The relationship between CUB and loglinear models with latent variables

DL Oberski* and JK Vermunt*

*Dept of Methodology and Statistics, Tilburg University

December 2, 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.

©Università del Salento
ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index
For an observed ordinal variable \(Y\) with \(K\) categories (\(K\) is known), 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”,

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

with \(0 < \xi < 1\), 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: \(\pi\) and \(\xi\). Since the latent class variable \(X\) is often referred to in this literature as the “uncertainty” component, the proportion of observations in this class, \(P(X = 2) = 1 - \pi\), is referred to as “uncertainty”. The \(\xi\) parameter (or \(1 - \xi\)) 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 a 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:

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

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

$$\xi = \frac{1}{1 + \exp(\beta)}.$$  \hfill (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, shown in Equation 2, equal to a linear function of the category number $k$,

$$\ln P(Y = k|X = 1) = \ln \binom{K - 1}{k - 1} + (K - k) \ln \xi + (k - 1) \ln(1 - \xi)$$

$$= a_k + \beta k,$$

can always be solved exactly for $a_k$ and $\beta$:

$$a_k = \ln \binom{K - 1}{k - 1} - \ln(1 - \xi) + K \ln(\xi)$$

$$\beta = \ln \frac{1 - \xi}{\xi}.$$  \hfill (8)\hfill (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 ($\eta_{1|1} = 0$),

$$
\eta_{k|1} = \ln \left[ \frac{P(Y = k|X = 1)}{P(Y = 1|X = 1)} \right] \\
= \ln P(Y = k|X = 1) - \ln P(Y = 1|X = 1) \\
= (a_k + \beta k) - (a_1 + \beta) \\
= \beta (k - 1) + \ln \left( \frac{K - 1}{k - 1} \right) .
$$

(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

$$
\eta_{1|1} = 0 \quad \eta_{2|1} = \beta + 1.0986 \quad \eta_{3|1} = 2\beta + 1.0986 \quad \eta_{4|1} = 3\beta \\
\eta_{1|2} = 0 \quad \eta_{2|2} = 0 \quad \eta_{3|2} = 0 \quad \eta_{4|2} = 0.
$$

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
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 (restricted) loglinear latent class models.

- 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., 2015).

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.

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


Iannario, M. (2014). Modelling uncertainty and overdispersion in ordinal data. *Com-


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*.


A. Examples of CUB extensions in Latent GOLD

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

<table>
<thead>
<tr>
<th>CUP-adjacent categories (Tutz et al., 2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>variables</strong></td>
</tr>
<tr>
<td>dependent fltdpr ordinal;</td>
</tr>
<tr>
<td>independent marsts nominal, ppltrst, agea, hinctnta;</td>
</tr>
<tr>
<td>latent Cluster nominal 2 coding=1;</td>
</tr>
<tr>
<td><strong>equations</strong></td>
</tr>
<tr>
<td>Cluster &lt;- 1;</td>
</tr>
<tr>
<td>fltdpr &lt;- (a) 1</td>
</tr>
<tr>
<td>a[1,]=0;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CUB regression with complex sampling weights (Gambacorta et al., 2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>variables</strong></td>
</tr>
<tr>
<td>dependent fltdpr ordinal;</td>
</tr>
<tr>
<td>samplingweight dweight;</td>
</tr>
<tr>
<td>independent marsts nominal, ppltrst, agea, hinctnta;</td>
</tr>
<tr>
<td>latent Cluster nominal 2 coding=1;</td>
</tr>
<tr>
<td><strong>equations</strong></td>
</tr>
<tr>
<td>Cluster &lt;- 1;</td>
</tr>
<tr>
<td>fltdpr &lt;- (a_wei) 1</td>
</tr>
<tr>
<td>a[2]={1 3 3 1};</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shelter option (Iannario, 2012b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>variables</strong></td>
</tr>
<tr>
<td>dependent fltdpr ordinal;</td>
</tr>
<tr>
<td>samplingweight dweight;</td>
</tr>
<tr>
<td>independent ppltrst;</td>
</tr>
<tr>
<td>latent Cluster nominal 3 coding=1; // Third class is shelter</td>
</tr>
<tr>
<td><strong>equations</strong></td>
</tr>
<tr>
<td>Cluster &lt;- 1;</td>
</tr>
<tr>
<td>fltdpr &lt;- (a_wei) 1</td>
</tr>
<tr>
<td>a[2]={1 3 3 1};</td>
</tr>
<tr>
<td>a[3]={1 0 0 0}; // First category is &quot;shelter option&quot;</td>
</tr>
<tr>
<td>b0[1,2]=0;</td>
</tr>
<tr>
<td>b1[1,2]=0;</td>
</tr>
</tbody>
</table>
### B. Normalization constants for different numbers of categories

Table 1 gives the normalization constant

\[
\ln \left( \frac{K-1}{k-1} \right)
\]

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

<table>
<thead>
<tr>
<th>Number of categories $K$</th>
<th>$k$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.6931</td>
<td>1.0986</td>
<td>1.3863</td>
<td>1.6094</td>
<td>1.7918</td>
<td>1.9459</td>
<td>2.0794</td>
<td>2.1972</td>
<td>2.3026</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1.0986</td>
<td>1.7918</td>
<td>2.3026</td>
<td>2.7081</td>
<td>3.0445</td>
<td>3.3322</td>
<td>3.5835</td>
<td>3.8067</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1.3863</td>
<td>2.3026</td>
<td>2.9957</td>
<td>3.5553</td>
<td>4.0254</td>
<td>4.4308</td>
<td>4.7875</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1.6094</td>
<td>2.7081</td>
<td>3.5553</td>
<td>4.2485</td>
<td>4.8363</td>
<td>5.3471</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1.7918</td>
<td>3.0445</td>
<td>4.0254</td>
<td>4.8363</td>
<td>5.5294</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1.9459</td>
<td>3.3322</td>
<td>4.4308</td>
<td>5.3471</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2.0794</td>
<td>3.5835</td>
<td>4.7875</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2.1972</td>
<td>3.8067</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2.3026</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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