LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN
INSTITUT FÜR STATISTIK

Gerhard Tutz, Micha Schneider, Maria Iannario, Domenico Piccolo

Mixture Models for Ordinal Responses to Account for Uncertainty of Choice

Technical Report Number 175, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Mixture Models for Ordinal Responses to Account for Uncertainty of Choice

Gerhard Tutz[1], Micha Schneider[1], Maria Iannario[2], Domenico Piccolo[2]

[1] Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München, D

[2] University of Naples Federico II, Via L.Rodinò 22, 80138 Naples, I

December 16, 2014

Abstract

In CUB models the uncertainty of choice is explicitly modelled as a Combination of discrete Uniform and shifted Binomial random variables. The basic concept to model the response as a mixture of a deliberate choice of a response category and an uncertainty component that is represented by a uniform distribution on the response categories is extended to a much wider class of models. The deliberate choice can in particular be determined by classical ordinal response models as the cumulative and adjacent categories model. Then one obtains the traditional and flexible models as special cases when the uncertainty component is irrelevant. It is shown that the effect of explanatory variables is underestimated if the uncertainty component is neglected in a cumulative type mixture model. Visualization tools for the effects of variables are proposed and the modelling strategies are evaluated by use of real data sets. It is demonstrated that the extended class of models frequently yields better fit than classical ordinal response models without an uncertainty component.

Keywords: Ordinal responses, rating analysis, CUP model, CUB model

1 Introduction

In many applications the responses are measured on an ordinal scale and given in categories. There is a considerable amount of literature devoted to the adequate modelling of such ordered categorical data. In particular the seminal paper of McCullagh (1980) stimulated research to find parametric models which should be both parsimonious and well fitted to real data. Overviews on recent research are found, for example, in Agresti (2010), Agresti (2013) and Tutz (2012).

Ordered categorical responses typically come in two forms, as *grouped continuous variables* and *assessed ordinal categorical variables* (Anderson, 1984). The first type is a mere categorized version of a continuous variable, which in principle can be observed itself. The second type of ordered variable arises when an assessor processes an unknown amount of information, leading to the judgement of the grade of the ordered categorical scale. This sort of variable is found, for example, in preference or evaluation studies and the assessment of pain.

With the focus on ordinal variables generated by judgements, a mixture type model that accounts for the psychological process of human choices has been introduced by Piccolo (2003) and developed in a series of papers by D'Elia and Piccolo (2005), Iannario and Piccolo (2010), Manisera and Zuccolotto (2014).

The basic concept of these so-called CUB models is that the choice of a response category is determined by a mixture of feeling and uncertainty. Feeling refers to the deliberate choice of a response category determined by the preferences of a person while uncertainty refers to the inherent individual's indecision. The first component is modelled by a binomial distribution, the latter by a discrete uniform distribution across response categories. These components, effectively parameterized in a parsimonious manner, allow CUB models to be extremely flexible for capturing the different shapes of ordinal data distributions; in addition, the parameters to be estimated are immediately related to the concept of uncertainty (indecision, fuzziness) and feeling (attraction, preference), which improves the simplicity of the interpretation and makes the comparison among subgroups easier. An introduction and overview was given by Iannario and Piccolo (2012) whereas several generalizations in different fields have been obtained to include objects' covariates multilevel data (Iannario, 2012a), and data surveys with *shelter* effects (Iannario, 2012b).

Alternative approaches to finite mixtures for ordinal data have been advanced by Wedel and DeSarbo (1995); Greene and Hensher (2003); Grün and Leisch (2008); Breen and Luijkx (2010), among others. These authors propose convex combinations of probability distributions belonging to the same class of models and assume the existence of subgroups whose responses should be differently modelled.

In the present paper the mixture approach is extended to include more traditional models for the modelling of preferences by including an uncertainty component. We consider distributions in which the preference part is determined by a cumulative or adjacent categories model, which yields more flexible models. The paper is organized as follows: in the next section we consider uncertainty as a relevant component quite often present in human choices; thus CUB models are briefly reviewed and a new class of models (called CUP) is introduced. For both of them a non-parametric measure of heterogeneity may help to understand the weights and the effect of introducing uncertainty in the mixture. In Section 3 a deeper discussion is given concerning the effects of the uncertainty component in the interpretation of the model whereas Section 4 deals with the problem of model selection by adequate fitting measures. Section 5 presents some empirical evidence on data sets of different scientific fields and compares standard approaches with mixtures that include an uncertainty component. Some concluding remarks and an appendix devoted to estimation problems end the paper.

2 Modelling Uncertainty by Mixtures

In the following we first sketch the CUB model, which is an abbreviation for Combination of discrete Uniform and shifted Binomial random variables. Then we consider an extended class that contains the CUB as well as standard models for ordinal data as special cases.

2.1 The CUB Model

Let in a regression model the response of an individual R_i given explanatory variables $\boldsymbol{z}_i, \boldsymbol{x}_i$ take values from ordered categories $\{1, \dots, k\}$. Then, the mixture distribution denoted as CUB as considered, for example, by Iannario and Piccolo (2012) has been defined for each subject by

$$Pr(R_i = r | \boldsymbol{z}_i, \boldsymbol{x}_i) = \pi_i \, b_r(\xi_i) + (1 - \pi_i) \, p_r^U \quad r \in \{1, \dots, k\}, \tag{1}$$

where the two components of the mixture are specified in the following way. The first component is a shifted binomial distribution given by

$$b_r(\xi_i) = \binom{k-1}{r-1} \xi_i^{k-r} (1 - \xi_i)^{r-1}, \quad r \in \{1, \dots, k\}.$$

It is a simple binomial distribution determined by the parameter ξ but shifted so that the support is $\{1, \dots, k\}$ instead of the usual support that includes zero. The component represents the preferences for specific categories, which is captured by the parameter ξ_i.

The second component is a uniform distribution across the response categories,

$$p_r^U = 1/k, \quad r \in \{1, \dots, k\}.$$

It represents the additional uncertainty arising from factors like amount of time devoted to the response, fatigue, partial understanding, etc. It is explicitly modelled as the indecision component related to the nature of human choices. Iannario and Piccolo (2012) discuss extensively the logical foundations and psychological motivations of the mixture.

In CUB models the parameters π_i and ξ_i are linked to the covariates $(\boldsymbol{z}_i^T, \boldsymbol{x}_i^T)$ by the logit links

$$\text{logit}\,(\pi_i) \;=\; \boldsymbol{z}_i^T \boldsymbol{\beta}\,; \qquad \text{logit}\,(\xi_i) \;=\; \boldsymbol{x}_i^T \boldsymbol{\gamma}\,; \qquad i = 1, 2, \dots, n\,. \tag{2}$$

In fact, alternative link functions representing a one-to-one mapping $\mathbb{R}^p \leftrightarrow [0, 1]$ between parameters and covariates are also legitimate. It should be noted that, given the parameterization (1), the covariates in $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ may coincide, overlap or be completely different.

The two components, preference/feeling represented by the binomial model, and uncertainty represented by the uniform model, are combined in a mixture with weights $\pi_i, 1-\pi_i$. The interpretation is that each interviewee has a *propensity* to adhere to a meditated choice (represented by the first component) and to a totally random decision (represented by the uniform distribution) and $\pi_i, 1-\pi_i$ are just the weights for those propensities. Thus, the quantity $1-\pi_i$ is interpreted as a measure of uncertainty whereas π_i is seen as a measure of adherence to the structured choice.

In the following we briefly investigate the uncertainty component, which is at the core of this paper. It is strongly related to heterogeneity and the related effects on the variability of the distribution. For simplicity we drop the index i for the individual. The first effect of the mixture is that for $\pi < 1$ the distribution of the CUB model is more spread out than the distribution of the binomial model. This can be seen by considering that the variance of the distribution of a CUB model is

$$\mathrm{var}(R) = (k-1)\left[\pi\,\xi\,(1-\xi)\,\{\pi\,(k-1)-(k-2)\} + (1-\pi)\,\frac{3\,\pi\,(k-1)+(k+1)}{12}\right].$$

It is immediate to show that $\mathrm{var}(R)$ is monotonically increasing in a linear way with respect to $1-\pi$ only for $\xi = 1/2$ (a symmetric CUB distribution) whereas it has a minimum for $\pi = 1$ (a shifted binomial model) and a relative maximum for $\pi = 0$ (a discrete uniform model). In fact, the absolute maximum of the parabolic shape happens at

$$\pi = \frac{(1-6\,\xi + 6\,\xi^2)\,(k-2)}{3\,(2\,\xi-1)^2\,(k-1)}, \qquad \text{if} \quad \xi \neq 1/2\,.$$

As a consequence, as shown in the left panel of Figure 1, although variance generally increases with uncertainty one cannot conclude that π is strictly related to this aspect of variability.

On the other side, according to the results of Iannario (2012c, pp.169-170;181), the normalized Gini heterogeneity index increases with uncertainty. It is defined for any discrete distribution $(p_r,\, r = 1, 2, \ldots, k)$ by $G = (1-\sum_{r=1}^{k} p_r^2)\,k/(k-1)$. For the CUB model one obtains

$$G_{CUB} = 1 - \pi^2\,(1 - G_{BIN}),$$

where G_{BIN} is the Gini index computed for the distribution of the binomial component. From this last result, one can derive that for $\pi < 1$ the Gini index for the mixture model is larger than the Gini index for the binomial model: $G_{CUB} > G_{BIN}$, that is, the heterogeneity of the mixture is greater than that of the binomial component, and heterogeneity is increased if the uniform component can not

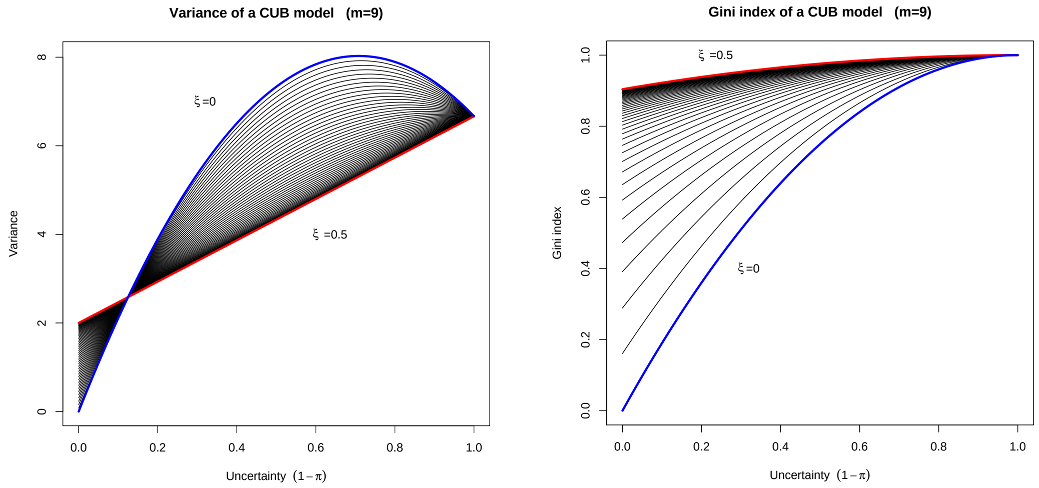

FIGURE 1: *Variance and Gini heterogeneity measures for CUB models as functions of* $1 - \pi$

be neglected. Differently from the variance, the Gini index is monotonically increasing with uncertainty as measured by $(1 - \pi)$ for any given ξ, and this confirms that one should interpret the π parameter as an inverse heterogeneity measure. The behaviour of G_{CUB} with respect to the uncertainty $(1 - \pi)$ is depicted in the right panel of Figure 1.

Some difficulties arise when the responses are more complex and do not follow a definite pattern as implied by the binomial component (which requires a single mode, for instance). Thus, it seems attractive to extend the standard models for ordinal models by including an uncertainty component, which is the added value of the CUB models framework.

2.2 An Extended Class of Models

In the CUB model the choice of a binomial distribution and a uniform distribution is mostly based on simplicity criteria although the binomial may be interpreted as a counting process of selection among the k categories and the uniform distribution may be introduced as the most extreme and uninformative case among all discrete alternatives. In a wider class of models proposed here the rather restrictive binomial model is replaced by more flexible ordinal models while the uniform distribution as an uninformative distribution is retained. The general mixture model we consider has the form

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i) P_U(U_i = r), \tag{3}$$

where R_i represents the observed response and Y_i, U_i are the unobserved random variables taking values from $\{1, \ldots, k\}$. The distribution of Y_i is determined by

$P_M(Y_i = r|\boldsymbol{x}_i)$, which can be any ordinal model M, whereas $P_U(U_i = r) = 1/k$ represents the uniform distribution.

For the specification of the latent variable Y_i one can use models that are in common use in ordinal regression, in particular, cumulative type and adjacent categories type models, which have already been considered by McCullagh (1980). The *cumulative model* has the general form

$$P(Y_i \leq r|\boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}), \quad r = 1, \ldots, k-1,$$

where $F(.)$ is a cumulative distribution function and $-\infty = \gamma_{00} < \gamma_{01} < \cdots < \gamma_{0k} = \infty$. This model is obtained by assuming that a latent regression model $\tilde{Y}_i = -\boldsymbol{x}_i^T \boldsymbol{\gamma} + \epsilon$ holds, where ϵ is a noise variable with distribution function F. If we consider the link between the observable categories and the latent variable given by

$$Y_i = r \;\Leftrightarrow\; \gamma_{0,r-1} < \tilde{Y}_i \leq \gamma_{0r}, \quad r = 1, 2, \ldots, k$$

it is straightforward to derive the model.

The most widely used model from this class of models is the cumulative logit model, which uses the logistic distribution $F(.)$ It is also called *proportional odds model* and has the form

$$\log \left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)} \right) = \gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

An alternative choice is the *adjacent categories model* given by

$$P(Y_i = r+1|Y_i \in \{r, r+1\}, \boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}), \quad r = 1, \ldots, k-1.$$

The specific model that uses the logistic distribution is the *adjacent categories logit model*

$$\log \left(\frac{P(Y_i = r+1|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)} \right) = \gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

Also sequential models or other ordinal models could be useful. For a discussion of these classes of ordinal models, see, for example, Tutz (2012).

We refer to the general model (3) as a CUP model for the Combination of Uniform and Preference structures. Of course, the CUB model is a special case that uses the binomial distribution in the preference part. The use of models like the cumulative or adjacent categories model is attractive because it adds flexibility to the model. For example, the probability distribution of the binomial model is strictly unimodal, in contrast to the cumulative and the adjacent categories model, which allow for all forms of distributions by including the flexible intercepts $\gamma_{01}, \ldots, \gamma_{0k}$. Moreover, cumulative and adjacent categories models are the most widely used models for ordinal data, but an additional uncertainty component seems not to have been used for these models before. As will be shown parameter estimates are biased if the uncertainty component is ignored. In the following we will use the abbreviations CUP(c) and CUP(a) if the structural response model in the mixture is the cumulative or the adjacent categories model, respectively.

In both models, the effect of the explanatory variables in the model that specifies preference is contained in the linear predictors, which have the form $\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}$. Therefore, the specification of the linear predictor replaces the assumption $\text{logit}\,(\xi_i) = \boldsymbol{x}_i^T\boldsymbol{\gamma}$, which specifies the dependence of CUB model parameters on covariates. The dependence of the uncertainty component on covariates is modelled in the same way as in CUB models, namely by $\text{logit}(\pi_i) = \boldsymbol{z}_i^T\boldsymbol{\beta}$, where $\boldsymbol{z}_i$ can be identical to $\boldsymbol{x}_i$.

For CUB models the link between the uncertainty component and heterogeneity measured by the Gini index was systematically investigated by Iannario (2012c). A similar link holds for the CUP model. The Gini index for any mixture model is given by $G_{MIX} = \pi^2 G_M + 1 - \pi^2$, where M denotes the ordinal model used in the mixture and the mixture model itself is given by (3). The maximal heterogeneity is obtained for the uniform distribution, that is, $G_{UNI} = 1$. Thus the Gini index can also be given by

$$G_{MIX} = G_{UNI} - \pi^2(1 - G_M).$$

Considered as a function with argument π it decreases quadratically with increasing probability π from the maximal value to G_M. Therefore, the mixture model has an heterogeneity index between the uniform model and model M, but for $\pi < 1$ is larger than the Gini index for the model M. That means, in the mixture model the probabilities of response categories are more evenly distributed than in model M. By assuming a mixture the basic ordinal model M is shrunk toward the uniform model.

For illustration Figure 2 shows the Gini index as a function of the weight of the uncertainty component $1 - \pi$. The underlying model is a simple cumulative model with ten categories and a binary predictor with coefficient γ. It is seen that the Gini index increases with growing uncertainty $(1 - \pi)$. The increase is strong for strong effects of the predictor and weak if the predictor is less influential.

When considering the effect of uncertainty on the variance it can not be recommended to examine the variance of $Y \in \{1, \dots, k\}$ itself if one takes the ordinal scale level of Y seriously. Therefore, we consider the variances of the binary variables $Y_r = I(Y \leq r), r = 1, \dots, k$, which are explicitly modelled within the cumulative model framework. Figure 3 shows the cumulative probabilities $P(Y \leq r)$ (left panel) and the variances of the corresponding variables Y_r (right panel) for an example with 10 categories. For increasing uncertainty $1 - \pi$ the cumulative probabilities tend to lie on a straight line. However, the effect on the variances is different. The curves are not monotone and as the uncertainty grows the variance decreases for categories smaller than 3 but increases for categories greater than 3.

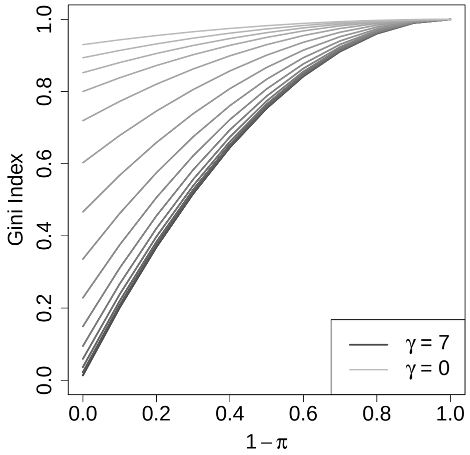

FIGURE 2: *Gini heterogeneity measures for CUP models as functions of* $1 - \pi$

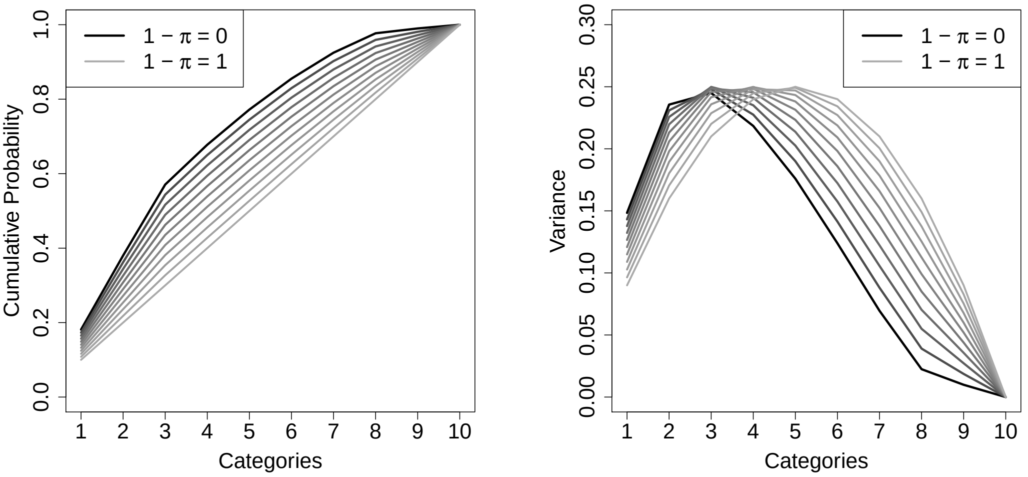

FIGURE 3: *Cumulative probabilities and variances for CUP models with ten categories*

2.3 Estimation

In mixture models, estimation issues can be pursued by exploiting the EM algorithm as proposed by Dempster et al. (1977) and used with special reference to mixtures by McLachlan and Peel (2000). In this context, estimation and tests are obtained by asymptotically efficient procedures based on maximum likelihood methods. For readability we give the used EM algorithm in the appendix.

Specific results for CUB models were given by Piccolo (2006).

3 Effect Strength in Mixture Models

If the response is affected by an additional random component that is modelled by a uniform distribution within the mixture framework, the effects of explanatory variables will differ from the effects found by the fitting of a traditional response model. For simplicity we consider in the following the binary logit model. In this case the cumulative and the sequential models are equivalent. Then the probability of response category 1, denoted by $p(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$, is given by

$$p(\boldsymbol{x}) = \pi p_M(\boldsymbol{x}) + (1 - \pi)/2,$$

where $p_M(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{x}^T\boldsymbol{\gamma})/(1 + \exp(\gamma_0 + \boldsymbol{x}^T\boldsymbol{\gamma}))$ denotes the probability of the logit model. If $\pi < 1$, that is, in the presence of the uncertainty component, one obtains

$$|p(\boldsymbol{x}) - 0.5| = |\pi p_M(\boldsymbol{x}) + (1 - \pi)/2 - 0.5| = \pi\,|p_M(\boldsymbol{x}) - 0.5| < |p_M(\boldsymbol{x}) - 0.5|. \quad (4)$$

That means the true probabilities $p(\boldsymbol{x})$ are closer to 0.5 than the probabilities in the structured component $p_M(\boldsymbol{x})$. This shrinkage toward 0.5 means that the effect strength $\boldsymbol{\gamma}$ tends to be underestimated if the uncertainty component is ignored. More concrete, equation (4) shows that the distance between the probability $p(\boldsymbol{x})$ of the data generating process and 0.5 is equal to $\pi|p_M(\boldsymbol{x}) - 0.5|$. Therefore, the distance reduces by the factor π. It is essential that the reduction is proportional to the distance between the probability and 0.5. That means that a value $p_M(\boldsymbol{x}_1)$ that is farer away from 0.5 changes stronger than a value $p_M(\boldsymbol{x}_1)$ that is closer to 0.5. The consequence is that one observes a weaker effect strength in the mixture model than is present in the model M. In the simplest case one has a binary explanatory variable $x \in \{0, 1\}$. Then both models M and the mixture model are saturated and one can compute the parameter β for the model M and the corresponding parameter $\tilde{\beta}$ that is found when using probabilities $p(x)$. Because $|p(x) - 0.5| = \pi|p_M(x) - 0.5|$ the increase (or decrease) from $p(0)$ to $p(1)$ is always larger than the increase (or decrease) from $p_M(0)$ to $p_M(1)$. Therefore, one obtains $|\tilde{\beta}| < |\beta|$. The case of binary explanatory variables is not interesting by itself, but the tendency to underestimate the effect strengths holds in the general case.

Before considering the effect in the general model with ordered categories it should be noted that the inclusion of an uncertainty component has one other effect. Since the probability $p(\boldsymbol{x})$ of the data generating process is closer to 0.5 than the probability $p_M(\boldsymbol{x})$ also the variance is larger than in the logit model M. Therefore, the inclusion of a uniform component is one way of modelling *overdispersion.*

For ordinal models with $k > 2$ and a cumulative logit model one gets similar results for the cumulative probability $p_r(\boldsymbol{x}) = P(Y \leq r|\boldsymbol{x})$, which is given by

$$p_r(\boldsymbol{x}) = \pi p_{M,r}(\boldsymbol{x}) + (1-\pi)r/k,$$

where $p_{M,r}(\boldsymbol{x}) = \exp(\gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma})/(1+\exp(\gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma}))$ specifies a binary logit model. That means one obtains a shrinkage toward r/k. It is easily derived that now $|p_r(\boldsymbol{x}) - r/k| = \pi|p_{M,r}(\boldsymbol{x}) - r/k|$ with the same consequence as in the binary model, namely that the effect strength γ tends to be underestimated if the mixture component is neglected. What differs from the binary model is that one does not necessarily model overdispersion. Of course, for k even and $r = k/2$ one has the same effect as in the binary model considered previously: one has stronger variability than assumed in the model M and therefore models overdispersion. But this has not to hold for all values of r. For example, if $k = 10$, one obtains for $r = 1$ shrinkage toward 0.1. If $p_{M,r}(x)$ is larger than 0.1, then the shrinkage toward 0.1 means that the variance is smaller than in the model without a uniform mixture component. Therefore, in terms of the cumulative probabilities one might model underdispersion in the sense that the mixture model allows to model smaller variance than the pure model with $\pi = 1$.

Although estimation procedures will be considered later, we consider a small example to illustrate the shrinkage effect. Table 1 shows data that have been analysed previously by Mehta et al. (1984). For patients with acute rheumatoid arthritis a new agent was compared with an active control. Each patient was evaluated on a five-point assessment scale ranging from ”much improved” to ”much worse.” Table 2 shows the corresponding estimates for the mixture model with a cumulative logit model as the structuring component CUP(c) and the simple cumulative logit model. It is seen that the effect strength is 0.291 for the cumulative model but 0.394 for the cumulative mixture model. Thus if the mixture component is omitted one obtains a weaker effect of treatment. The difference between effect strengths is rather large because the uniform distribution is included with the rather large probability 0.294.

TABLE 1: *Clinical trial of a new agent and an active control (Mehta et al., 1984)*

.

		Global Assessment			
Drug	Much Improvement	Improvement	No Change	Worse	Much Worse
New agent	24	37	21	19	6
Active control	11	51	22	21	7

To further investigate the bias of the estimate if the mixture component is neglected we give the results of a small simulation study. Let the data be generated from a mixture model with the cumulative model in the mixture given by

TABLE 2: *Model fits for arthritis data with explanatory variable drug; fitted models are mixture with a cumulative model (CUP(c)) and a simple cumulative model without uncertainty*

	CUP(c)	Cumulative Model
Intercept:1	-1.945	-1.802
Intercept:2	0.371	0.115
Intercept:3	1.385	1.008
Intercept:4	7.668	2.631
drug	0.394	0.291
Prob(uniform)	0.294	0

the model fitted for the clinical trial data. We use the thresholds given in Table 2 with effect strength 0.5 and vary the probability $1-\pi$, that is, the probability of uncertainty in the mixture. The left panel of Figure 4 shows the estimated parameters when a cumulative CUP model is fitted. The true parameter value is included as a horizontal line. It is seen that the estimates are almost unbiased. Of course, a small bias is not surprising for an ML estimate with finite sample size. Overall the estimation works well with increasing variability if the uncertainty component gets stronger. The results change dramatically if one fits a cumulative model and therefore ignores the uncertainty component (right panel of Figure 4). It is seen that the true parameter is strongly underestimated with the bias getting stronger with increasing importance of the uncertainty component.

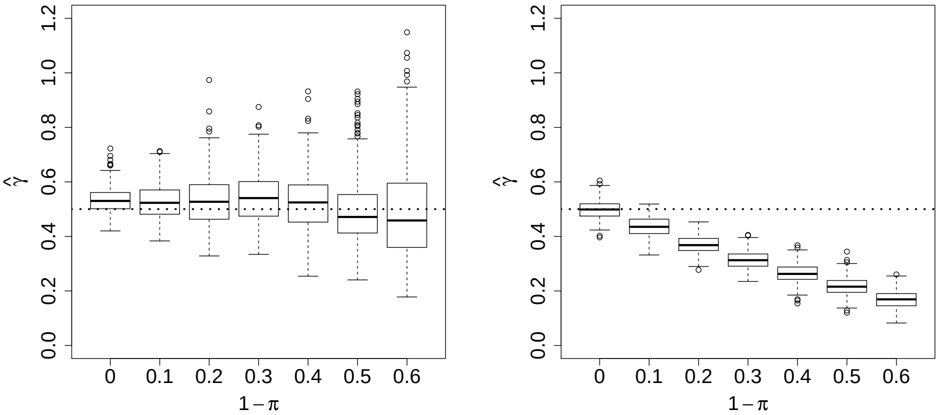

FIGURE 4: *Simulation with the data generating model being a cumulative mixture model; left panel shows the parameter estimates when a cumulative mixture model is fitted, right panel shows the estimates if a simple cumulative model is fitted.*

The main point of the illustrations is that when the data generating model is a mixture model of the form considered here one tends to underestimate the effects of explanatory variables. The effect is similar to what is found in binary (and ordinal) random intercept models. Let repeated measurements on individual i be given by $y_{i1}, \dots, y_{im}$, $y_{it} \in \{0, 1\}$ with covariates $\boldsymbol{x}_i$. Then the random intercept model assumes $P(y_{it} = 1|\boldsymbol{x}_i, b_i) = h(b_i + \boldsymbol{x}_i^T \boldsymbol{\gamma})$, where $h(.)$ is a response function and b_i is a subject-specific random effect, typically assumed to be normally distributed, $b_i \sim N(0, \sigma_{\boldsymbol{b}}^2)$. The parameter $\boldsymbol{\gamma}$ contains the *conditional* effect of the explanatory variable *given the random effect* b_i. If one considers the *marginal* model $P(y_{it} = 1|\boldsymbol{x}_{it}) = \int P(y_{it} = 1|\boldsymbol{x}_{it}, b_i)p(b_i)db_i$, effects tend to be weaker, see, for example, Caffo et al. (2007). Because the models are non-linear the omission of the random effects yields estimates of parameters that are closer to zero than the actual parameters. Marginal effects are attenuated as compared to the conditional effects.

Interpretation of the parameters in the mixture model is not so straightforward but effects can be interpreted in a similar way as conditional effects in random effects models. Let us consider the proportional odds model for the structured response in the mixture model. Let C denote the latent class; $C = 1$ denotes that the choice made by the individual is deliberate and determined by the proportional odds model M; $C = 0$ means that the choice is made in random mode, determined by the uniform distribution. The mixture is determined by the weights π, $1 - \pi$. Then the parameters of the proportional odds model determine the response given $C = 1$, that is, $P(Y \le r|\boldsymbol{x}, C = 1) = p_M(\boldsymbol{x}) = \exp(\gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma})/(1 + \exp(\gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma}))$. If one compares two individuals that differ in the variable x by one unit one obtains for the cumulative odds given $C = 1$

$$\frac{P(Y \le r|x+1, C=1)/P(Y > 0|x+1, C=1)}{P(Y \le r|x, C=1)/P(Y > 0|x, C=1)} = \exp(\gamma).$$

Thus the parameter contains the effect of explanatory variable x given both individuals make a deliberate choice, that is, $C = 1$. In that sense the effect is conditional on the action mode $C = 1$. Given this action mode the interpretation is the same as in the common proportional odds models. Ignoring the uncertainty component yields attenuated effects.

4 An Illustrative Example

To illustrate the effects in a cumulative CUP we consider data from the Survey on Household Income and Wealth (SHIW) by the Bank of Italy, for earlier use of the data see (Gambacorta and Iannario, 2013). In the analysis presented in Table 3 the response is the happiness index indicating the overall life well-being measured on a Likert Scale from 1 (very unhappy) to 10 (very happy). As covariates the following factors were chosen: the marital status, the place of living, the general degree of confidence in other people (1 to 10), the atmosphere the interview took place in (1 (low) to 10 (high)), the citizenship and the age. The respondents were also asked about their assessment if the household income is sufficient to see the family through to the end of the month rated from 1 (with great difficulty) to 5 (very easily). The analysis is based on a subset with 3816 respondents of the SHIW of 2010. We fitted a cumulative CUP model with explanatory variables in the cumulative part as well as in the logistic model that determines the mixture probability. In addition we fitted a simple cumulative model without a mixture component and the CUB model. The standard errors of the coefficients are obtained by 500 non-parametric bootstrap samples (Efron and Tibshirani, 1994). The results are given in Table 3.

Covariates	CUP(c)		Cumulative		CUB	
	est.	BS.se	est.	se	est.	BS.se
Constant (β_0)	0.375	0.146			0.419	0.460
Marital status: Single	0.579	0.182			0.604	0.214
Marital status: Separated	0.866	0.192			1.224	0.256
Marital status: Widow	0.954	0.177			1.261	0.212
Living: Centre of Italy	0.809	0.171			1.039	0.177
Living: South of Italy	0.425	0.132			0.487	0.156
Confidence in people	0.092	0.024			0.097	0.025
Interview atmosphere	-0.162	0.028			-0.185	0.054
Marital status: Single	1.208	0.173	0.356	0.089	0.460	0.066
Marital status: Separated	1.340	0.178	0.276	0.108	0.509	0.066
Marital status: Widow	1.442	0.168	0.327	0.085	0.567	0.057
Living: Centre of Italy	-0.585	0.140	-0.762	0.075	-0.240	0.050
Living: South of Italy	0.347	0.127	-0.087	0.068	0.124	0.047
Confidence in people	-0.107	0.044	-0.080	0.012	-0.041	0.012
Income sufficient	-0.301	0.050	-0.094	0.024	-0.110	0.017
Interview atmosphere	-0.277	0.044	-0.092	0.020	-0.094	0.014
Citizenship: Foreign	0.845	0.368	0.243	0.153	0.342	0.123
Age (centered)	0.019	0.005	0.004	0.002	0.006	0.002
Probability(uniform)	0.464		0		0.458	

TABLE 3: *Parameter estimates and standard errors based on bootstrap for the SHIW study.*

It is seen that the uncertainty component is very strong with $1 - \bar{\pi} = 0.458$ for the CUB model and $1 - \bar{\pi} = 0.464$ for the cumulative mixture model, where $\bar{\pi} = 1/n \sum_{i=1}^{n} 1/(1 + e^{-\boldsymbol{z}_i^T \boldsymbol{\beta}})$ is the mean value over all the probabilities of the observations. In the following tables $1 - \bar{\pi}$ is always denoted by Prob(uniform). As in the previous example it is seen that the estimated effects in the cumulative model part of the mixture model are much stronger than the effects found in the simple cumulative model. We only included effects that have been found to be influential in previous studies (Gambacorta and Iannario (2013)) and do not give the threshold parameters.

A tool that has also been used for CUB models is the visualization of effects of explanatory variables that are included in the preference part of the model and also determine the uncertainty. However, alternative specifications are needed to link the effects on the preference part with the effects on uncertainty. Therefore, a specific form of the cumulative logistic model that determines the preference component of the mixture has to be used. A form of the model that allows for easy interpretation of the effects on the response is

$$\frac{P(Y \leq r|\boldsymbol{x})}{P(Y > r|\boldsymbol{x})} = \exp(\gamma_{0r}) \exp(\boldsymbol{x}^T \boldsymbol{\gamma}) = e^{\gamma_{0r}} (e^{\gamma_1})^{x_1} \dots (e^{\gamma_p})^{x_p}, \quad r = 1, \dots, k-1.$$

That means that the odds of preferring categories $\{1, \dots, r\}$ over categories $\{r+1, \dots, k\}$ are modified by the factor e^{γ_j} if the jth variable is increase by one unit. It is important that the factor is the same for all categories and therefore characterizes the effect of the covariates in a unique way. This is essentially the proportional odds assumption that gives the model its name, see McCullagh (1980) for an extensive discussion of the proportional odds property. Since e^{γ_j} contains the effect of the preference part it can be used to visualize the effect together with the uncertainty, which is contained in the model $\text{logit}(\pi) = \boldsymbol{z}^T \boldsymbol{\beta}$. In Figure 5 the factor e^{γ_j} is plotted against the strength of the uncertainty $1 - \pi_j$ for the explanatory variables marital status and area of living. To obtain a scale for the uncertainty the other variables are set to fixed values, in particular, confidence and atmosphere are set to category 1, income to 3 and all other variables to zero. Since in the cumulative model large values of $\exp(\boldsymbol{x}^T \boldsymbol{\gamma})$ indicate preference for low response categories, large values indicate unhappiness. It is seen from Figure 5 that the marital status "widow" corresponds to high values of unhappiness and high certainty (small $1 - \pi_j$). In contrast, the status "married" indicates happiness but a large amount of uncertainty in the response. From the plot for the variable area it is seen that the people living in the north have large uncertainty and medium happiness whereas people from the south tend to categories that indicate unhappiness with a middle level of uncertainty.

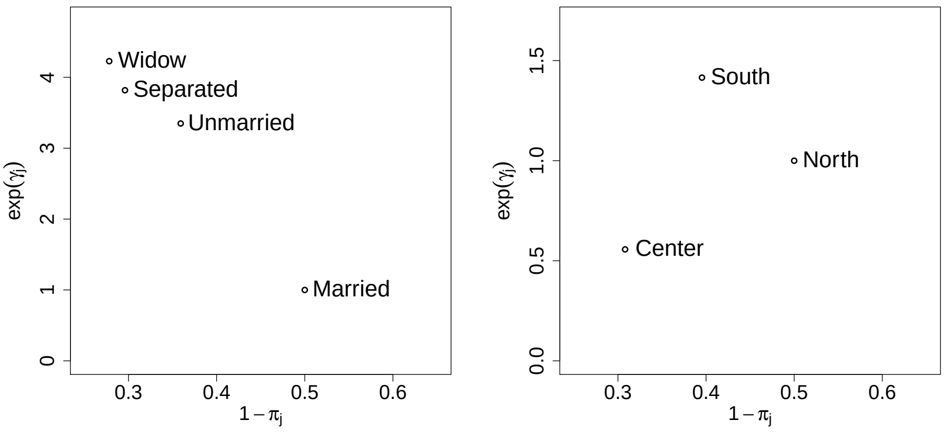

Figure 5: *Effects of the categorical covariates marital status (left) and area of living (right) in the structure and uncertainty component*

5 Comparison of Models

In this section we consider the usefulness of including the uncertainty component in the traditional cumulative models and also compare CUP and CUB proposals by use of real data sets. First we briefly discuss criteria for the comparison of models.

5.1 Criteria

Comparison of models is not straightforward since in general the models are not nested. But even for nested models, for example, when comparing a cumulative mixture model and a pure cumulative model, one can not simply use likelihood ratio tests because one is at the boundary of the parameter space. So one cannot expect the likelihood ratio tests to have the usual $\chi^2_{(1)}$-distribution, compare Böhning et al. (1994).

Alternative are information criteria as the AIC and BIC given by

$$AIC = -2l(\hat{\boldsymbol{\theta}}) + 2m\,; \qquad BIC = -2l(\hat{\boldsymbol{\theta}}) + m\,log(n),$$

where m is the number of model parameters, n is the number of observations and $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood function computed at the maximum of the estimated parameter vector $\boldsymbol{\theta}$. Information criteria are in common use in mixture models although no strong foundation seems available. Leroux (1992) gave some justification for the use of information criteria but it refers to very special cases only. Therefore alternative ways to compare models seem warranted.

A more data driven strategy is the evaluation of the predictive performance. In particular we will consider the deviance as a measure of the discrepancy between data and fit. For the multinomially distributed response one can distinguish two cases. One can group all observations for a fixed value of the explanatory variables obtaining the distribution $\boldsymbol{r}_i^T = (r_{i1}, \dots, r_{ik}) \sim \mathrm{M}(n_i, \boldsymbol{p}_i)$, $i = 1, \dots, N$, where N is the number of distinct values of the explanatory variables, n_i is the number of observations for the i-th value of the explanatory variables. The true underlying probabilities are $\boldsymbol{p}_i^T = (p_{i1}, \dots, p_{ik})$ and the corresponding estimates without assuming a model are the relative frequencies $(f_{i1}, \dots, f_{ik})$. Then the deviance for the multinomial distribution has the general form

$$D = 2\sum_{i=1}^{N} n_i \sum_{r=1}^{k} f_{ir} \log\left(\frac{f_{ir}}{\hat{p}_{ir}}\right).$$

In this grouped form it uses that n_i observations are available for a fixed value of the explanatory variable and for GLMs asymptotic distributions are available for $(n_i/N \to \lambda_i \in (0,1))$ (Fahrmeir and Tutz (2001)). If one does not group data, but works with single observations one uses $\boldsymbol{r}_i^T \sim \mathrm{M}(1, \boldsymbol{p}_i)$, $i = 1, \dots, n$ and obtains the form

$$D = 2\sum_{i=1}^{n}\sum_{l=1}^{k} r_{il}\log\left(\frac{r_{il}}{\hat{p}_{il}}\right) = -2\sum_{i=1}^{n}\log(\hat{p}_{iR_i}),$$

where R_i denotes the observation in the categories, that is, $R_i \in \{1,\ldots,k\}$.

In both forms, grouped or un-grouped, the deviance measures the discrepancy between data and fit. It can be as a predictive measure. Let the data be split into a learning set and a validation set. The model is fitted on the learning set and then one computes the deviance for all the observations in the validation set $(R_i^{(V)}, \boldsymbol{x}_i^{(V)})$, $i = 1,\ldots,n_V$. In the un-grouped form one obtains for the averaged deviance

$$D/n_V = -2\sum_{i=1}^{n_V}\log(\hat{p}_{iR_i^{(V)}})/n_v,$$

where $\hat{p}_{il}$ is the estimated probability of category l at value $\boldsymbol{x}_i^{(V)}$. It is also known as the logarithmic score. A criticism of scores like the logarithmic score is that the predictive distribution $\hat{\boldsymbol{p}}$ is only evaluated at the value of the observation. Therefore, it takes not the whole predictive distribution into account. In the case of an ordinal response measures that make use of the whole predictive distribution can be derived from the continuous ranked probability score approach discussed by Gneiting and Raftery (2007). For categorical responses one obtains the averaged value

$$L_{RPS}/n_V = \sum_{i=1}^{n_V}\sum_{r}(\hat{p}_i(r) - I(R_i \le r))^2/n_v, \tag{5}$$

where $\hat{p}_i(r) = \hat{p}_{i1} + \cdots + \hat{p}_{ir}$ is the estimated cumulative probability at value $\boldsymbol{x}_i^{(V)}$ and $I(.)$ is the indicator function. It is a sum over quadratic (or Brier) scores for binary data and takes the closeness between the whole estimated distribution and the observed value into account. For a discussion of measures for the closeness of data and fit see also, with the focus on categorical data, Tutz (2012), Chapter 15.

5.2 Empirical Studies

The models that are used in the applications are

- the cumulative model (without uncertainty),
- CUP(c): the cumulative model with uncertainty component,
- the adjacent categories model model (without uncertainty),
- CUP(a): the adjacent categories with uncertainty component,
- CUB: the binomial with uncertainty component.

Income and Wealth

For the Survey on Household Income and Wealth (SHIW) considered in the previous section the performance measures for selected models are given given in Table 4. It is seen that the cumulative mixture model performs best in terms of AIC, BIC, deviance and logScore. The ranked score is the same for CUP(c) and CUP(a). The relevance of the mixture component is underlined by the strong reduction of AIC; the value of the AIC for the mixture model (16218) is much smaller than the AIC for the simple cumulative model (16472). The same reduction is found when an uncertainty component is included in the adjacent categories model. For the ranked score the performance of all models is very similar. Therefore, in this application the cumulative mixture model with a substantial amount of uncertainty is to be preferred.

Covariates	CUP(c)	Cumulative	CUP(a)	Adjacent	CUB
Probability(uniform)	0.464	0	0.491	0	0.458
Deviance	16164	16434	16185	16497	16311
AIC	16218	16472	16239	16535	16349
BIC	16387	16591	16408	16654	16467
logScore	4.250	4.307	4.256	4.323	4.284
RankedScore	1.306	1.310	1.306	1.311	1.307

TABLE 4: *Results for the SHIW study*

PLUS Study

In the Participation, Labour and Unemployment Survey (PLUS) carried out by ISFOL (Institute for training of workers, Ministry of Labour and Welfare, Italy), the participants were asked to rate their probability to reach the age of 75. They chose a value between 0 for a impossible event and 100 for a certain event. Because of rounding effects ordered categories instead of the observed continuous values are to be preferred as suggested by Iannario and Piccolo (2010): see Table 5. The data consists of 20,184 individuals from the survey wave of 2006 and includes several more covariates such as gender (1: female, 0: male), age, marital status (widowed, divorced, married/single) and employment status (1: worker, 0: noworker). Table 6 shows the results with all of the explanatory variables. In this application the uncertainty component is rather weak ($1 - \bar{\pi} = 0.100$ for the cumulative mixture model, 0.099 for the adjacent categories mixture model and 0.137 for the CUB). Nevertheless the inclusion of the uncertainty component reduces the AIC and the BIC distinctly. It is seen that the cumulative and the adjacent categories mixture models perform better than the models without uncertainty and the CUB with regard to all performance measures.

Category	Expressed probability	Interpretation of the perception
1	$0.00 \leq \mathrm{Pr(S)} \leq 0.05$	Impossible/Almost Impossible
2	$0.05 < \mathrm{Pr(S)} \leq 0.25$	Low
3	$0.25 < \mathrm{Pr(S)} \leq 0.45$	Moderately Low
4	$0.45 < \mathrm{Pr(S)} \leq 0.55$	About Fifty/Fifty
5	$0.55 < \mathrm{Pr(S)} \leq 0.75$	Moderately High
6	$0.75 < \mathrm{Pr(S)} \leq 0.95$	High
7	$0.95 < \mathrm{Pr(S)} \leq 1.00$	Sure/Almost Sure

TABLE 5: *Qualitative assessment of subjective survival probabilities Pr(S)*

Covariates	CUP(c)		Cumulative		CUP(a)		Adjacent		CUB	
	est.	BS.se	est.	se	est.	BS.se	est.	se	est.	BS.se
Intercept(β_0)	1.811	0.085			1.823	0.082			1.511	0.078
Female	0.046	0.083			0.092	0.085			0.032	0.080
Age	-1.064	0.144			-1.044	0.132			-1.310	0.125
Age^2	2.713	0.344			2.550	0.334			2.534	0.350
Female	0.128	0.030	0.105	0.027	0.082	0.016	0.038	0.011	0.104	0.024
Divorced	0.308	0.103	0.276	0.079	0.159	0.052	0.100	0.032	0.246	0.076
Widowed	0.360	0.139	0.412	0.114	0.192	0.073	0.179	0.044	0.290	0.108
Work	-0.074	0.032	-0.049	0.028	-0.031	0.018	-0.013	0.012	-0.047	0.025
Age	-0.242	0.037	-0.085	0.033	-0.098	0.021	0.026	0.015	-0.226	0.033
Age^2	-0.716	0.099	-0.925	0.092	-0.426	0.056	-0.458	0.040	-0.559	0.085
Prob(uniform)	0.100		0		0.099		0		0.137	
Deviance	59601		59729		59612		59716		60381	
AIC	59633		59753		59644		59740		60403	
BIC	59759		59847		59771		59835		60490	
logScore	2.9545		2.9592		2.9551		2.9585		2.9927	
RankedScore	0.6726		0.6734		0.6727		0.6733		0.6757	

TABLE 6: *Results for the PLUS study*

Allbus

In the German General Social Survey ALLBUS data on behavior, attitudes and social structure in Germany are collected. 3480 persons answered the questionnaire in 2012. In the present study the respondents rated their trust in the health care system on a scale from 1 (not at all) to 7 (very much). In addition they give their assessment of the own state of health from 1 (very good) to 5 (poor) and their overall life satisfaction from 0 (very unhappy) to 10 (very happy). Other covariates are the responders age, net income (in 1000 Euros) and citizenship. The variable region specify if the interview took place in the former east part of Germany.

Table 7 shows the results for CUB and the CUP model using a cumulative model or an adjacent category model for the preference structure. It is again found that the inclusion of uncertainty reduces AIC and BIC strongly. Among the models with an uncertainty component there is not much difference in terms of AIC and BIC, BIC even favors the CUB. While the logscore profits from the inclusion of uncertainty the ranked score is very similar for all models.

Covariates	CUP(c)		Cumulative		CUP(a)		Adjacent		CUB	
	est.	BS.se	est.	se	est.	BS.se	est.	se	est.	BS.se
Intercept(β_0)	4.976	0.895			3.461	0.615			3.663	0.461
Poor Health	-1.051	0.222			-0.701	0.168			-0.749	0.144
German	1.116	0.236	0.850	0.162	0.598	0.135	0.321	0.072	0.542	0.111
Income	0.058	0.047	0.005	0.020	0.030	0.024	0.015	0.008	0.031	0.016
Age (centred)	-0.007	0.047	0.005	0.002	-0.004	0.001	-0.002	0.001	-0.004	0.001
Region: East	-0.372	0.093	-0.318	0.071	-0.208	0.050	-0.125	0.029	-0.190	0.042
Life Satisfaction	-0.193	0.034	-0.168	0.019	-0.104	0.021	-0.061	0.008	-0.092	0.015
Prob(uniform)	0.122		0		0.172		0		0.164	
Deviance	9925		9976		9928		9990		9942	
AIC	9951		9998		9954		10012		9958	
BIC	10029		10064		10032		10078		10005	
logScore	3.380		3.389		3.381		3.394		3.383	
RankedScore	0.751		0.752		0.751		0.752		0.751	

Table 7: *Model results for the Allbus data*

6 Concluding Remarks

It has been shown that the basic concept to include an uncertainty component in the model, as has been done in CUB models before, can be extended to the familiar classes of ordinal models yielding models that show better fit and better performance in terms of AIC, BIC and prognostic measures. If the uncertainty component is neglected the strength of the explanatory variables tends to be underestimated. An advantage of the models is that the effects of covariates can be easily visualized.

Acknowledgements. This work has been partially supported by FIRB2012 project (code RBFR12SHVV) at University of Perugia and the frame of Programme STAR (CUP E68C13000020003) at University of Naples Federico II, financially supported by UniNA and Compagnia di San Paolo. ISFOL survey data has been used under the agreement ISFOL/PLUS 2006/430.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd Edition.* New York: Wiley.

Agresti, A. (2013). *Categorical Data Analysis, 3d Edition.* New York: Wiley.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117–128.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society B 46*, 1–30.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics 46*, 373–388.

Breen, R. and R. Luijkx (2010). Mixture models for ordinal data. *Sociological Methods and Research 39*, 3–24.

Caffo, B., M.-W. An, and C. Rhode (2007). Flexible random intercept models for binary outcomes using mixtures of normals. *Computational Statistics & Data Analysis 51*, 5220–5235.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*, Volume 57. London: CRC Press.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models.* New York: Springer.

Gambacorta, R. and M. Iannario (2013). Measuring job satisfaction with cub models. *Labour 27*(2), 198–224.

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–376.

Greene, W. and D. Hensher (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Resesearch, Part B 39*, 681–689.

Grün, B. and F. Leisch (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification 25*, 225–247.

Iannario, M. (2012a). Hierarchical cub models for ordinal variables. *Communications in Statistics-Theory and Methods 41*, 3110–3125.

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

Iannario, M. (2012c). Preliminary estimators for a mixture model of ordinal data. *Advances in Data Analysis and Classification 6*, 163–184.

Iannario, M. and D. Piccolo (2010). Statistical modelling of subjective survival probabilities. *Genus 66*, 17–42.

Iannario, M. and D. Piccolo (2012). CUB models: Statistical methods and empirical evidence. In S. S. Kennett, R. (Ed.), *Modern Analysis of Customer Surveys: with applications using R*, pp. 231–258. New York: Wiley.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics 20*, 1350–1360.

Manisera, M. and P. Zuccolotto (2014). Modeling rating data with nonlinear cub models. *Computational Statistics & Data Analysis 78*, 100–118.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*, 109–127.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models.* New York: Wiley.

Mehta, C. R., N. R. Patel, and A. A. Tsiatis (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics 40*, 819–825.

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 in mub models. *Quaderni di Statistica 8*, 33–78.

Tutz, G. (2012). *Regression for Categorical Data.* Cambridge: Cambridge University Press.

Wedel, M. and W. DeSarbo (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification 12*, 21–55.

Appendix: Estimation

For simplicity, estimation of the mixture model is considered for the general case of m mixture components. Thus, the mass function of observation R given $\boldsymbol{x}$ has the form

$$f(r|\boldsymbol{x}) = \sum_{j=1}^{m} \pi_j f_j(r|\boldsymbol{x}, \boldsymbol{\gamma}_j), \tag{6}$$

$$\sum_{j=1}^{m} \pi_j = 1 \text{ and } 0 \leq \pi_j \leq 1,$$

where $f_j(r|\boldsymbol{x}, \boldsymbol{\gamma}_j)$ represents the density and π_j the mixing proportion of the jth component of the mixture. In the CUP model (3) one has only two components, $m = 2$, and the parameters $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ are reduced to only one parameter, namely $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}$. The second component is the uniform distribution, which is fixed and does not depend on covariates.

For given data $r_i|\boldsymbol{x}_i$, $i = 1, \ldots, n$, and collecting all parameters in the parameter $\boldsymbol{\theta}$, the log-likelihood to be maximized is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \pi_j f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j) \right).$$

Direct maximization of the log-likelihood is time-consuming and to obtain stable solutions starting values near the true values are required. A better way is to consider it as a problem with incomplete data and obtain estimates by using the EM algorithm. Therefore, let z_{ij} denote the unknown mixture components that indicate whether r_i belongs to component j

$$z_{ij} = \begin{cases} 1, \text{observation } r_i \text{ is from the } j\text{th mixture component} \\ 0, \text{otherwise.} \end{cases}$$

If the observation r_i is from the jth mixture component the vector $\boldsymbol{z}_i^T = (z_{i1}, \ldots, z_{im}) = (\ldots, 0, 1, 0, \ldots)$ contains only zeros, but one at the j-th position. One gets for one particular observation r_i

$$f(r_i|z_{ij} = 1, \boldsymbol{x}_i, \boldsymbol{\theta}) = f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j) = \prod_{l=1}^{m} f_l(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_l)^{z_{il}}.$$

Since $\boldsymbol{z}_i^T$ is multinomially distributed with probability vector $\pi^T = (\pi_1, \ldots, \pi_m)$ the complete density for $r_i, \boldsymbol{z}_i$ is

$$f(r_i, \boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = f(r_i|\boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{\theta}) f(\boldsymbol{z}_i|\boldsymbol{\theta}) = \prod_{j=1}^{m} f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j)^{z_{ij}} \prod_{j=1}^{m} \pi_j^{z_{ij}}$$

yielding the complete log-likelihood

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(f(r_i, \boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\left(\log(\pi_j) + \log(f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j)\right).$$

The EM algorithm treats z_{ij} as missing data and maximizes the log-likelihood iteratively by using an expectation and a maximization step. During the E-step the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{r}$ and the current estimate $\boldsymbol{\theta}^{(s)}$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_c(\boldsymbol{\theta})|\boldsymbol{r}, \boldsymbol{\theta}^{(s)})$$

has to be computed. Because $l_c(\boldsymbol{\theta})$ is linear in the unobservable data z_{ij}, it is only necessary to estimate the current conditional expectation of z_{ij}. From Bayes's theorem follows

$$\begin{aligned} E(z_{ij}|\boldsymbol{y}, \boldsymbol{\theta}) &= f(z_{ij} = 1|r_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= f(r_i|z_{ij} = 1, \boldsymbol{x}_i, \boldsymbol{\theta}) f(z_{ij} = 1|\boldsymbol{x}_i, \boldsymbol{\theta})/f(r_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_j f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\theta})/f(r_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_j f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\theta})/\sum_{l=1}^{m} \pi_l f_l(r_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \hat{z}_{ij}. \end{aligned}$$

This is the posterior probability that the observation r_i belongs to the $j-$th component of the mixture. For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{z}_{ij}^{(s)} \left(\log(\pi_j) + \log(f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j))\right) \\ &= \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{z}_{ij}^{(s)} \log(\pi_j)}_{M_1} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{z}_{ij}^{(s)} \log(f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j)}_{M_2}. \end{aligned}$$

Thus, for given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{z}_{ij}^{(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ (or rather M_1 and M_2), which yields the new estimates

$$\pi_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ij}^{(s)} \quad \text{and} \quad \boldsymbol{\gamma}_j^{(s+1)} = \mathrm{argmax}_{\boldsymbol{\gamma}_j} \sum_{i=1}^{n} \hat{z}_{ij}^{(s)} \log(f_j(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}_j)).$$

The E- and M-steps are repeated alternatingly until the difference $L(\boldsymbol{\theta}^{(s+1)}) - L(\boldsymbol{\theta}^{(s)})$ is small enough to assume convergence. Computation of $\boldsymbol{\gamma}_j^{(s+1)}$ can be based on familiar maximization tools, because one maximizes a weighted log-likelihood with known weights. In the case where only intercepts are component-specific, the derivatives are very similar to the score function used in a Gauss-Hermite quadrature and a similar EM algorithm applies with an additional calculation of the mixing distribution $\{\pi_1, \ldots, \pi_m\}$ (see Aitkin (1999)).

Dempster et al. (1977) showed that under weak conditions the EM algorithm finds a local maximum of the likelihood function $L(\boldsymbol{\theta})$. Hence it is sensible to use different start values $\boldsymbol{\theta}^{(0)}$ to find the solution of the maximization problem.

The log-likelihood of models with categorical response as proposed in the previous section is computed over the sum of k categories. For the M-step of the mixture of j categorical models follows

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{z}_{ij}^{(s)} \left(\log(\pi_j) + \sum_{r=1}^{k} \log(f_j(r|\boldsymbol{x}_i, \boldsymbol{\gamma}_j) \right),$$

where $\hat{z}_{ij}^{(s)} = \pi_j^{(s)} \sum_{r=1}^{k} f_j(r|\boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}) / \sum_{l=1}^{m} \pi_l^{(s)} \sum_{r=1}^{k} f_l(r|\boldsymbol{x}_i, \boldsymbol{\theta}^{(s)})$.
It is also possible to include covariates to determine the probability that observation i belongs to mixture component j according to characteristics of observation i. In this case π_j is replaced by

$$\pi_{ij} = 1/(1 + e^{-\boldsymbol{z}_i^T \boldsymbol{\beta}_j}).$$

Thus M_1 is the weighted log-likelihood of a multinomial logit model. M_1 and M_2 are maximized separately and provide the estimates. While in model (6) π_j is constant for all observations, now the π_{ij} vary depending on individual characteristics which gives the model more flexibility.