Generalized mixture models for the analysis of ordinal data*

Maria Iannario, Domenico Piccolo
Department of Political Sciences
University of Naples Federico II (IT)
e-mail: maria.iannario@unina.it; domenico.piccolo@unina.it

Abstract

A generalized mixture model for categorical ordinal data aimed at representing the composite nature of the elicitation mechanism in rating processes is proposed.

Models for ordinal data are based on two main approaches: methods concerning latent variables which are behind the ordered selection and methods which derive from a structured probability distribution. In the first instance a close relationship with the data generating process is required whereas in the second case more stringent fitting aptitudes are pursued. Thus, the main frameworks commonly used consist of cumulative functions and discrete mixtures: proportional odds models (McCullagh, 1980; Agresti, 2010) and `CUB` models (Piccolo, 2003; Iannario and Piccolo, 2012) are the simplest instances of these methodologies.

We look for a comprehensive approach which turns out to include most of the current proposals.

The main idea is that the stochastic mechanism of a discrete choice is a convex combination of attractiveness, satisfaction, awareness, etc. (a component denoted as *feeling*) and indecision, fuzziness, blurriness, etc. (a component denoted as *uncertainty*). In this respect, a distinctive feature of the paradigm is the role of an inherent uncertainty which is present in any human choice and whose inclusion improves both data fitting and model parsimony.

A focal point of the discussion concerns *unsupervised* and *supervised* methods for "cutpoints" (or *thresholds*), that is some real values able to transform the continuous phenomenon (whose existence is real or virtual) into a sequence of categories which should be in one-to-one correspondence with the first m integers.

Then, for a given m, if R_i is the rating of the i-th subject whose response is coded as j, the basic specification of the generalized mixture (GEM) is given by the *stochastic component*:

$$Pr\left(R_i = j \mid \boldsymbol{\theta}\right) = \pi_i \, Pr\left(Y_i = j \mid \boldsymbol{t}_i^{(\Psi)}, \boldsymbol{\Psi}\right) \; + \; (1 - \pi_i) \, Pr\left(V_i = j\right) \; , \tag{1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $\pi_i = \pi(\boldsymbol{t}_i^{(\pi)}, \boldsymbol{\beta}) \in (0, 1]$ and $\boldsymbol{t}_i^{(\pi)}, \boldsymbol{t}_i^{(\Phi)} \in \boldsymbol{T}^{(\pi)}$ are subsets of the subjects' information matrix $\boldsymbol{T}$. In addition, V_i is the uncertainty random variable.

Subsequently, a *systematic component* allows to relate subjects' characteristics and parameters by means of a convenient link.

Since an increase of π_i implies a reduced impact of the uncertainty component, the quantity $(1 - \pi_i)$ may be considered as a measure of the uncertainty implied by the model. Notice that

*A preliminary version of this research has been presented at the FIRB2012 project meeting held in Rome, Italy, January 23-24, 2015

$Pr(V_i = j)$ is assumed known on *a priori* basis: quite often, a discrete Uniform random variable is an adequate option.

The interpretation of (1) is twofold:

- according to the logic of latent class models, respondents split into two clusters in proportions (π_i) and $(1 - \pi_i)$, respectively, and one class consists of people whose selection is completely random;
- each respondent has a *propensity* to select an ordinal category with a meditated or a random choice and this happens with weights (π_i) and $(1 - \pi_i)$, respectively.

We strongly support the second interpretation although the first one is useful for estimation and simulation purposes of the approach.

Current models introduced for ordinal data may fit the paradigm implied by (1) and many others might be proposed by specifications of the *stochastic* and *systematic component*. This perspective has been recently adopted by Tutz et al. (2014), who generalized the cumulative models to take account of the uncertainty component.

A case study will support this aspect and synthesize the main results.

A generalized mixture with uncertainty is a useful framework to compare models, to discover unexpected similarities and to introduce new distributions. More specifically, the opportunity to inspect the same data from different points of view is an added value for the statistical analysis of ordinal data.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley, New York, 2^{nd} edition.

Iannario, M. and Piccolo, D. (2012). CUB models: Statistical methods and empirical evidence. In Kenett, R. S. and Salini, S., editors, *Modern Analysis of Customer Surveys*, pages 231–258. Wiley, New York.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42:109–142.

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

Tutz, G., Schneider, M., Iannario, M., and Piccolo, D. (2014). Mixture models for ordinal responses to account for uncertainty of choice. Technical Report 175, University of Munich, Department of Statistics.