Quaderni di Statistica
Vol. 10, 2008

A statistical procedure for clustering ordinal data

Marcella Corduas
Dipartimento di Scienze Statistiche,Università degli Studi di Napoli Federico II
E-mail: corduas@unina.it

Summary: This article presents a testing procedure for comparing ordinal data distributions which helps the interpretation of results in presence of complex surveys involving the evaluation of several items or the evaluations expressed by different groups of respondents about a given item. For this purpose, a mixture model (denoted CUB) for ordinal data is considered. Specifically, Kullback-Liebler divergence is used in order to measure the dissimilarity among ratings distributions and a clustering technique is applied for grouping distributions. A case study on university teaching evaluation is finally illustrated.

Keywords: Ordinal Data, CUB Models, Kullback-Liebler Divergence.

1. Introduction

Since 2000 Italian universities have implemented own procedures for evaluating teaching activities by means of extensive surveys concerning students' opinions on attended courses. This fact originated a consistent number of studies discussing the principles ruling an efficient evaluation system (Biggeri, 2000; Biggeri and Bini, 2001) and developing statistical techniques for the assessment of teaching quality and the evaluation of educational processes.

As often happens in complex surveys, the questionnaire involves several aspects of teaching activity and, moreover, interviewees are clustered in different groups according to various features (for instance, faculties or degree courses). For this reason, a statistical tool for the comparison

among rating distributions related to a number of items or to different bunch of respondents is needed in order to provide meaningful information to decision makers.

In this respect, a mixture distribution, recently introduced by D'Elia and Piccolo (2005a), provides a useful probabilistic model to describe ordinal data. In this article, we propose to detect significant similarities and differences between raters' overall judgements by comparing the estimated CUB models.

The article is organized as follows. In Section 2, the CUB distribution is briefly introduced. In Section 3, a testing procedure based on Kullback-Liebler (KL) divergence is discussed and a clustering technique is presented. Finally, a case study concerning a survey carried out at the University of Naples Federico II on teaching quality is illustrated.

2. The mixture distribution

According to D'Elia and Piccolo (2005a) the mechanism leading to the formulation of subject's judgement about a given item can be summarized by means of two random components *uncertainty* and *selectiveness*. The first one describes the fact that the rater, who is requested to assign a score on a discrete scale to a certain item, tends to hesitate before providing the answer since he/she has to force his/her mental construct about liking/disliking into a numerical value. The second one, instead, is related to the profound belief of the judge concerning the item object of evaluation.

The final rating is then described by means of the random variable R such that:

$$P(R=r) = \pi \left(\begin{array}{c} m-1 \\ r-1 \end{array} \right) (1-\xi)^{r-1}\xi^{m-r} + (1-\pi)\frac{1}{m}, \qquad r = 1, 2, ..., m \tag{1}$$

where $\xi \in [0,1]$, $\pi \in (0,1]$ and m is the number of grades of the evaluation scale. For given $m > 3$, then, R is a well defined Mixture of a *Uniform* and a (shifted) *Binomial* distribution. The *uncertainty* is repre-

sented by means of a Uniform distribution which assigns to each possible score the same probability. The parameter π determines the importance of uncertainty in the final judgment: the lower the weight $(1 - \pi)$ the smaller the contribution of the Uniform distribution in the mixture.

The parameter ξ, instead, characterizes the shifted Binomial distribution which may assume different shapes (symmetric, asymmetric, peaked or flat to a certain extent). This represents the subject's *selectivity*. Depending on the meaning of the highest score (positive or negative judgment), this parameter denotes the strength of 'liking' (or 'disliking') that the rater feels for the item.

In this sense, the model (1) is very flexible because it provides a continuum of alternative theoretical distributions to represent observed ratings: from the very extreme case of a completely random choice of the score (the Uniform distribution is dominant and the shifted Binomial is annihilated) to a very accurate and conscious choice (the shifted Binomial distribution is dominant and the Uniform is annihilated). Moreover, the model is further extended in order to take the influence of external factors into account (Piccolo, 2006; Piccolo and D'Elia, 2008). In particular, two relations, which connect the model parameters to significant *covariates* by means of a logistic link function, are added to (1). This final form fully justifies the acronym CUB given by the proposing authors. In the following section, this acronym will be used despite explanatory variables are not considered.

3. Comparing and clustering CUB models

First, we briefly recall some basic properties of KL divergence; in this respect, Pardo (2005) provides an extensive review of divergence measures and related generalizations, for studying different statistical problems.

In general, KL divergence provides a measure of dissimilarity between two probability distributions $f_1(x, \theta_1)$ and $f_2(x, \theta_2)$ characterizing a random variable X under two different hypotheses, respectively (Kullback, 1959).

Specifically, the KL divergence is defined as:

$$J(f_1, f_2) = I(f_1, f_2) + I(f_2, f_1), \tag{2}$$

where, assuming the case of a continuous random variable:

$$I(f_1, f_2) = \int_{-\infty}^{\infty} f_1(x, \theta_1) \ln \frac{f_1(x, \theta_1)}{f_2(x, \theta_2)} dx = E_1 \left(\ln \frac{f_1(x, \theta_1)}{f_2(x, \theta_2)} \right) \tag{3}$$

is the mean information, with respect to f_1, for discrimination in favor of the first hypothesis against the second one. The other term in (2), $I(f_2, f_1)$ is similarly defined. Of course, the case of a discrete random variable can be easily introduced by extending (3) accordingly.

Note that the KL divergence is a symmetric and almost positive definite, but it is not a metric since it doesn't satisfy the triangular inequality. However, due to its statistical properties, it represents a very interesting tool for establishing the comparison of CUB models as a problem of hypothesis testing.

For this aim, we illustrate a general result derived from Kupperman (1957). Consider two discrete populations each characterized by a probability distribution function having the same functional form $p(x, \boldsymbol{\theta}_i)$ with unspecified vector parameters $\boldsymbol{\theta}_i$, $i = 1, 2$. Also assume that, on the random variable support, $p(x, \boldsymbol{\theta}_i) > 0$. Suppose that we have two independent samples of N_1 and N_2 observations randomly drawn from the specified i-th population and we wish to decide if they were in fact generated from the same population. In order to test the hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ against $H_0 : \theta_{1j} \neq \theta_{2j}$, the KL divergence statistic is defined:

$$\hat{J} = \frac{N_1 N_2}{N_1 + N_2} \left[\sum_x (p(x, \boldsymbol{\theta}_1) - p(x, \boldsymbol{\theta}_2)) \ln \frac{p(x, \boldsymbol{\theta}_1)}{p(x, \boldsymbol{\theta}_2)} \right]_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2} \tag{4}$$

where the vector parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have been replaced by the maximum likelihood estimators. Then it can be shown that $\hat{J}$ is asymptotically distributed as a χ^2_g random variable when the null hypothesis is true, being g the common dimension of the vector parameter (Kullback, 1959). In the case under investigation, the probability distributions which

are object of comparison are CUB distributions, each characterized by two parameters as in (1), $\boldsymbol{\theta} = (\pi, \xi)'$, then $g = 2$.

In a previous contribution (Corduas, 2008), a strategy for clustering estimated CUB models has been proposed. We assume that observed rating distributions have been originated by opinions of a group of respondents about k items (or by k groups of respondents about a given item). The procedure consists of the following steps:

- a CUB model is fitted to each observed ratings distribution;
- the (k, k) matrix of KL divergences among fitted CUB models is evaluated;
- the test of hypotheses described in (4) is performed for each couple of CUB models. The results of testing is summarized into a binary matrix where the (i, j) element is 1 if the homogeneity hypothesis is not rejected and 0 otherwise;
- the Bond Energy Algorithm (namely BEA; McCormick *et al.*, 1972) is applied in order to rearrange rows and columns of the binary matrix into a diagonal block form.

Clusters are then identified by inspecting the rearranged matrix to find well separated unit diagonal blocks, and elongated clusters are recognized as blocks containing very few zero values. The method is quite flexible since it gives a visual representation of all data at the same time and it does not impose any general rule for clustering detection.

In the following section, by means of the analysis of an empirical data set, we will illustrate how hierarchical clustering techniques can provide a possible alternative to the above mentioned approach. The selection of clusters is performed by means of a widely applied method, the complete linkage method and using the selected critical value in the hypothesis testing as threshold for sectioning the dendrogram derived from the divergence matrix.

Complete linkage method imposes a fixed hierarchical rule in order to decide when to create a new cluster and is optimal for specific cluster shapes since it merges groups on the base of the farthest neighbour criterion.

Having imposed a fixed hierarchical rule reduces the flexibility of the approach with respect to the use of BEA algorithm where the user is left the possibility of looking at all data relationships (Tran-Luu and De Claris, 1997). However, using the significance value as threshold for sectioning the dendrogram leads to a meaningful data representation. As a matter of fact, a complete linkage method meets a more conservative strategy, in the sense that only elements whose reciprocal divergences are not significant belong to the same identified cluster. Of course, the two procedures provide the same final clusters whenever these clusters are well separated.

4. A case study

Numerous contributions in literature have analyzed data derived from periodic surveys carried out in Italian universities. Among others, we mention the work by Petrucci and Rampichini (2000), Rampichini *et al.* (2000), Capursi and Porcu (2001), Bernardi *et al.*(2004), Chiandotto *et al.* (2006), Fabbris (2007), Iannario and Piccolo (2008), Mignani and Cagnone (2008).

In this section, the proposed technique is applied in order to describe the results from the survey on students' opinions about teaching quality at the University of Naples Federico II in the 2005-06 academic year. According to the CNVSU (2002) guidelines, the questionnaire is aimed at assessing the students' opinions about various elements which characterize teaching activity: 1) quality of lecture halls and teaching equipments; 2) several features of the specific course the interviewees are attending; 3) teacher's ability.

The data set, object of this analysis, was gathered from the 13 Faculties [1] of the University Federico II and refers only to valid records, that is

[1] The Faculties are: Medicine, Veterinary medicine, Pharmacy, Agricultural Science, Biotechnology, Engineering, Architecture, Mathematics and Natural Science, Classics and Modern studies, Law, Economics, Political Sciences, Sociology. Statistical analysis at a lower level of aggregation (such as curricula) has not been carried out. As a matter of fact, the use of the database was restricted by the University of Naples Federico II which also requested that no identifier establishing the identity of the Faculties could be included in any publications.

34,507 students who completed the questionnaire. The students' ratings are expressed using a 7 point Likert scale where 7 relates to the highest positive judgement.

Regarding this aspect, the experience at the University of Naples is, in some ways, special since ratings are expressed by means of a 7 point scale and, so CUB models can be fruitfully applied for data modelling purposes.

Other universities have experienced different ordinal scales such as 4 point scales in order to force interviewees to express their positive/negative opinions (Chiandotto and Bertaccini, 2008).

In the rest of this section, our attention will be focussed on the ratings that the students express about the *overall quality* of courses and teachers' ability and a comparison between Faculties will be performed. The ratings expressed by students in each Faculty will be considered as independent samples. Faculties are denoted by numbers from (1) to (13).

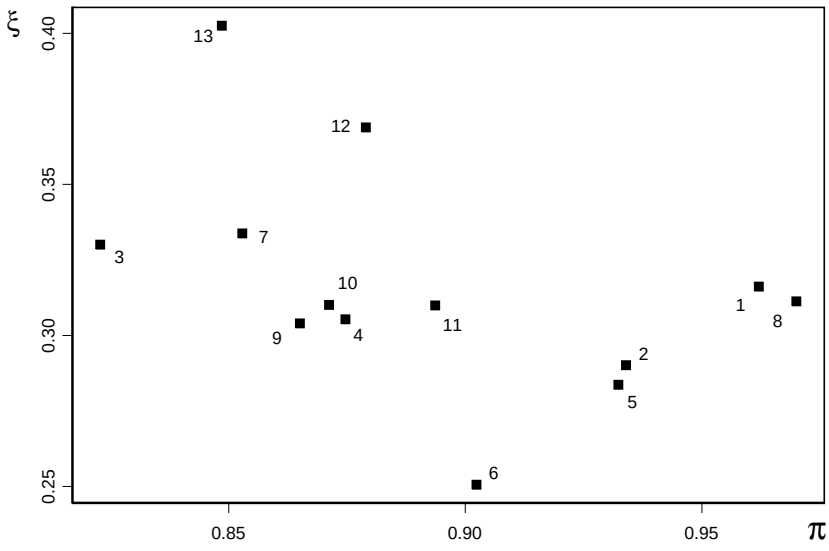

Figure 1. *Overall quality of courses: estimated CUB models*

The values of CUB estimated coefficients range in the lower right part of the unit square representing the parameter space (Figure 1): specifically, $\widehat{\xi} \in (0.25, 0.40)$ and $\widehat{\pi} \in (0.84, 0.98)$. This implies that CUB distributions are skewed to the left and characterized by rather low uncertainty.

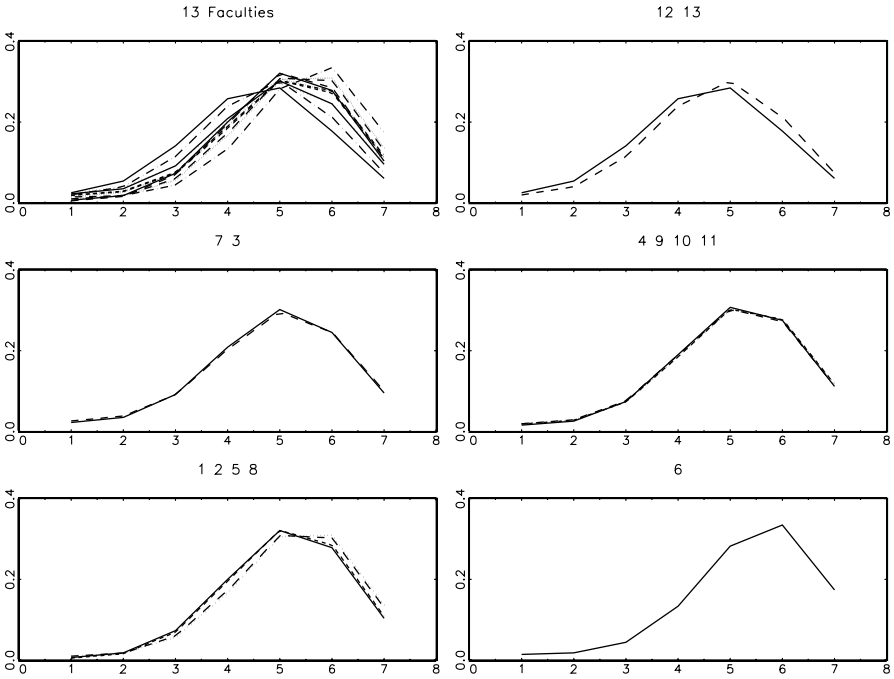

Figure 2. *Overall quality of courses: clustering by BEA algorithm*
panels by rows: 13 Faculties; clusters: (12,13); (7,3); (4,9,10,11); ((2,5),(1,8)); (6)

In Figure 2, the clustered CUB distributions, produced by means of BEA algorithm, are illustrated.

Note that the hypothesis testing is performed using 5% significance level. Only, (12) and (13) are merged reducing this level to 1%; the same consideration applies to the further merging of (2,5) with (1,8). The graphs highlights the fact that Faculties generally achieve satisfying performances in terms of overall quality. However, the divergence is able to discern differences among the Faculties. The elements which are merged

together show some common features: (7,3) share a strong technological background; (4,9,10,11) correspond to medium size Faculties; (1,2,5,8) have very specialized curricula.

Finally, students belonging to the Faculties (12,13) are characterized by a strong vocation for their future profession and generally are more demanding. This consideration probably justifies the lower ratings that they assign to overall teaching quality.

In Figure 3, the clustered estimated CUB models obtained by means of the complete linkage method from the divergence matrix is presented. The dendrogram has been truncated in order to provide a better view of the lower branches. The method recognizes the same clusters as the BEA algorithm; also, further merging obtained by decreasing the significance level are correctly identified.

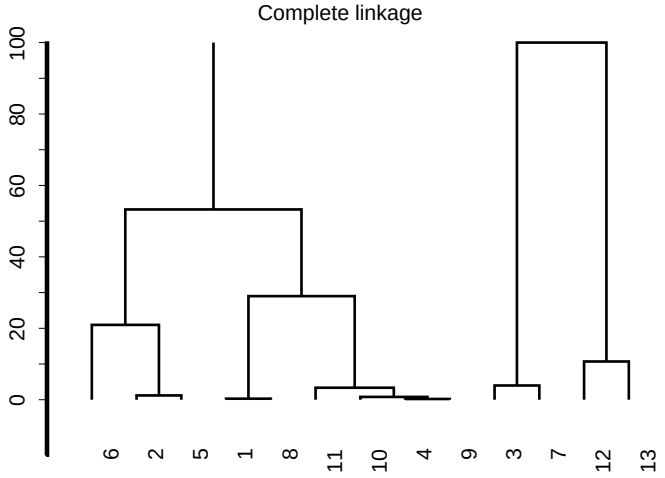

Figure 3. *Overall quality of courses: complete linkage method*

At this stage of the analysis it is worth noting that, apart from (12), (8) and (16) which consist of less than 800 students, the number of observations for each Faculty is very high, ranging from 1533 students of (10) to 9492 of (3). This fact makes the hypothesis testing very selec-

tive, and, therefore, depending on the objective of the analysis, a higher threshold may be needed. Moreover, the graph in Figure 2 (see first panel) would suggest that the overall teaching quality is judged by students in a very similar way in all Faculties. Instead, given the high number of respondents, the proposed procedure has revealed that there are differences among ratings that students express.

The method suggested here produces a classification which relies on an inferential approach. For this reason the results are more meaningful with respect to the traditional descriptive approach. The use of the test critical value for the identification of clusters allows for the pairwise test of homogeneity for the rating distribution. According to the complete linkage method, each time an element is joined to a given cluster, this implies that the related rating distribution is not significantly different from those already included in the group. The comparison is done pairwise with the existing merged elements. In the case under investigation, there are Faculties where the overall judgments expressed by students are similar. Then, those Faculties should be assigned the same rank.

Note, also, that in the situation under study the power of the statistical test is rather high. This fact justifies the evident selectiveness of the clustering technique.

Furthermore, in the presence of stable relationships among identified clusters, the single linkage method could be used, as a term of comparison, in order to find further elements that merge with existing clusters producing elongated shapes or elements that act as bridges between groups.

Table 1. Average rating of overall quality

Faculty	13	12	3	7	10	9	11	4	1	8	2	5	6
Average	4.50	4.69	4.84	4.85	4.99	5.02	5.02	5.02	5.06	5.10	5.18	5.21	5.35

Finally, Table 1 shows the Faculties ranked according to their average score. The ranking confirms the previous findings. Note that all rating distributions of overall teaching quality in the examined Faculties are skewed negatively, so the average still provides a meaningful ranking.

5. Final remarks

CUB models may describe ordinal data by means of a parsimonious representation and provide a useful starting point for comparing ratings expressed by different groups of respondents. This type of models has been successfully applied to data from various research areas, such as, medicine (D'Elia, 2007), food analysis (Piccolo and D'Elia, 2008), social analysis (D'Elia and Piccolo, 2005b; Iannario, 2007), linguistics (Balirano and Corduas, 2008). Moreover, it has proved to be flexible enough to summarize distributions with very different shapes.

In this article, we exploited this feature of CUB distribution for modelling the ratings expressed by students on teaching quality. Also, an inferential approach based on Kullback-Liebler divergence has been proposed for clustering ordinal data. The results are very encouraging since the technique is able to distinguish among distributions which apparently are quite similar in terms of statistical indices such as mode or average.

Acknowledgements: The author thanks the University of Naples Federico II, and especially the Nucleo di Valutazione di Ateneo and UPSV for kindly providing the data set which has been analyzed in this article. This work is part of PRIN 2006-08 project "Stima e verifica di modelli statistici per l'analisi della soddisfazione degli studenti universitari" and has benefited from support by CFEPSR (Portici, Italy).

References

Balirano G., Corduas M.(2008), Detecting semiotically expressed humor in diasporic TV productions, *Humor: international journal of humor research*, 3, 227-251.

Bernardi L., Capursi V., Librizzi L. (2004), Measurement awareness: the use of indicators between expectations and opportunities, *Atti della XLII Riunione Scientifica SIS*, Padova: CLEUP, 315-326.

Biggeri, L. (2000), Valutazione: idee, esperienze, problemi. Una sfida per gli statistici, *Atti della XL Riunione Scientifica SIS*, Firenze: CS2p, 31-48.

Biggeri L., Bini M. (2001), Evaluation at University and State level in Italy:

need for a system of evaluation and indicators, *Tertiary education and management*, 7, 149-162.

Capursi V., Porcu M. (2001), La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi, *Atti Convegno Intermedio della SIS "Processi e Metodi Statistici di Valutazione"*, Roma.

Chiandotto B., Bertaccini B., Bini M. (2007), Evaluating the quality of the university educational process: an application of the ECSI model., in Fabris L. (2006), Heidelberg: Springer- Verlag.

Chiandotto B., Bertaccini B. (2008), Sis-valdidat: a statistical information system for evaluating university teaching, *Quaderni di Statistica*, this issue.

CNVSU (2002), *Proposta di un insieme minimo di domande per la valutazione della didattica da parte degli studenti frequentanti*, 9/02, www.cnsvu.it.

Corduas M. (2008), A testing procedure for clustering ordinal data by CUB models, *Proceedings of Joint SFC-CLADAG 2008 meeting*, ESI, Napoli, 245-248.

D'Elia A. (2007), A statistical modelling approach for the analysis of TMD chronic pain data, *Statistical Methods in Medical Research*, 1-15.

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

D'Elia A., Piccolo D. (2005b), Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico, *Quaderni di Statistica*, 7, 121-161.

Fabbris L. (ed.), (2006), *Effectiveness of University Education in Italy: Employability, Competences, Human Capital*, Heidelberg: Springer-Verlag

Iannario M. (2007), A statistical approach for modelling Urban Audit Perception Surveys, *Quaderni di Statistica*, 9, 149-172.

Iannario M., Piccolo D. (2008), University teaching and students' perception: models and evidences of the evaluation process, *Proceedings of DIVAGO meeting*, University of Palermo, 10-12 July 2008.

Kullback S. (1957), *Information theory and statistics*. New York: Dover publ.

Kupperman M. (1957), *Further applications of information theory to multivariate analysis and statistical inference*, George Washington University.

McCormick W.T., Schweitzer P.J., White T.W. (1972), Problem decomposition and data reorganization by a clustering technique, *Operation Research*, 20, 993-1009.

Mignani S., Cagnone S. (2008), University formative process: quality of teaching versus performance indicators, *Quaderni di Statistica*, this issue.

Pardo L. (2005), *Statistical inference based on divergence measures*, Boca Raton: Chapman & Hall-CRC.

Petrucci A., Rampichini C. (2000), Indicatori statistici per la valutazione della didattica universitaria, in Civardi M. and Fabbris L. *Valutazione della didattica con sistemi computer-assisted*, Padova: Cleup.

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

Piccolo D., D'Elia A. (2008), A new approach for modelling consumers' preferences, *Food Preference and Quality*, 247-259.

Rampichini C., Grilli L., Petrucci A. (2000), Analisi della qualita' della didattica universitaria attraverso modelli multi-livello, in Civardi M. and Fabbris L.: *Valutazione della didattica con sistemi computer-assisted*, Padova: Cleup.

Tran-Luu T.D., DeClaris N. (1997), Visual heuristics for data clustering, *IEEE Transactions on Systems, Man and Cybernetics*, 1, 19-24.