Detecting latent components in ordinal data with overdispersion by means of a mixture distribution

Maria Iannario

Abstract The paper describes a mixture distribution generated by Beta Binomial and Uniform random variables to allow for a possible *overdispersion* in surveys when the response of interest is an ordinal variable. This approach considers the joint presence of feeling, uncertainty and a possible dispersion sometimes present in the evaluation contexts. After a discussion of the main properties of this class of models, asymptotic likelihood methods have been applied for efficient statistical inference. The implementation on the survey on household income and wealth (SHIW) will confirm the versatility of this distribution and the usefulness to distinguish the determinants of uncertainty and overdispersion in real data.

Keywords Latent components · Overdispersion · CUB model · CUBE model

1 Introduction

In categorical data analysis it is often found that data exhibit greater variability than predicted by the implicit mean-variance relationship. This phenomenon, denoted as *overdispersion*, is fairly common in qualitative data and it is usually associated with Poisson, multinomial and other exponential family models. Failure to take account of this process can lead to serious underestimation of standard errors and misleading inference for the regression parameters. Consequently, several approaches have been proposed for handling such a problem as discussed by Finney (1971), Cox (1983), McCullagh and Nelder (1989), among others. However, only few papers concern the overdispersion in ordinal data for which some generalizations of the binary logistic regression are usually introduced.

As a rule, overdispersion can be due to interviewer effects, hidden clusters, different number of categories for each response or, more generally, to the absence of relevant predictors in the model or design effects, to reasons related to scale usage heterogeneity (Greenleaf 1992; Rossi et al. 2001), or to a subjective interpretation of the wording of categories (Farewell 1982). Thus, Cochran (1977) considers how to adjust the variance of an estimator to account

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

for the sampling design and Hinde and Demétrio (1998) implement a two-stage model for the response belonging to the linear exponential family. However, the residual variation, beyond that predicted by the mean, may be so large that no model but the saturated one appears adequate for the data (Fitzmaurice et al. 1997).

In this context we propose a parametric approach based on a mixture distribution which extends CUB models, proposed by Piccolo (2003) and discussed by Iannario and Piccolo (2012). These models are a combination of a discrete Uniform and a shifted Binomial random variables to take uncertainty and feeling components of respondents into account. The generalized model has been denoted as CUBE (Iannario 2014a) for the presence of a Beta-Binomial distribution instead of the Binomial one in order to include the overdispersion effect.

The main goal of the paper is to discuss how overdispersion in ordinal data surveys has to be considered with attention since its omission induces an increase of the uncertainty and thus biases the interpretation of the possible determinants of the latent variables which characterize the process of response. Some empirical evidence, derived from both simulated and real data, will support this assertion and suggest to consider CUBE models as a consistent formulation for the analysis of ordinal data.

The paper is organized as follows. Section 2 describes the notation and the main characteristics of CUBE models, including some inferential issues. Section 3 is mainly devoted to the interpretation of parameters. Section 4 summarizes the main results obtained from the investigation of these models on a simulated data set and mainly on the SHIW, a periodic survey organized by the Bank of Italy to achieve information on income and wealth of Italian households. Some concluding remarks end the paper.

2 A finite mixture model for ordinal data

In a given survey, we interpret the ordinal response as a random variable R whose probability distribution may be defined as the convex combination of two components related to the personal *feeling* of the subject towards the item and an inherent *uncertainty* derived by a constrained choice among an ordered list of alternatives. Indeed, when a rater expresses his/her evaluation, several facts are involved: rater's personal attitudes, educational background, knowledge about the question, emotional, cognitive and behavioral aspects, such as the strategies developed to deal with uncertainty (Sartori and Ceschi 2011) and/or the tendencies to respond systematically to questionnaire items on some criterion other than what the items were specifically designed to measure (Baumgartner and Steenback 2001).

If we further assume that an extra dispersion is present among respondents as a consequence of personal attitude then a (shifted) Beta-Binomial may be introduced since this random variable assumes a changeable opinion among the subjects (Chatfield and Goodhart 1970). In this way, a more general approach would consider a *C*ombination of discrete *U*niform and *BE*ta-Binomial distributions (which produces the acronym CUBE) with probability mass function: $Pr(R = r) = \pi\, g_r(\xi, \phi) + (1 - \pi)\, p_r^U$, $r = 1, 2, \ldots, m$, where $g_r(\xi, \phi)$ is the shifted Beta Binomial distribution parameterized by:

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

and $p_r^U = 1/m$, $r = 1, 2, \ldots, m$, is the distribution of a discrete Uniform random variable. It is immediate to prove that if $\phi = 0$ a CUBE model reduces to a CUB model; thus, the latter is nested in the first one.

The parameter vector $\boldsymbol{\theta} = (\pi, \xi, \phi)'$ belongs to the parameter space $\Omega(\boldsymbol{\theta}) = \{(\pi, \xi, \phi) : 0 < \pi \leq 1; \ 0 < \xi < 1; 0 \leq \phi < \infty\}$. Thus, the parameters $1 - \pi$, $1 - \xi$ and ϕ may be related to uncertainty, feeling and overdispersion, respectively. Then, we could exploit the unit square for a visualization of both models; in fact, CUB models are shown as points in correspondence with $(1 - \pi, 1 - \xi)$ to visualize uncertainty and feeling, respectively. Similarly, CUBE models may be represented with points whose sizes become proportional to ϕ.

To relate the personal characteristics of the respondents to uncertainty, feeling and overdispersion, respectively, it is possible to include subjects' covariates in a CUBE model. Given a $\boldsymbol{T}$ matrix of observed k covariates with rows $\boldsymbol{t}_i$, $i = 1, 2, \ldots, n$, we assume a logistic link, for instance, among the parameters and the subject's covariates (Piccolo 2014):

$$logit(\pi_i) = \boldsymbol{y}_i \, \boldsymbol{\beta}; \quad logit(\xi_i) = \boldsymbol{w}_i \, \boldsymbol{\gamma}; \quad logit(\phi_i) = \boldsymbol{x}_i \, \boldsymbol{\alpha}; \qquad i = 1, 2, \ldots, n,$$

where $logit(p) = \log(p/(1-p))$ and the rows $\boldsymbol{y}_i$, $\boldsymbol{w}_i$, $\boldsymbol{x}_i$ are subsets of $\boldsymbol{t}_i$ (here, we assume that the first element of these vectors is 1). Hence, the parameter vectors $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}$ are able to interpret the direction and the weight of the covariates to explain their effect on uncertainty, feeling and overdispersion, respectively. A log function as a link for ϕ is also admissible.

For a sample of ordinal data $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)'$, the log-likelihood function of a CUBE model is: $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \pi \left[g_{r_i}(\xi, \phi) - \frac{1}{m} \right] + \frac{1}{m} \right\}$. Effective procedures for getting ML estimates are obtained by the EM algorithm (McLachlan and Krishnan 1997) and they have been detailed for CUB (Piccolo 2006) and CUBE models (Iannario 2014a).

The asymptotic theory of ML estimators allows to detect the significance of the parameters (Piccolo 2006). In this regard, a relevant issue is the check of a possible significant overdispersion by testing $H_0 : \phi = 0$ against $H_1 : \phi > 0$, that is if a CUB model is more adequate than a CUBE model. Empirical and simulated evidence strongly suggest a likelihood ratio test (Iannario 2014b), a quantity easily obtained by routinely estimating both CUB and CUBE models to given observations. It may be worthwhile to mention that in this case the LRT is testing an hypothesis on the boundary of the parameter space, thus the p-value must be halved. Finally, the BIC criterion is generally used to compare non-nested models.

The evaluation of the predictive ability of ordinal data models requires some critical discussion. In fact, the estimated model refers to a whole probability distribution (optimal in a likelihood sense) whereas observations are integers $r_i \in \{1, 2, \ldots, m\}$. As a consequence, to measure predictive ability of a CUBE model (without covariates) we need to synthesize the estimated $p_r(\hat{\boldsymbol{\theta}})$ by means of a single number $\hat{r}_i = \hat{r}$, $\forall i$: this predictor may be the estimated expectation, modal value or median, for instance. Then, given the sample mean $\bar{r}$ and variance s^2 of observed ratings, the root mean square error (RMSE) of prediction is equal to:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_i - \hat{r}_i)^2} = \sqrt{s^2 + (\hat{r} - \bar{r})^2},$$

which is minimized when we select as predictor the expectation from the estimated model. Thus, $RMSE \simeq s$ if we compare models whose expectations are sufficiently similar. A critical point is that RMSE is based on a point estimate and although helpful to measure predictive ability of models it cannot be used to discriminate among models since it is
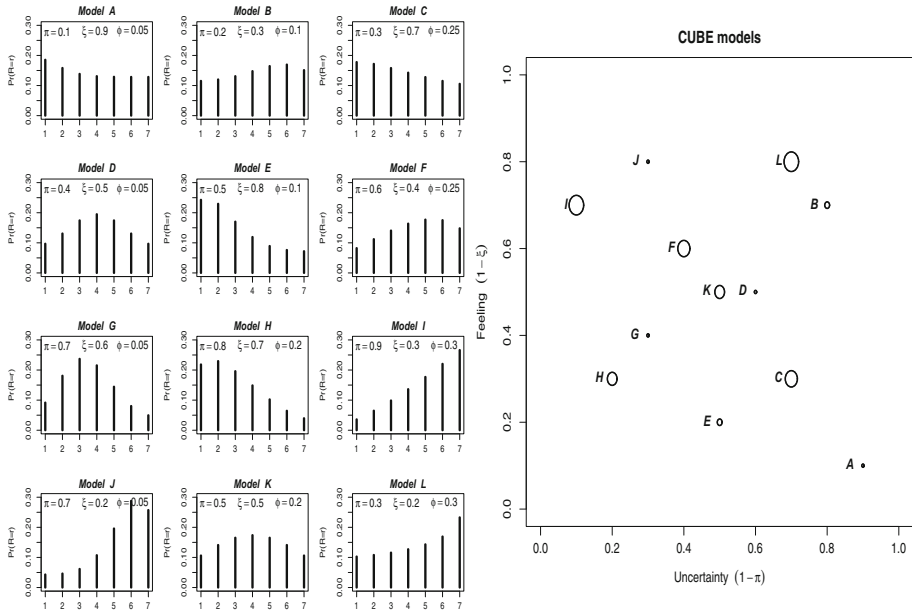
Fig. 1 CUBE probability distributions for varying parameters π, ξ, ϕ. Each model (left panel) is visualized as a point in the parameter space (right panel) with abscissa and ordinate related to uncertainty $(1 - \pi)$ and feeling $(1 - \xi)$, respectively. The size of the points is proportional to the overdispersion parameter (ϕ)

affected by possible misspecifications. More elaborate measures are necessary for models with covariates.

3 Interpretation of the CUBE model parameters

The general shape of the CUBE distribution derives from the corresponding properties of the Beta-Binomial random variable (Tripathi et al. 1994): unimodal models with an intermediate mode arise when $\phi < 0.5$. Instead, bimodal distributions with modal values at $R = 1$ and $R = m$ are obtained when $\phi > 0.5$. Finally, a CUBE random variable has positive, zero, negative asymmetry according to $\xi \gtreqless \frac{1}{2}$.

Figure 1, left panel, shows several CUBE probability distributions for $m = 7$ and for varying parameters π, ξ, ϕ. In the right panel these models are visualized in the parameter space: the size of the point is proportional to ϕ, thus greater points denote models with higher overdispersion.

To investigate the role of the CUBE model parameters we briefly consider central moments and variability measures.

First of all, the expectation of a CUBE random variable is given by $E(R) = (m + 1)/2 + \pi(m - 1)(1/2 - \xi)$ and it is unaffected by the overdispersion parameter ϕ; thus, it may be consistently related to the feeling which is invariant with respect to overdispersion.

Instead, the parameter π is related to heterogeneity concepts as shown by computing the normalized Gini index $G = (1 - \sum_{i=1}^{m} p_i^2) m/(m - 1)$. In fact, with obvious notation and according to Iannario (2012, pp. 169–170), we get:

$$G_{CUBE} = 1 - \pi^2 \left(1 - G_{BETA}\right).$$
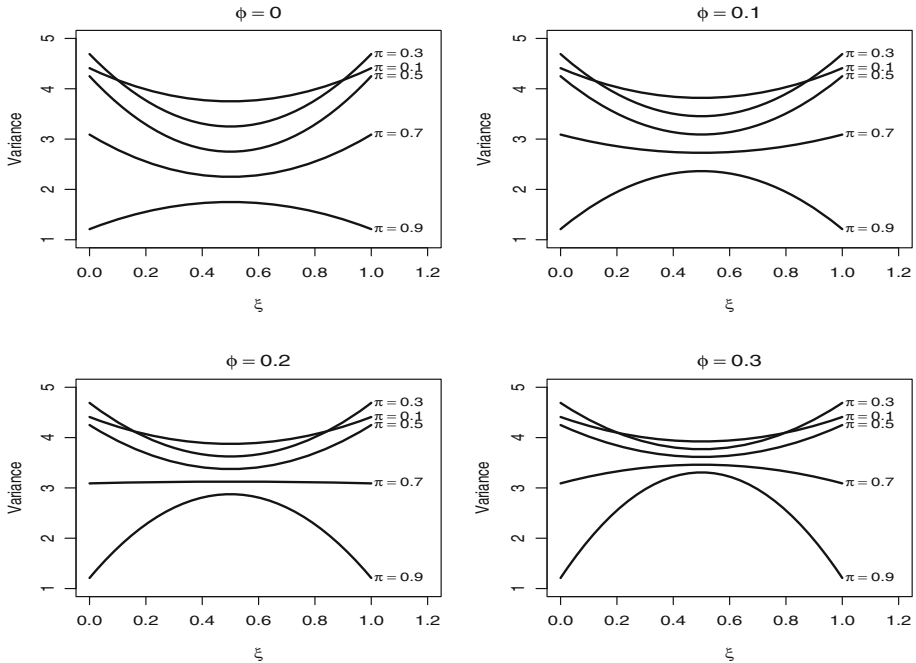
Fig. 2 Variance of CUBE models as a function of ξ for fixed ϕ and different levels of π

Then, for a given Beta-Binomial component (characterized by ξ and ϕ), heterogeneity as measured by the Gini index is inversely related to π and it increases with the uncertainty measure (that is, $1 - \pi$).

The variance of a CUBE random variable is:

$$Var(R) = Var(R^*) + \pi \, \xi \, (1 - \xi) \, (m - 1) \, (m - 2) \, \frac{\phi}{1 + \phi},$$

where $Var(R^*)$ is the variance of a CUB random variable with the same parameters (π, ξ). The additional contribution to the variance of a CUBE distribution increases with ϕ (and π) by a quantity which is maximized when $\xi = 1/2$ (Iannario 2014a). Instead, when $\xi \to 0$ or $\xi \to 1$ the overdispersion effect tends to 0.

Notice that ξ is mainly related to the categories of the response while π is mostly related to the probabilities of these categories as a whole. In the same line of reasoning it is possible to consider the relationship of the ϕ parameter with the shape of the distribution. In fact, we observe that the overdispersion parameter ϕ modifies also the shape of the distribution by changing the relative importance of a modal value: compare models J and I of Fig. 1, for instance, which have similar feeling and uncertainty but a quite different parameter of overdispersion.

A strong relationship among the parameters exists, especially with reference to the variability. Figure 2 shows that the variance is symmetric with respect to $\xi = 1/2$ and how it changes as a function of ξ when ϕ is given, for different levels of π. The behaviour of these curves confirms that $Var(R)$ may be a confusing measure to detect the overdispersion of this model.
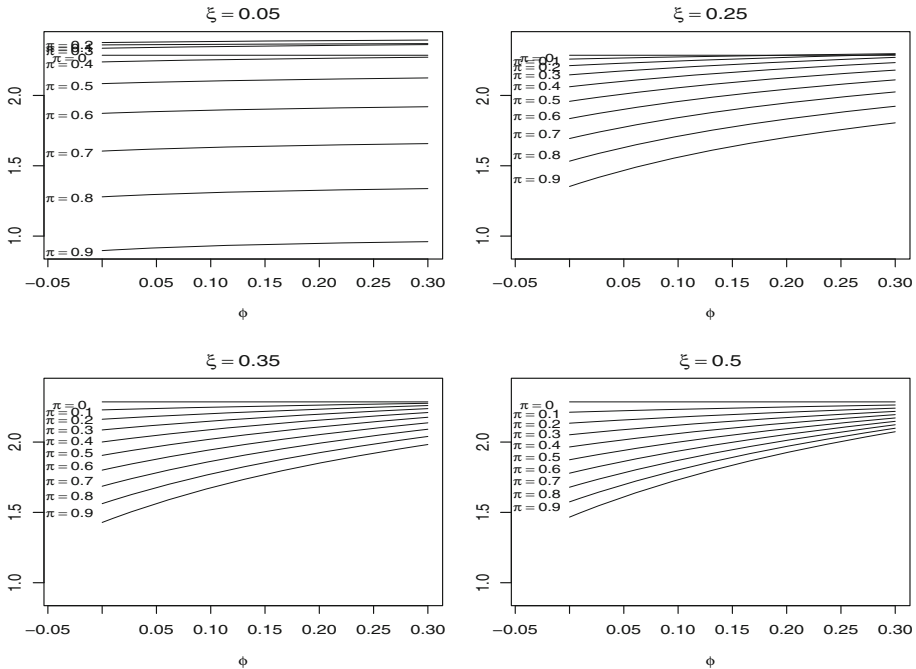
Fig. 3 Δ measures for varying parameters π, ϕ, fixed ξ

Finally, we consider the *mean difference* since it synthesizes the mutual variability of the responses. It has been used also to describe a possible inflated distribution (Gerstenkorn and Gerstenkorn 2003). Given a probability distribution $p_r = Pr(R = r|\boldsymbol{\theta})$ with distribution function $F_r = \sum_{j=1}^{r} p_r$, $r = 1, 2, \ldots, m$, the mean difference Δ is defined by:

$$\Delta = \sum_{r \neq s=1}^{m} |r - s| \, p_r \, p_s = 2 \sum_{r=1}^{m-1} F_r(1 - F_r),$$

where the second formulation is the Finetti and Paciello (1930) formula and it is computationally effective.

Since $\Delta(\xi) = \Delta(1 - \xi)$, it suffices to consider $\xi \in (0, 1/2)$. Then, we plot Δ as a function of $\phi \in (0, 0.3)$ for given values of $\xi = \{0.05, 0.25, 0.35, 0.50\}$ and different levels of $\pi \in (0, 1)$.

Figure 3 shows that the value of Δ monotonically increases with ϕ (abscissa) and, given ξ, for higher level of π. Thus, we may assume that in this model the overdispersion is related to the concept of *mutual variability* and it increases when respondents give ordinal responses mostly different each other.

4 Empirical evidence

We prove that the variability induced by the overdispersion may be wrongly detected as uncertainty if the statistical model is misspecified. To this end, we consider both a simulated set of ordinal data and some specific responses related to the SHIW data set: in the first
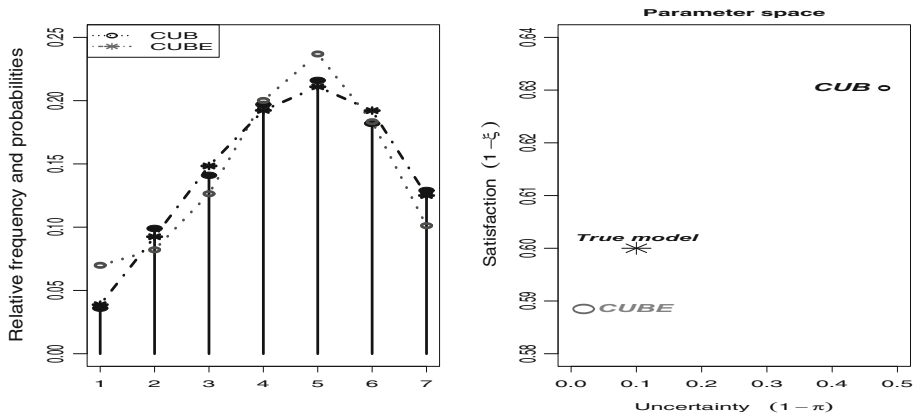
Fig. 4 Simulated ordinal distribution and fitted CUB and CUBE models

Table 1 Mixture models fitted to an ordinal data set with overdispersion

Models	π	ξ	ϕ	$Log - lik$	BIC
CUB	0.520	0.370		−1867.1	3748
CUBE	0.981	0.411	0.201	−1851.1	3723
True values	0.900	0.400	0.200		

case no covariates are involved whereas in the second case also covariates are found to be significant for all components.

4.1 Simulated data set

We generate a sample of $n = 1,000$ ratings from a CUBE model with $m = 7$, specified by $\pi = 0.9, \xi = 0.4$ and $\phi = 0.2$. We get the observed frequencies $\boldsymbol{n} = (36, 99, 141, 197, 216, 182, 129)'$ shown in Figure 4 (left panel). Table 1 lists the main results of the estimation process (all parameters are significant).

With respect to the true value of $\pi = 0.9$, the specification of a CUBE model estimates a closer value whereas CUB model provides a seriously biased estimate of $\hat{\pi}$ which implies a larger uncertainty.

Figure 4 compares the fitting of CUB and CUBE models (left panel) with the observed distribution of simulated data; it is evident that the latter sharply pinpoints the frequency distribution. Although the estimated CUB model is able to mimic the pattern of the frequency distribution it shows a poor fit at categories 1 and 5.

The visualization in the parameter space (Fig. 4, right panel) emphasizes how the misspecified CUB model includes a greater uncertainty which instead is correctly estimated by the corresponding CUBE model; then, notice how the level of ξ is barely affected by the presence of overdispersion since it is mainly related to location and skewness of the distribution.

This example (which we repeated with several other models) manifests how the presence of overdispersed data biases the evaluation of the uncertainty if we limit ourselves to consider only the class of CUB models; this misspecification causes inefficiency in the estimation and interpretation steps of the inferential process.

Table 2 Estimated CUB and CUBE models for the items in the two waves

Item	Waves	Model	log-lik	BIC	RMSE
compr	2008	CUB	−7274.9	14566	1.6711
		CUBE	−7166.4	14358	1.6701
	2010	CUB	−6844.6	13706	1.6225
		CUBE	−6733.4	13491	1.6218
klima	2008	CUB	−6875.0	13767	1.5412
		CUBE	−6768.4	13562	1.5409
	2010	CUB	−6443.5	12904	1.5076
		CUBE	−6319.3	12663	1.5066
facil	2008	CUB	−7255.6	14528	1.6590
		CUBE	−7177.3	14379	1.6584
	2010	CUB	−6912.5	13841	1.6307
		CUBE	−6806.2	13637	1.6300

4.2 SHIW data set

The SHIW data set is a large survey periodically accomplished by the Bank of Italy to measure income and wealth components of the Italian households. In addition, it collects several information on related issues as economic perceived conditions, capital gains, financial assets, health status, family choices, and so on. At the end, the interviewer asks some topics about the interview: "interviewee level of understanding of the questions" (*compr*), "general atmosphere in which the interview took place" (*klima*) and "easiness for the interviewee to answer the questions" (*facil*). The ordinal responses range from 1 (low) to 10 (high) and we will concentrate our analysis on the last two waves (2008 and 2010): the available observations consist of 3, 887 and 3, 816 interviews, respectively.

Table 2 summarizes the global fitting of the models and we deduce a better performance of the CUBE mixture. We present also the RMSE when $\hat{r} = \bar{r}$ and it barely improves for overdispersed models. Since all opinions about the quality of the interview show high levels of feeling and low levels of uncertainty only small differences among the three items are observed.

Figure 5 summarizes the estimated models of the ordinal variables (*compr, klima, facil*) in the parameter space and emphasizes how the introduction of an overdispersion parameter induces a regular decrease of the uncertainty. It is evident that satisfaction is always higher for the 2010 wave whereas the overdispersion is substantially unchanged between the waves. Thus, the respondents do not change their relative habits except for improving their feeling.

Then, we check the significance of explanatory covariates for each parameter in case of the response to *Global comprehension of questions* collected in 2010. We found a significant impact of *Education* for uncertainty, *Gender* for feeling and perceived *Health status* for overdispersion. The estimated CUB and CUBE models are reported in Table 3 and fitting measures clearly favour the model with overdispersion. Noticeably, the relationship between covariates and uncertainty and feeling, respectively, is maintained in both models. Then, by considering the sign of the parameters $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\alpha}$, we observe that the *Global comprehension of questions* is higher for men, the level of uncertainty increases for lower level of *Education* and overdispersion increases with a lower perception of own *Health status*.

Since the estimated CUBE models generate a probability distribution of responses conditional to subjects' covariates, it is possible to create profiles of expected responses for given
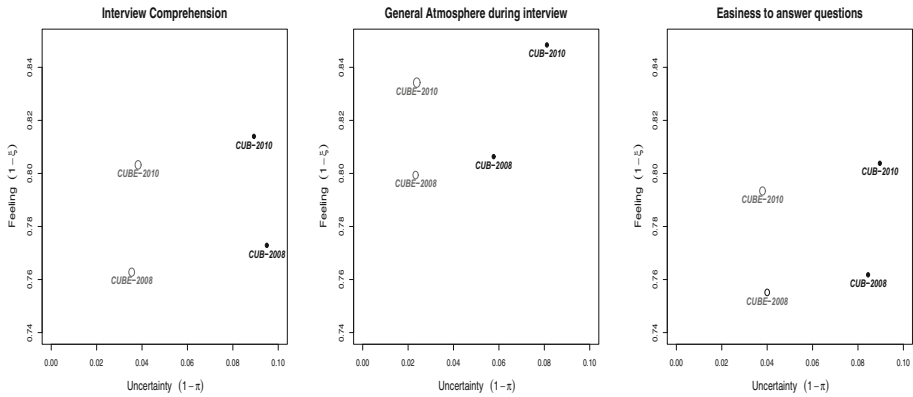
Fig. 5 Visualization of CUB and CUBE models for *compr, klima, facil* in both waves

Table 3 *Global comprehension of questions*: estimated models with covariates (wave 2010)

Components	Covariates	CUB model parameters	CUBE model parameters
Uncertainty	Constant	$\hat{\beta}_0 = -1.824\,(0.324)$	$\hat{\beta}_0 = -1.610\,(0.441)$
	Education	$\hat{\beta}_1 = 1.446\,(0.130)$	$\hat{\beta}_1 = 1.657\,(0.195)$
Feeling	Constant	$\hat{\gamma}_0 = -1.593\,(0.023)$	$\hat{\gamma}_0 = -1.536\,(0.025)$
	Gender	$\hat{\gamma}_1 = 0.133\,(0.033)$	$\hat{\gamma}_1 = 0.139\,(0.036)$
Overdispersion	Constant		$\hat{\alpha}_0 = -3.541\,(0.259)$
	Health status		$\hat{\alpha}_1 = 0.284\,(0.096)$
	$\ell(\hat{\boldsymbol{\theta}})$	-6737.1	-6660.5
	BIC	13507.0	13370.5
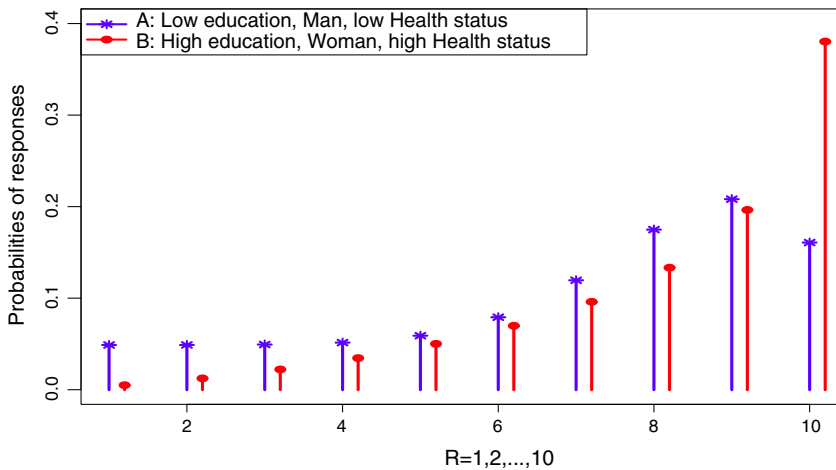
Fig. 6 CUBE models: expected profiles of probability for *Global comprehension of questions*

levels of covariates. Specifically, we compare the probabilities of responses of a low educated man with a low level of perceived health (A: *Gender* = Man, *Education* = compulsory class, *Health* = 1) with those of a high educated woman, with a high level of perceived health status

(B: *Gender* = Woman, *Education* = post graduate, *Health* = 10). Figure 6 shows how the effect of covariates appreciably changes the expected distribution of the responses.

5 Concluding remarks

We discussed the class of CUBE models to explicitly take a possible overdispersion into account in case of ordinal responses; such a component is interpreted as generated by an interpersonal variability of subjects which are requested to make a choice among ordinal categories. This enriched model may be estimated with a limited number of parameters since -with respect to cumulative models- we do not estimate cutpoints. In addition, it turned out that the omission of a component related to the overdispersion of the responses can cause an undue increase of the estimated uncertainty as confirmed by a simulated data analysis and a real case study.

Given the availability of free and effective software for estimating both CUB and CUBE models, a convenient strategy is to fit both of them to the available observations. Since CUB models are nested into CUBE models the comparison of log-likelihood functions computed at their maxima leads to a likelihood ratio test which is able to confirm or reject the need for an overdispersion component.

In this respect, further studies are required to select significant variables for CUBE models and also to implement faster procedures to estimate these models in presence of several covariates.

References

Baumgartner, H., Steenback, J.B.: Response styles in marketing research: a cross-national investigation. J. Mark. Res. **38**, 143–156 (2001)

Chatfield, C., Goodhart, G.J.: The beta-binomial model for consumer purchasing behavior. Appl. Statist. **19**, 240–250 (1970)

Cochran, W.G.: Sampling techniques, 3rd edn. John Wiley & Sons, New York (1977)

Cox, D.R.: Some remarks on overdispersion. Biometrika **70**, 269–274 (1983)

De Finetti, B., Paciello, U.: Calcolo della differenza media. METRON **8**, 1–6 (1930)

Farewell, V.T.: A note on regression analysis of ordinal data with variability of classification. Biometrika **69**, 533–538 (1982)

Finney, D.J.: Probit Analysis. Cambridge University Press, Cambridge (1971)

Fitzmaurice, G.M., Heath, A.F., 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)

Gerstenkorn, T., Gerstenkorn, J.: Gini's mean difference in the theory and application to inflated distributions. Statistica **LXIII**, 469–488 (2003)

Greenleaf, E.: Improving rating scale measures by detecting and correcting bis components in some response styles. J. Marketing Res. **29**, 176–188 (1992)

Hinde, J., Demétrio, C.G.B.: Overdispersion: Models and Estimation. ABE, Sao Paulo (1998)

Iannario, M.: Preliminary estimators for a mixture model of ordinal data. Adv. Data Anal. Classif. **6**, 163–184 (2012)

Iannario, M.: Modelling uncertainty and overdispersion in ordinal data. Comm. Stat. Theory Methods **43**, 771–786 (2014a)

Iannario, M.: Testing overdispersion in CUBE models. Comm. Stat. Simul. Comput. **43**, 771–786 (2014b)

Iannario, M., Piccolo, D.: cub models: statistical methods and empirical evidence. In: Kenett, R.S., Salini, S. (eds.) Modern Analysis of customer surveys: with applications using R, pp. 231–258. Wiley, Chichester (2012)

McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edition. Chapman & Hall, London (1989)

McLachlan, G., Krishnan, T.: The EM algorithm and extensions. Wiley, New York (1997)

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

Piccolo, D.: Observed information matrix for MUB models. Quad. Stat. **8**, 33–78 (2006)

Piccolo, D.: Inferential issues on $CUBE$ models with covariates. Comm. Stat. Theory Methods, 44, forthcoming (2014)

Rossi, P.E., Gilula, Z., Allenby, G.M.: Overcoming scale usage heterogeneity: a Bayesian hierachical approach. J. Amer. Statist. Assoc. **96**, 20–31 (2001)

Sartori, R., Ceschi, A.: Uncertainty and its perception: experimental study of the numeric expression of uncertainty in two decisional contexts. Qual. Quant. **45**, 187–198 (2011)

Tripathi, R.C., Gupta, R.C., Gurland, J.: Estimation of parameters in the Beta Binomial model. Ann. Inst. Statist. Math. **46**, 317–331 (1994)