A finite mixture distribution for modelling overdispersion

Maria Iannario

Abstract This paper describes a strategy for modelling overdispersion in survey data concerning marginal ranking variables. We assume that ordinal data have been generated by a mixture distribution, defined as CUBE model, where a specific parameter explains a possible presence of overdispersion. Asymptotic likelihood methods will be applied for efficient statistical inference. The method is primarily motivated and then implemented by a sample on the emergencies in a metropolitan area. The proposed model improves both fitting and interpretation for the social and urban problems involved.

Key words: Overdispersion, CUBE models, Beta Binomial distribution, Discrete Uniform distribution

1 Introduction

Overdispersion is fairly common in categorical data analysis and it is generally associated with Poisson, multinomial and other exponential family models. McCullagh and Nelder [13] remarked that it is generally convenient to assume that overdispersion is present to some extent unless the data or prior information suggest otherwise.

Single-parameter distributions, in which the variance is a function of the mean value, often do not fit well since observed data require more involved parameterizations. In fact, the residual variation, beyond that predicted by the mean, is so large that no model but the saturated one appears to fit the data [6]; [3].

Generally, overdispersion can be due to design effects, hidden clusters, interviewer effects, number of modalities for each response or, more generally, to the absence of relevant predictors in the model which do not present correlation with

Maria Iannario
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22, Naples, Italy, e-mail: maria.iannario@unina.it

the included ones. Whatever be the underlying cause, overdispersion can be represented either by a positive correlation between the responses or by variation in the response probabilities.

There is an extensive literature that considers how to adjust the variance of an estimator to account for the sampling design [1]. It assumes a general form for the variance function by including additional parameters or implementing a two-stage model for the response in which the basic response model parameter has some distribution [5].

Thus, we focus on general problems where overdispersion might arise because of the absence of relevant predictors that could account for the additional source of random variability, for reasons related to scale usage heterogeneity [4]; [17] or for a subjective interpretation of the wording of modalities [2].

When data derive from rating or ranking responses it is possible to assume a finite mixture distribution [9] since this random variable more reasonably may emulate the generating process of respondents when respondents select a score or a position.

This class of models allows to put ordinal data in a more general framework [9]. In fact, the approach we will adopt is quite different from the standard one since the overdispersion is ascribed to a variability among the personal feeling of each respondent [12]. As a consequence, this extra-variability is not necessary generated by a difference among clusters.

The paper is organized as follows. Section 2 describes a new mixture model for the overdispersion in ordinal data. Some inferential issues were reported in Section 3. In Section 4 we summarise empirical results for diagnostic assessment of overdispersion models and compare them with a standard framework. A discussion concludes the paper.

2 Overdispersion in a finite mixture model

Our framework interprets an ordinal response as a random variable R generated by the combination of a *feeling* component towards the item and an *uncertainty* caused by the modality to submit the survey (CUB model: [15, 11, 8]). If we further assume that each respondent changes the way by which he/she chooses the category (among m given ordinal alternatives), then the feeling component may be conveniently modelled as a Beta Binomial random variable. This approach implies that the parameter p of the Binomial component has *a priori* distribution over $0 < p < 1$. This extra assumption solves in a very flexible distribution with respect to location, variability and shape. Notice that a CUBE model solves in a CUB model if overdispersion is absent. In this way, we are introducing a more general class which encompasses the previous one.

Formally, a CUBE (Combination of *U*niform and *Be*ta Binomial) random variable has the following probability mass function:

$$Pr(R = r) = \pi \beta e_r(\xi, \phi) + (1 - \pi) U_r, \quad r = 1, 2, \ldots, m, \tag{1}$$

where $\beta e_r(\xi,\phi)$ has been specified by:

$$\beta e_r(\xi,\phi) = \binom{m-1}{r-1} \frac{\prod_{k=1}^{r}[1-\xi+\phi(k-1)] \prod_{k=1}^{m-r+1}[\xi+\phi(k-1)]}{[1-\xi+\phi(r-1)]\,[\xi+\phi(m-r)] \prod_{k=1}^{m-1}[1+\phi(k-1)]},$$

whereas $U_r = 1/m, \forall r$, is the discrete Uniform random variable. The parameter vector $\boldsymbol{\theta} = (\pi,\xi,\phi)'$ belongs to the parameter space

$$\Omega(\boldsymbol{\theta}) = \{(\pi,\xi,\phi):\ 0<\pi\leq 1;\ 0<\xi<1; 0\leq\phi<\infty\}.$$

The variance of a CUBE distribution increases with ϕ (and π) by a quantity which, *ceteris paribus*, is maximized when $\xi = 1/2$. Otherwise, when $\xi \to 0$ or $\xi \to 1$ the overdispersion effect tends to disappear [10].

A CUB model ($\phi = 0$) is nested into a CUBE model ($\phi > 0$). Thus, we could exploit the unit square for representing both models; CUB models are shown as points in correspondence with (π,ξ) whereas CUBE models maybe represented by means of points whose size is proportional to ϕ.

A CUBE model allows for the inclusion of covariates for each latent component (feeling, uncertainty, overdispersion). For this purpose, we assume a $\boldsymbol{T}$ matrix of observed k covariates and a logistic link among the parameters and the subject covariates [16]. For the overdispersion parameter, for instance, we have:

$$\phi_i = 1/[1+exp(-t_i\alpha)]; \qquad i = 1,2,\ldots,n.$$

3 Inferential issues

For a sample of ordinal data $\boldsymbol{r} = (r_1,r_2,\ldots,r_n)'$, the log-likelihood function is:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{\pi\left[\beta e_{r_i}(\xi,\phi) - \frac{1}{m}\right] + \frac{1}{m}\right\}, \tag{2}$$

and all inferential issues are derived by the maximum likelihood approach. The estimation process relies on the EM algorithm which is characterized by an almost sure convergence process [14].

Local and global validation measures are necessary for cheeking the usefulness of the estimated model. According to [3] and [18], it is possible to discover several other approaches to detect the overdispersion; some of them could be implemented for this mixture allowing for a better analysis of the validation step.

For measuring global fitting we usually implement a normalized fitting measure $\mathscr{I}$ [7]. It is obtained by comparing the estimated *saturated* log-likelihood of a CUBE model (ℓ_{sat}) and the log-likelihood of an uninformative model (ℓ_0). In addition, we

also compute the index $\mathscr{F}^2$ which compares observed relative frequencies f_r and expected probabilities $p_r(\hat{\boldsymbol{\theta}})$:

$$\mathscr{I} = \frac{\ell(\hat{\boldsymbol{\theta}}) - \ell_0}{\ell_{sat} - \ell_0}; \qquad \mathscr{F}^2 = 1 - \frac{1}{2}\sum_{r=1}^{m} \mid f_r - p_r(\hat{\boldsymbol{\theta}}) \mid .$$

Finally, we consider the Bayesian information criterion (BIC) as a further index for comparing models.

4 Some empirical evidence

We will check the usefulness of the proposed mixture on a real case study.

During several years, a sample has been selected in order to comparatively rank the main $m = 9$ emergencies perceived by people living in the metropolitan area of Naples. They are listed as: *CLI=political patronage and corruption; CRI=organized crime; DIS=unemployment; INQ=envi-ronmental pollution; MAL=public health shortcomings; MIC=petty crimes; IMM= immigration; PUL= streets cleanness and waste disposal; TRA=traffic and local transport.*

Since the vector of ranks expressed by a single respondent is a collection of non-independent observations, we consider the marginal distribution of a given emergency as ordinal variable to be modelled by CUBE random variables. In fact, the marginal ranking is a rated evaluation of the personal worry: a low ranking position of an item implies a serious worry about it, and vice versa.

The *feeling* parameter ξ measures the perceived risk for each emergency whereas the uncertainty in the responses is summarised by $(1-\pi)$ parameter. Finally, the parameter ϕ expresses the level of overdispersion of the responses.

In the following, we first analyse the data for the waves 2004, 2007 and then we will check the time stability of the responses during the waves 2004-2007.

4.1 Overdispersion for the wave 2004

We first consider all the items for the Emergency data set of the 2004 wave. Thus, we would consider the effect of overdispersion parameter on the displacement of both uncertainty and feeling. Results may be assessed in a single picture by plotting in the parameter space (π, ξ) all the estimated CUB and CUBE models and, then, by denoting the shift induced on these parameters when the estimated ϕ is significant. The size of the points is proportional to the value of $\hat{\phi}$.

Figure 1 shows that a CUBE model is better than a CUB one for 6 out of 9 items. In fact, the representation supports significant overdispersion for those emergencies which are located at the extreme of the worry scale; on the contrary, problems located as intermediate (*CLI, MAL, PUL*) are well fitted by CUB models. Specifically,

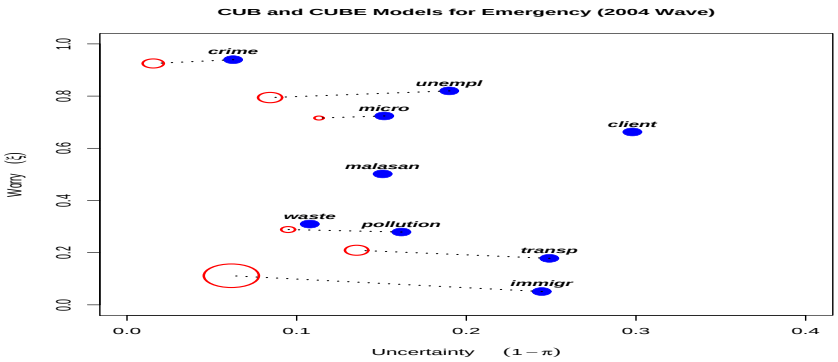

Fig. 1 Visualization of CUB (bullet) and CUBE models (circle) for Emergency (2004 wave). The amount of overdispersion is related to the dimension of circle. On the left is possible to notice the reduction of uncertainty in the same responses for the additional introduction of the parameter which captures overdispersion.

"Immigration" (which is registered as the last in the ranking) requires an overdispersion parameter as large as $\phi = 0.251$. A close examination of this data discloses more carefully the significant improvement achieved by the introduction of the parameter ϕ (the Pearson fitting measure X^2 reduces from 45.583 to 6.3061).

This empirical analysis confirms the role of ϕ in our modelling framework. When people have a definite opinion (as it mostly manifests itself when emergencies are considered as extreme in the scale) the selection of a modality happens with constant probability since it is oriented with high confidence towards the largest or the smallest categories of the scale. On the contrary, when people rank a problem as intermediate the assessment of probability of selecting a given category is not so resolute.

For this data set, the added ϕ parameter improves the flexibility of the maintained distribution although it is not strictly related to a real "overdispersion" of the observed distributions. In fact, the variance of the observed rankings is regularly smaller than the average; thus, any standard test should reject the presence of overdispersion.

4.2 Overdispersion for the wave 2007

In this subsection we perform another aspect of overdispersion related to the design effect. Specifically, the sample size of this survey has been increased during the waves: from $n = 354$ (wave 2004) and $n = 419$ (wave 2006) up to $n = 2381$ (wave 2007). This circumstance suggests an high variability among responses of the last wave and, in fact, we found the significance of the extra-parameter ϕ in all the perceived emergencies, with the exception of *PUL*, as reported in Table 1: the first and

second rows are referred to estimated CUB and CUBE models, respectively. Notice that, when significant, the estimated ϕ range in $[0.13, 0.38]$; its role is relevant in lowering an undue uncertainty (expressed by a smaller π) but also in partially modifying the feeling parameter.

Generally, the reduction of uncertainty due to overdispersion underlines the role of heterogeneity for the standard framework and the ability of the Beta Binomial component to capture such a variability.

Table 1 Estimated CUB and CUBE models for the wave 2007

Emergencies	$\hat{\pi}$	$\hat{\xi}$	$\hat{\phi}$	*LRT*	*BIC*	$\mathscr{F}^2$	$\mathscr{I}$
CLI	0.350	0.679			10236.1	0.954	0.892
CLI	0.538	0.645	0.132	27.452	10216.4	0.988	0.993
CRI	0.689	0.918			8174.1	0.895	0.932
CRI	0.849	0.863	0.193	166.958	8014.9	0.990	0.999
DIS	0.525	0.809			9796.1	0.859	0.737
DIS	0.910	0.708	0.256	229.615	9574.3	0.968	0.985
INQ	0.411	0.348			10169.2	0.930	0.822
INQ	0.730	0.383	0.173	62.119	10114.8	0.983	0.987
MAL	0.399	0.544			10206.4	0.958	0.898
MAL	0.694	0.533	0.149	28.151	10186.1	0.985	0.991
MIC	0.517	0.707			9937.7	0.902	0.816
MIC	0.781	0.651	0.135	75.707	9869.7	0.947	0.930
IMM	0.498	0.087			9312.8	0.845	0.774
IMM	0.945	0.241	0.380	315.943	9004.7	0.959	0.984
PUL	0.480	0.279			9955.2	0.954	0.945
PUL	0.505	0.286	0.017	1.435	9961.6	0.955	0.948
TRA	0.702	0.124			8651.5	0.895	0.926
TRA	0.849	0.171	0.145	143.782	8515.5	0.986	0.999

Table 1 confirms that in all cases, except for *PUL*, CUBE models improve the fitting measures by a significant amount.

4.3 Time stability of marginal rankings

We will check if the expressed worry is stable during the waves of 2004, 2006 and 2007 by considering models without and with an overdispersion parameter. Specifically, we do not linger on the details of the several estimated models because in a

large majority of cases CUBE models are preferable to a simpler CUB model. Thus, we summarize the main results in a graphical way.

In this regard, if we estimate CUB models to the marginal ranking of the emergencies (Figure 2, left panel), we register a strong constancy over time of the ranked items (a modest exception is *TRA*) according to a general pattern.

On the contrary, if we estimate CUBE models and we represent them in the same parameter space, we get a different picture (Figure 2, right panel). It seems that most of the uncertainty shown in these waves is indeed the consequence of an ubiquitous overdispersion, as already discussed in the previous subsection.

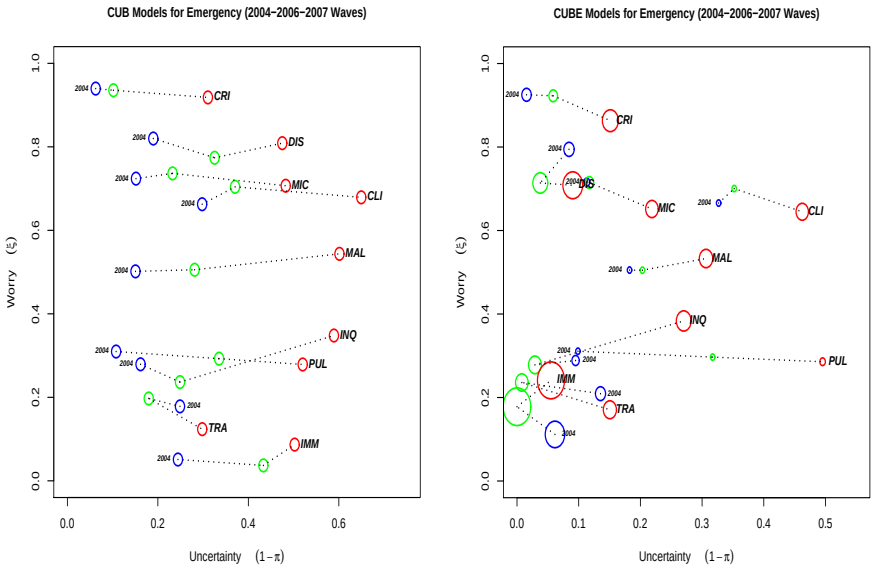

Fig. 2 CUB and CUBE models for all Emergencies and waves: 2004 (blue), 2006 (green) and 2007 (red) waves. There is a regular increase of the uncertainty component of the responses even if the amount of worry does not change in a significant measure (left panel). Notice the different patterns of CUBE models -circle size is proportional to $\hat{\phi}$ (right panel).

5 Discussion

In this paper we described a strategy for adjusting for overdispersion in surveys concerning ordinal data. We motivated a finite mixture distribution where the Beta Binomial replaces the shifted Binomial introduced in the common CUB framework. The maximum likelihood approach with common fitting measures has been reported to underline the improvement of the additional parameter ϕ as in the above men-

tioned case study. It captures the extra-variability motivated not only by sampling design or probability of responses but also by the structure of the scale.

Generalizations in several directions could be obtained. A comparison with the standard models for detecting overdispersion represents the first aim of future analyses.

Acknowledgements This research has been partly supported by FIRB 2012 project on "Mixture and latent variable models for causal inference and analysis of socio-economic data" at University of Perugia.

References

1. Cochran, W. G.: Sampling techniques (3rd ed.). New York: John Wiley & Sons (1977)
2. Farewell, V.T.: A note on regression analysis of ordinal data with variability of classification. Biometrika, **69**, 533–538 (1982)
3. Fitzmaurice, G. M., Heath, A. F. and Cox, D. R.: Detecting Overdispersion in Large Scale Surveys: Application to a Study of Education and Social Class in Britain. J. Roy. Statist. Soc. Ser. C, **46**, 415–432 (1997)
4. Greenleaf, E.: Improving rating scale measures by detecting and correcting bis components in some response styles, J. Market Res., **29**, 176–188 (1992)
5. Hinde, J., Demétrio, C. G. B.: Overdispersion: Models and Estimation. Sao Paulo: ABE (1998)
6. Morel, J.G., Koehler, K.J.: A one-step Gauss-Newton estimator for modeling categorical data with extraneous variation, J. Appl. Statist., **44**, 187–200 (1995)
7. Iannario, M.: Fitting measures for ordinal data models, Quad. Stat., **11**, 39–72 (2009)
8. Iannario, M.: On the identifiability of a mixture model for ordinal data, METRON, **LXVIII**, 87–94 (2010)
9. Iannario, M.: CUBE models for interpreting ordered categorical data with overdispersion, Quad. Stat., **14**, 137–140 (2012)
10. Iannario, M.: Modelling Uncertainty and Overdispersion in Ordinal Data. Submitted (2013)
11. Iannario, M., Piccolo, D.: CUB models: Statistical methods and empirical evidence, in: Kenett R. S. and Salini S. (eds.), *Modern Analysis of Customer Surveys: with applications using R*. Chichester: J. Wiley & Sons, pp.231–258 (2012)
12. Jansen, J.: On the statistical analysis of ordinal data when extravariation is present. J. Appl. Statist., **39**, 75–84 (1990)
13. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*, 2^{nd} edition. London: Chapman & Hall. (1989)
14. McLachlan, G., Krishnan, T.: *The EM algorithm and extensions*. New York: J.Wiley & Sons. (1997)
15. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables, Quad. Stat., **5**, 85–104 (2003)
16. Piccolo, D.: Inferential issues on *CUBE* models with covariates. Submitted (2013)
17. Rossi, P.E., Gilula, Z., Allenby, G.M. Overcoming scale usage heterogeneity: a Bayesian hierachical approach. J. Amer. Statist. Assoc., **96**, 20–31 (2001)
18. Wilson, J. R.: Chi-Square Tests for Overdispersion with Multiparameter Estimates. J. Roy. Statist. Soc. Ser. C, **38**, 441–453 (1989)