesian networks, Evaluation procedures.

1. Problem, data and model

Nowadays, various statistical (probabilistic and algorithmic) models are widely used to predict one of the three possible results of a soccer match: loss, draw or win of the home team. This ambitious goal can be pursued using information available before the match starts, as players performance statistics or expert judgements. In the case of models with players performance predictors, draw is the most difficult outcome to forecast (Carpita *et al.*, 2015, 2018). Also for models with experts, the prediction accuracy of draw is considerably worse than that for win and loss (Strumbelj and Sikonja, 2010, Franck *et al.*, 2010). The difficulty in predicting the draw result can be due to the fact that its probability is lower than the probabilities of loss and win, so that the classification models underestimate the matches resulting in draw. In fact, the standard predictive criterium of classification is the majority rule, so that the mode result (i.e. the result with higher predicted probability) is used. In this study, other predictive criteria are proposed and compared with the modal one. Taking into account the ordinal nature of the result of a match (loss $\prec$ draw $\prec$ win), it has been codified as 0-1-2 and considered as

[*] University of Brescia, silvia.golia@unibs.it

[**] University of Brescia, maurizio.carpita@unibs.it

a numerical random variable with the predicted probabilities as its frequency distribution. The dataset used in this paper comes from the Kaggle European Soccer database (Carpita *et al.*, 2018) and contains the matches results reported in terms of goals scored by the home and away teams and the overall performance indicators (averaged according to the coach decisions before each match, for the four roles in soccer team: goalkeeper, defender, midfielder and forward role) used as predictors, for the seasons from 2009/2010 to 2015/2016 of the Italian League Serie A (Carpita and Golia, 2018). From the goals scored by the two teams during the match, it is possible to determine the outcome of the match from the home team point of view, *result*, classified as win (W), draw (D) and loss (L).

The model used in this study to predict the probability distribution of *result* is the Bayesian Network (BN). The BNs belong to the class of probabilistic networks which are graphical models that explicit through a graph, the interactions among a set of variables represented as nodes of the graph. A BN is given by the pair (G, P), where G is a directed acyclic graph (DAG), and P is a probability distribution which factorizes according to G. The DAG G is composed by a set of nodes V, which correspond to a set of random variables X_V indexed by V, and a set E of directed edges between pairs of nodes in V. The joint probability distribution P over the set of variables X_V is factorized as follows:

$$P(X_V) = \prod_{\nu \in V} P(X_\nu | X_{pa(\nu)}) \tag{1}$$

where $X_{pa(\nu)}$ denotes the set of parent variables of variable X_ν for each node $\nu \in V$.

Once a BN is identified and estimated, it can be used to evaluate the effect of new evidence Ev on one or more target variables X' using the knowledge encoded in the BN and computing the posterior distribution $P(X'|Ev)$ (Koller and Friedman, 2009).

2. *Proposed classifiers*

When evaluating new evidence, BN gives the probability of each possible result, so it is necessary to fix a criterium to transform the set of three probabilities into a number which correspond to the predicted result. In this paper five different methods, reported in Table 1, have been considered.

Table 1. The proposed classifiers

Classifiers	Criterium	Threshold	$result$'s code
M1	Mode	–	–
M2	Median	–	–
M3	Max. Dist.	–	–
M4	Expected Value	–	0, 1, 2
M5.1	Mode + Expected Value	0.4	0, 1, 2
M5.2	Mode + Expected Value	0.5	0, 1, 2
M5.3	Mode + Expected Value	0.6	0, 1, 2
M5.4	Mode + Expected Value	0.7	0, 1, 2

Rounding rules for classifiers M4 and M5: ceiling (c), floor (f), round (r)

The first one (M1) uses the Mode criterium, which corresponds to the majority rule in the binary case, whereas the second one (M2) involves the median, taking advantage of the ordinal scale of the variable *result*. The third method (M3) evaluates the difference between the predicted probabilities of the three possible results of the match and the corresponding sample frequencies and takes the result corresponding to the maximum difference. The forth method (M4) involves the coding of the *result* variable as L=0, D=1, W=2 and the use of the expected value rounded following the ceiling (c), round (r) or floor (f) rule. Other 12 classification methods (M5.j) are obtained making different use of the modal criterium, transforming in numerical the ordinal match results and using different rounding methods: the predicted result is simply the modal one if the corresponding probability is bigger than a given threshold (0.4, 0.5, 0.6, 0.7), otherwise it is the expected value of a random variable with predicted probabilities and values obtained coding the match results in the same way of M4 and rounded following the ceiling (c), round (r) or floor (f) rule.

3. Predictive performance indices

The predictive performance of a classifier can be summarized using the so-called Confusion Matrix; the elements of this matrix report the count of how many units that truly belong to each class (rows) were predicted by the model to belong to that class (columns); a unit belonging to class A and predicted to belong to class P is counted in n_{AP}. Starting from this matrix, it is possible to compute some indicators that allow one to evaluate the goodness of the predictive performances of a classifier.

The first indicator is the Sensitivity for class C:

$$Sens_C = \frac{n_{CC}}{n_{C\bullet}}$$

which expresses how well the classifier recognizes a unit belonging to the class C. Connected with it there is the Positive Predictive Value of class C:

$$PPV_C = \frac{n_{CC}}{n_{\bullet C}}$$

that gives a measure of the probability that a unit truly belongs to the class C given that its prediction is the class C. It can be of interest to compare the sensitivities of the classes computing their difference in absolute value; a useful indicator from these differences should be their maximum (Maximum Distance Between Sensitivities - MDBS), that is:

$$MDBS = \max_{i \neq j} |Sens_{C_i} - Sens_{C_j}|$$

The lower the MDBS, the better the classification. Considering only the the cases of correct classification for class C, it is possible to compute the Precision for the classifier as:

$$Prec = \frac{\sum_C n_{CC}}{n}$$

where n is the sample size. Rearranging the confusion matrix as a 2×2 matrix with reference of each class, it is possible to compute the Accuracy for the class C:

$$ACC_C = \frac{n_{CC} + n_{\overline{C}\overline{C}}}{n}$$

where $\overline{C}$ is the complement of class C, and then their average obtaining the Average Accuracy (AA), which is the average per-class effectiveness of a classifier. The last indicator considered here, called Mean Weighted Classification Error (MWCE), originates from the idea of a different impact of different misclassifications and it attributes score 0 to perfect classification, score 1 to the case in which the classifier predict win (loss) and the actual value is loss (win), and score 1/3 otherwise. MWCE is a weighted average of the previous three scores, $S_j = 0, 1/3, 1$:

$$MWCE = \sum_{j=1}^{3} S_j \cdot V_j$$

with weights respectively equal to:

$$\begin{aligned} V_1 &= \frac{n_{LL} + n_{DD} + n_{WW}}{n} \\ V_2 &= \frac{n_{DW} + n_{WD} + n_{DL} + n_{LD}}{n} \\ V_3 &= \frac{n_{LW} + n_{WL}}{n} \end{aligned}$$

The lower the MWCE, the better the classification.

4. Results and conclusions

The predictive performance indices described in Section 3 were evaluated taking a random samples of 2,087 matches form the 2,587 available as training set, used to estimate the BN, and the remaining 500 matches as test set. All the variables in the test set, except for the match result which plays the rule of the target variable X', were considered as new evidence Ev to be used to compute the posterior distribution $P(X'|Ev)$. In order to obtain bootstrap standard errors, the 500 matches results were randomly sampled 1,000 times. Table 2 reports the mean values of all the predictive performance indices except for MDBS and AA (standard errors in parenthesis), whereas Figure 1 compares MDBS with AA.

The modal classifier M1, which is the standard one, has the highest precision, even if is not different from the one of other classifiers such as M3 or M5.1, a high average accuracy, but also the highest MWCE and a high MDBS

Table 2. Mean values of the predictive performance indices of the considered classifiers based on the prediction of 500 matches results randomly sampled 1,000 times (standard errors are in parenthesis)

Classifier	$Prec$	MWCE	$Sens_L$	$Sens_D$	$Sens_W$	PPV_L	PPV_D	PPV_W
M1	0.502	0.314	0.458	0.020	0.806	0.451	–	0.533
	(0.02)	(0.02)	(0.07)	(0.03)	(0.05)	(0.04)	–	(0.03)
M2	0.443	0.244	0.212	0.499	0.551	0.536	0.282	0.602
	(0.02)	(0.01)	(0.06)	(0.07)	(0.06)	(0.07)	(0.03)	(0.03)
M3	0.480	0.297	0.565	0.230	0.573	0.426	0.318	0.598
	(0.02)	(0.02)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)
M4.c	0.455	0.280	–	0.287	0.823	–	0.266	0.531
	(0.02)	(0.02)	–	(0.06)	(0.04)	–	(0.04)	(0.03)
M4.r	0.340	0.234	0.004	0.853	0.246	–	0.273	0.663
	(0.02)	(0.01)	(0.02)	(0.05)	(0.06)	–	(0.02)	(0.05)
M4.f	0.315	0.279	0.445	0.729	–	0.456	0.264	–
	(0.02)	(0.02)	(0.07)	(0.06)	–	(0.05)	(0.02)	–
M5.1.c	0.501	0.297	0.353	0.094	0.823	0.497	–	0.531
	(0.02)	(0.02)	(0.05)	(0.06)	(0.05)	(0.05)	–	(0.03)
M5.1.r	0.494	0.267	0.353	0.260	0.713	0.497	0.297	0.574
	(0.02)	(0.02)	(0.05)	(0.06)	(0.04)	(0.05)	(0.04)	(0.03)
M5.1.f	0.499	0.285	0.445	0.182	0.713	0.456	–	0.574
	(0.02)	(0.02)	(0.07)	(0.07)	(0.04)	(0.05)	–	(0.03)
M5.2.c	0.481	0.285	0.204	0.174	0.823	0.536	0.257	0.531
	(0.02)	(0.02)	(0.06)	(0.06)	(0.04)	(0.07)	(0.05)	(0.03)
M5.2.r	0.442	0.244	0.204	0.502	0.550	0.536	0.281	0.602
	(0.02)	(0.01)	(0.06)	(0.07)	(0.06)	(0.07)	(0.03)	(0.04)
M5.2.f	0.466	0.274	0.445	0.345	0.550	0.456	0.297	0.602
	(0.02)	(0.02)	(0.07)	(0.08)	(0.06)	(0.05)	(0.03)	(0.04)
M5.3.c	0.455	0.280	0.003	0.284	0.823	–	0.265	0.531
	(0.02)	(0.02)	(0.02)	(0.06)	(0.04)	–	(0.04)	(0.03)
M5.3.r	0.373	0.230	0.005	0.785	0.357	–	0.278	0.655
	(0.02)	(0.01)	(0.02)	(0.05)	(0.04)	–	(0.02)	(0.04)
M5.3.f	0.424	0.264	0.445	0.518	0.357	0.456	0.287	0.655
	(0.02)	(0.02)	(0.07)	(0.07)	(0.04)	(0.05)	(0.03)	(0.04)
M5.4.c	0.455	0.280	–	0.287	0.823	–	0.266	0.531
	(0.02)	(0.02)	–	(0.06)	(0.04)	–	(0.04)	(0.03)
M5.4.r	0.340	0.234	0.004	0.853	0.246	–	0.273	0.663
	(0.02)	(0.01)	(0.02)	(0.05)	(0.06)	–	(0.02)	(0.05)
M5.4.f	0.325	0.279	0.445	0.703	0.037	0.456	0.265	–
	(0.02)	(0.02)	(0.07)	(0.07)	(0.05)	(0.05)	(0.02)	–

The symbol – corresponds to the case of an index equal to 0 or infinity

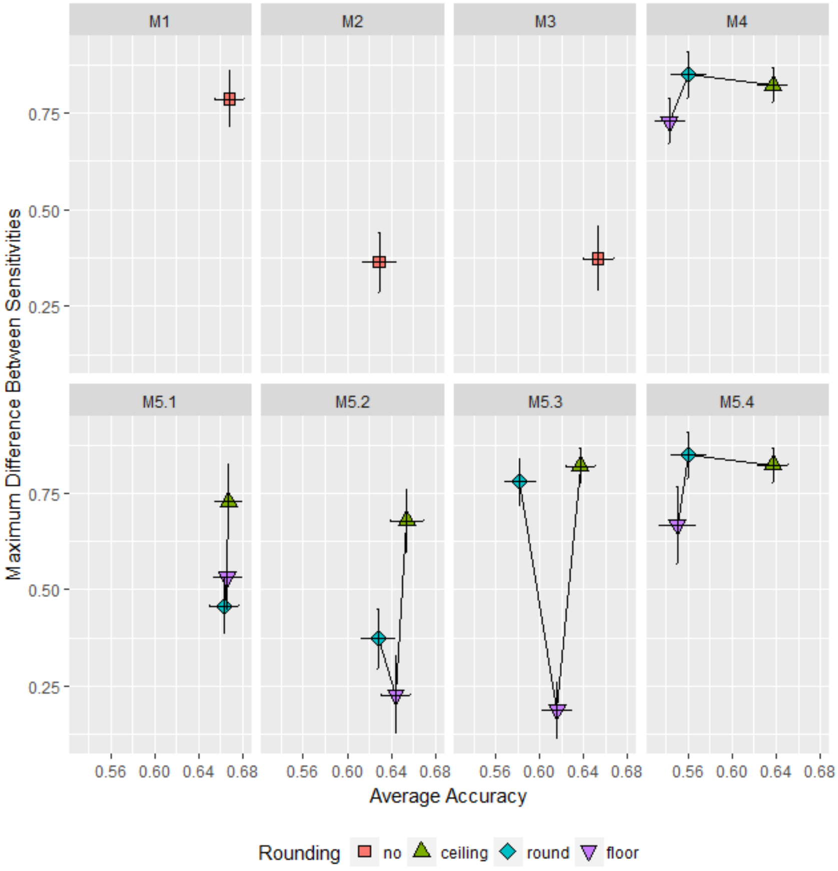


Figure 1. MDBS versus AA for the considered classifiers

due to the fact that mainly it is not able to predict the draw result. Classifiers M2 and M3 allow to increase the ability in predicting the draw, but at the same time classifier M2 partially looses the ability to predict the loss.

Let one consider the classifiers M4 and M5, it has to be noted the role played by the rounding procedure. For example, let one consider the classifier M4, the ceiling and round roundings reduce the possibility to predict the loss, whereas the floor rounding reduces the possibility to predict the win.

From Figure 1, the best combinations of low MDBS and high AA are shown by classifiers M2, M3, M5.2.f and M5.3.f, and between them, M3 and M5.2.f exhibit better combinations of $Prec$, MWCE, $Sens$ and PPV.

The results highlighted the sensibility of the prediction performances to the choice of the classifier: there are classifiers that are unbalanced towards one of the three results of a match, other more balanced.

These results could be related to the dataset in use: an interesting development of the present work is to apply the proposed classifiers, to the leagues of other countries in Europe and verify if similar results are obtained.

References

Carpita M., Golia S. (2018) Exploring the Kaggle European Soccer database with Bayesian Networks: the case of the Italian League Serie A. In *Proceedings of SIS2018 - 49th Meeting of the Italian Statistical Society*, University of Palermo.
Web: meetings3.sis-statistica.org/index.php/sis2018/49th/paper/viewFile/1429/128

Carpita M., Ciavolino E., Pasca P. (2018) Exploring and modelling team performances of the Kaggle European Soccer Database. To appear in *Statistical Modelling*.

Carpita M., Sandri M., Simonetto A., Zuccolotto P. (2015) Discovering the drivers of football match outcomes with data mining, *Quality Technology & Quantitative Management*, 12, 561-577.

Franck E., Verbeek E., Nuesch S. (2010) Prediction accuracy of different market structures - bookmakers versus a betting exchange, *International Journal of Forecasting*, 26, 448-459.

Koller D., Friedman N. (2009) *Probabilistic graphical models: principles and techniques*, MIT Press.

Strumbelj E., Sikonja M.R. (2010) Online bookmakers' odds as forecasts: The case of European soccer leagues, *International Journal of Forecasting*, 26, 482-488.

Multiple imputation and selection of ordinal level-2 predictors in multilevel models

Leonardo Grilli [*], Maria Francesca Marino[**], Omar Paccagnella [***], Carla Rampichini [****]

Abstract: We devise a strategy to handle ordinal level-2 predictors of a two-level random effect model in a setting characterized by two nontrivial issues: (*i*) level-2 predictors are severely affected by missingness; (*ii*) there is redundancy in both the number of predictors and the number of categories of their measurement scale. We tackle the first issue by considering a multiple imputation strategy based on information at both level-1 and level-2. We tackle the second issue by means of regularization techniques for ordinal predictors, also accounting for the multilevel data structure. The work is motivated by a case study at the University of Padua about the relationship between student ratings of a course and several characteristics of the course, including teacher feelings (ordinal predictors) and practices (binary predictors) collected by a specific survey with nearly half missing respondents.

Keywords: Multilevel models, Multiple imputation, Variable selection.

1. Case study

We analyse student satisfaction, as measured by student evaluation of teaching (SET). The peculiarity of the study lies in the availability of many variables measuring teacher characteristics and beliefs, and teaching practices. Indeed, this work exploits a dataset of the University of Padua for academic year 2012/13, merging three different sources: (*i*) the traditional SET survey with 18 items, measured on a ten-point scale (1: low, 10: high); (*ii*) administrative data on students, teachers and didactic activities; (*iii*) a survey carried out by the PRODID project on teacher beliefs and practices.

[*] University of Florence, leonardo.grilli@unifi.it

[**] University of Florence, mariafrancesca.marino@unifi.it

[***] University of Padua, omar.paccagnella@unipd.it

[****] University of Florence, carla.rampichini@unifi.it

Data have a two-level hierarchical structure, with $56,775$ student ratings at level-1 and $1,016$ classes at level-2. The average class size is 79 (min 5, max 442). Our aim is that of analyzing student's satisfaction about two key aspects: teacher ability to involve students (item D06 of the SET questionnaire) and teacher clarity (item D07).

The analysis is based on the following bivariate 2-level linear model for item m (m: 1 for D06, 2 for D07) of student i in class j:

$$Y_{mij} = \alpha_m + \boldsymbol{\beta}'_m \mathbf{x}_{ij} + \boldsymbol{\gamma}'_m \mathbf{w}_j + u_{mj} + e_{mij} \tag{1}$$

where $\mathbf{x}_{mij}$ is the vector of student covariates (level-1) and $\mathbf{w}_{mj}$ is the vector of teacher and class covariates (level-2). Level-1 errors, e_{mij}, are assumed to be independent across students, while level-2 errors (the random effects), u_{mj}, are assumed to be independent across classes and independent from level-1 errors. We make standard assumptions for the distributions of the model errors, including homoscedasticity (within each outcome) and normality. Therefore, the response vector $\mathbf{Y}_{ij} = (Y_{1ij}, Y_{2ij})'$ has residual variance-covariance matrix equal to $Var(\mathbf{Y}_{ij}) = \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$, where $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are the covariance matrices of the errors at level-2 and level-1, respectively.

The survey on teacher beliefs and practices has about fifty percent of missing questionnaires, posing a serious issue of missing data at level-2. An analysis based on list-wise deletion would discard the entire set of student ratings for non responding teachers, causing two main problems: (*i*) a dramatic reduction of the sample size, and thus of the statistical power, and (*ii*) possibly biased estimates if the missing data mechanism is not MCAR.

2. Handling missing data at level-2

In multilevel models, the treatment of missing data requires special techniques. In this framework, data have a hierarchical structure and, thus, missing values can occur at any level of the hierarchy. Furthermore, missing values can alter variance components and correlations, thus, leading to possible misleading inferential conclusions. Multiple imputation (MI) is the most flexible approach to handle missing data. It has been extended to the multilevel setting following two main approaches: a fully conditional specification, also known as multivariate imputation by chained equations (MICE), and a joint modelling (Mistler and Enders, 2017; Grund *et al.*, 2018).

In our case study, although the substantive model (1) is multilevel in nature, missing data are only at level-2. This feature makes the imputation simpler as we directly can apply standard MI techniques to level-2 data and, then, merge level-1 and level-2 data to obtain *complete* datasets. On the other hand, the MI step remains a challenging task as we have to deal (and, thus, impute) a high number of categorical variables with a high percentage of missing information. In particular, about 50% of the teachers did not respond to the whole questionnaire, producing missing values on 10 binary items (teacher practices) and 20 ordinal items (teacher beliefs on a 7-point scale.

According to the available literature, the imputation model at level-2 should include both all the level-2 covariates and information on level-1 variables, in particular the response variables. Several strategies may be adopted to summarize information from level-1 variables (Erler *et al.*, 2016; Grund *et al.* 2017). Here, we consider the cluster mean, which is effective in general and easy to implement in our case, where level-1 variables (including the response) are completely observed. We perform multivariate imputation by chained equations based on binary logit models for the 10 binary items (teacher practices) and cumulative logit models for the 20 ordinal items (teacher beliefs). The imputation model for a given item includes the following covariates: the fully observed class and teacher characteristics, the cluster means of level-1 variables (covariates and outcomes), and the cluster size.

3. Results

The bivariate two-level model in equation (1) is fitted by maximum likelihood on $M = 10$ imputed data sets, and the results are combined with the standard MI rules. The analysis is conducted using the `gsem` and `mi` commands of Stata, version 15. We first fit the model without covariates, to explore the correlation structure of the two outcomes. We find out that the Intraclass Correlation Coefficient (ICC) is about 30% for both the teacher's ability to involve students and teacher's clarity. The two outcomes are highly correlated (0.83), especially at level-2 (0.933).

Then, we add the available covariates in the model. In particular, teacher practices are included as binary indicators, while teacher beliefs are summarized into 6 indicators averaging the relevant seven-point ordinal items, i.e. *passion for teaching* (2 items), *passion for research* (2 items), *need for teaching support* (4 items), *care about student needs* (3 items), *role of active learning* (4 items), *interest in innovative teaching methods* (3 items). The final model has a total of 6 student characteristics and 22 covariates at the second level (5 class variables, 3 teacher characteristics, 8 teacher practices, and 6 teacher beliefs). To keep the number of parameters reasonably small, we start treating the ordinal predictors as if they were effectively continuous, thus assuming they linearly influence the two outcomes under investigation.

Results from the analysis show that, among the objective traits of the teacher included in the model, only age and gender are significantly related with SET ratings on teacher ability. On the other hand, several subjective traits of the teachers (available thanks to the PRODID survey) are significantly related with SET ratings. In particular, teacher beliefs, such as feelings about teaching and need of support to improve teaching activities, seems to be strongly related with the outcomes, while practices turn out to be less relevant.

To quantify the influence of missingness on the sampling variance of a parameter estimates, we can consider the Fraction of Missing Information (FMI - Rubin, 1987). For imputed covariates, FMI ranges from 0.15 to 0.68, with a median value of 0.44, indicating that 44% of the sampling variance is ascribable to missing data. For the imputed covariates with FMI $< 50\%$ (the fraction of missing data), the trade-off between the increase of the standard errors due to MI and its reduction due to data augmentation is favorable, i.e. the relative efficiency is high.

4. Ongoing research: selecting the covariates with regularization techniques

As described above, teacher beliefs are measured by 20 ordinal items based on a 7-point Likert scale. To account for the ordinal nature of such variables in model (1), we should include in the linear predictor 6 dummies per variable that would lead to disproportionate number of parameters, difficult to be estimated and interpreted. A simple way to avoid the issue is to consider ordinal variables as if they were measured on a continuous scale. The results discussed in Section 3 are based on the assumption that the ordinal items affect the outcomes through a set of scales with linear effects. Apart from linearity, the use of scales raises issues of validity.

An alternative, more flexible approach to model the effect of ordinal items is to rely on regularization methods which allow us to explicitly consider the ordinal nature of the predictors while ensuring model parsimony. This latter goal is achieved by jointly identifying the predictors to be included in the model and the categories of each predictor to be distinguished. Clearly, in our case study, standard methods need to be extended to deal with both the hierarchical structure of the data and the missingness.

As regards the former aspect, we aim at considering the penalty term introduced by Gertheiss and Tutz (2010) to handle ordinal predictors

$$J(\boldsymbol{\gamma}) = \sum_{s=1}^{p} \sum_{r=1}^{k_s} |\gamma_{sr} - \gamma_{sr-1}|,$$

where p denotes the number of parameters in $\boldsymbol{\gamma}$ and k_s the number of categories of the s-th predictor (7 in our case). This is combined with the estimation approach based on a gradient ascent algorithm introduced by Groll and Tutz (2014) in the context of generalized linear mixed model for variable selection. As regards the missing data issue, we aim at adopting a strategy based on the selection of the optimal predictors on each imputed dataset; at last, we retain those predictors which result to be significant in at least one of the imputed dataset.

Acknowledgements: The authors gratefully acknowledge the support of the University of Padua project Advances in Multilevel and Longitudinal Modelling, principal investigator Omar Paccagnella, grant no. SID2016

References

Ahrens A., Hansen C.B., Schaffer M.E. (2018) LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation. *Statistical Software Components S458458*, Boston College Department of Economics, revised 07 Apr 2018.

Carpenter J., Kenward M. (2013) *Multiple imputation and its application*. Chichester, United Kingdom: John Wiley & Sons, Ltd.

Erler N.S., Rizopoulos D., van Rosmalen J., Jaddoe V.W.V., Franco O.H., Lesaffre E.M.E.H. (2016) Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian, *Statistics in Medicine*, 35, 2955-2974.

Gertheiss J., Tutz G. (2009) Penalized regression with ordinal predictors, *International Statistical Review*, 77, 345-365.

Gertheiss J., Tutz G. (2010) Sparse modeling of categorical explanatory variables, *The Annals of Applied Statistics*, 4, 2150-2180.

Groll, A., Tutz G. (2014) Variable selection for generalized linear mixed models by L 1-penalized estimation, *Statistics and Computing* 24, 137-154.

Grund S., Ludtke O., Robitzsch A. (2017) Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs, *Journal of Educational and Behavioral Statistics*, XX, 1-38.

Tutz G., Gertheiss J. (2016) Regularized regression for categorical data, *Statistical Modelling*, 16, 161-200.

van Buuren S. (2012) *Flexible Imputation of Missing Data*, Chapman & HallCRC Interdisciplinary Statistics. CRC Press Taylor & Francis Group: Boca Raton, FL.

Vermunt J.K., Van Ginkel J.R., Van der Ark L.A., Sijtsma K. (2008) Multiple imputation of incomplete categorical data using latent class analysis, *Sociological Methodology*, 38, 369-397.

Vidotto D., Kaptein M.C., Vermunt J.K. (2015) Multiple imputation of missing categorical data using latent class models: State of art, *Psychological Test and Assessment Modeling*, 57, 542-576.

Zhao Y., Long Q. (2017) Variable selection in the presence of missing data: imputation-based methods, *WIREs Comput Stat*, 9:e1402. doi: 10.1002/wics.1402.

Why the number of response categories in rating scales should be large

Maria Iannario[*], Anna Clara Monti[**], Pietro Scalera[***]

Abstract: The paper investigates the impact of the number of response categories m on the efficiency of the estimator of the regression coefficients in cumulative models for ordinal data with proportional link. Results point out that efficiency is an increasing function of m.

Keywords: Efficiency, Ordinal data, Cumulative models.

1. Introduction

A crucial point in the collection of ordinal responses is the number of categories m to be made available to the respondents.

This issue has been extensively dealt with in the literature though the proposals are not completely consistent, mostly because of the variety of measurement contexts (e.g., medicine, marketing, psychology, etc.) and optimization criteria (e.g., reliability, validity, sensitivity, information processing, ease-of-use, response time, and so forth) (see the seminal papers of Cox (1980) and Preston and Colman (2000); an updated list of references can be found in Lewis and Erdinç (2017)). Even if there is no unanimous consensus on the choice of m, there is a wide agreement that the scale should be refined enough to collect all the available information without being so refined to encourage response errors. In fact, although the information transmission capacity of a scale is improved by increasing the number of response alternatives, response error seems to increase concurrently. Frequently recommended reference criteria are respondent preference; reliability; validity; need of "uncertain" category; information theoretic measures and statistical efficiency of estimators (see Benson (1971) and Ramsey (1973), among others).

[*] University of Naples Federico II, maria.iannario@unina.it

[**] University of Sannio, acmonti@unisannio.it

[***] University of Naples Federico II, pietroscalera@hotmail.it

Within the statistical literature the choice of m is discussed by Agresti (2010) who points out that a large m allows a more powerful detection of associations between variables; a result confirmed by Allahyari *et al.* (2016) with reference to test on differential item functioning. Furthermore Iannario *et al.* (2016) show that a larger value of m reduces the impact of response errors on the local robustness of the estimators in the modelling framework for ordinal data denoted as CUB models.

The current paper investigates the effect of varying m on the efficiency of the estimators by exploiting the properties of the cumulative models with proportional link (Agresti, 2013). Here it is assumed that the observable ordinal variable Y is linked to an underlying latent variable Y^* through the relationship

$$Y = j \qquad \Longleftrightarrow \qquad \alpha_{j-1} < Y^* \leq \alpha_j, \qquad j = 1, 2, \ldots, m, \tag{1}$$

where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = +\infty$ are the thresholds of the continuous support of Y^*. The latent variable Y^*, in turn, depends on $p \geq 1$ covariates, so that for the i-th statistical unit we have the latent regression model

$$Y_i^* = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + \epsilon_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ and ϵ_i is a random variable whose distribution function is indicated by $G(\epsilon)$.

Relationship (1) yields the following probability mass function for Y_i, conditionally on $\boldsymbol{X}_i = \boldsymbol{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ip})'$,

$$\begin{aligned} P\left(Y_i = j \mid \boldsymbol{x}_i\right) &= P\left(\alpha_{j-1} < Y_i^* \leq \alpha_j\right) \\ &= G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}) \end{aligned}$$

for $j = 1, 2, \ldots, m$.

Formula (1) implies that from the same latent variable Y^*, a countable set of ordinal variables $\left\{Y^{(m)}\right\}$ can be generated by allowing m to vary in $\{3, 4, \ldots\}$. These variables differ from each other for the number of categories. Nevertheless all of them refer to the same latent regression model and therefore the different estimators of the regression coefficient, which are obtained by varying m, estimate always the same $\boldsymbol{\beta}$. This feature of the cumulative model with proportional link can be exploited to analyze how the choice of m affects the efficiency of the estimator and the reliability of the derived inferential procedures.

The paper considers cumulative models with logit, probit and complementary log-log link and it is organized as follows. The next Section provides a brief overview of the likelihood inference, whereas Section 3 describes the design of the experiment. The efficiency of the estimators is investigated in Sections 4 while hypothesis testing is considered in Section 5. Final remarks end the paper.

2. Likelihood inference

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ be the parameter vector, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m-1})'$ is the vector of the thresholds. Given an observed random sample $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{X}$ be the matrix whose rows are given by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The log-likelihood function is $\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$ with individual term

$$\ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) \quad = \quad \sum_{j=1}^{m} I[y_i = j] \log Pr(Y_i = j | \boldsymbol{x}_i) \tag{3}$$

where $I[\omega]$ is an indicator function which takes value 1 when ω holds and 0 otherwise. The Maximuml Likelihood Estimator (MLE) $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is obtained by maximizing (3) (see Iannario *et al.* (2017) for details).

The generic term of the information matrix $\mathcal{I}(\boldsymbol{\theta}, \boldsymbol{X})$ for a single observation, conditionally on $\boldsymbol{X} = \boldsymbol{x}$, is given by

$$\mathcal{I}_{rs}(\boldsymbol{\theta}, \boldsymbol{x}) = E_Y \left\{ \frac{\partial\, \ell(\boldsymbol{\theta}, Y, \boldsymbol{X})}{\partial\, \theta_r} \frac{\partial\, \ell(\boldsymbol{\theta}, Y, \boldsymbol{X})}{\partial\, \theta_s} \middle| \boldsymbol{X} = \boldsymbol{x} \right\}$$

for $(r, s) = 1, 2, \ldots, m + p - 1$. The elements of the unconditional information matrix $\mathcal{I}(\boldsymbol{\theta})$ are given by $\mathcal{I}_{rs}(\boldsymbol{\theta}) = E_{\boldsymbol{X}}\big\{\mathcal{I}_{rs}(\boldsymbol{\theta}, \boldsymbol{X})\big\}$. The asymptotic variance-covariance matrix of the MLE is obtained by inverting the information matrix, i.e. $\mathcal{I}(\boldsymbol{\theta})^{-1}$.

Asymptotically we have $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to N\left(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1}\right)$. In particular for the estimator $\hat{\beta}_k$ of the single regression coefficient β_k we have

$$\sqrt{n}\left(\hat{\beta}_k - \beta_k\right) \to N\left(0, \mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}\right) \tag{4}$$

where $\mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}$ is the element on the diagonal of $\mathcal{I}(\boldsymbol{\theta})^{-1}$ corresponding to β_k.

3. The design of the experiment

To investigate the efficiency of the estimators when m varies, the following regression models for the latent variable Y^* are considered.

- **Model 1** *(with one continuous covariate).* The latent variable depends on a continuous covariate

  $$Y^* = X\beta + \epsilon\,,$$

  where $X \sim N(0,1)$ and $\beta = 1.5$.

- **Model 2** *(with dichotomous covariates).* The latent variable depends on two dichotomous covariates

  $$Y^* = X_1\beta_1 + X_2\beta_2 + \epsilon\,,$$

  where $X_1 \sim Ber(0.5)$, $X_2 \sim Ber(0.25)$ and X_1 and X_2 are mutually independent. The regression coefficients are $\beta_1 = 1.5$ and $\beta_2 = 0.7$.

- **Model 3** *(with mixed covariates).* The latent variable depends on a continuous covariate, a dichotomous one and their interaction. The regression model is

  $$Y^* = X_1\beta_1 + X_2\beta_2 + X_1X_2\beta_3 + \epsilon\,,$$

  where $X_1 \sim N(0,1)$, $X_2 \sim Ber(0.5)$ and X_1 and X_2 are mutually independent. The regression coefficients are $\beta_1 = 2.7$, $\beta_2 = 1.5$ and $\beta_3 = 0.7$.

The probit link assumes $\epsilon \sim N(0,1)$, the logit link assumes a logistic distibution for ϵ and the complementary log-log link is such that $G(\epsilon) = 1 - \exp\{-\exp(\epsilon)\}$ (McCullagh and Nelder, 1989).

The thresholds are assumed to be equidistant, that is they satisfy the constraint $\alpha_j - \alpha_{j-1} = h$. The distance is $h = (Y^*_{0.975} - Y^*_{0.025})/m$, where $Y^*_{0.025}$ e $Y^*_{0.975}$ are the 2.5% and 97.5% percentiles of the distribution of Y^* simulated from 5 millions of observations. The thresholds are centered around the median.

The efficiency of the MLE is assessed through a Monte Carlo experiment. Initially, for the three models, $10,000$ samples of the response variable Y^* are generated from the latent regression model (2). Then, for any values of m between 3 and 11, the thresholds are identified and the values of Y^* are transformed into ordinal responses by applying (1). The estimates are computed by means of the R package `ordinal` under the constrain $\alpha_j - \alpha_{j-1} = h$.

4. The efficiency of the estimators

The efficiency of the estimators of the regression coefficients is summarized in Table 1. In Model 1 there is a single regression coefficient, so that the efficiency is measured by the Mean Square Error $MSE(\hat{\beta})$. In case of Models 2 and 3, where there is a vector of regression coefficients, the efficiency is measured by the trace of the portion of the MSE matrix related to $\hat{\boldsymbol{\beta}}$, that is $\sum_{k=1}^{p} MSE(\hat{\beta}_k)$. The sample size is $n = 500$.

Table 1. Efficiency of the estimators of the regression coefficients ($MSE \times 10$*)*

		Model 1			Model 2			Model 3	
m	Probit	Logit	Cloglog	Probit	Logit	Cloglog	Probit	Logit	Cloglog
3	0.094	0.151	0.111	0.305	0.779	0.385	1.451	1.998	1.563
4	0.072	0.127	0.086	0.268	0.699	0.321	0.960	1.447	1.036
5	0.061	0.113	0.072	0.253	0.666	0.286	0.724	1.204	0.801
6	0.056	0.110	0.067	0.244	0.651	0.268	0.593	1.080	0.687
7	0.053	0.105	0.063	0.238	0.642	0.259	0.530	1.007	0.609
8	0.051	0.104	0.06	0.236	0.632	0.249	0.493	0.968	0.553
9	0.049	0.102	0.058	0.234	0.631	0.245	0.457	0.927	0.523
10	0.049	0.100	0.057	0.232	0.625	0.242	0.428	0.909	0.487
11	0.048	0.100	0.055	0.231	0.626	0.238	0.415	0.887	0.468

The results clearly point out that the efficiency of $\hat{\boldsymbol{\beta}}$ is an increasing function of m. When m increases, each category of Y corresponds to a smaller class on the support of the latent variable Y^*. The larger amount of information available on the latent variable, provided by a finer categorization, yields more efficient estimators.

In particular the decrease of the MSEs is especially marked for low values of m, and tapers off around $m = 10$. This behavior is shared by the three models and by the three link functions and it is consistent with the main findings of the psychometric literature, although derived through different optimization criteria (see Nunnally (1978), among others).

Notice that the efficiency of $\hat{\boldsymbol{\beta}}$ affects the efficiency of the estimators of the odds ratio (OR) which, therefore, depends on m too. For instance Figure 1 shows, for Model 3 with the logit link, the boxplot of the estimator of $OR(X_2|X_1 = 1.5)$ when m varies. Here it is possible to observe - when m increases - a reduction of both the interquartile distance and the length of the whiskers. In addition the occurrence of anomalous data decreases.

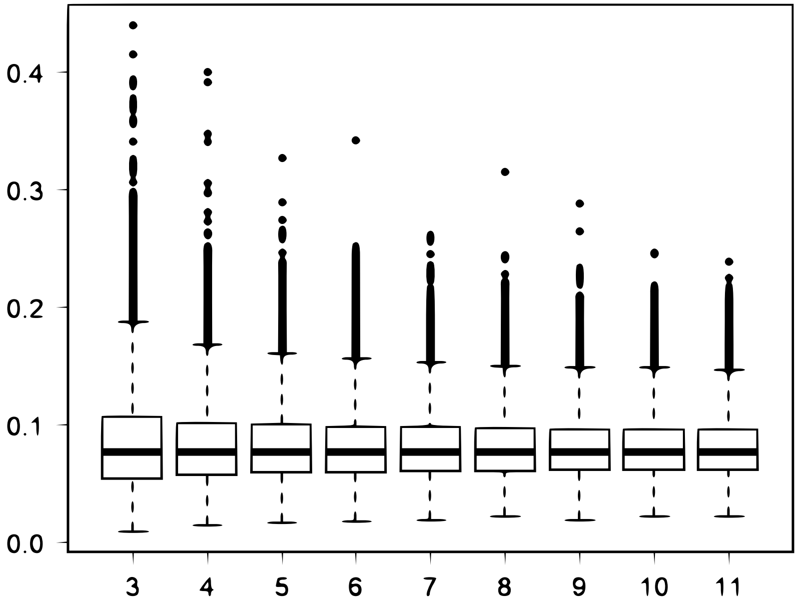


Figure 1. Boxplot of the estimators of the $OR(X_2|X_1 = 1.5)$ in Model 3 with logit link

5. Hypothesis testing

The impact of the choice of m on the efficiency of the estimators affects also the power of the test. Consider the hypothesis on a single regression coefficient $H_0 : \beta_k = \beta_k^0$ versus $H_1 : \beta_k \neq \beta_k^0$. It can be tested through a t-type statistic $t = (\hat{\beta}_k - \beta_k^0)/SE(\hat{\beta}_k)$ where the standard error is given by $SE(\hat{\beta}_k) = \sqrt{n\mathcal{I}(\hat{\boldsymbol{\theta}})^{\beta_k\beta_k}}$. By (4), under H_0, the test statistic is asymptotically $N(0,1)$ distributed. The null hypothesis is rejected when $|t| > z_{1-\alpha/2}$ where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the standard normal distribution function.

To investigate the effect of varying m on the power of the test we consider the null hypothesis $H_0 : \beta_3 = 0$ in Model 3 (which implies that the interaction between X_1 and X_2 is omitted from the latent model). Table 2 shows the power of the test at the 5% significance level, for the sample sizes $n = 250, 500$ and for the three link functions. The power, computed as percentage of rejection of H_0, clearly increases with m. A large m is especially recommended when the cumulative logit model is adopted since the power can be very low for small m.

Table 2. Power of the test on the null hypothesis $\beta_3 = 0$ in Model 3 at the 5% significance level

	Probit		Logit		Cloglog	
m	$n = 250$	$n = 500$	$n = 250$	$n = 500$	$n = 250$	$n = 500$
3	63.1	90.4	39.5	71.0	59.7	87.9
4	83.0	98.6	54.2	85.9	79.5	97.5
5	91.8	99.7	63.8	91.5	89.1	99.5
6	95.6	100.0	69.2	94.5	93.5	99.9
7	97.6	100.0	72.5	95.9	96.5	100.0
8	98.3	100.0	75.6	96.6	97.5	100.0
9	98.7	100.0	77.0	97.3	98.1	100.0
10	99.1	100.0	77.9	97.6	98.6	100.0
11	99.3	100.0	79.1	97.8	98.9	100.0

6. Final remarks

By increasing m more information become available on the latent variable underlying the observable response which allow more efficient estimation and more powerful tests.

The gain in efficiency is especially marked for small values of m while it gets smaller as m increases. These findings are in agreement with the recommendations from psychometric literature though the latter are obtained with different criteria (see Preston and Coleman (2000); Fox and Jones (1998), among others).

References

Agresti A. (2013) *Categorical Data Analysis*, 3^{nd} edition, J.Wiley & Sons, Hoboken.

Agresti A. (2010) *Analysis of Ordinal Categorical Data*, 2^{nd} edition, J.Wiley & Sons, Hoboken.

Allahyari E., Jafari P., Bagheri Z. (2016) A Simulation Study to Assess the Effect of the Number of Response Categories on the Power of Ordinal Logistic Regression for Differential Item Functioning Analysis in Rating Scales, *Computational and Mathematical Methods in Medicine*, http://dx.doi.org/10.1155/2016/5080826.

Benson P.H. (1971) How many scales and how many categories shall we use in consumer research? A comment, *Journal of Marketing*, 35, 59-61.

Cox E.P. (1980) The Optimal Number of Response alternatives for a scale: A review. *Journal of Marketing Research*, 17, 407-422.

Fox C.M., Jones J.A. (1998) Use of Rasch modeling in counseling psychology research, *Journal of Counseling Psychology*, 45, 30-45.

Iannario M., Monti A.C., Piccolo D. (2016) Robustness Issues for CUB Models, *TEST*, 25, 731-750.

Iannario M., Monti A.C., Piccolo D., Ronchetti E. (2017) Robust inference for ordinal response models, *Electronic Journal of Statistics*, 11, 3407-3445.

Lewis, J.R., Erdinç O. (2017) User Experience Rating Scales with 7, 11, or 101 Points: Does It Matter? *Journal of Usability Studies*, 12, 73-91.

McCullagh P., Nelder J.A. (1989) *Generalized Linear Models*, 2^{nd} edition, Chapman & Hall, London.

Nunnally J.C. (1978) *Psychometric theory*, New York, NY: McGrawHill.

Preston C.C., Colman A.M. (2000) Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences, *Acta Psychologica*, 104, 1-15.

Ramsey J.O. (1973) The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values, *Psychometrika*, 38, 513-532.

Solutions to issues with partial proportional odds models

Altea Lorenzo-Arribas*, Mark J. Brewer **, Antony M. Overstall***

Abstract: Proportional Odds Models (POMs) are still the most commonly used cumulative link models despite clear arguments by some authors supporting the fact that Partial Proportional Odds Models (PPOMs) are "often a superior alternative [to POMs]" (Williams, 2016) and a particularly better and more accurate alternative when the PO assumption is violated by some or all of the explanatory variables". These authors also state however that "the use of PPOMs has itself been problematic or at least sub-optimal." In this paper we focus on one of he most common drawbacks of PPOMs as reported in the literature, alas PPOMs can produce negative predicted probabilities (Hedeker et al., 2006). We propose a reparameterisation that solves this issue and is computationally more efficient than previously proposed Lasso penalisations, and compare the results for both approaches.

Keywords: Partial proportional odds models, Lasso penalisation, Reparameterisation.

1. Partial proportional odds models

Given an ordinal response variable Y_i with C ordered categories, we define a Partial Proportional Odds Model (PPOM) as follows:

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k X_{ik} + \sum_{k=1}^{q} \gamma_{jk} Z_{ik} \tag{1}$$

with $i = 1, \ldots, n$; $j = 1, \ldots, C-1$; $-\infty < \alpha_1 < \alpha_2 < \ldots < \alpha_{C-1} < \infty$; $\gamma_{jk} = \gamma_k + u_{jk}$ with $\gamma_{1k} = 0$ for all $k = 1, \ldots, q$, and $q \leq p$ (Peterson & Harrell, 1990). β_k corresponds to the covariates for which the PO holds and are also referred to as 'global effects' (Poβnecker & Tutz, 2016) , while we get parameters γ_{jk} for the covariates for which we relax the PO assumption, and are also known as 'category-specific effects' (Poβnecker & Tutz, 2016). Each individual parameter γ_{jk} can be additively broken down into a fixed

[*] Biomathematics and Statistics Scotland, altea.lorenzo-arribas@bioss.ac.uk

[**] Biomathematics and Statistics Scotland, mark.brewer@bioss.ac.uk

[***] University of Southampton, A.M.Overstall@soton.ac.uk

component γ_k and an individual component u_{jk}, i.e., $\gamma_{jk} = \gamma_k + u_{jk}$ where u_{jk} represents the deviation of γ_{jk} from the "typical" value γ_k in the population for individual j.

Peterson & Harrell (1990) also define the corresponding log-likelihood as:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{C} I_{ij} \log(\pi_{ij}) \tag{2}$$

with $i = 1, \ldots, n$; $j = 1, \ldots, C$. where an indicator variable is defined as:

$$I_{ij} = \begin{cases} 1 & if \quad Y_i = j \\ 0 & if \quad Y_i \neq j \end{cases} \tag{3}$$

and the probabilities π_{ij} are defined as follows:

$$\pi_{ij} = P(Y_i = j) = \begin{cases} P(Y_i \leq 1) & if \quad Y_i = 1 \\ P(Y_i \leq j) - P(Y_i \leq j-1) & if \quad 1 < Y_i < C \\ 1 - P(Y_i \leq C-1) & if \quad Y_i = C \end{cases} \tag{4}$$

where $P(Y_i \leq j)$ is the cumulative probability that a given observation is less than the j-th level and for $j = 1, \ldots, C$ we have that $P(Y_i \leq C) = 1$. Our initially proposed solution is to add a penalty J to the original log-likelihood L (defined in Formula (5)) via Lasso (Tibshirani, 1996):

$$\hat{\gamma} = argmin(-L + J) = argmin(-\sum_{i=1}^{n} \sum_{j=1}^{C} I_{ij} \log(\pi_{ij}) + \sum_{k=1}^{q} \lambda_k \sum_{j=1}^{C-1} |u_{jk}|) \tag{5}$$

where the penalisation only applies to u_{jk} - the category-specific component of $\gamma_{jk} = \gamma_k + u_{jk}$ - , and $\lambda_k \geq 0$ are the tuning or shrinkage parameters that we will determine by cross-validation.

2. Re-parameterisation

While Lasso penalisation performs well for small samples, its model selection capabilities are limited (Zhao & Yu, 2006). We propose as an alternative a geometric reformulation of the model which also guarantees that class probabilities will be non-negative. The proposed parameterisation for the log-likelihood is derived from two straight lines for which we impose a restriction so that they do not overlap within the stated limits. In order to find the parameter values, we re-express the original definition of PPOMs as:

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \gamma_j Z_i, \tag{6}$$

where $p = 0$ and $q = 1$, and $i = 1, \ldots, n$.

2.1. One covariate

In order to avoid the crossing of regression lines, for an ordinal response variable with 3 categories, we would require that $P(Y_i \leq 1) \leq P(Y_i \leq 2)$, that is:

$$\alpha_1 + \gamma_1 z_i \leq \alpha_2 + \gamma_2 z_i \tag{7}$$

or

$$\begin{cases} \gamma_2 \geq (\alpha_1 - \alpha_2)/z_i + \gamma_1 & for \quad z_i \geq 0 \\ \gamma_2 \leq (\alpha_1 - \alpha_2)/z_i + \gamma_1 & for \quad z_i \leq 0 \end{cases} \tag{8}$$

When we assume a minimum for our data $z_{min} = 0$ and a maximum of $z_{max} = 1$, we find that:

$$\gamma_2 \geq (\alpha_1 - \alpha_2) + \gamma_1 \tag{9}$$

where $(\alpha_1 - \alpha_2) \leq 0$. (This is one of the many possible combinations).

We could have the following parameterisation for a 2 categories response variable:

$$\begin{cases} \alpha_1 = \alpha_1^*; & \alpha_2 = \alpha_1 + \alpha_2^* & where \quad \alpha_1^*, \alpha_2^* \geq 0 \\ \gamma_1 = \gamma_1^*; & \gamma_2 = \gamma_1 + (\alpha_1 - \alpha_2) + \gamma_2^* & where \quad \gamma_1^*, \gamma_2^* \geq 0 \end{cases} \tag{10}$$

Or we could define more conveniently: $a_2 = log(\alpha_2^*)$ and $g_2 = log(\gamma_2^*)$ and apply this new parameterisation to the corresponding log-likelihood.

This method imposes the constraint systematically rather than arbitrarily. This parameterisation could be extended to models with more covariates.

2.2. Two covariates

This parameterisation could be extended to models with more covariates (see Figure 1 for the case of two covariates).

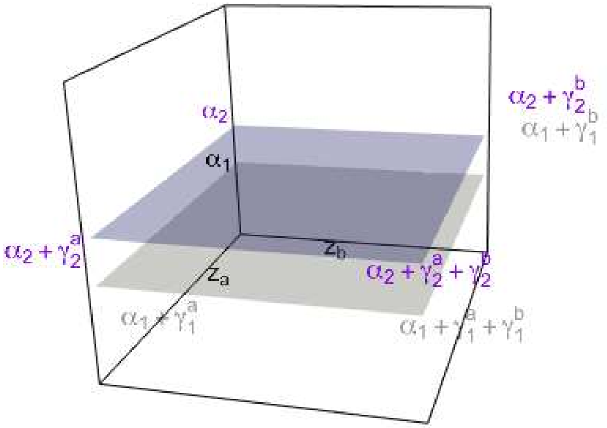


Figure 1. Re-parameterisation for a PPOM with two covariates z_a and z_b.

In order to avoid the crossing of the planes, we would have the following conditions for the four corners:

$$\begin{cases} \alpha_2 > \alpha_1 \\ \alpha_2 + \gamma_2^a > \alpha_1 + \gamma_1^a \\ \alpha_2 + \gamma_2^a + \gamma_2^b > \alpha_1 + \gamma_1^a + \gamma_1^b \\ \alpha_2 + \gamma_2^b > \alpha_1 + \gamma_1^b \end{cases} \tag{11}$$

In summary,

$$\alpha_0 + \mathbf{z}^{\mathbf{t}}\gamma_{\mathbf{0}}{}' \leq \ldots \leq \alpha_{C-1} + \mathbf{z}^{\mathbf{t}}\gamma_{\mathbf{C-1}} \tag{12}$$

For further dimensions, a similar approach would be necessary (e.g., 8 corners for cube in 3-D).

3. Case studies

3.1. Environmental attitudes

We apply our new parameterisation to data from the Scottish Environmental Attitudes and Behaviours Survey (Scottish Government, 2008) which evaluates respondents' awareness of environmental issues and their greener behaviours including; knowledge and attitudes towards climate change, travel behaviour, and eco-friendly purchasing. The study found that high environmental engagement is more concentrated among certain groups in the population, with educational attainment, social class, and age being the strongest predictors. We have assessed different ordinal models including those variables. We have focused on one where we found crossing of regression lines. It models *educational attainment* versus *age* via a PPOM for which the PO assumption is relaxed for the continuous variable *age* (Figure 3). Although we acknowledge that for an appropriate analysis of the data, we would need to control for other covariates (e.g., sex), for the purposes of this methodological study, we start with one covariate only. Both the Lasso and the reparameterisation fix the problem (Figure 4).

3.2. Eye disease risk factors

We then look at data from the Wisconsin Epidemiological Study of Diabetic Retinopathy (Agresti, 2010). The primary outcome is *severity of retinopathy* which was measured in the left and right eye of every subject (ordinal variable with categories; none, mild, moderate, and proliferative). For the sake of simplicity, we restrict our data to the left eye and model it as a function of the *left eye refraction index* and *systolic blood pressure* (both continuous variables) and ignore subject effect (Figure 4). The Lasso penalisation does

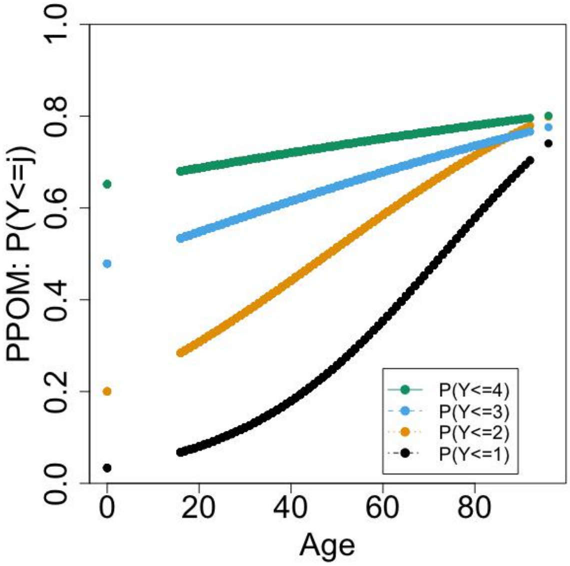


Figure 2. Environmental predictions. Predictions from PPOM versus age.

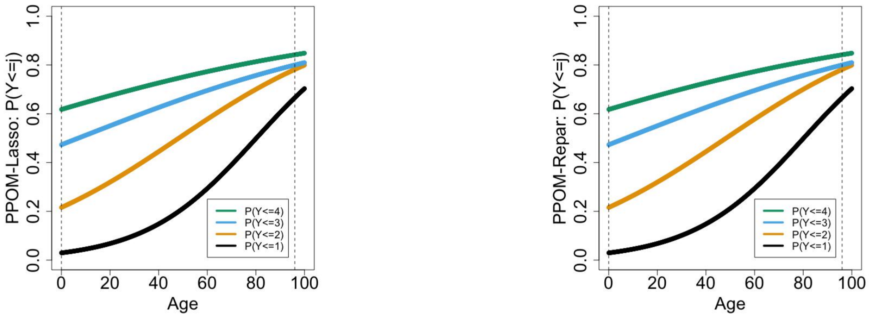


Figure 3. Environmental attitudes. Predictions from Lasso penalised (left) and reparameterised (right) PPOM versus age.

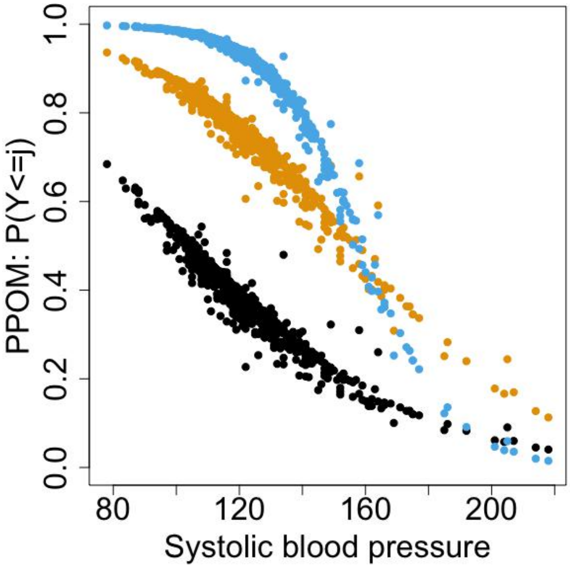


Figure 4. Eye disease risk factors. Predictions from PPOM versus systolic blood pressure.

not fully fix the issue while the reparameterisation does not show any overlap within the range of the covariate (Figure 5).

4. Conclusion

Issues associated to PPOMs can be overcome by different methods. Lasso penalisation requires a choice of shrinkage parameter, which can be challenging. In addition to this limitation, it does not necessarily fix the issue in all cases. Our proposed reparameterisation is a more systematic and computationally efficient approach and has proven to work consistently for the two examples under study.

Acknowledgements: This work was partially supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS).

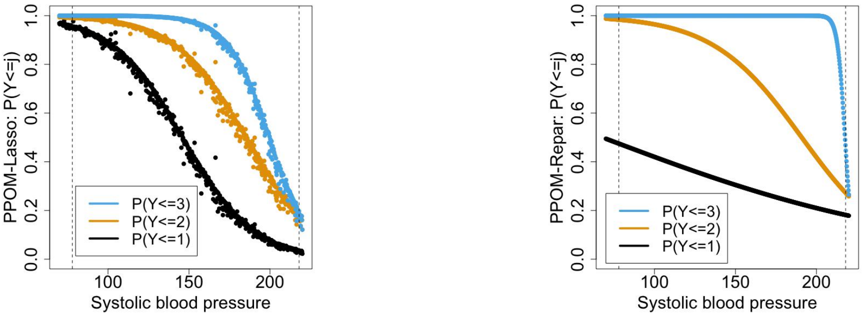


Figure 5. Eye disease risk factors. Predictions from Lasso penalised (left) and reparametersided (right) PPOM versus systolic blood pressure.

References

Agresti A. (2010) *Analysis of ordinal categorical data*, Wiley, New Jersey.

Hedeker D., Berbaum M., Mermelstein R. (2006) Location-scale models for multilevel ordinal data: between- and within-subjects variance modelling, *Journal of Probability and Statistical Science*, 4, 1-20.

Peterson B., Harrell F.E. (1990) Partial proportional odds models for ordinal response variables, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 39, 205-217.

Poβnecker, W., Tutz G. (2016) A general framework for the selection of effect type in ordinal regression, *Ludwig-Maximilians-Universitat Munchen Technical report*, 186.

Scottish Government (2008) The Scottish environmental attitudes and behaviours survey 2008-2009. IPSOS Mori.

Tibshirani R. (1996) Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B*, 40(1), 267-288.

Williams R. (2016) Understanding and interpreting generalized order logit models, *The Journal of Mathematical Sociology*, 58, 7-20.

Zhao P., Yu B. (2006) On model selection consistency of Lasso, *Journal of Machine Learning Research*, 7, 2541-2563.

A latent variable model for a derived ordinal response accounting for sampling weights, missing values and covariates

Fulvia Pennoni*, Miki Nakai**

Abstract: We consider a latent class model especially tailored for an ordinal response derived by comparing two continuous variables. We propose a general method to estimate the model parameters with survey data when there are missing responses and survey weights. First, we estimate the model with the missing responses without covariates with a weighted likelihood function maximised through the Expectation-Maximization algorithm. In order to determine the suitable number of latent classes we rely on the Akaike Information Criterion. Second, by fixing the parameters of the measurement model we estimate the remaining parameters by adding the full set of covariates. We make predictions on the basis of the maximum a posteriori probability. In the application, we consider Japanese survey data collected at four waves covering 40 years with the aim to study changes on couples' breadwinning patterns.

Keywords: Akaike information criterion, Expectation-Maximization algorithm, Gender inequality, Household income composition.

1. Introduction

The latent class model (Lazarsfeld and Henry, 1968) has been considered for the analysis of data arising in different contexts by many authors since it is a flexible model to account for the heterogeneity among responses provided by different individuals which cannot be explained by means of the observable covariates. This model is especially tailored for an ordinal response variable when it has been derived for example by comparing values of two or more continuous variables. It is a model-based approach that properly accounts for the underlying latent continuous responses and allow us to investigate the associations with the covariates as well as to dispose of data driven typologies of individuals (see, among others, Pennoni, 2014). Another advantage is that

*Department of Statistics and Quantitative Methods, University of Milano-Bicocca, fulvia.pennoni@unimib.it

**College of Social Sciences, Ritsumeikan University, mnakai@ss.ritsumei.ac.jp

it is possible to elaborate the model in many ways and to assess the tenability of the underlying hypothesis.

Maximum likelihood estimation of the model parameters is well established and it is carried out through the Expectation-Maximization algorithm (see, among others Bartolucci *et al.*, 2013). However, the use of weighted methods for the estimation of the parameters with missing responses and covariates still deserves research. In the current proposal, instead of performing listwise deletion we rely on the missing at random assumption and we retain the missing responses for the outcome, while the values of the missing covariates are imputed through multivariate imputation by chained equations. In this way, we allow the allocations on the latent variables at individual level also for individuals not providing a response.

In Section 2 we introduce the model and the steps of the maximum likelihood estimation. In Section 3 we describe the data collected within the Japanese Stratification and Social Psychology Survey and in Section 4 we show the main results.

2. *The proposed model*

In the following, we deal with a derived response variable and we introduce the latent class model to account for the missing responses assuming that they are conditionally independent given the latent variable and the observed covariates as well as for survey weights for the representativeness of each unit in the population.

With reference to a random unit drawn from the population of interest let Y_{ij} be the observed derived variable with $j, j = 1, \dots, r$ ordered categories for individual $i, i = 1, \dots, n$. This response is obtained by comparing two or more continuous variables. We assume that the observed response depends on the underlying unobserved latent variable denoted as U_i which has a distribution with k support points assuming finite discrete values. The observed responses are independent one another conditionally to this latent variable.

The first set of parameters in the model is related to the probability to belong to each latent class. These probabilities may be influenced by time-specific individual covariates arranged in the vector $\boldsymbol{X}$ where $\boldsymbol{x}$ is a corresponding realization. We use a baseline category logit model for the following parameters

$$\log\frac{p(U=u|\boldsymbol{X}=\boldsymbol{x})}{p(U=1|\boldsymbol{X}=\boldsymbol{x})} \quad = \quad \log\frac{\pi_{u|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \beta_{0u} + \mathbf{x}'\beta_{\mathbf{1u}}, \quad u = 2,\dots,k, \; (1)$$

where β_0 is an intercept specific of each latent class and $\boldsymbol{\beta}_{1u}$ is the vector of parameters that define the influence of the covariates on the distribution of the latent variable.

Another set of parameters is referred to the manifest part of the model and is given by the conditional probability of each response category given the latent variable denoted as

$$\phi_{j|u} = p(Y_j = y|U = u), \quad u = 1,\dots,k,\; j = 1,\dots,r.$$

To account for individual sampling weights denoted as w_i, $i = 1,\dots,n$ such as that provided with survey data we propose to estimate the model through a weighted log-likelihood. The latter is determined given a sample of n independent individuals for which we observe the responses $y_1,\dots,y_n$ as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i\ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \log p(y_i,\dots,y_n),$$

where $\boldsymbol{\theta}$ denotes the overall vector of free parameters arranged in a suitable way. The above quantity is maximized through the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). It is based on the *complete data log-likelihood* and it represents the main tool to estimate the models with latent variables. For more details see Bartolucci, Farcomeni and Pennoni (2013).

To avoid that parameters referred to the covariates are biased we perform a two step estimation procedure. First, we estimate the model with the missing responses and sampling weights excluding the covariates and we execute a model selection strategy to choose the proper number of latent classes. We perform the model estimation several times to check for local maxima and we rely on the AIC criterion (Akaike, 1973). The latter is a measure of the relative goodness of it of a model, accounting simultaneously for the accuracy and complexity of the model since it is defined on the basis of the following index

$$\mathrm{AIC} = -2\,\hat{\ell}(\boldsymbol{\theta}) + 2\#\mathrm{par},$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood and $\#\mathrm{par}$ denotes the number of free parameters of the model. Then, by fixing the parameters of the measurement model we estimate the remaining parameters by considering the full set of covariates. Standard errors for the parameters estimates are obtained according to the observed information matrix computed through numerical methods.

Once all the parameters have been estimated, the estimated *a-posteriori* probability to be assigned to a latent class is determined as

$$\hat{q}_u = \frac{\prod_{j=1}^{r} \hat{\phi}_{j|u}\hat{\pi}_{u|\boldsymbol{x}}}{\sum_{u=1}^{k}\prod_{j=1}^{r} \hat{\phi}_{j|u}\hat{\pi}_{u|\boldsymbol{x}}}, \quad u = 1,\ldots,k, \quad j = 1,\ldots,r. \qquad (2)$$

In this way, we dispose of a suitable allocation rule for each individual to be assigned to the latent class having the maximum *a-posteriori* probability.

3. Data

The proposed model is applied to explore the coherent breadwinning arrangement classes and to estimate the effects of the covariates on the underlying latent variable. The data are related to spouses within the households and were obtained from the past three decades (1985, 1995, 2005) of Japanese cross-sectional data of the Social Stratification and Social Mobility (SSM) surveys, and the last decade (2015) of the Japanese Stratification and Social Psychology (SSP) survey.

The respondents are interviewed and asked a wide range of questions such as respondents' socioeconomic background. The derived response variable of interest is couple's income provision-role type consisting of five ordinal categories obtained by comparing the declared incomes (earned and investment incomes) and it has been constructed on whether a dominant provider exists and who s/he may be. This response is of primary importance since marriage between man and women in Japan has been considered the only way to form a family until recently, and a necessary way for women's financial sur-

Table 1. Observed and missing frequencies in 2015 for the response variable weighted with the survey weights: (1) "husband sole provider", (2) "husband provides majority", (3) "equal providers", (4) "wife provides majority", (5) "wife sole provider".

Response categories (%)	1	2	3	4	5	Missing
Income provision-role type	22.9	42.2	11.8	5.3	0.6	17.2

vival, social interaction and personal well-being. Moreover, the trend towards dual-earner families can be detected in recent years but gender division of labor has been accepted as "normal" and still strong in Japan. Many studies, see Sorensen and McLanahan (1987), argued that women's economic dependency on men is an important attribute of stratification systems and essential force in the maintenance of gender inequality.

In Table 1 we report the observed frequencies for the last wave concerning 2,497 couples.[5] We notice that despite the continuing rise in Japanese women's participation in the economy as well as in many Western societies, husbands until recently have been the sole or the primary breadwinner in 65% of the couples and equal-provider couples have been only 11.8%.

The available covariates are chosen according to subject matter knowledge for example couples' relative education-level between spouse has been considered to measures whether wife has equal, higher or lower education level than husband.[6]

4. Results

We report the results of the model estimated on the data collected in 2015 due to space limitations. First, we performed a multivariate imputation for the missing values reported for age and husband income by a using weighted

[5] The missing values for the response are due to *missing household* and/or wife's income information.

[6] List of covariates and corresponding categories: *wife's age*: $\leq$ 32, (32,37], (37,40]; (40,44]; (44,47]; (47,51]; (51,55]; (55,58]; (58,61]; $>$ 61; *husband's age:* $\leq$ 34; (34,39]; (39,43]; (43,46]; (46,50]; (50,54]; (54,58]; (58,61]; (61,64]; $>$ 64; *husband's income in ten thousands yen*: $\leq$ 175, (175,275], (275,325]; (325,375]; (375,425]; (425,500]; (500,600]; (600,700]; (700,900]; >900; *size of the place of living*: major cities; $\geq$ 200,000; [100,000,200,000]; $<$ 100,000; small towns and villages; number of children: 0,1,2,$>$ 3; *preschool children*: no, yes; *wife's educational level*: less than high school, high school, college degree, higher; *wife's relative education*: equal, lower, higher than the husband level.

Table 2. Estimated conditional probabilities ($\hat{\phi}_{j|u}$) under the selected model of the responses given the latent classes.

Conditional Probabilities ($\hat{\phi}_{j\|u}$)	1	2	3	4	5
latent class 1 (U_T)	1.000	0.000	0.000	0.000	0.000
latent class 2 (U_N)	0.179	0.578	0.161	0.074	0.008

mean matching method according with the sampling weights and with the other covariates as predictive variables. Then, we estimated the latent class model without covariates with a number of latent classes ranging from 1 to 4 by accounting for different initializations of the EM algorithm[7]. The model with two latent classes has the highest maximum log-likelihood at convergence equal to $\hat{\ell} = -2,456.7$, and a lowest AIC value equal to 4,935.4 with 11 free parameters. The two latent subpopulations are disentangled on the basis of the estimated probabilities of the manifest model that are reported in Table 2. According to the results we define the first latent class as that of Traditional couple (U_T) and the second latent class as that of New couple (U_N). The first one is characterized by a high degree of gender role specialization, strong gender based division of work where the husband specialize in market-work and wife in domestic work and caregiver. The second one is comprised mainly of couples where the husband is not the unique provider. A small proportion of the class is husband sole provider 17$\%$ and 58$\%$ are the couples where the husband provide majority. It is important to note that only 16$\%$ is the probability of equal providers.

The other step of the analysis is performed by adding the covariates. The estimates of the logit regression parameters as in Equation (1) affecting the transition from the traditional couples (U_T) to the new couples (U_N) are reported in Table 3 only for the coefficients which resulted to be significant due to space limitations (see footnote 2 for the categories of each covariate). The estimated intercept is positive indicating a general tendency towards the new type of family U_N. We observe that having preschool children shows the highest estimated coefficient whose negative sign indicates that wife hav-

[7] The model is estimated by adapting the functions of the R package `LMest` (Bartolucci, Pandofi and Pennoni, 2017)

Table 3. Estimates of significant logit regression parameters. (Income in ten thousands yen; significance at 10%(†), 5%*,1%**).

Estimates	U_N
$\hat{\beta}_0$	3.777^{**}
wife age ≤ 32	-1.039^{*}
husband's income (600,700]	-1.479^{**}
husband's income (700,900]	-1.221^{**}
husband's income $>$900	-1.265^{**}
wife'education less than high school	$-1.085^{\dagger}$
wife'education higher than husband'edu.	$0.887^{\dagger}$
preschool children	-2.485^{**}
one child	$-0.479^{\dagger}$

ing preschool children tend to belong to the cluster of traditional couples (the estimated odd ratio for them is equal to 0.08). Interestingly, husband's top incomes determine a lower probability towards the U_N.

The spouses's allocation to each latent class is performed through the estimated maximum *a-posteriori* probability, determined as in Equation (2). The percentage of couples predicted in the traditional family structure U_T is 11.21%. For this subpopulation in Table 4 we show the covariates configuration that can be compare with that obtained for the couples assigned to latent class U_N. We notice that none has husband's income less then 1,750,000 yen a year and that 61% of them has preschool children. We expected that the younger couples support gender egalitarian values more and this would be reflected in gender equality in couples earnings structures. However, we found negative association between age and the probability of being in U_N. It is still not normative for young married women to share equal financial responsibilities within household. This is partly because of the chronic shortages of regular childcare arrangements.

Comparing these results with those obtained for the data collected at previous years we found an increase in the proportion of couples in non-traditional families. One of the reasons why the new families has been more represented in the last past decade is that being a conventional single-income household has becoming more difficult due to the recent financial crisis.

Table 4. Weighted frequencies with survey weights of the relevant covariates for couples allocated in latent class U_T (h.s. high school, h.e. husband'education, see footnote 2 for categories).

Covariates (%)	1	8	9	10
wife'age	29.4			
husband's income	0.0	12.5	11.9	15.2
wife'edu. < h.s.	8.4			
wife'edu. > h.e.	6.4			
preschool	61.1			
one child	40.9			

Acknowledgements: Fulvia Pennoni acknowledges the financial support of the grant "Finite mixture and latent variable models for causal inference and analysis of socio-economic data" (FIRB – Futuro in ricerca) funded by the Italian Government (RBFR12SHVV_004). Miki Nakai acknowledges the financial support of the JSPS Grant-in-Aid for Scientific Research (No. 26380658, No. 17K04103 and No. 16H02045 as part of the SSP project). She thanks both the SSM and the SSP committee for the permission to use their data.

References

Akaike H. (1973) Information Theory as an Extension of the Maximum Likelihood Principle. In BN Petrov, C F (Eds.), *Second International Symposium on Information Theory*, 267-281. Budapest: Akademiai Kiado.

Bartolucci F., Farcomeni A., Pennoni F. (2013) *Latent Markov Models for Longitudinal Data*, Boca Raton: Chapman and Hall/CRC press.

Bartolucci F., Pandolfi S., Pennoni F. (2017) `LMest`: An `R` Package for Latent Markov Models for Longitudinal Categorical Data, *Journal of Statistical Software*, 81, 1-38.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society B*, 39, 1-38.

Lazarsfeld P.F., Henry N.W. (1968) *Latent Structure Analysis, Houghton Mifflin, Boston.*

Nakai M. (2017) Changes in couples' breadwinning patterns and wife's economic role in Japan. In: Greselin, F. *et al.* (Eds.), *Proceedings of the conference of the CLAssification and Data Analysis Group*, Universitas Studiorum, Mantova, 1-6.

Pennoni F. (2014) *Issues on the Estimation of Latent Variable and Latent Class Models*, Scholars'Press, Saarbücken.

Sorensen A., McLanahan S. (1987) Married women economic dependency, 1940-1980, *American Journal of Sociology*, 93, 659-687.

Permutation tests for stochastic ordering with ordinal data

Fortunato Pesarin*, Luigi Salmaso**, Huiting Huang***,
Rosa Arboretti****, Riccardo Ceccato*****

Abstract: This article deals with testing for stochastic dominance and for monotonic stochastic ordering. Several solutions to the univariate case based on restricted maximum likelihood ratio tests have been proposed in the literature. These solutions are generally criticized since their asymptotic null distributions are mixtures of chi-squared variables with weights depending on the underlying population distribution F and so the related accuracy is difficult to assess. Further, testing for stochastic dominance and stochastic ordering in multivariate cases by likelihood approach is known to be a more difficult problem. By working within the conditioning on a set of sufficient statistics in the null hypothesis and the nonparametric combination of dependent permutation tests it is possible to find exact solutions to that kind of problems. Some solutions, guided by one medical application example, are provided.

Keywords: Conditional inference, Multivariate permutation testing, Stochastic ordering.

1. Introduction and motivating application

Testing of hypotheses for ordinal data is known to be quite a difficult problem when testing for stochastic dominance and for monotonic stochastic ordering, that is for a set of restricted alternatives. We considered data coming from Katery and Agresti (2013) concerning a survey on subarachnoid hemorrhage measured by Glasgow outcome scale, where 210 subjects received a Placebo, 190 a Low dose of a treatment, 207 a Medium dose and 195 a High dose. Response data, related to the extent of trauma, reported in the $(C = 4) \times (K = 5)$ Table 1, are classified according to 4 increasing doses of a treatment, $v = \{Placebo, Low, Medium, High\}$, each with 5

*University of Padova, pesarin@stat.unipd.it

**University of Padova, luigi.salmaso@unipd.it

***University of Padova, huang@stat.unipd.it

****University of Padova, rosa.arboretti@unipd.it

*****University of Padova, ceccato@gest.unipd.it

ordered categories $k = \{Death, Veget, Major, Minor, Recov\}$. It is expected that patients taking increasing dose of a drug present non-decreasing responses X. So, it is required to statistically establish if there is a monotonic stochastic ordering according to dose on related data. That is, to see whether $X_P \overset{d}{\leq} X_L \overset{d}{\leq} X_M \overset{d}{\leq} X_H$, with at least one strict inequality. If responses, instead of ordinal, were quantitative this problem is also termed of *isotonic regression*. By defining the analogue of a cumulative distribution function for responses X at ordered categories $c_1 \prec \ldots \prec c_K$ as $F_X(c_k) = \Pr\{X \leq c_k\}$, it is required to test for the null hypothesis $H_0 : F_{X_P} = F_{X_M} = F_{X_L} = F_{X_H}$ against the set of restricted alternatives $H_1 : F_{X_P} \geq F_{X_M} \geq F_{X_L} \geq F_{X_H}$,with at least one strict inequality. As dose is ordered, here onwards we will use the notation: $X_P = X_1$, $X_L = X_2$, $X_M = X_3$, and $X_H = X_4$. For $C = 2$, that is with a 2 × K table, this problem has one quite difficult solution within the likelihood theory, as that discussed in Colombi and Forcina (2016).

Table 1. Dose and Extent of trauma due to subarachnoid hemorrhage

	Death	Veget	Major	Minor	Recov	Total
Placebo	59	25	46	48	32	210
Low	48	21	44	47	30	190
Medium	44	14	54	64	31	207
High	43	4	49	58	41	195
Total	194	64	193	217	134	802

2. The two-sample dominance problem

In order to find suitable general solutions to the testing problems raised by the medical example, our proposal is to stay within the theory of conditional inference where conditioning is on a set of sufficient statistics in the null hypothesis for the underlying unknown distribution F. This implies to stay within the permutation theory and the NPC of dependent tests. Indeed, with clear meaning of the symbols, indicating with $p_F(X)$ the probability density associated with the population distribution F, the likelihood of any two independent samples of IID data, sized n_1 and n_2, $\mathbf{X} = (X_{11}, \ldots, X_{1n_1}; X_{21} \ldots,$

$X_{2n_2})$ is $p_F(\mathbf{X}) = \prod_{i=1}^{n_1} p_{F_1}(X_{1i}) \prod_{i=1}^{n_2} p_{F_2}(X_{2i})$. This, when $F_1 = F_2$, as stated by the null hypothesis, is invariable with respect to any data rearrangements $\mathbf{X}^*$, i.e. permutations, of observed data $\mathbf{X}$. So, $p_F(\mathbf{X}) = p_F(\mathbf{X}^*)$ is a permutation invariable likelihood. Of course, such a property is not true under the alternative where $F_1 \neq F_2$. Moreover, the observed data $\mathbf{X}$ in the null hypothesis is always a set of sufficient statistics for every underlying F. Thus, the act of conditioning on $\mathbf{X}$ makes any inference to be independent of F. As a matter of fact, *the null conditional probability, given* $\mathbf{X}$, *of every event Amember of a suitable family of events* $\mathcal{A}$, *is independent of* F; *indeed*: $\forall F$ *and* $\forall A \in \mathcal{A}$, $\Pr\{\mathbf{X}^* \in A; F|\mathbf{X}\} = \Pr\{\mathbf{X}^* \in A|\mathbf{X}\}$. This makes permutation inferences distribution-free and nonparametric. In practice, indicating by $\mathbf{X} = \{X(i), i = 1, \ldots, n; n_1, n_2\}$ the $n_1 + n_2 = n$ data, where it is intended that the first n_1 data in the list are from the first sample and the rest from the second, a random permutation $\mathbf{X}^* \in \mathbf{\Pi}(\mathbf{X})$ can be obtained as $\mathbf{X}^* = \{X(u_i^*), i = 1, \ldots, n; n_1, n_2\}$, where $\mathbf{u}^* = \{u_1^*, \ldots, u_n^*\}$ is any random permutation of unit labels $\mathbf{u} = \{1, \ldots, n\}$. Thus, the permuted table associated with $\mathbf{X}^*$ is computed as $\{f_{jk}^* = \#(X_{ji}^* \in c_k), k = 1, ..., K, j = 1, 2\}$, where: $\#(\cdot)$ is the number of elements of $\mathbf{X}^*$ that satisfy $(\cdot)$; $X_{1i}^* = X(u_i^*)$ for $i \leq n_1$ and $X_{2i}^* = X(u_i^*)$ for $n_1 < i \leq n$. Of course, each permuted table is such that $f_{\cdot k} = f_{1k} + f_{2k} = f_{1k}^* + f_{2k}^* = f_{\cdot k}^*$, $k = 1, \ldots, K$, and so marginal frequencies $f_{\cdot k}$, as well as cumulative marginal frequencies $N_{\cdot k} = N_{1k} + N_{2k} = N_{\cdot k}^*$ with $N_{jk} = \sum_{s \leq k} f_{js}$, are permutation invariable quantities.

2.1. The one-dimensional dominance problem

For $C = 2$, the testing problem related to Example 1, i.e. $H_0 : X_1 \overset{d}{=} X_2$ against $H_1 : X_1 \overset{d}{<} X_2$, can be equivalently written as $H_0 : F_1 = F_2 \equiv \bigcap_{k=1}^{K-1} [F_1(c_k) = F_2(c_k)]$ against $H_1 : F_1 > F_2 \equiv \bigcup_{k=1}^{K-1} [F_1(c_k) \geq F_2(c_k)]$, with at least one strict inequality. It is worth noting that: i) since CDFs at last category c_K are $F_1(c_K) = F_2(c_K) = 1$, such category does not contain information for discriminating between H_0 and H_1, so it can be ignored without loss of generality; ii) according to Roy's (1953) *union-intersection* methodology, the testing problem has been broken-down into $K - 1$ simpler sub-problems; iii) clearly, that problem can be properly solved by the joint analysis of $K-1$ dependent partial tests and by their NPC; iv) the nonparameric property of permutation tests is of great practical importance since it is not required to specify and to estimate the unknown dependence coefficients involved on partial tests into which the problem is broken-down. The $K-1$ partial test statistics we propose are:

$$T_k^* = (\hat{F}_{1k}^* - \hat{F}_{2k}^*)\left[\bar{F}_{\cdot k}(1-\bar{F}_{\cdot k})\right]^{-\frac{1}{2}}, \; k = 1, \ldots, K-1. \tag{1}$$

where: $\hat{F}_{jk}^* = \hat{F}_j^*(c_k) = N_{jk}^*/n_j$, $j = 1,2$, $\bar{F}_{\cdot k} = N_{\cdot k}/n$ are permutation and marginal empirical distribution functions (EDFs); N_{1k}^* and N_{2k}^*, $k = 1, ..., K-1$ are permutation cumulative frequencies obtained from the permuted table $\{f_{jk}^*, \, k = 1, ..., K, \, j = 1,2\}$. Note that EDFs $\hat{F}_{jk}$ are maximum likelihood unbiased estimates of population CDFs $F_j(c_k)$, $k = 1, ..., K-1$, $j = 1,2$. It is noticeable to observe that: i) each partial test T_k^* is a standardized comparison of two relative frequencies and so corresponding to a reformulation of Fisher's exact probability test; ii) large values of each T_k^* are significant; iii) the T_k^* are positively dependent; iv) 0 is assigned to expressions with the form $0/0$; v) each T_k^* is conditionally optimal with conditional variance $\sigma^2_{T_k^*} = 4n_2/[n_1(n-1)]$ not dependent on k; vi) each T_k^* is asymptotically normally distributed. Their NPC can be done according to the methods discussed in Pesarin and Salmaso (2010). The simplest way of combination is by their direct sum:

$$T_{AD}^* = \sum_{k=1}^{K-1} \left(\hat{F}_{1k}^* - \hat{F}_{2k}^*\right)\left[\bar{F}_{\cdot k}(1-\bar{F}_{\cdot k})\right]^{-\frac{1}{2}}. \tag{2}$$

Such T_{AD} looks like the Anderson-Darling goodness-of-fit test for directional (dominance) alternatives. T_{AD} is provided with some nice properties (Pesarin and Salmaso, 2010): i) as all T_k^* are unbiased, it is unbiased; ii) as at least one of T_k^* is consistent, it is consistent; iii) as its combined acceptance region is convex, it is admissible; iv) it is an admissible combination of conditionally optimal partial tests. The admissibility of a test T means that there does not exist any other test G, for the same hypotheses and within the same conditions, that is uniformly better than T. The p-value statistic related to the pair $(T_{AD}, \mathbf{X})$ is defined as $\lambda_{AD} = \Pr\{T^*_{AD} \geq T^o_{AD}|\mathbf{X}\}$, where the conditioning on actual data set $\mathbf{X}$ is emphasized and $T^o_{AD} = T_{AD}(\mathbf{X})$ represents the observed value of T_{AD} on data $\mathbf{X}$. According to the general testing rule, if $\lambda_{AD} \leq \alpha$ the null hypothesis is rejected at significance level $\alpha > 0$.

To justify the NPC solution, let us consider the representation, related to a general problem with K partial tests, R random permutations and combining function ψ, that follows: where: the first column of first sub-table contains the observed values of K partial tests calculated on the given data set $\mathbf{X}$, i.e. $T^o_k(\mathbf{X})$, $k = 1, \ldots, K$; the r-th column contains the values of the K partial tests at the r-th random permutation $\mathbf{X}^*_r$, $r = 1, \ldots, R$. The first element of second sub-table contains the observed value of a combined tests obtained by the combining function ψ, i.e. $T^o_\psi = \psi(T^o_1, \ldots, T^o_K)$, and the r-th element is the permutation value of combined test ψ at the r-th data permutation.

$\mathbf{X}$	$\mathbf{X}^*_1$		$\mathbf{X}^*_r$		$\mathbf{X}^*_R$
T^o_1	T^*_{11}	$\cdots$	T^*_{1r}	$\cdots$	T^*_{1R}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
T^o_K	T^*_{K1}	$\cdots$	T^*_{Kr}	$\cdots$	T^*_{KR}
			$\downarrow$		
T^o_ψ	$T^*_{\psi 1}$	$\cdots$	$T^*_{\psi r}$	$\cdots$	$T^*_{\psi R}$

If the null hypothesis would be true the sub-matrix $\{T^*_{kr}\}_{K\times R}$ simulates the K-dimensional null distribution of K partial permutation tests. Accordingly, the sub-vector $\{T^*_{\psi r}\}_R$ simulates the null permutation distribution of combined test ψ. Thus, the statistic $\hat{\lambda}_\psi = \#(T^*_\psi \geq T^o_\psi)/R$ gives an unbiased and strongly consistent estimate, as R diverges, of the p-value statistic of combined test T_ψ. If the null hypothesis would not be true, at least one of partial tests would give larger observed values than in H_0 and so, if combining function ψ is non-decreasing in each argument, the p-value statistic satisfies the relation: $\hat{\lambda}_{\psi H_1} \leq \hat{\lambda}_{\psi H_0}$ uniformly for every data set $\mathbf{X}$ and for every underlying distribution F. The latter implies the unbiasedness property and so justifies that if $\hat{\lambda}_\psi \leq \alpha$ then H_0 is rejected.

The same unidimensional problem (Pesarin and Salmaso, 2010) can even be tackled by considering the *comparison of two probability generating func-*

tions. A more practical solution is by assigning non-decreasing W scores to ordered classes, e.g. as $c_k \rightarrow w_k$, with $w_1 \leq w_2 \leq \ldots \leq w_K$, and at least one strict inequality. In such a case the data are transformed into $w_{ki} = w_k \cdot \mathbf{1}(X(i) = c_k)$, $i = 1, \ldots, n$, where $\mathbf{1}(\cdot)$ is the counting function. Thus, the permutation solution is nothing else than a comparison of sample means of scores: $T_W^* = \bar{w}_2^* - \bar{w}_1^*$. One further solution is by transforming data X_{ji} into ranks $R_{ji} = \#(X \leq X_{ji})$ or mid-ranks $M_{ji} = \#(X < X_{ji}) + \#(X = X_{ji})/2$, $i = 1, \ldots, n_j$, $j = 1, 2$, and then to proceed, in the spirit of Mann and Withney, by comparing mean of ranks and of mid-ranks: $T_R^* = \bar{R}_2^* - \bar{R}_1^*$ and $T_M^* = \bar{M}_2^* - \bar{M}_1^*$, respectively. Clearly, although unbiased, consistent and easy to interpret, these last three solutions suffers from the arbitrary substitution of categorical data with numerical quantities.

2.2. The multidimensional dominance problem

Let us again be guided by the two-sample V-dimensional problem. In that problem, to test for the multidimensional hypotheses we used the formulation: $H_0 : \mathbf{X}_1 \overset{d}{=} \mathbf{X}_2$ against the set of restricted alternatives $H_1 : \mathbf{X}_1 \overset{d}{<} \mathbf{X}_2$, where the latter is equivalent to $\bigcup_{v=1}^{V} \bigcup_{k=1}^{K-1} [F_{1v}(c_k) \geq F_{2v}(c_k)]$, with at least one strict inequality in at least one of 24 points. So a simple extension of the Anderson-Darling goodness-of-fit type solution, shown for the unidimensional case, with clear meaning of the symbols, leads to the test statistic:

$$T_{AD}^* = \sum_{v=1}^{V} \sum_{k=1}^{K-1} \left(\hat{F}_{1vk}^* - \hat{F}_{2vk}^* \right) \left[\bar{F}_{\cdot vk} (1 - \bar{F}_{\cdot vk}) \right]^{-\frac{1}{2}}, \tag{3}$$

where $V \geq 2$ is the number of variables under study and $K \geq 2$ is the number of ordered categories for responses. According to the unidimensional formulation, if the alternative is true then at least one summand assumes values not smaller than in H_0. So, that test is unbiased, consistent and admissible. In place of the direct combination of V partial tests T_{ADv}^*, i.e. one Anderson-Darling test for each variable, it is possible to think of a more general combination like $T_\psi^* = \psi(T_{AD1}^*, \ldots, T_{ADV}^*)$. The most commonly used of combining functions ψ is Fisher's $T_F = -2 \sum_v \log(\lambda_{ADv})$, where

λ_{ADv} is the p-value statistic of T^*_{ADv}. Similarly to the unidimensional setting, the multidimensional problem (Pesarin and Salmaso, 2010) can, however, be tackled by assigning non-decreasing W_t scores to ordered classes, e.g. as $c_{tk} \to w_{tk}$, $k = 1, \ldots, K$, $t = 1, \ldots, V$, where $w_{t1} \leq \ldots \leq w_{tK}$, with at least one strict inequality $\forall t$. In such a case the data are transformed into $w_{tki} = w_{tk} \cdot \mathbf{1}(X_{tji} = c_{tk})$. Thus, the permutation solution is nothing else than a comparison of sample means of scores: $T^*_{\psi W} = \psi[(\bar{w}^*_{12} - \bar{w}^*_{11}), \ldots, (\bar{w}^*_{V2} - \bar{w}^*_{V1})]$. And so on also for $T^*_{\psi R}$ and $T^*_{\psi M}$ with ranks and mid-ranks, respectively.

2.3. The C-sample stochastic ordering problem

With reference to the first example and in accordance with the Jonckheere-Terpstra idea, we may equivalently break down the 4-sample testing problem into three two-sample ones. To be specific, let us imagine that for any $j \in \{1, \ldots, C-1\}$, the whole data set is divided into two pooled pseudo-groups, where the first is obtained by pooling together data of the first j ordered groups and the second by pooling the rest. To this end, we define the first pooled pseudo-group as $\mathbf{Y}_{1(j)} = \mathbf{X}_1 \biguplus \ldots \biguplus \mathbf{X}_j$ and the second as $\mathbf{Y}_{2(j)} = \mathbf{X}_{j+1} \biguplus \ldots \biguplus \mathbf{X}_C$, $j = 1, \ldots, C-1$, where $\biguplus$ is the symbol for pooling data into one pseudo-group and $\mathbf{X}_j = \{X_{ji}, i = 1, \ldots, n_j\}$ is the data set in the jth group.

In the null hypothesis, data from every pair of pseudo-groups are exchangeable because related pooled variables satisfy the relationships $Y_{1(j)} \overset{d}{=} Y_{2(j)}$, $j = 1, \ldots, C-1$. In the alternative we see that $Y_{1(j)} \overset{d}{\leq} Y_{2(j)}$, for each j, which corresponds to the stochastic dominance between each pair of pseudo-groups. This suggests that we express the monotonic stochastic ordering hypothesis into the equivalent form $H_0 : \{\bigcap_{j=1}^{C-1}(Y_{1(j)} \overset{d}{=} Y_{2(j)})\}$ and $H_1 : \{\bigcup_j (Y_{1(j)} \overset{d}{\leq} Y_{2(j)})\}$, emphasizing a break-down into a set of sub-hypotheses. So this problem is solved by combining the $C-1$ partial tests:

$$T^*_{(j)} = \sum_{k=1}^{K-1} \left(\hat{F}^*_{1(j)k} - \hat{F}^*_{2(j)k} \right) \left[\bar{F}_{\cdot(j)k}(1 - \bar{F}_{\cdot(j)k}) \right]^{-\frac{1}{2}}, \ j = 1, \ldots, C-1. \quad (4)$$

According to our experience, the most suitable combining functions for this problem are Fisher's and Liptak's. Since in the stochastic ordering alternative all $C-1$ partial tests contain a positive non-centrality quantity, i.e. are all in their respective sub-alternative, Tippett's combination is less sensitive than others.

Of course, if $V>1$ variables were involved, the multivariate stochastic ordering solution would require one stochastic ordering partial test for each variable $v=1,\ldots,V$. So the global test would be the NPC combination of:

$$T^*_{(j)V}=\sum_{v=1}^{V}\sum_{k=1}^{K-1}\left(\hat{F}^*_{1v(j)k}-\hat{F}^*_{2v(j)k}\right)\left[\bar{F}_{\cdot v(j)k}(1-\bar{F}_{\cdot v(j)k})\right]^{-\frac{1}{2}},\ j=1,\ldots,C{-}1. \tag{5}$$

Table below shows the results of our analyses with the motivating example described in the first paragraph, based on $R=100\,000$ random permutations, for tests: Anderson-Darling T^*_{AD}, on scores T^*_W, and on mid-ranks T_M, and their combinations: T''_D direct, T''_F Fisher's, T''_L Liptak's and T''_T Tippett's. Note that W scores are stated as $(w_1=1, w_2=2, w_3=3, w_4=4, w_5=5)$. The NPC of dependent tests method is suitable and effective for many multivariate testing problems which are very difficult or even impossible to solve within likelihood parametric frameworks.

	$T^*_{(1)}$	$T^*_{(2)}$	$T^*_{(3)}$	T''_D	T''_F	T''_L	T''_T
$\hat{\lambda}_{AD(j)}$	0.0141	0.0025	0.0074	0.0012	0.0015	0.0012	0.0068
$\hat{\lambda}_{W(j)}$	0.0131	0.0021	0.0076	0.0010	0.0012	0.0010	0.0053
$\hat{\lambda}_{M(j)}$	0.0144	0.0024	0.0062	0.0011	0.0014	0.0011	0.0068

References

Colombi R., Forcina A. (2016) Testing under order restrictions in contingency tables, *Metrika*, 79, 73-90.

Kateri K., Agresti A. (2013) Bayesian inference about odds ratio structure in ordinal contingency tables, *Environmetrics*, 24, 281-288.

Pesarin F., Salmaso L. (2010) *Permutation Tests for Complex Data: Theory, Applications and Software*, Wiley & Sons, Chichester, UK.

Consensus measures for preference rankings with ties: an approach based on position weighted Kemeny distance

Antonella Plaia[*], Mariangela Sciandra [**], Simona Buscemi [***]

Abstract: Preference data are a particular type of ranking data where some subjects (voters, judges, ...) give their preferences over a set of alternatives (items). It happens, in most of the real cases, that some items receive the same preference by a judge, giving raise to a ranking with ties. The purpose of our paper is to investigate on the consensus between rankings with ties taking into account the importance of swapping elements belonging to the top (or to the bottom) of the ordering (position weights). Combining the structure of the τ_x proposed by Emond and Mason and the class of weighted Kemeny-Snell distances, we propose a position weighted rank correlation coefficient to compare rankings with ties. The one-to-one correspondence between the weighted distance and the rank correlation coefficient holds, analytically speaking, using both equal and decreasing weights.

Keywords: Weighted rank correlation, Weighted Kemeny distance, Position weights.

1. Introduction

Ranking is one of the most simplified cognitive processes useful for people to handle many aspects in their life. When some subjects are asked to indicate their preferences over a set of alternatives (items), ranking data are called preference data. Therefore, ranking data arise when a group of n individuals (judges, experts, voters, raters etc) shows their preferences on a finite set of items (m different alternatives of objects, like movies, activities and so on). If the m items, labeled $1, \dots m$, are ranked in m distinguishable ranks, a complete ranking or linear ordering is achieved (Cook, 2006): this ranking a is a mapping function from the set of items $\{1, \dots, m\}$ to the set of ranks

[*] University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, antonella.plaia@unipa.it

[**] University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, mariangela.sciandra@unipa.it

[***] University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, simona.buscemi@unipa.it

$\{1, \ldots, m\}$, endowed with the natural ordering of integers, where $a(i)$ is the rank given by the judge to item i. The ranking a is, in this case, one of the $m!$ possible permutations of m elements, containing the preferences given by the judge to the m items. When some items receive the same preference, then a tied ranking or a weak ordering is obtained. In real situations, it can happen that not all items are ranked: partial rankings, when judges are asked to rank only a subset of the whole set of items, and incomplete rankings, when judges can freely choose to rank only some items. In order to get homogeneous groups of subjects having similar preferences, it's natural to measure the spread between rankings through dissimilarity or distance measures among them. Distances between rankings have received a growing consideration in the past few years. Usual examples of metrics in this framework are Kendall's and Spearman's. In 1962 Kemeny introduced a metric defined on linear orders, known as Kemeny distance (or metric), later generalized to the framework of weak orders by Cook et al in 1986, which satisfies the constraints of a distance measure suitable for rankings. The Kemeny distance (d_K) between two rankings a and b is a city-block distance defined as:

$$d_K(a, b) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} |a_{ij} - b_{ij}| \tag{1}$$

where a_{ij} and b_{ij} are the generic elements of the $m \times m$ score matrices associated to a and b respectively, assuming the following values:

$$a_{ij}, b_{ij} = \begin{cases} 1 & \text{if i is preferred to j} \\ 0 & \text{if i = j or if i is tied with j} \\ -1 & \text{if j is preferred to i} \end{cases} \tag{2}$$

Considering the usual relation between a distance d and its corresponding correlation coefficient $\tau = 1 - 2d/D_{max}$, where D_{max} is the maximum distance, d_K is in a one-to-one correspondence to the rank correlation coefficient τ_x

proposed by (Emond and Mason, 2002), defined as:

$$\tau_x(a,b) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}^{'} b_{ij}^{'}}{m(m-1)} \tag{3}$$

where $a_{ij}^{'}$ and $b_{ij}^{'}$ are the generic elements of the $m \times m$ score matrices associated to a and b respectively, assuming the following values

$$a_{ij}^{'}, b_{ij}^{'} = \begin{cases} 1 & \text{if i is preferred to or tied with j} \\ 0 & \text{if i = j} \\ -1 & \text{if j is preferred to i} \end{cases} \tag{4}$$

Distances and correlations are the two possible approach to a consensus ranking problem: given n rankings, full or weak, of m items, what best represents the consensus opinion? This consensus is the ranking that shows the maximum correlation, or equivalently the minimum distance, with the whole set of n rankings.

2. *Weighted distances*

Kumar and Vassilvitskii (2010) introduced two aspects essential for many applications involving distances between rankings: positional weights and element weights. In short, i) the importance given to swapping elements near the head of a ranking could be higher than the same attributed to elements belonging to the tail of the list or ii) swapping elements similar between themselves should be less penalized than swapping elements which aren't similar. In this paper, we deal with case i) and consider the weighted version of the Kemeny metric, since the Kemeny metric is not sensitive towards where the disagreement between two rankings occurs. For measuring the weighted distances, the non-increasing weights vector $w = (w_1, w_2, ..., w_{m-1})$ constrained to $\sum_{i=1}^{m-1} w_i = 1$ is used, where w_i is the weight given to position i in the ranking. Given two generic rankings of m elements, a and b, the weighted Kemeny distance is defined by García-Lapresta and Pérez-Román (2010) as follows:

$$d_K^w(a,b) = \frac{1}{2}\left[\sum_{\substack{i,j=1\\i<j}}^{m} w_i|a_{ij}^{(\sigma_1)} - b_{ij}^{(\sigma_1)}| + \sum_{\substack{i,j=1\\i<j}}^{m} w_i|b_{ij}^{(\sigma_2)} - a_{ij}^{(\sigma_2)}|\right], \qquad (5)$$

where (σ_1) states to follow the a ranking and (σ_2), similarly, orders according to b. More specifically, $b_{ij}^{(\sigma_1)}$ is the score matrix of the ranking b reordered according to a, $a_{ij}^{(\sigma_2)}$ is the score matrix of the ranking a reordered according to b and $a_{ij}^{(\sigma_1)} = b_{ij}^{(\sigma_2)}$ is the score matrix of the linear order $1, 2, ..., m$ (see Plaia and Sciandra, 2017 for more details).

3. A new weighted rank correlation coefficient

Recently we proposed a new rank correlation coefficient (Plaia et al, 2018), suitable for position weighted rankings which handles linear orders. In this paper, we propose its generalization to cope with the presence of ties. Combining the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010) and the extension of τ_x provided by Emond and Mason (2002), we propose a new rank correlation coefficient working with a couple of score matrices. Let's define the generic (i, j) element of the score matrices $a_{ij}^{'}$ and a_{ij}^{*} related to a ranking a as follows:

$$a_{ij}^{'}, b_{ij}^{'} = \begin{cases} 1 & \text{if i is preferred to or tied with j} \\ 0 & \text{if i = j} \\ -1 & \text{if j is preferred to i} \end{cases} \qquad a_{ij}^{*}, b_{ij}^{*} = \begin{cases} 1 & \text{if i is preferred to j} \\ 0 & \text{if i = j} \\ -1 & \text{if j is preferred to or tied with i} \end{cases} \qquad (6)$$

Our new rank correlation coefficient uses both these score matrices (the corresponding element of the score matrices are equal to 1 and to -1 according to the considerations in Emond and Mason (2000), secc. 38, 39) and is defined as:

$$\tau_x^w(a,b) = \frac{\sum_{i<j}^{m}(a_{ij}^{'\sigma_1} b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2} b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1} b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2} b_{ij}^{*\sigma_2})w_i}{2Max[d_K^w]}, \qquad (7)$$

where the denominator represents twice the maximum value of the Kemeny weighted distances (García-Lapresta and Pérez-Román, 2010), equal to:

$$Max[d_K^w(a,b)] = 2\sum_{i=1}^{m-1}(m-i)w_i. \tag{8}$$

4. Correspondence between distance and correlation

We will demonstrate that eq. (7) is the correlation coefficient corresponding to the distance (5) through the straightforward linear transformation:

$$\frac{\sum_{i<j}^{m}(a_{ij}^{'\sigma_1}b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2}b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1}b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2}b_{ij}^{*\sigma_2})w_i}{2Max[d_K^w]} = 1 - \frac{2d_K^w}{Max[d_K^w]}$$

or equivalently

$$\sum_{i<j}^{m}(a_{ij}^{'\sigma_1}b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2}b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1}b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2}b_{ij}^{*\sigma_2})w_i = 2Max[d_K^w] - 4d_K^w \tag{9}$$

where $Max[d_K^w]$ and d_K^w are defined in Eq. (8) and in Eq. (5) respectively, and we use the matrix representation of a ranking a of m objects as in Eq. (2) for computing d_K^w and the two different score matrices of Eq. (6) for calculating τ_x^w. According to Emond and Mason (2002), if two rankings a and b agree except for a set S of k objects, which is a segment of both, then $d_K^w(a,b)$ may be computed as if these k objects were the only objects being ranked. As a consequence, to prove the equality in (9) we will show that for each pair of objects i and j:

$$a_{ij}^{'\sigma_1}b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2}b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1}b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2}b_{ij}^{*\sigma_2} = 4(m-i) - 2\big[|a_{ij}^{\sigma_1} - b_{ij}^{\sigma_1}| + |b_{ij}^{\sigma_2} - a_{ij}^{\sigma_2}|\big] \tag{10}$$

In Eq. (10) the weights w_i have been omitted from both the sides. There are nine possible combinations of orderings for item i and j between voters A and B, but only four distinct cases must be considered. The other five are equivalent to one of these four through a simple relabeling of the rankers and/or the objects. (Emond and Mason, 2002).

Case 1. Both A and B prefer object i to j. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_1} = a_{ij}^{\sigma_2} = b_{ij}^{\sigma_2} = 1$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_1} = a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_1} = a_{ij}^{*\sigma_2} = b_{ij}^{*\sigma_2} = 1$. Hence, the equality in

equation (10) holds:
$1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 = 4 - 2[|1-(1)| + |1-(1)|]$.
Case 2. A prefers object i to j and B prefers the two objects as tied. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = a_{ij}^{\sigma_2} = 1$ and $b_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = 0$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_1} = a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_1} = a_{ij}^{*\sigma_2} = 1$ and $b_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_2} = -1$. Hence, the equality in equation (10) holds:
$1 \cdot 1 + 1 \cdot 1 + 1 \cdot (-1) + 1 \cdot (-1) = 4 - 2[|1-0| + |1-0|]$.
Case 3. A prefers object i to j and B prefers j to object i. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = 1$ and $a_{ij}^{\sigma_2} = b_{ij}^{\sigma_1} = -1$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_2} = 1$ and $a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_2} = b_{ij}^{*\sigma_1} = -1$. Hence, the equality in equation (10) holds:
$1 \cdot (-1) + (-1) \cdot 1 + 1 \cdot (-1) + (-1) \cdot (1) = 4 - 2[|1-(-1)| + |1-(-1)|]$.
Case 4. Both A and B rank the objects i and j as tied. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = a_{ij}^{\sigma_2} = b_{ij}^{\sigma_1} = 0$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_1} = a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = 1$ and $a_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_1} = a_{ij}^{*\sigma_2} = b_{ij}^{*\sigma_2} = -1$. Hence, the equality in equation (10) holds:
$1 \cdot (1) + (1) \cdot 1 + 1 \cdot (1) + (1) \cdot (1) = 4 - 2[|0-0| + |0-0|]$.

5. Minimum and Maximum values of τ_x^w

From the demonstrations in sec. 4 τ_x^w can be maximum, and equal to 1, if and only if for all i and j only *Case 1* or only *Case 4* are observed. Therefore, differently from what happens with Kendall τ_b (see Emond and Mason, 2002, sect 3.3), τ_x^w is maximum even when a generic all tied ranking is compared with itself. Analogously, τ_x^w can be minimum, and equal to -1, if and only if for all i and j only *Case 3* occurs.

6. Correspondence between weighted and unweighted measures

For equal weights assigned to the items ($w_i = \frac{1}{m-1}$, for each $i = 1, 2, ..., m-1$) the weighted distance is proportional to the classical Kemeny distance, according to the number of items:

$$d_x^w = \frac{d_x}{m-1}$$

Proof. Referring to the cases listed in Section 4:

Case 1. $d_x^w = \frac{1}{2}[|1-(1)|+|1-(1)|]w_i = 0$ and $d_x = \frac{1}{2}[|0-0|+|0-0|] = 0$

Case 2. $d_x^w = \frac{1}{2}[|1-0|+|1-0|]w_i = \frac{1}{m-1}$ nd $d_x = \frac{1}{2}[|1-0|+|1-0|] = 1$

Case 3. $d_x^w = \frac{1}{2}[|1-(-1)|+|1-(-1)|]w_i = \frac{2}{m-1}$ nd $d_x = \frac{1}{2}[|1-(-1)|+|1-(-1)|] = 2$

Case 4. $d_x^w = \frac{1}{2}[|0-0|+|0-0|]w_i = 0$ and $d_x = \frac{1}{2}[|0-0|+|0-0|] = 0$

Corollary Since $\tau_x \leftrightarrow d_K$ and $\tau_x^w \leftrightarrow d_K^w$, then the weighted rank correlation coefficient is equivalent to the rank correlation coefficient defined by Emond and Mason, when equal importance is given to the positions occupied by the items:

$$\tau_x^w = \tau_x, \quad \text{with } w_i = \frac{1}{m-1} \quad \forall i = 1, 2, ..m-1$$

7. Consensus ranking

The proposed weighted correlation coefficient can be used to deal with a consensus ranking problem: given n rankings, full or weak, of m items, what best represents the consensus opinion? This consensus is the ranking that shows the maximum correlation, with the whole set of n rankings. Given a nxm matrix $\mathbf{X}$, whose l-th row represents the ranking associated to the l-th judge, the consensus ranking, i.e. the ranking c that best represents the matrix $\mathbf{X}$, is that ranking that maximizes the following expression:

$$Max \sum_{l=1}^{n} \frac{\sum_{i<j}^{m}(x_{ij}^{'\sigma_l} c_{ij}^{'\sigma_l} + x_{ij}^{'\sigma_c} c_{ij}^{'\sigma_c} + x_{ij}^{*\sigma_l} c_{ij}^{*\sigma_l} + x_{ij}^{*\sigma_c} c_{ij}^{*\sigma_c})w_i}{2Max[d_K^w]}$$

8. Conclusions

In this paper, we provided a rank correlation coeffient τ_x^w for weak orderings, as an extension of τ_x^w for linear orderings (Plaia et al, 2018). We demonstrated the correspondence between τ_x^w and the weighted Kemeny distance and, finally, we showed that the weighted rank correlation coefficient τ_x^w is equal to the Emond and Mason rank correlation coefficient τ_x in the case of tied rankings and $w_i = \frac{1}{m-1}$ for all i. Our future purpose is the extension and the implementation in R of the branch and bound algorithm proposed in (Plaia et al 2018) for linear orders to the case of weak orderings.

References

Cook W.D. (2006) Distance based and ad hoc consensus models in ordinal preference ranking, *European Journal Operation Research*, 172, 369-385.

Cook W.D., Kress M., Seiford L.M. (1986) An axiomatic approach to distance on partial orderings, *Operations research*, 20, 115-122.

Emond E.J., Mason D.W. (2000) *A new technique for high level decision support.* Department of National Defence Canada, Operational Research Division, Directorate of Operational Research (Corporate, Air & Maritime).

Emond E.J., Mason D.W. (2002) A new rank correlation coefficient with application to the concensus ranking problem, *Journal of Multi-criteria decision analysis*, 11, 17-28.

García-Laprest J.L., Pérez-Román D. (2010) Consensus measures generated by weighted Kemeny distances on weak orders. In: *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications*, Cairo.

Kemeny J. G. (1962) *Mathematical models in the social sciences*, Ginn and Company.

Kumar R., Vassilvitskii S. (2010) Generalized Distances Between Rankings. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, 571-580, New York, NY, USA. ACM.

Plaia A., Sciandra M. (2017) Weighted distance-based trees for ranking data, *Advances in Data Analysis and Classification*, 1-18. Springer, https://doi.org/10.1007/s11634-017-0306-x.

Plaia A., Buscemi S., Sciandra M. (2018) A proper correlation coefficient for weighted position rankings in a consensus ranking process, *DEASS Working papers*, submitted.

Simultaneous clustering and dimensional reduction of mixed-type data

Monia Ranalli[*], Roberto Rocci[**]

Abstract: In real applications, it is very common to have the true clustering structure masked by the presence of noise variables and/or dimensions. A mixture model is proposed for simultaneous clustering and dimensionality reduction of mixed-type data: the continuous and the ordinal variables are assumed to follow a Gaussian mixture model, where, as regards the ordinal variables, it is only partially observed. To recognize discriminative and noise dimensions, the variables are considered to be linear combinations of two independent sets of latent factors where only one contains the information about the cluster structure while the other one contains noise dimensions. In order to overcome computational issues, the parameter estimation is carried out through an EM-like algorithm maximizing a composite log-likelihood based on low-dimensional margins.

Keywords: Mixture models, Composite likelihood, EM algorithm.

1. Introduction

The aim of cluster analysis is to partition the data into meaningful groups which should differ considerably from each other. The cluster analysis is made more difficult by the presence of mixed-type data (ordinal and continuous variables) combined by the presence of dimensions (named noise) that are uninformative for recovering the groups and could obscure the true cluster structure. It follows that there are two main points to be addressed: combining continuous with ordinal variables; taking into account the presence of noise variables/dimensions. As regards the first point, the literature on clustering for continuous data is rich and wide; the most commonly clustering model-based used is the finite mixture of Gaussians (McLachlan et al., 2016). Differently, that one developed for categorical data is still limited. Models used for ordinal data mainly adopt two approaches developed in the factor

[*] University of Tor Vergata, monia.ranalli@uniroma2.it

[**] University of Tor Vergata, roberto.rocci@uniroma2.it

analysis framework: Item Response Theory (IRT) (see e.g. Bartholomew et al. (2011), Bock and Moustaki (2007)), and the Underlying Response Variable (URV) (see e.g. Jöreskog, 1990; Lee et al., 1990; Muthén, 1984). In the URV approach, the ordinal variables are seen as a discretization of continuous latent variables jointly distributed as a finite mixture; examples are: Everitt (1988), Lubke and Neale (2008), Ranalli and Rocci (2016a, 2017a, 2017b). However, this makes the maximum likelihood estimation rather complex because it requires the computation of many high dimensional integrals. The problem is usually solved by approximating the likelihood function. In this regard we mention some useful surrogate functions, such as the variational likelihood (Gollini and Murphy, 2014) or the composite likelihood (Ranalli and Rocci (2016a, 2017a, 2017b)). Although it is possible to cluster via a model based approach continuous or ordinal variables separately, combining both into a common framework may raise some issues. Following the URV approach, Everitt (1988) and Ranalli-Rocci (2017a) proposed a model according to which both the continuous and the categorical ordinal variables follow a Gaussian mixture model, where the ordinal variables are only partially observed through their ordinal counterparts. This satisfies the two main requirements: dealing with ordinal data properly and modelling dependencies between ordinal and continuous variables. As regards the presence of noise variables, different approaches exist in literature. Several techniques for simultaneous clustering and dimensionality reduction (SCR) have been proposed in a non-model based framework for quantitative (e.g.: Rocci et al., 2011; Vichi and Kiers, 2001) or categorical data (e.g.: Hwang et al., 2006; Van Buuren et al., 1989). There are also approaches based on a family of mixture models which fits the data into a common discriminative subspace (see e.g. Bouveyron and Brunet, 2012; Kumar and Andreou, 1998; Ranalli and Rocci, 2017b). The key idea is to assume a common latent subspace to all groups that is the most discriminative one. This allows to project the data into a lower dimensional space preserving the clustering characteristics in order to improve visualization and interpretation of the underlying structure of the data. The model can be formulated as a finite mixture of Gaussians with a particular set of constraints on the parameters. Combining all pieces together, following the URV approach, in our proposal the continuous and the ordinal variables are assumed to follow a heteroscedastic Gaussian mixture model, where, as regards the ordinal variables, it is only partially observed. To recognize discriminative and noise dimensions, these variables are considered to be linear combinations of two independent sets of latent factors where only one contains the information about the cluster structure, defining a discriminative subspace, distributed as a finite mixture of Gaussians. The other one contains noise dimensions distributed as a multivariate normal. The model specification is parsimonious and is able to identify a reduced set of discriminative latent factors/dimensions even when there are no noise variables to be detected. The main drawback of this model is that, in practice, it cannot be estimated through a full maximum likelihood approach, due to the presence of multidimensional integrals making the estimation time consuming. To overcome this issue, we propose to replace this cumbersome likelihood with a surrogate objective function, easier to maximize, that is the product of marginal likelihoods. It is a composite likelihood method (Lindsay, 1988; Varin et al., 2011) where surrogate functions are defined as the product of marginal or conditional events. In particular, our proposals is based on the existing results within a mixture model framework Ranalli-Rocci (2016a, 2017a, 2017b). It consists of replacing the joint likelihood with all possible marginals, like bivariate marginal distributions of ordinal variables and the marginal distributions of one ordinal variable and all continuous variables.

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, respectively. 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 (URV) developed within the SEM framework (see e.g. Jöreskog, 1990; Lee et al., 1990; Muthén, 1984), the ordinal variables $\mathbf{x}$ are considered as a categorization of a continuous multivariate latent variable $\mathbf{y}^O = [y_1, \ldots, y_O]'$. 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, $x_i = c_i \Leftrightarrow \gamma^{(i)}_{c_i-1} \leq y_i < \gamma^{(i)}_{c_i}$, 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 $\mathbf{\Gamma}$ whose elements are given by the vectors $\boldsymbol{\gamma}^{(i)}$. To accommodate both cluster structure and dependence within the groups, we assume that $\mathbf{y} = [\mathbf{y}^{O\prime}, \mathbf{y}^{\bar{O}\prime}]'$ follows a heteroscedastic Gaussian mixture, $f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi_p(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where the p_g's are the mixing weights and $\phi_p(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a P-variate normal distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Let us set $\boldsymbol{\psi} = \{p_1, \dots, p_G, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G, \boldsymbol{\Gamma}\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space. For a random i.i.d. sample of size N, $(\mathbf{x}_1, \mathbf{y}_1^{\bar{Q}}), \dots, (\mathbf{x}_N, \mathbf{y}_N^{\bar{Q}})$, the log-likelihood is

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

where, with obvious notation

$$\begin{aligned}
\pi_n\left(\boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \boldsymbol{\Gamma}\right) &= \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_O-1}^{(O)}}^{\gamma_{c_O}^{(O)}} \phi_O(\mathbf{u}; \boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}) d\mathbf{u} \\
\boldsymbol{\mu}_{n;g}^{O|\bar{O}} &= \boldsymbol{\mu}_g^{O} + \boldsymbol{\Sigma}_g^{O\bar{O}} (\boldsymbol{\Sigma}_g^{\bar{O}\bar{O}})^{-1} (\mathbf{y}_n^{\bar{O}} - \boldsymbol{\mu}_g^{\bar{O}}) \\
\boldsymbol{\Sigma}_g^{O|\bar{O}} &= \boldsymbol{\Sigma}_g^{OO} - \boldsymbol{\Sigma}_g^{O\bar{O}} (\boldsymbol{\Sigma}_g^{\bar{O}\bar{O}})^{-1} \boldsymbol{\Sigma}_g^{\bar{O}O},
\end{aligned}$$

$\pi_n\left(\boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \boldsymbol{\gamma}\right)$ is the conditional joint probability of response pattern $\mathbf{x}_n = (c_1^{(1)}, \dots, c_O^{(O)})$ given the cluster g and the continuous variables $\mathbf{y}_n^{\bar{O}}$. Finally p_g is the probability of belonging to group g subject to $p_g > 0$ and $\sum_{g=1}^{G} p_g = 1$. In order to identify the discriminative dimensions, it is assumed that there is a set of P latent factors $\tilde{\mathbf{y}}$, 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 $E(\tilde{\mathbf{y}}^{Q} \mid g) = \boldsymbol{\eta}_g$ and $\mathrm{Cov}(\tilde{\mathbf{y}}^{Q} \mid g) = \boldsymbol{\Omega}_g$, 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: $E(\tilde{\mathbf{y}}^{\bar{Q}} \mid g) = \boldsymbol{\eta}_0$ and $\mathrm{Cov}(\tilde{\mathbf{y}}^{\bar{Q}} \mid g) = \boldsymbol{\Omega}_0$. 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.

2.1. Construction of surrogate functions

The presence of multidimensional integrals makes the maximum likelihood estimation computationally demanding and infeasible as the number of observed ordinal variables increases. To overcome this, a composite likelihood approach is adopted (Lindsay, 1988). It allows us to simplify the problem by replacing the full likelihood with a surrogate function. As suggested in Ranalli-Rocci (2016a, 2017a, 2017b) within a similar context, the full log-likelihood could be replaced by the sum of two estimating-block functions: $O(O-1)/2$ bivariate marginals of ordinal variables; O marginal distributions each of them composed of one ordinal variable and the $\bar{O}$ continuous variables. This leads to the following surrogate function

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

where now, after the reparameterization induced by the reduction model, the set of parameters is $\boldsymbol{\psi} = \{p_1, \ldots, p_G, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_G, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_G, \mathbf{A}, \boldsymbol{\gamma}\}$, $\delta_{nc_ic_j}^{(ij)}$ 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; similarly $\delta_{nc_j}^{(j)}$ is a dummy variable assuming 1 if the n-th observation presents category c_j for variable x_j, 0 otherwise; $\pi_{c_ic_j}^{(ij)}(\boldsymbol{\mu}_g^{(ij)}, \boldsymbol{\Sigma}_g^{(ij)}, \boldsymbol{\Gamma}^{(ij)})$ is the probability under the model obtained by integrating the density of a bivariate normal distribution with parameters $(\boldsymbol{\mu}_g^{(ij)}, \boldsymbol{\Sigma}_g^{(ij)}, \boldsymbol{\Gamma}^{(ij)})$ between the corresponding threshold parameters. On the other hand, $\pi_{c_j}^{(j|\bar{Q})}(\mu_{n;g}^{(j|\bar{Q})}, \sigma_g^{(j|\bar{Q})}, \boldsymbol{\Gamma}^{(j)})$ is the conditional probability of variable x_j of being in category c_j given all the continuous variables $\mathbf{y}^{\bar{Q}}$. Finally, $\boldsymbol{\mu}_g = E(\mathbf{y} \mid g) = \mathbf{A}E(\tilde{\mathbf{y}} \mid g)$, while $\boldsymbol{\Sigma}_g = \text{Cov}(\mathbf{y} \mid g) = \mathbf{A}\text{Cov}(\tilde{\mathbf{y}} \mid g)\mathbf{A}'$, as specified previously. The parameter estimates are carried out through an EM-like algorithm, that works in the same manner as the standard EM.

2.2. Classification model selection and identifiability

When we adopt a composite likelihood approach, since we do not compute the joint distribution for each observation, it is not possible anymore to assign the observation to the component with the maximum a posteriori probability (MAP criterion) without further computations. To solve the problem we follow the CMAP criterion (Ranalli-Rocci, 2017a, 2017b), according to which the observation is assigned to the component with the maximum scaled composite fit (scaled by the corresponding mixing weight). As regards model selection, the best model is chosen by minimizing the composite version of penalized likelihood selection criteria like BIC or CLC (see Ranalli-Rocci, 2016b and references therein). Finally, as regards identifiability, within a full maximum likelihood approach, it is well known that a sufficient condition for local identifiability is given by the non singularity of the information matrix; while a necessary condition is that the number of parameters must be less than or equal to the number of canonical parameters. Adopting a composite likelihood approach, the sufficient condition should be reformulated by investigating the Godambe information matrix, that is, the analogous of the information matrix in composite likelihood estimation. However, as far as we know, such modification has not been formally investigate yet. About the necessary condition, we note that the number of essential parameters in the block of ordinal variables equals the number of parameters of a log linear model with only two factor interaction terms. Thus it means that we can estimate a lower number of parameters compared to a full maximum likelihood approach. Furthermore, under the underlying response variable approach, the means and the variances of the latent variables are set to 0 and 1, respectively, because they are not identified. This identification constraint individualizes uniquely the mixture components (ignoring the label switching problem), as well described in Millsap and Yun-Tein (2004). This is sufficient to estimate both thresholds and component parameters if all the observed variables have three categories at least and when groups are known. Given the particular structure of the mean vectors and covariance matrices, it is preferable to adopt an alternative, but equivalent, parametrization. This is analogous to that one used by Jöreskog and Sörbom (1996); it consists in setting the first two thresholds to 0 and 1, respectively. This means that there is a one-to-one correspondence between the two sets of parameters. If there is a binary variable, then the variance of the corresponding latent variable is set equal to 1 (while its mean should be still kept free). Finally, we note that the model has the same rotational freedom that characterizes the classical factor analysis model. In other words, writing $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$ according to $\tilde{\mathbf{y}} = [\tilde{\mathbf{y}}^{Q\prime}, \tilde{\mathbf{y}}^{\bar{Q}\prime}]'$, only the subspaces generated by the columns of $\mathbf{A}_1$ and $\mathbf{A}_2$ are identified. In order to estimate such subspaces, we impose some constraints on the model parameters, in complete analogy with what is usually done in the factor analysis model. In this way, we select a particular solution, one which is convenient to find, and leave the experimenter to apply whatever rotation he thinks desirable, as suggested by Lawley and Maxwell (1962). In particular, we require a spherical distribution for the noise factors, i.e. $\boldsymbol{\Omega}_0 = \mathbf{I}$, and informative factors in the first cluster, i.e. $\boldsymbol{\Omega}_1 = \mathbf{I}$. Such constraints still allow a rotational freedom by orthonormal matrices. This can be eliminated by requiring a “lower" triangular form for the two loading matrices. In general, $\mathbf{A}_1$ and $\mathbf{A}_2$ have a lower triangular matrix in the first Q and $(P - Q)$ rows, respectively. Of course, after the estimation the parameter matrices can be rotated to enhance the interpretation.

Further details will be given in the extended version of the paper along with simulation and real data results to show the effectiveness of the proposal.

References

Bartholomew D.J., Knott M., Moustaki I. (2011) *Latent Variable Models and Factor Analysis: A Unified Approach.* Wiley Series in Probability and Statistics. Wiley, third edition, 2011.

Bock D., Moustaki I. (2007) *Handbook of Statistics on Psychonometrics*, chapter: Item response theory in a general framework. Elsevier.

Bouveyron C., Brunet C. (2012) Simultaneous model-based clustering and visualization in the fisher discriminative subspace, *Statistics and Computing*, 22, 301-324.

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

Gollini I., Murphy T.B. (2014) Mixture of latent trait analyzers for model-based clustering of categorical data, *Statistics and Computing*, 24, 569-588.

Hwang H., Montréal H., Dillon W.R., Takane Y. (2006) An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents, *Psychometrika*, 71, 161-171.

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

Jöreskog K.G., Sörbom D. (1996) *LISREL 8: User's Reference Guide*. Scientific Software.

Kumar N., Andreou A.G. (1998) Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, *Speech Communication*, 26, 283 - 297.

Lawley D.N., Maxwell A.E. (1962) Factor analysis as a statistical method, *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12, 209-229.

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, 91-97.

Lindsay B.G. (1988) Composite likelihood methods, *Contemporary Mathematics*, 80, 221-239.

Lubke G., Neale M. (2008) Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models, *Multivariate Behavioral Research*, 43, 592-620.

McLachlan G., Rathnayake S.I. (2016) Mixture models for standard p-dimensional Euclidean data. *Handbook of cluster analysis*, CRC Press, 145-172.

Millsap R.E., Yun-Tein J. (2004) Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39, 479-515.

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

Ranalli M., Rocci R. (2016a) Mixture models for ordinal data: a pairwise likelihood approach, *Statistics and Computing*, 26, 529-547.

Ranalli M., Rocci R. (2016b) Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. *Analysis of Large and Complex Data. Studies in Classification,Data Analysis and Knowledge Organization. Editors: Adalbert F.X. Wilhelm Hans A. Kestler.*.

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

Ranalli M., Rocci R. (2017b) A model-based approach to simultaneous clustering and dimensional reduction of ordinal data, *Psychometrika*, 82, 1007-1034.

Rocci R., Gattone S.A., Vichi M. (2011) A new dimension reduction method: Factor discriminant k-means. *Journal of classification*, 28, 210-226.

Van Buuren S. and Heiser W.J. (1989) Clustering objects into *k* groups under optimal scaling of variables, *Psychometrika*, 54, 699-706.

Varin C., Reid N., Firth D. (2011) An overview of composite likelihood methods, *Statistica Sinica*, 21, 1-41.

Vichi M., Kiers H.A.L. (2001) Factorial k-means analysis for two-way data, *Computational Statistics & Data Analysis*, 37, 49-64.

Bi-Factor MIRT Observed-Score Equating under the NEAT design for tests with several content areas

Valentina Sansivieri[*], Mariagiulia Matteucci[**], Stefania Mignani[***]

Abstract: Traditional item response theory (IRT) equating procedures are based on unidimensionally scored test forms. In this work, we propose an observed-score equating procedure based on the bi-factor extension of the three-parameter logistic (3PL) model under the nonequivalent groups with anchor test (NEAT) design. The bi-factor 3PL model is chosen because it allows for the presence of overall and specific traits and the guessing parameter which are especially important in educational assessment. The NEAT design is chosen because, when an anchor test is available, it allows to work with nonequivalent groups. The results obtained by using both simulated and real data show that, in presence of bidimensionality, the proposed equating procedure is more efficient than the unidimensional observed-score equating.

Keywords: Equating, Bi-factor, NEAT design.

1. Introduction

Test equating is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen and Brennan, 2014). Several designs can be used in test equating: single-group (Kolen and Brennan, 2014), equivalent groups (Kolen and Brennan, 2014), nonequivalent groups with anchor test (Kolen and Brennan, 2014) and nonequivalent groups with covariates (Wiberg and Bränberg, 2015; Sansivieri, 2017). In this study we will use the the nonequivalent groups with anchor test (NEAT) design, under which we assume that two nonequivalent groups of examinees take two test forms (one form for each group) which have a set of common items (the anchor).

Lord (1980) defined item response theory (IRT) as a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments

[*] University of Bologna, valentina.sansivier2@unibo.it

[**] University of Bologna, m.matteucci@unibo.it

[***] University of Bologna, stefania.mignani@unibo.it

measuring abilities, attitudes, or other variables and he introduced IRT equating to compare different test scores from different forms when IRT is used to assemble tests. By using IRT we obtain item parameters and examinee ability estimates and it could be necessary to put these estimates on the same scale: this process is called scale linking (Mislevy, 1992). Kolen and Brennan (2014) underline that scale linking is necessary when we have two nonequivalent groups of examinees.

If we assume that multiple hypothetical factors influence the performance on test items, we are within a multidimensional item response theory (MIRT) framework (Reckase, Ackerman, and Carlson, 1988). MIRT scale linking adjust for differences in traslation, dilation, rotation and correlation (Reckase, 2009; Brossman and Lee, 2013). In this work we will focus on scale linking under the NEAT design.

Although several MIRT models exist (Reckase, 2009), the BF model is one of the most general, and, for this reason, we choose to work with it.

Current research is an extension of the work by Lee and Lee (2016), who developed a bi-factor MIRT (BF-MIRT) observed-score equating procedure for mixed-format tests.

2. BF-MIRT model for a test with several content areas

The three-parameter logistic BF-MIRT model (Cai et al., 2011) is defined as follows

$$P(y = 1 \mid \theta_0, \theta_s) = c + (1 - c)\frac{1}{1 + \exp(-[d + a_0\theta_0 + a_s\theta_s])}, \tag{1}$$

where $P(y = 1 \mid \theta_0, \theta_s)$ is the probability of item endorsement given the general dimension θ_0 and the specific factor θ_s, c is the “guessing" probability, d the item intercept, a_0 the item slope on the primary factor, and a_s the item slope on specific factor s.

3. Scale linking

MIRT scale linking procedures are the instrument which we use to adjust for differences in translation, dilation, rotation and correlation (Brossman and

Lee, 2013). Translation implies that the origin of the $\boldsymbol{\theta}$-space changes. Dilation occurs when the units of the coordinate axes of the $\boldsymbol{\theta}$-space change. Rotation means that one can change the orientation of the coordinate axes of the $\boldsymbol{\theta}$-space to obtain an orientation for the axes which is easier to interpret. Finally, correlation occurs when the coordinate axes of the $\boldsymbol{\theta}$-space are nonorthogonal (Reckase, 2009). The initial test form X is used as the base coordinate system for the testing program: the linking wants to transform the item parameter estimates from the second test form Y so they are as similar as possible to those in the base coordinate system. Under the NEAT design, we can use the common items to calculate the transformation equations. The transformation equation for the $\boldsymbol{a}$-parameters is (Reckase, 2009, p.267)

$$\boldsymbol{C}'\boldsymbol{Rot}' = (\boldsymbol{v}'\boldsymbol{v})^{-1}\boldsymbol{v}'\boldsymbol{a}, \tag{2}$$

where $\boldsymbol{C}$ is the scaling matrix which is used to correct dilation (Reckase, 2009, p.254); $\boldsymbol{Rot}$ is the rotation matrix (Reckase, 2009, p.244); $\boldsymbol{v}$ is the matrix of common item discrimination estimates in the base coordinate system; $\boldsymbol{a}$ is the matrix of common item discrimination estimates from the calibration of form Y. The full set of discrimination parameters from test Y can be transformed to the base coordinate system by postmultiplying it by the matrix in Equation (2). To obtain the transformed $\boldsymbol{d}$-parameters, which we indicate with $\tilde{\boldsymbol{d}}$, we can use the following transformation (Reckase, 2009, p.268)

$$\tilde{\boldsymbol{d}} = \boldsymbol{a}\boldsymbol{\Omega} + \boldsymbol{d}, \tag{3}$$

where $\boldsymbol{a}$ is the matrix of discrimination estimates in the base coordinate system, $\boldsymbol{d}$ is the matrix of intercept estimates from the calibration of form Y and $\boldsymbol{\Omega}$ is estimated by using common items as follows (Reckase, 2009, p.269)

$$\boldsymbol{\Omega} = (\boldsymbol{a}'\boldsymbol{a})^{-1}\boldsymbol{a}'(\tilde{\boldsymbol{d}} - \boldsymbol{d}), \tag{4}$$

where $\boldsymbol{a}$, $\boldsymbol{d}$ and $\tilde{\boldsymbol{d}}$ have been defined previously. Finally, we can calculate the matrix of coordinates for the examinees in the base coordinate system, which we indicate with $\boldsymbol{\theta}$, by using the following equation (Reckase, 2009, p.298)

$$\boldsymbol{\theta}' = (\boldsymbol{a}_{\boldsymbol{b}}'\boldsymbol{a}_{\boldsymbol{b}})^{-1}\boldsymbol{a}_{\boldsymbol{b}}'\boldsymbol{a}_{\boldsymbol{Y}}\boldsymbol{v}'(\boldsymbol{a}_{\boldsymbol{b}}'\boldsymbol{a}_{\boldsymbol{b}})^{-1}\boldsymbol{a}'(\boldsymbol{d}_{\boldsymbol{Y}} - \boldsymbol{d}_{\boldsymbol{b}})\mathbf{1}, \tag{5}$$

where $\boldsymbol{a}_{\boldsymbol{b}}$ indicates the item discrimination parameters after transformation to the base coordinate system, $\boldsymbol{a}_{\boldsymbol{Y}}$ indicates the item discrimination parameters for the same items from the calibration of form Y, $\boldsymbol{d}_{\boldsymbol{b}}$ and $\boldsymbol{d}_{\boldsymbol{Y}}$ are the corresponding intercept terms, $\boldsymbol{v}$ is the matrix of coordinates for the examinees from the calibration of form Y, and $\mathbf{1}$ is a vector containing all 1s.

4. Observed-score equating

To conduct observed-score equating, the first step after estimating the probabilities by using the model in Equation (1) is calculating conditional observed-score distributions $f_t(x \mid \theta_0, \theta_1, \theta_2)$ (where θ_1 and θ_2 are the two specific factors) over the first t items by using an extended version of the recursive formula (Lord and Wingersky, 1984). After we need to aggregate these conditional distributions to obtain a marginal observed-score distribution, as follows (Lee and Lee, 2016)

$$f(x) = \int\int\int f(x \mid \theta_0, \theta_1, \theta_2) g(\theta_0, \theta_1, \theta_2)\, d\theta_0 d\theta_1 d\theta_2, \tag{6}$$

where $g(\theta_0, \theta_1, \theta_2)$ is the trivariate θ distribution. The previous steps are repeated on the new form Y to find $g(y)$. Finally, to find the equating relationships, the traditional equipercentile equating is conducted, as follows

$$e_Y(x) = G^{-1}[F(x)], \tag{7}$$

where $e_Y(x)$ is the equivalent on Form Y of the score x on Form X, F is the cumulative distribution function of X and G^{-1} is the inverse of the cumulative distribution function G of Y.

5. Simulation study

To implement the simulation study we initially manipulated the data coming from the large scale standardized test administered by the National Evaluation Institute for the School System (Invalsi) to fifth grade Italian students.

This test showed a clearly bidimensional structure due to the presence of two different sets of items involving ability in reading and grammar (Matteucci and Mignani, 2015). The Invalsi test is composed by dichotomously scored items and it is given once a year with new test forms. The 2015 Invalsi fifth grade Italian test consisted of 33 items about the content area “Reading" and of 10 items about the content area “Grammar". We used a random sample of the national data of size 5000. In order to obtain two test forms and an anchor test, we manipulated the data by following Holland et al. (2008).

The simulation study is conducted by following these steps:

1. By using a BF model we estimate the item parameters on the two test forms. These parameters are considered to be the *true* item parameters.

2. By using a standard bivariate normal distribution we generate the examinees'abilities. The correlation between the latent abilities assume the values 0.2, 0.5, 0.7 and 0.9 in the different simulation conditions, while the number of examinees is 600, 2000 and 6000.

3. By using the item parameters estimated at the step 1 and the examinees'abilities generated at the step 2 we simulate two response dataset.

4. For each response dataset obtained at the step 3, we estimate the item parameters by using, in the unidimensional case, the three-parameter IRT model (Birnbaum, 1968) and, in the multidimensional case, the three-parameter logistic BF-MIRT model defined in Equation (1). After, we conduct our equating and we also conduct traditional equipercentile equating (Kolen and Brennan, 2014) to check the results obtained by using IRT equating (Kolen and Brennan, 2014; Lee and Lee, 2016). After conducting equating, we calculate the following weighted root mean squared differences (WRMSD1 and WRMSD2) indices to evaluate our results (Lee and Lee, 2016)

$$WRMSD1 = \{\sum_{i=1}^{k} w_i(EEQ_i - TEQ_i)^2\}^{\frac{1}{2}}, \tag{8}$$

where w_i is the frequency of the score i in the the test form Y, K is the maximum score, EEQ_i is the equating equivalent estimated by using either the UIRT or BF-MIRT and TEQ_i is the equating equivalent estimated by using the traditional equipercentile equating; and

$$WRMSD2 = \{\sum_{i=1}^{k} w_i(BFEEQ_i - UIRTEQ_i)^2\}^{\frac{1}{2}}, \tag{9}$$

where $BFEEQ_i$ and $UIRTEQ_i$ are the equating equivalents estimated by using the BF-MIRT and the UIRT, respectively, and all the other quantities have been defined previously.

The quantities defined in Equation (8) and in Equation (9) are averaged over 200 replications. We use the software R (R Core Development Team, 2016) to conduct the whole simulation study and also the empirical example.

Table 1. WRMSD1 of BF-MIRT and UIRT equating methods and WRMSD2 for the simulation study, n=6000

	WRMSD1 BF-MIRT	WRMSD1 UIRT	WRMSD2
R=0.2	0.197847	0.228975	0.161134
R=0.5	0.202119	0.210682	0.167186
R=0.7	0.211382	0.194221	0.185070
R=0.9	0.200157	0.184088	0.171951

Table 1 shows that the the BF-MIRT method provided WRMSD1 values lower than those provided by the UIRT method when the correlation between the specific factors is low (0.2 and 0.5) and the number of examinees is 6000. WRMSD2 is also low: this means that the results obtained by using the two methods are not so different. When the number of examinees is 2000 the two methods showed more or less the same WRMSD1 values through the different correlation levels; finally, when the number of examinees is 600 the UIRT method provided WRMSD1 values lower than those provided by the BF-MIRT method through the different correlation levels.

6. Real test data study

The 2012 and 2013 administrations of the Invalsi fifth grade Italian test were used in the real test data study. The 2012 Invalsi fifth grade Italian test consisted of 32 items about the content area “Reading" and of 11 items about the content area “Grammar"; the 2013 Invalsi fifth grade italian test was composed by 42 items of the content area “Reading" and by 10 items about the content area “Grammar". The sample sizes of the 2012 and 2013 administrations are 2767 and 2346, respectively.

We conduct UIRT observed-score equating, BF-MIRT observed-score equating and traditional equipercentile equating on the two forms described above and after we calculate the quantities WRMSD1 and WRMSD2 defined in Equation (8) and in Equation (9).

Table 2. WRMSD1 of BF-MIRT and UIRT equating methods and WRMSD2 for the Real test data study

WRMSD1 BF-MIRT	WRMSD1 UIRT	WRMSD2
1.170206	1.193503	0.1983227

The Invalsi 2012 and 2013 administrations are bidimensional, even if the correlation between the specific factors is about 0.8.

Table 2 shows that the the BF-MIRT method provided WRMSD1 value lower than the one provided by the UIRT method: this means that the equivalent scores of the BF-MIRT method were closer to those of the equipercentile method than under the UIRT method. WRMSD2 is also low: this means that the results obtained by using the two methods are not so different.

7. Discussion with concluding remarks

Both the simulation study and the real test data study show that when there is bidimensionality and a good number of examinees is available we obtain more accurate results by using the BF-MIRT observed-score equating than by using the UIRT observed-score equating.

A future interesting topic would be the sensitivity analysis to detect the effect of the test forms length and anchor length. Finally, future research could extend the proposed model to the case of three or more specific factors.

References

Birnbaum A. (1968) Some latent trait models and their use in inferring an examinee's ability in Lord F.M. and Novick M.R. (eds.), *Statistical theories of mental test scores*, chaps. 17-20. Addison-Wesley, Reading, MA.

Brossman B.G., Lee W. (2013) Observed score and true score equating procedures for multidimensional item response theory, *Applied Psychological Measurement*, 37, 460-481.

Cai L., Yang J.S., Hansen M. (2011) Generalized full-information item bifactor analysis, *Psychological Methods*, 16, 221-248.

Holland P.W., Sinharay S., von Davier A. (2008) An Approach to Evaluating the Missing Data Assumptions of the Chain and Post-stratification Equating Methods for the NEAT Design, *Journal of Educational Measurement*, 45, 17-43.

Kolen M.J., Brennan R.L. (2014) *Test equating, scaling and linking: Methods and practices (3rd ed.)*, Springer, New York, NY.

Lee G., Lee W.-C. (2016) Bi-factor MIRT observed-score equating for mixed-format tests, *Applied Measurement in Education*, 29, 224-241.

Lord F.M. (1980) *Applications of item response theory to practical testing problems*, Erlbaum, Hillsdale, NJ.

Lord F.M., Wingersky M.S. (1984) Comparison of IRT true-score and equipercentile observed score "equatings", *Applied Psychological Measurement*, 8, 452-461.

Matteucci M., Mignani S. (2015) Multidimensional IRT models to analyze learning outcomes of Italian students at the end of lower secondary school in Millsap R.E., Bolt D.M., van der Ark L.A., Wang W.-C. (eds.), *Quantitative Psychology Research, Springer Proceedings in Mathematics & Statistics*, Springer International Publishing, Switzerland, 89, 91-111.

Mislevy R.J. (1992) *Linking educational assessment: Concepts, issues, methods and prospects*, ETS policy information center, Princeton, NJ.

R Core Development Team (2016) *R: A language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for statistical computing. http://www.R-project.org/.

Reckase M.D., Ackerman T.A., Carlson J.E. (1988) Building a unidimensional test using multidimensional items, *Journal of Educational Measurement*, 25, 193-204.

Reckase M.D. (2009) *Multidimensional item response theory*, Springer, New York, NY.

Sansivieri V. (2017) *Item response theory equating with the non-equivalent groups with covariates design*. PhD thesis, Alma Mater Studiorum Università di Bologna, Bologna.

Wiberg M., Bränberg K. (2015) Kernel equating under the non-equivalent groups with covariates design, *Applied Psychological Measurement*, 39, 1-13.

On the choice of splitting rules for model-based trees for ordinal responses

Rosaria Simone[*], Francesca Di Iorio[**], Carmela Cappelli[***]

Abstract: The focus of the contribution is on the splitting criterion for model-based tree procedure based on the class of CUB mixture models for the analysis of ordinal scores. The flexibility of the chosen modelling framework allows to select the splitting criterion to grow the tree according to the purposes of the study and the available data. In particular, the selection of variables yielding to the best partitioning results can be driven by fitting measures or classical likelihood and deviance measures. The contribution proposes to investigate the features of the available decision rules by a set of Montecarlo experiments, thus implicitly facing the problem of selecting the model-based tree to obtain an adequate and satisfying overview of response profiles.

Keywords: Rating Data, Model-based trees, Splitting criterion.

1. Introduction

In the last decades tree based methods have proven to be a useful non-parametric approach for high dimensional data analysis. In a nutshell, the process of growing trees relies on a recursive binary splitting that allows to choose at each tree node (i.e. a subset of observations), the best split, i.e. the optimal binary division into two subgroups of observations according to a certain rule. All the covariates considered in the procedure, irrespective of their original scale of measurements, are dichotomized for the identification of the optimal split that achieves the highest reduction in impurity when dividing the parent node into its child nodes.

Following the model-based approach to classification trees introduced in Zeileis *et al.* (2008), Cappelli *et al.* (2017) proposed a model-based partitioning algorithm focussing on CUB models (D'Elia and Piccolo 2005, Piccolo

[*] University of Napoli Federico II, rosaria.simone@unina.it
[**] University of Napoli Federico II, fdiiorio@unina.it
[***] University of Napoli Federico II, carcappe@unina.it

and D'Elia 2008), a class of models that has received an increasing attention in recent years due to successful applications to the analysis of judgements, evaluations and perceptions, in various fields of research. Those models aim to disentangle the individual answer into the personal feeling, usually related to the subjects' motivations, and the inherent uncertainty in choosing the ordinal response. The model can be employed without covariates to estimate the expected distribution given a sample of n observed ordinal values. However the introduction of covariates greatly improves its usefulness and relevance. If we consider the model parameters as functions of subjects' covariates we get a CUB regression model, i.e., a regression model for an ordinal response in which the selection of the covariates for uncertainty and/or feeling, that mostly explain the response and improve the fitting, is a relevant issue.
The procedure for growing trees for ordinal responses in which every node is associated with a CUB regression model is known as CUBREMOT (CUB REgression MOdel Trees). So far, two splitting criteria have been implemented for node partitioning with CUBREMOT : the first considers the log-likelihood increment from the father node to the child nodes for each possible split, and then chooses the one that maximizes such deviance, the second focuses on the dissimilarity between child nodes, aiming at generating child nodes as far apart as possible with respect to the probability distributions estimated by CUB models. Both splitting criteria generate a model-based tree whose terminal nodes provide different profiles of respondents, which are classified into nodes according to levels of feeling and/or uncertainty conditional to the splitting covariates. In this way the most explanatory covariates are automatically selected in the partitioning process and the terminal nodes in the tree provide alternative profiles of respondents based on the covariates values. This contribution investigates the relative performance of the two splitting criteria by a set of Monte Carlo experiments on a given tree structure, providing support to the application of CUBREMOT while outlining further splitting rule that could overcome possible issues.

2. CUBREMOT

In CUBREMOT, the top-down partitioning algorithm that grows the tree is based on the estimation, at each tree node, of CUB models (D'Elia and Piccolo, 2005) where the discrete choice on a rating scale is assumed as the combination of a *feeling* and an *uncertainty* components. A shifted Binomial distribution describes the feeling component (to account for substantial likes and agreement) and a discrete Uniform distribution describes uncertainty to shape heterogeneity. Let R_i denotes the response of the i-th subject to a given item of a questionnaire collected on a m-point scale, the model is given by:

$$Pr(R_i = r|\pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} + (1-\pi_i)\frac{1}{m}, \quad r = 1, \dots, m,$$

where the parameters π_i and ξ_i are called uncertainty and feeling parameter, respectively. Covariates can be included in the model in order to relate feeling and/or uncertainty to respondents' profiles. Customarily, a logit link is considered:

$$logit\,(\pi_i) = \boldsymbol{y}_i\boldsymbol{\beta}\,; \qquad logit\,(\xi_i) = \boldsymbol{w}_i\boldsymbol{\gamma}\,, \tag{1}$$

where $\boldsymbol{y}_i, \boldsymbol{w}_i$ are the values of selected explanatory variables for the i-th subject. When no covariate is considered for feeling and uncertainty, the $\pi_i = \pi$ and $\xi_i = \xi$ are constant among subjects. Likelihood methods, and the implementation of the Expectation-Maximization (EM) algorithm, are the preferred estimation procedures for CUB models.

According to the binary tree recursive approach, the CUBREMOT procedure sequentially transformes and uniquely associates with dichotomous factors each available covariates in order to yield the set of candidate splitting variables. In particular the procedure can be summarize as follows. Let $\mathcal{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k)$ the log-likelihood associated to the final ML estimates $(\hat{\pi}_k, \hat{\xi}_k)$ obtained by a CUB mobel without covariates for n_k individual observation a given node $k \geq 1$. Then the statistical significance of the parameter associated to a given splitting variable s in CUB regression model is tested. If the parameter is significant for at least one component, it implies a split into a left $(2k)$ and a right $(2k + 1)$ child nodes. The nodes will be associated with the conditional distributions $R|s = 0$ with parameter values $(\hat{\pi}_{2k}, \hat{\xi}_{2k})$ and $R|s = 1$ with parameter values $(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})$, respectively. This procedure selects a set of significant candidate splitting variables of node k, $\mathcal{S}_k = \{s_{k,1}, \dots, s_{k,l}\}$and the binary variable in $\mathcal{S}_k$ associated to the best split can be chosen according to two alternative *goodness of split* criteria.

- *Log-likelihood splitting criterion.* The best split maximizes the deviance:

$$\Delta\,\mathcal{L}_k = \big[\mathcal{L}_{n_{2k}}(\hat{\pi}_{2k}, \hat{\xi}_{2k}) + \mathcal{L}_{n_{2k+1}}(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})\big] - \mathcal{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k). \quad (2)$$

  This criterion considers the improvement in log-likelihood yielded by the inclusion of the significant splitting variable and the best split, being associated with the maximum log-likelihood increment, provides the child nodes characterized by the most plausible values for CUB parameters.

- *Dissimilarity measure splitting criterion.* This criterion selects for the k-th node the split that maximizes the distance between the estimated CUB probability distributions $\hat{p}^{(2k)}$ and $\hat{p}^{(2k+1)}$ for the child nodes:

$$\mathit{Diss}(2k, 2k+1) = \frac{1}{2}\sum_{r=1}^{m} |\hat{p}_r^{(2k)} - \hat{p}_r^{(2k+1)}|. \quad (3)$$

  The choice of this normalized index entails that the resulting terminal nodes determine well-separated profiles of respondents, in terms of feeling (agreement, preferences, and so on) and/or uncertainty (indecision, heterogeneity). This criterion considers a proper version of the normalized index proposed by Leti(1983) that compares an estimated probability distribution with the observed relative frequencies and it is generally considered in the framework of CUB models as a goodness of fit measure.

In both cases, the node partitioning process stops and a node is declared terminal if none of the available covariates is significant (neither for feeling nor for uncertainty), or if the sample size is too small to support a CUB model fit.

3. Monte Carlo Design

The contribution of this paper is to verify by a set of Monte Carlo experiments the performance of the proposed criteria with respect to a given rating response and some respondent characteristics. To this aim we simulate $nsimul = 100$ replicates of a sample of $n = 2000$ individuals ratings with $m = 7$, 1400 of them associated to male ($G = 1$). We also assume that patterns for the males' responses with $Age < 35$ ($A_i = 0$) is different from males' answers with $Age \geq 35$ ($A_i = 1$). Then the data are generated to have the tree structure displayed in Figure 1:

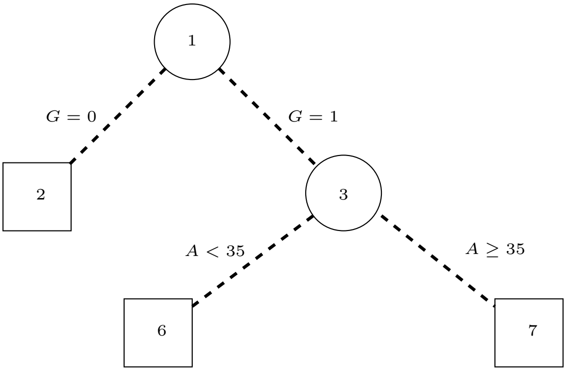


Figure 1. CUBREMOT *: Tree of simulated data*

Then, we consider a dummy variable G and a continuous variable for Age split at 35 (A), respectively, which are maintained constant while sampling only the rating response: for each simulation plan, Age is generated both from a Gaussian distribution $\mathcal{N}(\mu = 37, \sigma^2 = 4)$ and from a Uniform distribution over the interval $(30, 40)$; parameter values for terminal nodes are reported in Table 1.

Specifically, different scenarios have been designed:

1. Plan 1 corresponds to the case in which overall the uncertainty level is low and the distribution at the terminal nodes are very different;

Table 1. Parameter values at terminal nodes for the simulation plans

	Node 2		Node 6		Node 7	
	π_2	ξ_2	π_6	ξ_6	π_7	ξ_7
Plan 1	0.7	0.8	0.9	0.6	0.5	0.4
Plan 2	0.7	0.2	0.5	0.4	0.2	0.3
Plan 3	0.4	0.6	0.3	0.6	0.1	0.6

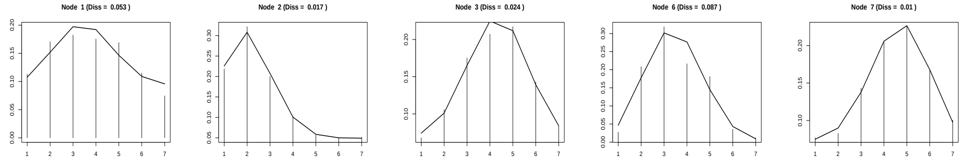


Figure 2. Nodes distribution for simulation plan 1

2. Plan 2 corresponds to the case in which the dummy split A at age 35 is significant for both feeling and uncertainty, but being more important for the latter component and feeling being almost homogeneous at the terminal nodes.

3. Plan 3 corresponds, instead, to data generated with a constant feeling and splits are to be called significant only with respect to the uncertainty components, with heterogeneity in the data being high or fairly high.

Figure 2-4 displays instances of simulated and estimated distributions at the tree nodes for the different plans.

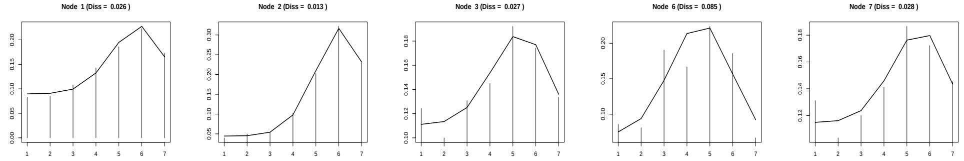


Figure 3. Nodes distribution for simulation plan 2

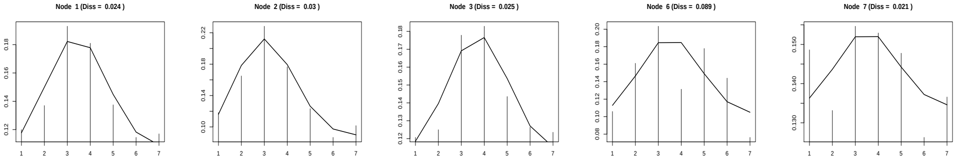


Figure 4. Nodes distribution for simulation plan 3

4. Results and Comments

The performances of each splitting criterion for the given tree structure are measured via mis-classification errors (that is, the proportion of cases in which, for the given node, a different split has been selected for partitioning), which are reported in Table 1. For node 3, the proportion of cases in which the procedure splits according to A or not is adjusted by including also the proportion of cases in which the selection is the split at Age 34 or Age 36 to control for uncertainty of classification (squared brackets).

Table 2. Miss-classification errors for splits at nodes 1 and 3 for different criteria

		Plan 1		Plan 2		Plan 3	
Age		$\Delta\mathcal{L}_k$	$Dissim$	$\Delta\mathcal{L}_k$	$Dissim$	$\Delta\mathcal{L}_k$	$Dissim$
$\sim \mathcal{N}(37, 4)$	Node 1	0	0	0	0	0.07	0.11
	Node 3	0	0	0.081 [0.031]	0.041 [0.031]	0.329 [0.151]	0.287 [0.178]
$\sim \mathcal{U}(30, 40)$	Node 1	0	0	0	0	0.368	0.347
	Node 3	0	0	0.18 [0.05]	0.28 [0.01]	0.643 [0.443]	0.647 [0.456]

From Table 1 we can claim that, if the parent distributions at the terminal nodes are very different, both splitting rules ensure very satisfactory performances. In general, the chosen partitioning rules behave in a comparable way. With reference to the simulation Plan 3, it is worth to report that the mis-classification error for Node 3 drops to 0.328 and 0.309 for the log-likelihood and the dissimilarity splitting criterion, resp., when enlarging the neighboring ages to the interval $(33, 37)$.

5. *Further developments*

The next step in the sensitivity analysis for the splitting criteria currently available for CUBREMOT is to assess the stability of the procedure with respect to the conditional distribution given the covariates as a sort of *fluctuation tests*: then, at each simulation run, also covariates should be generated from the parent distribution in order to have some natural noise. This issue is particularly important when checking the stability of a classification with respect to a continuous covariate; in that case, indeed, the model-based partitioning rule prescribes that a dummy split is tested for every possible covariate value. More interestingly, we have seen from our Monte Carlo experiment that, as the overall uncertainty increases, the mis-classification error for the split of continuous covariates also increases. This leads us to envisage a splitting criterion to grow a tree in which each partitioning corresponds to the split that most reduces the uncertainty in the data. This work is the subject of ongoing research.

Acknowledgements: The contribution is framed within the CUBREMOT project of University of Naples Federico II for findings and results.

References

Cappelli C., Simone R., Di Iorio F. (2017) Growing happiness: a model-based tree, in: Petrucci A., Verde R. (eds), *Statistics and Data Science: new challenges, new generations Proceedings of the Conference of the Italian Statistical Society 2017*, Firenze University Press, Firenze, 261-266.

D'Elia A., Piccolo D. (2005) A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917-934.

Leti G. (1983) *Statistica descrittiva*, Il Mulino, Bologna.

Piccolo D., D'Elia A. (2008) A new approach for modelling consumers' preferences, *Food Quality & Preference*, 19, 47-259.

Zeileis A., Hothorn T., Hornik K. (2008) Model-Based Recursive Partitioning, *Journal of Computational and Graphical Statistics*, 17, 492-514.

Robustness issues for categorical data

Francesca Torti[*], Silvia Salini[**], Marco Riani[***]

Abstract: Correspondence Analysis (CA) is a popular method to analyse relationships between categorical variables. Classically, the procedure involves the decomposition of Pearson residuals using singular value decomposition, thereby allowing the user to view the correspondence between categories in low-dimensional space. The aim in Correspondence Analysis is to find *k* dimensional coordinate matrices **X** and **Y**, for row and column points. The graphical representation depends on the parameter α_C that determines the type of coordinates in X and Y. This contribution considers the robustness of the correspondence map when the parameter α_C varies across a set of values. In this initial exploration, we address the problem empirically with an example from prices in international trade. In order to apply the Correspondence Analysis, the prices, which are naturally continuous, are robustly clustered in a discrete number of homogeneous groups.

Keywords: Biplot, Confidence ellipse, Adjusted residuals.

1. Introduction

Correspondence Analysis is a statistical technique that provides a graphical representation of the contingency tables. The literature is vast, but major references can be found in Lebart et al. (1984) or Greenacre (1984 and 2017). The typical plot in Correspondence Analysis visualizes the data in a two dimensional space using the first two extracted coordinates **X** and **Y** from both rows and columns. The objective is to get an idea of the association between two variables and understand which categories of rows and columns determine the structure of the dependence. To do this we consider the elliptically confidence regions proposed by Beh (2010). Confidence ellipses, and therefore the significant, or not, associations between categories of rows and columns, depend on the type of coordinates of **X** and **Y**.

[*] European Commission, Joint Research Centre, francesca.torti@ec.europa.eu

[**] University of Milan, silvia.salini@unimi.it

[***] University of Parma, marco.riani@unipr.it

There are different representations of the correspondence map that depend on the parameter α_C, as shown by Lorenzo-Seva et al. (2009). Flexible and informative representations that can be controlled over a set of values for the parameter $alpha$ have been recently introduced in the Flexible Statistics for Data Analysis Toolbox for MATLAB (FSDA, Riani et al. 2012). The toolbox is available at the address `http://rosa.unipr.it/fsdadownload.html` of the University of Parma, also via the European Commission's website `http://fsda.jrc.ec.europa.eu`.

The main function, `CorAnaplot`, produces elliptical confidence regions and the correspondence map for any level of α_C; therefore in this function the parameter `alpha` can be set with values that are not limited to the set $\{0, 0.5, 1\}$, typical of this context.

This extension allows to understand if an optimal level of α_C exists, which allows to obtain the map minimizing the distance between the row and column points, enriched by confidence ellipses highlighting robust row-column associations.

We present the problem of the choice of the α_C level through a real data example in the field of international trade. The trade prices, which constitute a continuous variable, before being analysed with the Correspondence Analysis, are clustered in homogeneous groups with a robust clustering method.

This preliminary exploration will be complemented, in future work, by an automatic procedure to select the optimal level of α_C and the corresponding robust map and an intensive simulation study for assessing the procedure.

Again in the direction of robustness, we also envisage to decompose the adjusted residuals instead of Pearson residuals.

More precisely, we will follow Fuchs and Kenett (1980), who have proposed the M-test based on adjusted residuals to detect outlying cells in the two-way contingency tables, and Beh (2012) who studied the impact of the application of the adjusted residues in the analysis of correspondence.

The introduction of the Correspondence Analysis is left out from this short and preliminary discussion: we refer, for example, to Lorenzo-Seva et al. (2009) for the theory on the topic. Similarly, for the mathematics of the elliptical confidence regions we refer to Beh (2010). Instead, in Section 2 we will illustrate and discuss the use of the tools with reference to a real application in the filed of trade prices in the European Union.

2. Application

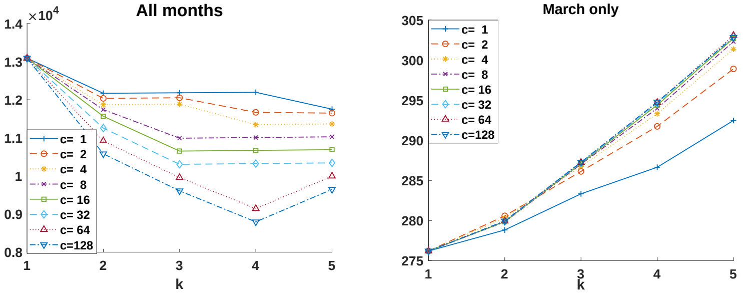
The regulatory framework of the European Union (EU) reserves to the EU institutions the responsibility of trade relations between the EU Member States (MSs) and the non-EU countries. The target is the development of statistical methods for the analysis of international trade data for the detection of Customs frauds (e.g. under-valuation of import duties), the defence against anti-competitive conducts and for facilitating price convergence among EU MSs.

The data considered are trade values and quantities aggregated monthly according to the product, the country of origin and the country of destination. The monthly aggregates are downloaded from the COMEXT database of the European Statistical Office, Eurostat. In COMEXT the product codes are classified at the detailed 8-digits level of the Combined Nomenclature (CN8); therefore, data for the same product code are reasonably comparable. The quantities are given in tons, or supplementary units if foreseen, and the values, in thousands of Euros.

The Joint Research Centre of the European Commission provides estimation of fair (import) prices from COMEXT data. They can be used for the determination of the customs value at the moment of the customs formalities (to establish how much duty the importer must pay) and the auditing activities in post-clearance checks of individual import or export transactions.

In this work we start from a set of these fair prices, concerning imports of a given product from one Third Country into EU during the year 2016. The set is formed by 28 (number of MSs) $\cdot 12$ (number of months) $= 336$ monthly fair prices. We cluster them in k homogeneous groups and then verify with the Correspondence Analysis if there are significant associations among MSs and the assignment to low-price or high-price clusters.
For doing this we have used the Flexible Statistics for Data Analysis (FSDA) MATLAB toolbox, introduced by Riani at al. (2012).

For clustering the fair prices, we have used the robust method TCLUST, introduced by Garcia-Escudero et al. (2008). TCLUST, as most of clustering



```
%%% Run tclustIC on all MS prices to estimate k    %%%
outIC=tclustIC(price_2016,'plots',0,'whichIC','MIXMIX');
%%%    Run tclustIC on one MS prices to estimate c    %%%
outIC=tclustIC(price_2016March,'plots',0,'whichIC','MIXMIX');
%%%%%%%    plot the monitoring of the likelihood %%%%%%%
tclustICplot(outIC);
```

Figure 1. Monitoring of the log-likelihood obtained by applying TCLUST on all monthly fair prices (left panel) and on March fair prices only (right panel), for different number of groups (x-axis) and restriction factors c (different curves). The box at the bottom of the figure contains the code used to generate the plots.

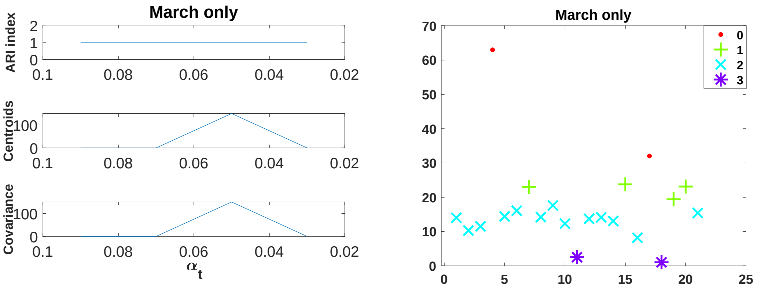
methods, determines a priori a number of parameters: the optimal number of groups, the restriction factor c, the trimming percentage α_t.

For the choice of the optimal number of groups, we have used the recent results of Cerioli et al. (2017). In particular we have applied the `tclustIC` function of FSDA in a univariate context to all monthly fair prices, i.e. estimated prices for all the 28 MSs.

Differently from what we will do for estimating the restriction factor and the percentage of trimming, we estimate the number of groups using prices of all MSs. In fact, for a reasonable application of the Correspondence Analysis, the number of groups have to be the same in all months.

The main output of the method of Cerioli et al. (2017) is reported in the left panel of Figure 1.

Each curve refers to a different restriction factor c, i.e. the parameter of TCLUST which allows to identify groups with a shape more (large values of c) or less (values of c closed to 1) elliptical. Each curve represents the



```
%%%%%%%%%          monitoring alpha     %%%%%%%%%%
c = 1; k = 3; alpha_vec = 0.01:0.02:0.09;
outeda = tclusteda(price_2016March,k,alpha_vec,c,'plots',1);
%%%%%%%%%     identify groups with TCLUST    %%%%%%%%%
k = 3; alpha = 0.05; c = 1;
out = tclust(price_2016March,k,alpha,c);
```

Figure 2. For March estimated prices, on the left panel monitoring the variation in the Adjusted Rand Index, centroids and covariance matrix for increasing values of the trimming percentage α_t (x-axis); the second and third measures indicate $\alpha_t = 0.05$ as the best percentage of trimming. On the right panel, TCLUST classification of March prices (the two red points labelled with "0" are trimmed units). The box at the bottom of the figure contains the code used to generate the plots.

log-likelihood obtained for different number of groups k (on the x-axis). The log-likelihood decreases when the number of groups and restriction factor approach the optimal ones. Of course, more are the number of groups, lower is the log-likelihood. Figure 1 shows that for intermediate values of c $(8, 16, 32)$ a good choice of k could be 3: the log likelihood improves compared to the log-likelihood for $k = 2$, while does not change compared to the log-likelihood for $k = 4$.

Set the number of groups to $k = 3$, we identify the optimal restriction factor c following Cerioli et al. (2017), as just described, but month by month. To give an example, in the right panel of Figure 1, we have reported the monitoring of the log-likelihood only for the month of March. The best restriction factor is 1, which corresponds to spherical clusters. Set the number of groups and the restriction factor, we choose the percentage of trimming α_t by monitoring with the function `tclusteda`, month by month, the variation in the Adjusted

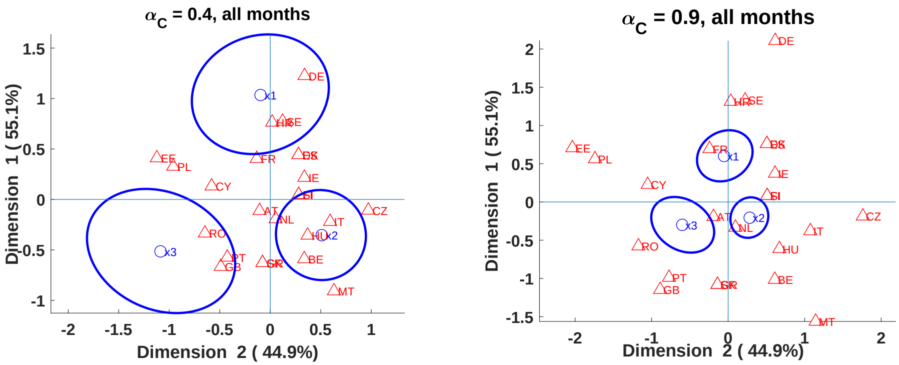
Rand Index (ARI), centroids and covariance matrix for increasing values of α_t.

To give an example, in the left panel of Figure 2 we have reported the monitoring of the three measures for the month of March. Both the monitoring of the centroids and of the covariance matrices indicate that a good percentage of trimming could be $\alpha_t = 0.05$.

Having identified the three parameters, we run TCLUST with the function `tclust` month by month; each estimated price (with the exception of the trimmed ones)is therefore assign to a cluster.

To give an example, in the right panel of Figure 2, we have reported the TCLUST assignments for March prices.

In the resulting TCLUST classification "1" is the group characterized by the highest prices, "2" with the medium prices, "3" with the lowest prices. On the described classification, we apply the Correspondence Analysis (Figure 3) to study the significant association between the TCLUST classification (blue circles labelled as 1, 2 and 3) with the MS of destination (red triangles). In particular we have reported the results for two levels of the parameter α_C which determines the type of coordinates: 0.4 (left panel) and 0.9 (right panel). The two plots show different significant associations, which are represented by MSs falling inside the confidence ellipses of the three TCLUST groups of prices. With $\alpha_C = 0.9$ there is a strong association of the group number 1 (of the high prices) with France; with $\alpha_C = 0.4$ there is a strong association of the group number 1 with Germany, Sweden and Croatia, and of the group number 3 (of the lowest prices) with Romania, Portugal and Great Britain. It is evident the need of an automatic procedure to select the optimal level of α_C.



```
%%%%%%%%%%%%%%%%%%% INPUT PARAMETERS %%%%%%%%%%%%%%%%%%%%
plots = struct; plots.alpha='bothprincipal'; %symmetrical
%plot or French symemtrical model.
labels_rows  = {'1' '2' '3' };
labels_columns= {'AT' 'BE' 'CY' 'CZ' 'DE' 'DK' 'EE' 'ES'
 'FI' 'FR' 'GB' 'GR' 'HR' 'HU' 'IE' 'IT' 'MT' 'NL' 'PL'
 'PT' 'RO' 'SE' 'SI' 'SK'};

%%% Cross tabulation of vector of MS and of assignements.%%
N=crosstab(groups,MS);
%%%%%%%% Correspondence Analysis execution %%%%%%%
out=CorAna(N,'Lr',labels_rows,'Lc',labels_columns,
'plots', plots,'d1',2,'d2',1);

%%%%%%%%%%%%%%%%%%% INPUT PARAMETERS %%%%%%%%%%%%%%%%%%%%
confellipse = struct;
confellipse.conflev = 0.95;
confellipse.method = {'multinomial'};
confellipse.AxesEllipse = false;
confellipse.selRows=[1 2 3]; confellipse.selCols=[];
plots = struct; plots.alpha=0.4; %or 0.9

%%%%%%%%%% Correspondence Analysis plot %%%%%%%%%
CorAnaplot(out,'plots',plots,'confellipse',confellipse,
'd1',2,'d2',1);
```

Figure 3. Main output produced by the Correspondence Analysis. The plot on the left is obtained with $\alpha_C = 0,4$, the plot on the right with $\alpha_C = 0,9$. The box at the bottom of the figure contains the code used to generate the plots.

References

Beh E.J. (2010) Elliptical confidence regions for simple correspondence analysis, *Journal of Statistical Planning and Inference*, 140, 2582-2588.

Beh E.J. (2012) Simple correspondence analysis using adjusted residuals, *Journal of Statistical Planning and Inference*, 142, 965-973.

Cerioli A., Garcia-Escudero L.A., Mayo-Iscar A., Riani M. (2018) Finding the number of normal groups in model-based clustering via constrained likelihoods, *Journal Computational Graphical Statistics*, 11, 404-416.

Fuchs C., Kenett R. (1980) A test for detecting outlying cells in the multinomial distribution and two-way contingency tables, *Journal of the American Statistical Association*, 75, 395-398.

Garcia-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.

Greenacre M. (2017) *Correspondence analysis in practice*, Chapman and Hall/CRC.

Greenacre M. (1984) *Theory and applications of correspondence analysis*, Academic Press London

Lebart L., Morineau A., Warwick K.M. (1984) *Multivariate descriptive analysis and related techniques for large matrices*, John Wiley and Sons, New York.

Lorenzo-Seva U., van de Velden M., Kiers H. (2009) Car: A matlab package to compute correspondence analysis with rotations, *Journal of Statistical Software*, 31, 1-14.

Riani M., Perrotta D., Torti F. (2012) FSDA: a matlab toolbox for robust analysis and interactive data exploration, *Chemometrics and Intelligent Laboratory Systems*, 116, 17-32.

Applications and theoretical results of association rules and compositional data analysis: a contingency table perspective

Marina Vives-Mestres*, Josep Antoni Martín-Fernández**,
Santiago Thió-Henestrosa***, Ron S. Kenett****

Abstract: Association rule mining was originally developed for basket analysis. To generate an association rule, the collection of more frequent itemsets must be detected. The association rules are then ranked using measures of interestingness. Using the associaton rule expression as a contingency table a representation on the unit simplex is appropiate. Compositional data analysis provides nice properties such as subcompostional coherence and scalability. We explore here the implication of compositional data analysis to association rule mining in large databases and big data and propose compositional measures of interestingness. Visualization of compositional measures on a simplicial representation of the itemsets gives new insights in association rule mining. The case study used here to demonstrate our approach is derived from a medical data set of side effects from Nicardipine.

Keywords: Aitchison geometry, Isometric logratio coordinates, Measures of interestingness.

1. Introduction

Commonly a large database with unstructured semantic data (Agrawal et al., 1993) is formed by a set of n binary variables or attributes $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_n\}$ called *items*; and a set of m rows $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m\}$ known as *transactions*. Let $\mathbf{S}_a$, $\mathbf{S}_c \subseteq \mathbf{I}$ be two sets of items (itemsets) with $\mathbf{S}_a \cap \mathbf{S}_c = \varnothing$, that is with empty intersection of items. An implication of the form $\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$ is called a *rule*, where $\mathbf{S}_a$ and $\mathbf{S}_c$ are respectively the antecedent and consequent itemsets. These rules that are "important" express that the itemsets are associated, having an association rule (AR).

*University of Girona, marina.vives@udg.edu

**University of Girona, josepantoni.martin@udg.edu

***University of Girona, santiago.thio@udg.edu

****KPA Group and Samuel Neaman Institute, ron@kpa-group.com

Let $\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$ be the AR of interest. Let n_1 be the absolute frequency of occurrence of both $\mathbf{S}_a$ and $\mathbf{S}_c$; n_2 the frequency of only $\mathbf{S}_a$; n_3 the frequency of only $\mathbf{S}_c$; and n_4 the number of transactions where neither $\mathbf{S}_a$ or $\mathbf{S}_c$ occur. In other words, let x_k be the relative frequency (support) which satisfy the conditions in n_k, ($x_k = n_k/m$, $k = 1, \ldots, 4$), and the total number of transactions is $\sum n_k$=m. Consequently, $\sum x_k$=1 and **x**= (x_1, x_2, x_3, x_4) can be considered as a *composition*. Compositions are vectors whose elements, called parts, provide relative information about a whole (Aitchison, 1986). Table 1 shows that x_k respectively estimates P(S$_a \cap$ S$_c$), P(S$_a \cap$ ^{c}S$_c$), P(^{c}S$_a \cap$ S$_b$), P(^{c}S$_a \cap$ ^{c}S$_c$).

Table 1. AR contingency table for the AR $\{S_a \Rightarrow S_c\}$

	$\boldsymbol{S}_c$	$^c\boldsymbol{S}_c$
$\mathbf{S}_a$	x_1	x_2
$^c\mathbf{S}_a$	x_3	x_4

When compositions are represented by vectors of constant sum, its sample space, the *simplex*, is $S^D = \{\mathbf{x} \in R_+^D : \sum_{j=1}^{D} x_j = k\}$, where *D* is the number of parts and the value of *k* is irrelevant, and a popular choice is *k*=1.

Nowadays there is a general agreement (Pawlowsky-Glahn and Buccianti, 2011) that the geometry of the simplex is based on log-ratio coordinates. This particular geometry has three basic elements: the operations perturbation, powering, and inner product, that provide an Euclidean structure to the simplex. This allows applying all the multivariate methods to analyse CoDa sets. An important step to use these statistical techniques is to build orthonormal bases in the simplex and to express any composition **x** in its corresponding coordinates, obtained using the *isometric log-ratio* function *ilr*(**x**). A Sequential Binary Partition (SBP) (Pawlowsky-Glahn and Buccianti, 2011, Chapter 2) is an easy and interpretable way to define a function ilr. A SBP of the parts of a composition consists of $D-1$ steps, where an orthonormal coordinate is built in each step of the partition. In a first step, a SBP consists of splitting parts of the composition **x** into two groups, which are indicated by +1 and -1. In consecutive steps, each previously created group of parts is split again into two groups. The partition ends when the groups are made up of a unique part. In the j^{th} step of a SBP, denoting by $\mathbf{x}^+$ the group of *r* parts marked with a +1 and by $\mathbf{x}^-$ the group of *s* parts marked with a -1, the corresponding coordinate, $ilr_j(\mathbf{x})$, is

$$ilr_j(\mathbf{x}) = \sqrt{\frac{r.s}{r+s}}\ln(\frac{g_m(\mathbf{x}^+)}{g_m(\mathbf{x}^-)})$$

where $\mathrm{g}_m(\cdot)$ is the geometrical mean of involved parts of **x**. Let **T** be the table of an AR (Table 1) identified by the composition **x**. Using a SBP, this table **T** can be expressed in terms of ilr-coordinates of **x**

$$ilr(\mathbf{x}) = (\frac{1}{2}\ln(\frac{x_1x_4}{x_2x_3}), \frac{\sqrt{2}}{2}\ln(\frac{x_1}{x_4}), \frac{\sqrt{2}}{2}\ln(\frac{x_2}{x_3})). \tag{1}$$

Despite the basis selected for the ilr-coordinates is not unique, this basis is useful for interpretation purposes and to define CoDa-AR measures of interestingness.

2. *CoDa-AR measures of interestingness*

Measures of interestingness are appropriate indices for measuring the strength of an AR. We present below four measures known as *support*, *confidence*, *lift* and *RLD*:

- support(AR)= x_1= n_1/m, informs of the proportion of transactions that verify the AR.
- confidence(AR) = $n_1/N\{\mathbf{S}_a\}$, where $N\{\mathbf{S}_a\}$ is the number of transactions containing the antecedent. Because confidence $\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$ = support$\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$/support$\{\mathbf{S}_a\}$, it can be interpreted as a conditional probability.
- lift(AR) = confidence$\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$/support$\{\mathbf{S}_c\}$. Following Kenett and Salini (2011), since lift$\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$ = support$\{\mathbf{S}_a \Rightarrow \mathbf{S}_c\}$/(support$\{\mathbf{S}_a\}$ support$\{\mathbf{S}_c\}$), this measure is a deviation under independence of the itemsets. For lift = 1 there is no association between the itemsets. When lift are respectively smaller and greater than 1, the knowledge that $\mathbf{S}_a$ holds causes a negative and positive effect on the probability of $\mathbf{S}_c$.

- RLD(AR): this measure of interestingness, called relative linkage disequilibrium, introduced for AR in Kenett and Salini (2008), assesses the relative distance of the AR from its projection on a surface with lift= 1. It captures the level of dependence of the AR, normalised by geometrical constraints on a simplex representation.

The compositional nature of an AR (Table 1) and the geometrical interpretation of measure RLD suggest to adapt its definiton to the Aitchison geometry (Pawlowsky-Glahn and Buccianti, 2011). According this approach Kenett et al. (2018) define the compositional measure of interestingness (C) as

$$C(AR) = ilr_1(\mathbf{x}) \tag{2}$$

The relation between C (first ilr-coordinate) and the classical odds ratio (OR) measure (Tan et al., 2004) is evident: OR(AR) = odds($\mathbf{S}_c$/$\mathbf{S}_a$)/odds($\mathbf{S}_c$/${}^c\mathbf{S}_a$) = (x_1x_4)/(x_2x_3). It therefore consists of a measure of dependence. In addition, the second ilr-coordinate (Eq. 1) is about the relationship between the estimates of the probabilities P($\mathbf{S}_a \cap \mathbf{S}_c$) and P(${}^c\mathbf{S}_a \cap\ {}^c\mathbf{S}_c$). Whereas the third coordinate represents the relationship between P($\mathbf{S}_a \cap\ {}^c\mathbf{S}_c$) and P(${}^c\mathbf{S}_a \cap \mathbf{S}_b$). The value OR(AR) = 1 indicates independence, OR(AR) $>$ 1 a positive effect and OR(AR) $<$ 1, a negative effect. Note that C(AR)=1/2·ln(OR(AR)) and OR(AR)=e$^{2\cdot \mathrm{C(AR)}}$. This monotonic functional relation indicates that both values have the same ranking. Moreover, when a measure is unbounded, some practical normalization to interval [-1, +1] is advisable. For example, Yule's Q (Tan et al., 2004) defined as OR*(AR)=$\frac{x_1x_4-x_2x_3}{x_1x_4+x_2x_3}$ is the normalized version of the OR. Note that the normalized version of C(AR) holds C*(AR)= tanh(C(AR))= OR*(AR).

On the other hand, The measure RLD for a table **T** (Table 1) measures the similarity between the value x_1 and the product (x_1+x_2)(x_1+x_3) via the subtraction D(AR)= $x_1 - (x_1 + x_2)(x_1 + x_3) = x_1x_4 - x_2x_3$ which measures disequilibrium (D). With no disequilibrium, or independence, we have D(AR)=$x_1x_4 - x_2x_3 = 0$. The comparison D(AR)= 0 can be formulated as

$$\frac{x_1x_4}{x_2x_3} = 1 \Leftrightarrow \ln(\frac{x_1x_4}{x_2x_3}) = 0 \Leftrightarrow \mathrm{C(AR)} = 0.$$

Therefore, using the relationship between C(AR) and D(AR), one can show that

- C(AR) <0 : negative repelling effect between itemsets ($\mathbf{S}_a$ true, $\mathbf{S}_c$ less likely true)
- C(AR) = 0 : independence
- C(AR) >0 : positive attractive effect ($\mathbf{S}_a$ true, $\mathbf{S}_c$ more likely true)

To determine if an AR is statistically different from random noise Kenett et al. (2018) introduce a parametric test based on the significance of an odds-ratio (Haldane, 1995). A 95% confidence interval for an odds-ratio is

$$\left(e^{\ln(OR)-1.96\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}}}, e^{\ln(OR)+1.96\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}}}\right).$$

Using this formula one can define the corresponding test (=0.05) for C(AR). With this approach, ARs where

$$\left|\frac{2\cdot \mathrm{C(AR)}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}}}\right| > 1.96, \tag{3}$$

are considered statistical significant and relevant for the study.

3. CoDa-AR measure applied to Nicardipine database

We focus on side effects of Nicardipine, a medication used to treat high blood pressure and angina that belongs to the dihydropyridine class of calcium channel blockers. The data is based on patient reports in blogs obtained through a website such as https://treato.com/Nicardipine/?a=s. The specific data analyzed consists of 6074 side effects reported by 882 different patients. Our objective is to identify interesting patterns in AR of side effects that state what goes with what. We performed on this data an AR analysis and selected ARs with a minimum support and confidence of 0.1 and 0.4 respectively. This process gives us a total of 62 ARs, each AR with its corresponding contingency table. Among these 62 ARs we selected only the associations that are statistically significant. According to Haldane's test in Eq. 3 only 30 ARs are significant. Using the CoDaPack package (Comas-Cufí and Thió-Henestrosa 2011) we calculated the ilr-coordinates of all the ARs with the basis defined in Eq. 1. Table 2 shows the mean and standard deviation (in parenthesis) of the measures: support, confidence, lift, C and C*. Both statistics are calculated respectively for all the 62 selected ARs and the 30 ARs that are significant by the Haldane test. For the measures support and confidence, the mean values decrease. On the other, the mean values for lift, C and C*, increase. A t-test for comparing the mean of the significant and non-significant ARs is applied to confirm that these variations are significant. If the significance level is the usual 0.05, the p-values in Table 2 suggest that only the variations in lift, C and C* are significant. In other words, both CoDa-AR criteria have no effect on the measures of support and confidence.

Table 2. Mean and standard deviation (in parenthesis) of support, confidence, lift, C and C respectively for all the 62 selected ARs, for the 30 ARs found significant by the Haldane test. The p-value corresponds to a t-test comparing the means of the 62 ARs and the ARs defined by Haldance significance criteria.*

	support	confidence	lift	C	C^*
All	0.135	0.554	1.458	0.446	0.401
n= 62	(0.036)	(0.124)	(0.323)	(0.232)	(0.184)
Haldane	0.134	0.517	1.616	0.587	0.511
n= 30	(0.043)	(0.125)	(0.386)	(0.197)	(0.140)
p-value	0.672	0.138	5.27e-07	1.51e-08	8.28e-09

The measure C(AR) in Eq. (2), that is, the first ilr-coordinate, has a mean of 0.587 and a standard deviation of 0.197, for the Haldane group. When the normalized measure C*(AR) is calculated, these values of mean and standard deviation transform respectively to 0.511 and 0.140, suggesting that, on average, the rules have a medium level of association. Importantly, all ARs have a C(AR) greater than 0, that is, in every AR, a positive effect exists because the product $x_1 \cdot x_4$ is greater than $x_2 \cdot x_3$. Importantly, once the ARs statistically significant and relevant for the study have been detected one can analyze them using all the multivariate techniques for compositional data. For example, one can plot the ARs using a CoDa-biplot (Pawlowsky-Glahn and Buccinati, 2011) where the relations among the ARs and between each AR and the estimates of probabilities (Table 1) can be interpreted.

Acknowledgements: This work has been partially financed by the project "CODA-RETOS" (MCI; Ref: MTM2015-65016-C2-1-R).

References

Agrawal R., Imielienski T., Swami A. (1993) Mining Association Rules between Sets of Items in Large Databases, in: *Proceedings of the Conference on Management of Data*, ACM Press, New York, 207-216.

Aitchison J. (1986) *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press), London, UK.

Comas-Cufí M., Thió-Henestrosa S. (2011) CoDaPack 2.0: a stand-alone, multi-platform compositional software in: Egozcue JJ, Tolosana-Delgado R, Ortego MI (eds) *CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Sant Feliu de Guíxols*. http://www.compositionaldata.com/codapack.php. Accessed 13 March 2018

Haldane J.B.S. (1955) The estimation and significance of the logarithm of a ratio of frequencies, *Ann Hum Genet*, 20, 309-311.

Kenett R.S., Salini S. (2008) Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks, *T Mach Learn and Data Min*, 1, 83-96.

Kenett R.S., Salini S. (2011) Measures of Association Applied to Operational Risks (Chapter 9), in: Kenett, R.S., Raanan, Y. (eds.) *Operational Risk Management*, 149-167. John Wiley & Sons.

Kenett R.S., Martín-Fernández J.A., Thió-Henestrosa S., Vives-Mestres M. (2018) Association rules and compositional data analysis: implications to big data, *Advances in Data Analysis and Classification* (*submitted*).

Pawlowsky-Glahn V., Buccianti A. (eds.) (2011) *Compositional Data Analysis: Theory and Applications*, John Wiley & Sons, Ltd., Chichester, UK.

Tan P.N., Kumar V., Srivastava J. (2004) Selecting the Right Objective Measure for Association Analysis, *Inform Syst*, 29, 293-313.

Scientific Sponsorship

SIS Società Italiana di Statistica

ClaDAG

Classification and Data Analysis (SIS-)Group

Financial Support

- University of Naples Federico II
- Department of Political Sciences, UNINA
- CUBREMOT research project, UNINA

Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali

This volume collects the peer-reviewed contributions presented at the 2nd International Conference on "Advances in Statistical Modelling of Ordinal Data" - ASMOD 2018 - held at the Department of Political Sciences of the University of Naples Federico II (24-26 October 2018). The Conference brought together theoretical and applied statisticians to share the latest studies and developments in the field. In addition to the fundamental topic of latent structure analysis and modelling, the contributions in this volume cover a broad range of topics including measuring dissimilarity, clustering, robustness, CUB models, multivariate models, and permutation tests. The Conference featured six distinguished keynote speakers: Alan Agresti (University of Florida, USA), Brian Francis (Lancaster University, UK), Bettina Gruen (Johannes Kepler University Linz, Austria), Maria Kateri (RWTH Aachen, Germany), Elvezio Ronchetti (University of Geneva, Switzerland), Gerhard Tutz (Ludwig-Maximilians University of Munich, Germany). The volume includes 22 contributions from scholars that were accepted as full papers for inclusion in this edited volume after a blind review process of two anonymous referees.

ISBN: 978-88-6887-042-3
DOI: 10.6093/978-88-6887-042-3
Online ISSN: 2532-4608

9 788868 870423