Electronic Journal of Applied Statistical Analysis
Vol. 00, Issue 00, August 2013, 1-35
DOI: DOI–xxxx

The CUB model and its variations are restricted loglinear latent class models

DL Oberski*[a] and JK Vermunt[a]

[a]*Dept of Methodology and Statistics, Tilburg University*

March 18, 2015

The "combination of uniform and shifted binomial" (CUB) model is a distribution for ordinal variables that has received considerable recent attention and specialized development. This article notes that the CUB model is a special case of the well-known loglinear latent class model, an observation that is useful for two reasons. First, we show how it can be used to estimate the CUB model in familiar standard software such as Mplus or Latent GOLD. Second, the mathematical equivalence of CUB with this well-known model and its correspondingly long history allows well-known results to be applied straightforwardly, subsuming a wide range of specialized recent developments of CUB and suggesting several possibly useful future ones. Thus, the observation that CUB and its extensions are restricted loglinear latent class models should be useful to both applied practitioners and methodologists.

1. Introduction

Ordinal variables are commonly found across a range of disciplines. They are especially common in the social sciences where they typically represent respondents' answers to questions in questionnaires–other examples being peoples' education levels, business size categories, disease progression levels, and so on. When such variables are used as outcomes to be predicted, their discrete and ordered nature must be taken into account in some manner. To do this, a range of models exist (see Agresti, 2002, for an overview); Piccolo (2003) and D'Elia and Piccolo (2005) introduced one such model, which they called the "combination of uniform and shifted binomial" (CUB) model.

*Corresponding author: doberski@uvt.nl.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

For an observed ordinal variable Y with K categories, the probability of observing $Y = k$ under the CUB model is a mixture of two components, or, equivalently, of two classes of a discrete latent variable X:

$$P(Y = k) = \pi P(Y = k|X = 1) + (1 - \pi)P(Y = k|X = 2). \tag{1}$$

The first class (mixture component), $X = 1$, follows the "shifted binomial",

$$P(Y = k|X = 1) = \binom{K-1}{k-1} \xi^{K-k}(1 - \xi)^{k-1} \tag{2}$$

and the second, $X = 2$, a uniform distribution over the K categories,

$$P(Y = k|X = 2) = \frac{1}{K}. \tag{3}$$

Thus, the CUB model has two parameters: π and ξ. Since the latent class variable X is often referred to in this literature as the "uncertainty" component, the proportion of observations in this class, $\mathrm{P}(X = 2) = 1 - \pi$, is referred to as "uncertainty". The ξ parameter is then referred to as the "feeling". Whether this terminology is substantively warranted will depend on the application and is beyond the scope of this article.

As also noted by Tutz et al. (2014, p. 5), the CUB model bears resemblance to a more general class of mixture models, in which the conditional probabilities can be modeled using standard models for ordinal variables familiar from GLM modeling (Agresti, 2002). Many such models can be subsumed in the loglinear latent class model (LCM),

$$P(Y = k|X = x) = \frac{\exp(\eta_{k|x})}{\sum_{k'=1}^{K} \exp(\eta_{k'|x})}, \tag{4}$$

where $\eta_{k|x}$ is the linear predictor for category k in class x. The uniform distribution in class 2 is obtained by setting $\eta_{k|2} = 0$; without covariates, $\eta_{k|1}$ would simply be an intercept parameter for each category. However, while the CUB model is identifiable without covariates (Iannario, 2010), this more general formulation is not. Underidentification can be easily verified by noting that the unrestricted model has K parameters but only $K - 1$ unique data patterns to estimate them from, leading to -1 degree of freedom. Tutz et al. (2014) did not impose further restrictions on $\eta_{k|1}$ but introduced covariates instead to resolve the identification issue.

The advantage of the approach suggested by Tutz et al. (2014) is that a standard modeling framework is used that allows practitioners to use standard software. Furthermore, technical development of the model is greatly facilitated by the ready availability of existing results on issues such as identification, variance estimation, model fit evaluation, and so on. A disadvantage of this approach, however, is that it is not mathematically equivalent to the CUB model and non-identifiable without covariates or other additional data. This paper shows that this last disadvantage can be removed by imposing certain restrictions on the standard loglinear latent class model in Equation 4. These restrictions lead to a mathematically equivalent reparameterization of the CUB model that can readily be estimated in standard software such as Mplus (Muthén and Muthén, 2012) or Latent GOLD (Vermunt and Magidson, 2013b,a).

2. Equivalence of loglinear LCM and CUB models

The loglinear latent class model in Equation 4 is an less restrictive version of the CUB model. This section demonstrates that fact by deriving the restrictions necessary to obtain a model that is mathematically equivalent to the CUB model.

First, as remarked above, the uniform distribution in class 2 is obtained simply by restricting all linear predictors in that class to zero, $\eta_{k|2} = 0$. The "shifted binomial" distribution in class 1, meanwhile, is obtained by setting the linear predictors proportional to the category number:

$$\eta_{k|1} = \beta(k-1) + \ln\binom{K-1}{k-1}, \tag{5}$$

where k is the category number, the final term is a normalization constant that does not depend on unknown parameters, and β is a reparameterization of the CUB location parameter ξ:

$$\xi = \frac{1}{1+\exp(\beta)}. \tag{6}$$

In some software it is also possible to specify a multiplicative offset w_k ("cell weight"), i.e. a factor by which $\exp(\eta_{k|x})$ is multiplied (e.g. Vermunt and Magidson, 2013b, p. 130–1); in that case the multiplicative offset is simply $w_k = \binom{K-1}{k-1}$.

2.1. Derivation

The derivation of the equivalence of CUB and LCM starts by observing that setting the log-probability in the CUB model's "shifted binomial" class 1 equal to a linear function of the category number k,

$$\begin{aligned}\ln P(Y=k|X=1) &= \ln\binom{K-1}{k-1} + (K-k)\ln\xi + (k-1)\ln(1-\xi) \\ &= a_k + \beta k,\end{aligned} \tag{7}$$

can always be solved exactly for a_k and β:

$$a_k = \ln\binom{K-1}{k-1} - (K-1)\ln(\xi), \tag{8}$$

$$\beta = \ln\left(\frac{1-\xi}{\xi}\right). \tag{9}$$

Note that the linear predictor can also be interpreted as the log-odds ratio of category k

versus a reference category. That is, choosing the first category as a reference ($\eta_{1|1} = 0$),

$$\begin{aligned}\eta_{k|1} &= \ln\left[\frac{P(Y=k|X=x)}{P(Y=1|X=x)}\right] \\ &= \ln P(Y=k|X=x) - \ln P(Y=1|X=x) \\ &= (a_k + \beta k) - (a_1 + \beta) \\ &= \beta(k-1) + \ln\binom{K-1}{k-1}.\end{aligned} \tag{10}$$

This proves that the restricted linear predictor in Equation 5 combined with the standard loglinear latent class model in Equation 4 is indeed mathematically equivalent to the CUB model.

2.2. Example

Consider a hypothetical observed variable with four categories. To estimate the CUB model without covariates using loglinear LCM, set

$$\begin{array}{llll} \eta_{1|1} = 0 & \eta_{2|1} = \beta + 1.0986 & \eta_{3|1} = 2\beta + 1.0986 & \eta_{4|1} = 3\beta \\ \eta_{1|2} = 0 & \eta_{2|2} = 0 & \eta_{3|2} = 0 & \eta_{4|2} = 0. \end{array}$$

Figure 1 demonstrates the implementation of the loglinear latent class specification of CUB models in two standard software packages. In Mplus syntax, shown on the left-hand side of Figure 1, this is achieved by creating an additional parameter with the `NEW` command (note that Mplus uses the last category as the reference). Latent GOLD allows the "ordinal" specification as well as a the multiplicative offset w_k (`~wei`), leading to the syntax on the right-hand side of Figure 1. For practitioners who wish to apply the syntax in Figure 1, some common values of the normalization constants for different numbers of categories are given in Appendix B.

3. Extensions of the CUB model

Recently, a number of extensions to the CUB model have been proposed. The following of these extensions are directly subsumed by the approach suggested here:

- Standard errors for CUB via analytic information (Piccolo, 2006);
- CUB models with covariates (Iannario, 2008);
- Hierarchical (random effect) CUB models (Iannario, 2012a);
- CUB with "shelter choice" (Iannario, 2012b);
- Latent class-CUB models (Grilli et al., 2014);
- Design-based inference for CUB in complex samples (Gambacorta et al., 2014).

For instance, LCM with covariates (Dayton and Macready, 1988; Huang and Bandeen-Roche, 2004) is already a standard feature of such software; the "shelter choice" yields an

Mplus syntax

```
VARIABLE:
  NAMES ARE Y;
  NOMINAL = Y;
  CLASSES = c (2);

ANALYSIS:   TYPE = MIXTURE;

MODEL:
%C#1%  ! Uniform
  [ Y#1@0 Y#2@0 Y#3@0 ];

%C#2%  ! Shifted Binomial
  [ Y#1 ] (eta11) ;
  [ Y#2 ] (eta21) ;
  [ Y#3 ] (eta31) ;

MODEL CONSTRAINT:
  NEW(b);

  eta11 =        - 3*b;
  eta21 = 1.0986 - 2*b;
  eta31 = 1.0986 -   b;
```

Latent GOLD syntax

```
variables
   dependent Y       ordinal;
   latent    Cluster nominal 2 coding=1;

equations
  Cluster <- 1;
  Y <- (a~wei) 1 | Cluster + (b) Cluster;
  a[2]={1 3 3 1};
```

Figure 1: CUB model for $K = 4$ formulated in Mplus and Latent GOLD syntax.

additional latent class in which a particular response is given with certainty; the LCM-CUB model yields an additional discrete latent variable (e.g. Hagenaars, 1990); design-based inference for complex sampling in latent class models is typically achieved using pseudo-ML (Skinner et al., 1989; Patterson et al., 2002; Asparouhov, 2005); and hierarchical models as well as analytic standard errors are available in the literature. Some examples of formulating these extended CUB models in Latent GOLD are given in Appendix A.

We did not study the following developments of CUB in sufficient detail to evaluate whether these are also mathematically equivalent or directly implementable in standard software:

- Bivariate (Corduas, 2011) or multivariate CUB using copulas (Andreis and Ferrari, 2013);
- The fit measures and residuals proposed by Iannario (2009); Di Iorio and Iannario (2012);
- Overdispersed CUB model (Iannario, 2013, 2014);
- "Nonlinear" CUB model (Manisera and Zuccolotto, 2014).

However, other solutions to the issues of multivariate modeling, overdispersion, and model fit evaluation than those proposed for CUB are readily available. For example, multivariate distributions with a wide range of dependencies can easily be modeled using loglinear models (Hagenaars, 1988), and fit measures and residuals for categorical LCM are available, including full- and limited information omnibus tests (e.g. Maydeu-Olivares and Joe, 2005), bivariate residuals (Vermunt and Magidson, 2013b; Oberski et al., 2013), score tests (Oberski et al., 2013), expected parameter change (Oberski and Vermunt, 2014), and sensitivity measures (Oberski and Vermunt, 2013; Oberski et al., frth).

4. Conclusion

This article showed that the CUB model, for which specialized software and developments have recently been proposed, is a restricted loglinear latent class model that falls within the standard framework adopted by commonly used software such as Mplus and Latent GOLD. This observation should prove useful for practitioners who wish to apply the CUB model, as well as for methodologists who are seeking to extend it. Some work remains concerning CUB extensions whose equivalence is not obvious (to the authors), but for which other solutions may be available.

The scope of this article was limited to showing that CUB is an LCM that has been highly restricted. It would not be surprising then to find in empirical applications that these restrictions lead to rejection of the CUB model. Moreover, since, without covariates, the identification of the latent classes $X = 1$ and $X = 2$ depends entirely on these restrictions (Iannario, 2010), their interpretation as representing a substantive "feeling" and "uncertainty" component may not always be tenable. In such cases, the more flexible model of Tutz et al. (2014) or a multiple indicator approach (e.g. Oberski, frth) may be more reasonable. Discovering the cases in which this may be an issue remains a topic for further study, however.

Acknowledgements

The first author was supported by the Netherlands Organization for Scientific Research (NWO) [Veni grant number 451-14-017] and the second author by [Vici grant number 453-10-002] from the same organization.

References

Agresti, A. (2002). *Categorical data analysis, 2nd ed.* Wiley-Interscience, New York.

Andreis, F. and Ferrari, P. (2013). On a copula model with cub margins. *Quaderni di Statistica*, 15.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural equation modeling*, 12(3):411–434.

Corduas, M. (2011). Modelling correlated bivariate ordinal data with cub marginals. *Quaderni di statistica*, 13(13).

Dayton, C. and Macready, G. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178.

D'Elia, A. and Piccolo, D. (2005). A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49(3):917–934.

Di Iorio, F. and Iannario, M. (2012). Residual diagnostics for interpreting cub models. *Statistica*, 72(2):163–172.

Gambacorta, R., Iannario, M., and Valliant, R. (2014). Design-based inference in a mixture model for ordinal variables for a two stage stratified design. *Australian & New Zealand Journal of Statistics*, 56(2):125–143.

Grilli, L., Iannario, M., Piccolo, D., and Rampichini, C. (2014). Latent class CUB models. *Advances in Data Analysis and Classification*, 8(1):105–119.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage, Newbury Park.

Hagenaars, J. A. P. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, 16(3):379–405.

Huang, G. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Iannario, M. (2008). A class of models for ordinal variables with covariates effects. *Quaderni di Statistica*, 10:53–72.

Iannario, M. (2009). Fitting measures for ordinal data models. *Quaderni di Statistica*, 11:39–72.

Iannario, M. (2010). On the identifiability of a mixture model for ordinal data. *Metron*, 68(1):87–94.

Iannario, M. (2012a). Hierarchical CUB models for ordinal variables. *Communications in Statistics-Theory and Methods*, 41(16-17):3110–3125.

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

Iannario, M. (2013). A finite mixture distribution for modelling overdispersion. In *Advances in Latent Variables-Methods, Models and Applications*.

Iannario, M. (2014). Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics-Theory and Methods*, 43(4):771–786.

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

Maydeu-Olivares, A. and Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables. *Journal of the American Statistical Association*, 100(471):1009–1020.

Muthén, L. K. and Muthén, B. (2012). *Mplus User's Guide, Seventh Edition*. Muthén & Muthén, Los Angeles, CA.

Oberski, D. (frth). The latent class MTMM model. *Psychological Methods*.

Oberski, D., Van Kollenburg, G., and Vermunt, J. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3).

Oberski, D. and Vermunt, J. (2013). A model-based approach to goodness-of-fit evaluation in item response theory. *Measurement: Interdisciplinary Research & Perspectives*, 11:117–122.

Oberski, D. and Vermunt, J. (2014). The Expected Parameter Change (EPC) for local dependence assessment in binary data latent class models. *Accepted for publication in Psychometrika*.

Oberski, D., Vermunt, J., and Moors, G. (frth). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest.

Patterson, B. H., Dayton, C. M., and Graubard, B. I. (2002). Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association*, 97(459):721–741.

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

Skinner, C., Holt, D., and Smith, T. (1989). *Analysis of Complex Surveys*. John Wiley & Sons, New York.

Tutz, G., Schneider, M., Iannario, M., and Piccolo, D. (2014). Mixture models for ordinal responses to account for uncertainty of choice.

Vermunt, J. and Magidson, J. (2013a). *LG-Syntax User's Guide: Manual for Latent GOLD 5.0 Syntax Module*. Statistical Innovations Inc., Belmont, MA.

Vermunt, J. K. and Magidson, J. (2013b). *Technical guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc., Belmont, MA.

A. Examples of CUB extensions in Latent GOLD

Please see the online appendix for this program input and data.

CUP-adjacent categories (Tutz et al., 2014)

```
variables
   dependent fltdpr ordinal;
   independent marsts nominal,  ppltrst, agea, hinctnta;
   latent      Cluster nominal 2 coding=1;

equations
  Cluster <- 1;

  fltdpr <- (a) 1 | Cluster + ppltrst Cluster +  marsts Cluster +
                        agea Cluster + hinctnta Cluster;

  a[1,]=0;
```

CUB regression with complex sampling weights (Gambacorta et al., 2014)

```
variables
   dependent fltdpr ordinal;
   samplingweight dweight;
   independent marsts nominal,  ppltrst, agea, hinctnta;
   latent      Cluster nominal 2 coding=1;

equations
  Cluster <- 1;

  fltdpr <- (a~wei) 1 | Cluster +  Cluster + ppltrst Cluster +  marsts Cluster +
                        agea Cluster + hinctnta Cluster;

  a[2]={1 3 3 1};
```

Shelter option (Iannario, 2012b)

```
variables
   dependent fltdpr ordinal;
   samplingweight dweight;
   independent  ppltrst;
   latent       Cluster nominal 3 coding=1; // Third class is shelter

equations
  Cluster <- 1;

  fltdpr <- (a~wei) 1 | Cluster +  (b0) Cluster + (b1) ppltrst Cluster +
       (s~nom) Cluster ; // Shelter option

   a[2]={1 3 3 1};

   b0[1,2]=0;
   b1[1,2]=0;

   // Shelter option is category 1
   s[1,]=0;
   s[1,4]=-99;
   s[1,5]=-99;
   s[1,6]=-99;
```

B. Normalization constants for different numbers of categories

Table 1 gives the normalization constant

$$\ln \binom{K-1}{k-1}$$

from Equation 5 needed for some commonly found numbers of categories.

	Number of categories K								
k	3	4	5	6	7	8	9	10	11
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.6931	1.0986	1.3863	1.6094	1.7918	1.9459	2.0794	2.1972	2.3026
3	0.0000	1.0986	1.7918	2.3026	2.7081	3.0445	3.3322	3.5835	3.8067
4		0.0000	1.3863	2.3026	2.9957	3.5553	4.0254	4.4308	4.7875
5			0.0000	1.6094	2.7081	3.5553	4.2485	4.8363	5.3471
6				0.0000	1.7918	3.0445	4.0254	4.8363	5.5294
7					0.0000	1.9459	3.3322	4.4308	5.3471
8						0.0000	2.0794	3.5835	4.7875
9							0.0000	2.1972	3.8067
10								0.0000	2.3026
11									0.0000

Table 1: The normalization constant for different numbers of categories K.