Lecture Notes in Statistics

Latent budget analysis

Peter G.M. van der Heijden*, Ab Mooijaart* and Jan de Leeuw**
***Department of Methodology and Psychometrics, University of Leiden, The Netherlands**
****Department of Psychology and Department of Statistics, U.C.L.A., CA**

1. Introduction

In this paper we will present a class of latent models for the analysis of two-way contingency tables. The basic model was discussed by Clogg (1981) as a reparametrization of the latent class model for social mobility tables. Independently, this model was proposed by De Leeuw and van der Heijden (1988) and van der Heijden (1987) for the analysis of time budget data. In the present paper we discuss the model in the more general context of contingency tables. In section 2 properties of this model are summarized, and the objective of the model is made clear. In section 3 relations with latent class analysis and correspondence analysis are discussed. In section 4 we extend the model for the analysis of higher-way contingency tables.

2. The latent budget model

Consider a two-way table with probabilities π_{ij} with rows i (i=1,...,I) and columns j (j=1,...,J). Assume that the row and column variable play a different role, for example, the row variable is the explanatory variable and the column variable the response variable. An example is shown in table 1 (see Srole, 1962). The rows stand for parental socioeconomic statuses from A (high) to F (low), the columns for mental health statuses from "well" to "impaired". Consider the following question: how can we characterize the rows in terms of a weighted sum of *typologies* defined by the column categories? Due to the *asymmetric* role of the variables we are interested in conditional probabilities π_{ij}/π_{i+}, i.e. the probabilities that, given someones parent has socioeconomic status i, he/she will have mental health status j. We will refer to a row of values π_{ij}/π_{i+} as a *budget*, since it shows how persons having parental socioeconomic status i are distributed in proportions that add up to 1 over the J mental health statuses in proportions that add up to 1 (compare financial budgets, where money can be spent only once).
The latent budget model aims to approximate the I observed budgets by a weighted sum of K latent budgets, indexed by k. We will denote *theoretical* budget i with vector $\boldsymbol{\pi}_i$, *observed* budget i with vector $\mathbf{p}_i$, and *latent* budget k with $\boldsymbol{\beta}_k$. The elements of each budget add up to one. The model for the theoretical budgets $\boldsymbol{\pi}_i$ is

$$\pi_{ij}/\pi_{i+} = \Sigma_{k=1}^{K} \alpha_{ik}\beta_{jk}, \qquad (1)$$

Table 1: Srole (1962) data with LBA solution. Row variable is "parental socioeconomic status stratum" with levels A to F, and column variable is "mental health category" with levels 'I=well', 'II=mild symptom formation', 'III=moderate symptom formation' 'IV=impaired'.

		β_{j1}	.29	.39	.21	.11	1.00	
		β_{j2}	.07	.34	.23	.36	1.00	
α_{i1}	α_{i2}		**I**	**II**	**III**	**IV**		**n**
.76	.24	**A**	.24	.36	.22	.18	1.00	262
.78	.22	**B**	.23	.38	.22	.16	1.00	245
.60	.40	**C**	.20	.37	.23	.21	1.00	287
.50	.50	**D**	.19	.37	.20	.24	1.00	384
.29	.71	**E**	.14	.37	.20	.29	1.00	265
.11	.89	**F**	.10	.33	.25	.33	1.00	217
		p_{+j}	.19	.36	.22	.23		

with restrictions

$$\alpha_{i+} = 1 \tag{2a}$$
$$\beta_{+k} = 1 \tag{2b}$$
$$0 \leq \alpha_{ik} \leq 1 \tag{2c}$$
$$0 \leq \beta_{jk} \leq 1. \tag{2d}$$

This model approximates the observed budgets $\mathbf{p}_i$ by K latent budgets $\boldsymbol{\beta}_k$. The parameters α_{ik} show how each theoretical budget $\boldsymbol{\pi}_i$ is built up from the $\boldsymbol{\beta}_k$. Due to restrictions (2a) and (2c), α_{ik} can be interpreted as a probability. Just like the $\boldsymbol{\pi}_i$, each latent budget $\boldsymbol{\beta}_k$ is also built up from probabilities β_{jk} that add up to one.

When we assume that the observations are collected under the multinomial, the product-multinomial or the Poisson distribution, we can use the EM-algorithm to obtain ML estimates. Let observed proportions be denoted by p_{ij}. The current estimates are denoted by $\underline{\alpha}_{ik}$ and $\underline{\beta}_{jk}$, giving current theoretical budgets $\underline{\pi}_{ij}/\underline{\pi}_{i+}$ using (1). Start with trial values for $\underline{\alpha}_{ik}$ and $\underline{\beta}_{jk}$. Then calculate the first cycle of the algorithm

$$p_{ijk} = p_{ij}\,\underline{\alpha}_{ik}\,\underline{\beta}_{jk}/(\underline{\pi}_{ij}/\underline{\pi}_{i+}) \tag{3}$$
$$\alpha^{+}_{ik} = p_{i+k}/p_{i++} \tag{4a}$$
$$\beta^{+}_{jk} = p_{+jk}/p_{++k} \tag{4b}$$

Since $\Sigma_k \underline{\alpha}_{ik}\,\underline{\beta}_{jk}/(\underline{\pi}_{ij}/\underline{\pi}_{i+}) = 1$, the observed proportions p_{ij} are distributed in (3) over K layers, hence $p_{ij}=p_{ij+}$. For the next iteration go to (3), replace $\underline{\alpha}_{ik}$ and $\underline{\beta}_{jk}$ by the new estimates obtained in (4a) and (4b), and start the second cycle. After each cycle the current fit can be evaluated using the likelihood ratio chi-square statistic G^2 where $np_{i+}\Sigma_k\alpha^{+}_{ik}\beta^{+}_{jk}$ gives current estimates of expected frequencies, n being the sample size. The model has (I-K)(J-K) degrees of freedom. When the difference between subsequent G^2 values is smaller than a prespecified criterion, iterating can stop, and we can test the fit of the model using the current chi-square value. In de Leeuw & van der Heijden (1988) it is proved that the algorithm

converges. It can converge to a non-global maximum. In order to be sure that one has reached a global maximum, and not a local maximum, different sets of trial values should be used. In the sequel we will not introduce special notation to distinguish theoretical values from estimates of theoretical values: it will be clear from the context what is meant.

For the data in table 1, for K=1 G^2=47.42 (df=15), and for K=2 G^2=2.75 (df=8). The observed budgets of the I parental socioeconomic status groups are adequately approximated by estimated budgets using 2 latent budgets. Though the solution is not identified, we will interpret it, because substantively the interpretation will not change in different solutions. The estimated budget π_1 and π_2 for persons in group A and B is built up for .76 and .78 from the first latent budget β_1, and for .24 and .22 from β_2, whereas at the other extreme π_6 for group F is built up for .11 from β_1 and .89 from β_2. In order to understand the difference between these groups, we have to study the latent budgets β_k. These are most easily studied by comparing proportions β_{jk} with the marginal proportions p_{+j}. This shows that β_1 specifies the situation that one is far more likely than average to have the status "well" and far less likely than average to have the status "impaired", whereas the parameters $\hat{\beta}_{j1}$ for the statuses inbetween do not differ much from the average. On the other hand, for β_2 we find the opposite: there it is less likely than average to have status "well" and more likely than average to have the status "impaired". By relating the interpretation of the β_k to the group parameters α_{ik}, we conclude that, in going from groups A and B to F, persons become much more likely to be "impaired", and much less likely to be "well"; on the whole the mental health status proportions β_{2k} and β_{3k} do not differ much from their averages p_{+2} and p_{+3}.

In order to find a parsimonious description of the data, the objective is to describe the observed budgets $\mathbf{p}_i$ by as few latent budgets β_k as possible. For the example we needed only two latent budgets. However, it could have been that we needed K=3 latent budgets, for example, one "being more than average impaired", one "being more than average well" and one "being more than average in between". On the other hand, it could also have been that each $\mathbf{p}_i$ was described by K=1 β_k only. In that case all expected budgets π_i would be identical, and the latent budget model would be identical to the independence model: then α_{i1}=1 and $\beta_{j1}=\pi_{+j}$.

Identification

The latent budget model is not identified. This is easy to see by rewriting (1) in matrix terms as

$$\Pi = \mathbf{A}\mathbf{B}' \tag{5}$$

where Π is of order IxJ, **A** of order IxK, and **B** of order JxK. Using a square non-singular matrix **T** we find

$$\Pi = (\mathbf{A}\mathbf{T})(\mathbf{T}^{-1}\mathbf{B}') = \mathbf{A}^*\mathbf{B}^{*\prime}$$

We obtain the same expected budgets, finding different parameters in $\mathbf{A}^*$ and $\mathbf{B}^*$. This was already noticed by Clogg (1981), and studied in detail by De Leeuw et al. (1989). Here we summarize their

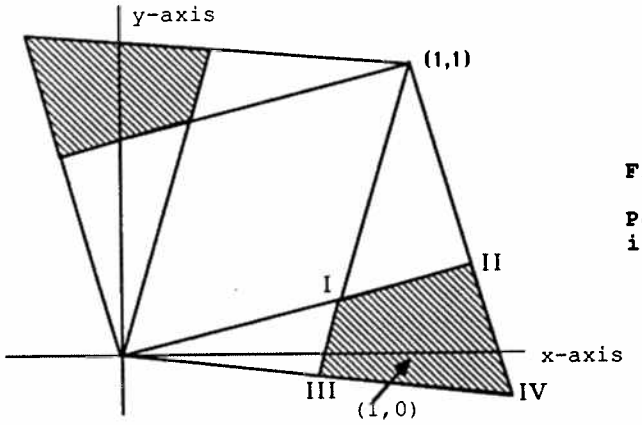

Figure 1:

Permissible region is striped

results. De Leeuw et al. show that T cannot be any arbitrary square nonsingular matrix since the parameter estimates in **A*** and **B*** should fulfil restrictions (2a) to (2d). If $\mathbf{uT} = \mathbf{1}$, where **u** is a unit column vector, then **T** fulfils (2a) and (2b). For K=2 restrictions (2c) and (2d) can be displayed graphically in the following way. Let elements of **T** be t_{11}=x, t_{12}=1-x, t_{21}=y, t_{22}=1-y. Then we can set out the permissible elements for **T** in a two dimensional (x,y) plane using restriction (2c) for **A***=**AT** and restriction (2d) for $\mathbf{B}^* = \mathbf{B}(\mathbf{T}^{-1})'$. For example, when we use the estimates in table 1 for **A** and **B**, then figure 1 shows the two regions with permissible (x,y) values. We can restrict attention to the lower right area since the upper left area corresponds to the situation that the parameters for k=1 and k=2 are interchanged. Notice that (x,y)=(1,0) is necessarily included since then **T**=**I**. In figure 1 each of the four lines circumventing the permissible region corresponds with a parameter set to zero. In the corner points two parameters are set to zero. In I $\beta_{41}^*=\beta_{12}^*=0$, i.e. $\boldsymbol{\beta}_1^*$ is made as distinct as possible from $\boldsymbol{\beta}_2^*$, making the column vector of α_{i1}^*-parameters $\boldsymbol{\alpha}_1^*$ less distinct from $\boldsymbol{\alpha}_2^*$. For IV $\alpha_{61}^*=\alpha_{12}^*=0$, and we find the reverse situation that $\boldsymbol{\alpha}_1^*$ is made as distinct as possible from $\boldsymbol{\alpha}_2^*$, at the same time making $\boldsymbol{\beta}_1^*$ more similar to $\boldsymbol{\beta}_2^*$. In II $\alpha_{22}^*=\beta_{12}^*=0$, making k=1 as important as possible. In III $\alpha_{61}^*=\beta_{41}^*=0$, making k=2 as important as possible. The extreme α_{ik}^* parameters are the lowest elements of vectors $\boldsymbol{\alpha}_1^*$ and $\boldsymbol{\alpha}_2^*$; the extreme β_{jk}^* parameters are β_{j2}^* for the highest ratio $\beta_{j1}^*/\beta_{j2}^*$ and β_{j1}^* for the lowest ratio $\beta_{j1}^*/\beta_{j2}^*$. Other parameters cannot be restricted to zero without a resulting drop in fit, i.e., using **T** to make other parameters equal to zero will result in a violation of (2c) or (2d) for the extreme parameters: **T** will lie outside the area described in figure 1.

For K>2 the identification problem becomes much more complicated, since the number of free elements of **T** increases rapidly. Therefore we will discuss another way to understand identification in section 4 (see also de Leeuw et al, 1989). The number of degrees of freedom of the model can be found as follows. The number of independent cells is I(J-1), the number of independent row parameters seems I(K-1) due to (2a), the number of independent column parameters seems K(J-1) due to (2b). These parameters are not

independent, but they can be identified by restriction K(K-1) values of **T** due to **uT=1**. Therefore #df = I(J-1) - [I(K-1)+K(J-1)+K(K-1)] = (I-K)(J-K).

We can identify the model by constraining parameters to specific values. After having imposed constraints the model is identified when **T=I** is the only admissible **T** (compare Mooijaart, 1982)

Constraining parameters

There are two main reasons why we might want to constrain the parameters of the model. Firstly, the model is unidentified, and by constraining parameters we can identify the model. Secondly, we might want to compare two identified models, where one model is a constrained version of the other, for substantive reasons: in this way we can test whether some parameter differs significantly from some specific value, or whether some parameters differ significantly from each other.

Specific value constraints and equality constraints to α_{ik} and β_{jk} can be built in easily. Specific value constraints should be imposed in each iteration after step (4a) or (4b) of the algorithm. Equality constraints can be built in in a similar way by calculating weighted averages. The α_{ik} should be weighted by p_{i++}, and the β_{jk} should be weighted by p_{++k}, where p_{ijk} is derived in (3). After having imposed constraints on some parameters the remaining parameters should be adjusted (if necessary: iteratively) so that (2a) and (2b) hold. We will now discuss some aspects of equality constraints not considered thus far in the context of LBA.

A test for collapsibility of rows and test for indifference of columns

If $\alpha_{ik} = \alpha_{i'k}$ for each k then the theoretical budgets for i and i' are built up in the same way from the latent budgets, i.e. $\pi_i=\pi_{i'}$. This can be interpreted as a *test for collapsibility* of rows i and i' (compare also Gilula, 1986). As an example, in table 1 the α_{ik} for row 1 and 2 are approximately equal. We can test this by imposing the constraint $\alpha_{11}=\alpha_{21}$ (due to (2a), this also implies $\alpha_{12}=\alpha_{22}$). When we fit the model with these constraints, we find a fit of $G^2=2.77$ (df=9). This model fits well, and the difference with the unrestricted model is not significant, namely $G^2_{diff}=.02$ (df=1). We conclude that parental socioeconomic statuses A and B can be collapsed.

If we constrain all parameters β_{jk} to be equal for some j, then the value found will be equal to p_{+j}, and *we test whether the differences between the observed budgets are due to differences in other categories than category j*. For our example, if we impose $\beta_{21}=\beta_{22}=p_{+2}$ and $\beta_{31}=\beta_{32}=p_{+3}$, we find a fit of $G^2=3.60$ (df=10), and the difference with the unrestricted model is not significant: $G^2_{diff}=.85$ (df=2). We can conclude that the 6 groups only differ as far as their conditional proportions for the first and fourth mental health status category are concerned. This simplifies interpretation considerably.

Conclusion

One of the advantages of LBA is that its parameters are easily interpretable, and the results are easily communicated to non-statisticians. The K *latent budgets* β_k can often be interpreted as *typologies*. So the example showed that there are basically two types of mental health statuses, and the *characterizing parameters* α_{ik} show how individual groups can be characterized in terms of these types. It is in many

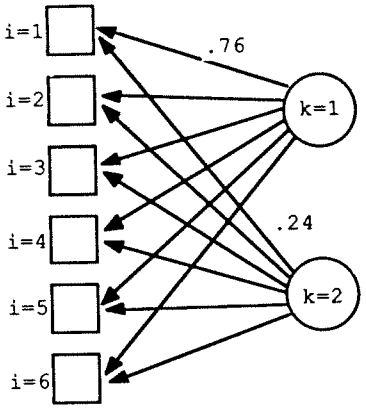

Figure 2: LBA as factor analysis

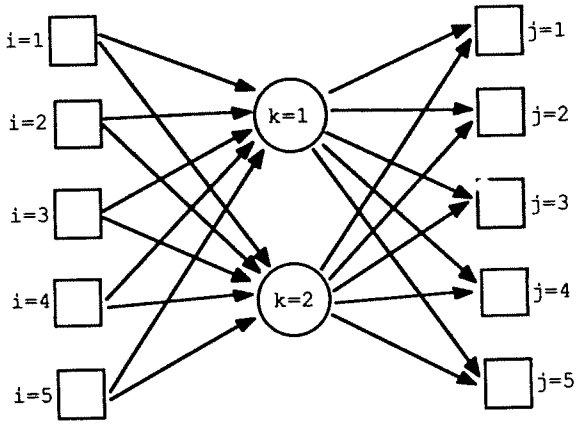

Figure 3: LBA as MIMIC model

situations also a very realistic model, in the sense that the model does not suppose typical groups to exist in real life situations, but instead the budgets of the groups are *built up* from the extreme typologies.

LBA is very similar to factor analysis, denoted as $\mathbf{X=VW'}$. LBA can be written as $\Pi = \mathbf{AB'}$ (see (5)), and it follows that we can interpret LBA is a form of factor analysis with restrictions defined in (2). The interpretation we give to LBA is also very similar to factor analysis, see figure 2: there K=2 latent budgets determine the estimated budgets of the six groups. This is very similar to factor analysis, where, for example, two latent factors determine the observed scores on a group of manifest variables. Notice also that factor analysis has similar identification problems.

3. Relation with other models

LBA is closely related to other models for the analysis of contingency tables, formally to latent class analysis and simultaneous latent class analysis, and conceptually to correspondence analysis and factor analysis. Due to space limitations we do not discuss relations to other models to which LBA is also similar, like association models (Goodman, 1986; Gilula & Haberman, 1986), ideal point discriminant analysis (Takane, 1987), and asymmetrical correspondence analysis (Lauro & D'Ambra, 1984; Lauro & Siciliano, 1988).

Relation with latent class analysis and simultaneous latent class analysis

For the two-variable case latent class analysis (LCA) can be denoted as

$$\pi_{ij}=\Sigma_k \pi_k^X \pi_{ik}^{AX} \pi_{jk}^{BX} \text{ with } \Sigma_k \pi_k^X = 1;\ \Sigma_i \pi_{ik}^{AX} = 1;\ \Sigma_j \pi_{jk}^{BX} = 1 \qquad (6)$$

(see, for example, Haberman, 1979). LBA is very similar to LCA: it is possible to derive the LBA parameter estimates from the LCA parameter estimates, and vice versa. Suppose we have estimates under the LCA model. Define $\beta_{jk}=\pi_{jk}^{BX}$, and $\alpha_{ik}=\pi_k^X\pi_{ik}^{AX}/\Sigma_l\pi_l^X\pi_{il}^{AX}$. Thus we find the LBA estimates. Suppose we have estimates under LBA. Define $\pi_k^X=\Sigma_i\pi_{i+}\alpha_{ik}$, $\pi_{jk}^{BX}=\beta_{jk}$, and $\pi_{ik}^{AX}=\pi_{i+}\alpha_{ik}/\Sigma_l\pi_{l+}\alpha_{lk}$. Thus we find the LCA estimates. This was first noticed by Clogg (1981).

We can draw the following conclusions. First, LCA and LBA have the same estimates of expected frequencies, the same fit and number of degrees of freedom. Secondly, since LCA is symmetrical in the sense that π_{ik}^{AX} and π_{jk}^{BX} have similar restrictions, and LCA can be reparametrized into LBA, it follows that LBA with restrictions (2) has the same estimates of expected frequencies as LBA of the transposed matrix

$$\pi_{ij}/\pi_{+j}=\Sigma_k\alpha^*_{ik}\beta^*_{jk} \text{ where } \Sigma_i\alpha^*_{ik}=1,\ \Sigma_k\beta^*_{jk}=1. \qquad (7)$$

Model (1) with restrictions (2) can answer questions different from those for (7). We can derive parameters with restrictions (2) from (7) and vice versa in a simple way: for example, $\alpha^*_{ik}=(\pi_{i+}\alpha_{ik})/(\Sigma_l\pi_{l+}\alpha_{lk})$ and $\beta^*_{jk}=(\beta_{jk}\Sigma_i\pi_{i+}\alpha_{ik})/\pi_{+j}$.

Usually, LCA is used in situations with more than two variables, each having a small number of categories, for example two. The objective is then to investigate whether the observed relations between the variables can be explained by one unobserved latent variable. LBA starts from a different interpretation. In LBA we want to characterize I groups in terms of K latent budgets. The characterization is given by the α_{ik}, and the latent budgets are given by β_k. So in LBA the variables have an asymmetric role, whereas in LCA they usually have a symmetric role. However, Clogg (1981), Gilula (1979, 1983, 1984), Goodman (1987) and Luijkx (1987) discuss LCA in the two variable case, and also provide us with asymmetric interpretations. For example, Clogg, using parametrization (1), analyses a social mobility table where the row variable is measured earlier than the column variable, and the latent classes are supposed to intervene between the two time points. See figure 3 for a graphical representation, which is similar to that of a MIMIC model. In this context the matrix Π can be interpreted as a transition matrix. Like us, Goodman (1987) discusses the identification problem for the Srole data (table 1) and shows for LCA how to find the solutions I to IV of figure 1 (see also Clogg, Gilula and Luijkx for a discussion of the identification problem in LCA of two-way tables).

LBA is also closely related to *restricted simultaneous* LCA (SLCA; see Clogg & Goodman, 1985). Let π_{ijl} be probabilities in a three-way table where i indexes groups, and j and l two categorical variables. The SLCA model is

$$\pi_{ijl}/\pi_{i++}=\Sigma_k\pi_{ik}^{AX}\pi_{jik}^{BAX}\pi_{lik}^{LAX} \text{ where } \Sigma_k\pi_{ik}^{AX}=1;\ \Sigma_j\pi_{jik}^{BAX};\ \Sigma_l\pi_{lik}^{LAX} \qquad (8)$$

It is easy to see how SLCA is related to LBA: first, as in LBA, in SLCA the interest is also in modelling of conditional probabilities. Second, LBA can be considered as *restricted* SLCA with *one* observed variable,

say B. So first the parameter π_{lik}^{LAX} can be dropped. If we further assume that $\pi_{jik}^{BAX} = \pi_{ji'k}^{BAX}$, we can ommit index i, so that (8) reduces to $\pi_{ij}/\pi_{i+} = \Sigma_k \pi_{ik}^{AX}\pi_{jk}^{BX}$, which is (1) in different notation.

Relation with loglinear analysis

LCA can be interpreted as a conditionally independent loglinear model for an unobserved contingency table (see, for example, Haberman, 1979). This must also holds for LBA due to its relation with LCA. Let the row variable be denoted by 1, the column variable by 2, and the latent budget variable by B. At a stationary point ML estimates of expected probabilities π_{ijk} are equal to estimates under the loglinear model where 1 and 2 are conditionally independent given B: (4a) and (4b) show that

$$\pi_{ij}/\pi_{i+} = \Sigma_k \alpha_{ik}\beta_{jk} = \Sigma_k (p_{i+k}/p_{i++})(p_{+jk}/p_{++k})$$

and we find for the elements π_{ijk} in the unobserved three-way matrix

$$\pi_{ijk} = p_{i++}\pi_{ijk}/\pi_{i++} = p_{i+k}p_{+jk}/p_{++k} \qquad (9)$$

Denoting loglinear models by placing variables corresponding with the highest fitted margins between square brackets (see Fienberg, 1980), estimates of expected frequencies under (9) follow model [1B][2B].

Relation with correspondence analysis

Recently Goodman (1986) and Gilula & Haberman (1986) proposed an ML-version of correspondence analysis that they named, respectively, the RC-correlation model and canonical analysis. The model is

$$\pi_{ij} = \pi_{i+}\pi_{+j}(1 + \Sigma_r^R \lambda_r r_{ir} c_{jr}) \text{ with}$$
$$\Sigma_i \pi_{i+} r_{ir} = \Sigma_j \pi_{+j} c_{jr} = 0 \text{ and } \Sigma_i \pi_{i+} r_{ir} r_{is} = \Sigma_j \pi_{+j} c_{jr} c_{js} = \delta^{rs} \qquad (10)$$

where δ^{rs} is Kronecker delta. If R<min (I-1,J-1), then (10) is unsaturated and can be tested with (I-R)(J-R) degrees of freedom. Gilula (1979, 1983, 1984) studied the relation between (10) and LCA in detail, see also Gilula & Haberman (1986) and Goodman (1987). They show that ML-estimates of LCA for K latent classes can be transformed into ML-estimates of (10) with R=(K-1) dimensions. It follows that LBA with K latent budgets can be derived from ML-CA with R=K-1 dimensions, and vice versa.

In the French correspondence analysis literature (see Greenacre, 1984) much emphasis is given to graphical representations made with the row and column scores. For example, in the representation for the rows each row is represented by a point in multi-dimensional space, and the distance between row points i and i' is derived from the two budgets $\mathbf{p}_i$ and $\mathbf{p}_{i'}$. It is possible to make graphical representations using the LBA parameters that are very similar to the graphical representations of correspondence analysis. We will illustrate this in section 4.

4. A generalization to higher-way tables

Consider a higher-way contingency table of more than two variables. Let each of the variables be either an explanatory variable or a response variable. We can analyze the higher-way table with (1) by first coding it into two-way form. The row variable of this two-way form is a composite variable constructed from the set of explanatory variables E, and the column variable is a composite variable constructed from the set of response variables R. The total number of rows (and columns) is equal to the product of the number of categories of variables in E (and in R).

Thus we can analyse higher-way tables with LBA. This approach to the analysis of higher-way tables with models for two-way tables is also adopted by van der Heijden et al. (1989) and Gilula and Haberman (1988). For LBA of *two-way* tables we found in section 3 that in the unobserved three-way table the estimates of expected frequencies follow loglinear model [1B][2B]. For *higher-way* tables the estimates of expected frequencies in the unobserved table follow model [EB][RB], i.e. the explanatory variables are conditionally independent from the response variables. For the baseline model where K=1 [EB][RB] reduces to the independence model [E][R].

We will now use the fact that the two-way coding stems from a higher-way table explicitly in constructing models. To simplify the discussion, we will introduce a second example earlier analyzed in van der Heijden et al. (1989) with correspondence analysis. In this example there are 18 rows (groups), cross-classified by sex (S) and age (A, 9 levels), and there are 13 columns, namely goods (G) that are stolen from shops. Frequencies specify the number of persons with a certain sex and age that are suspected of stealing a specific good. If we want to know how the budgets of groups can be characterized by a number of underlying latent budgets, LBA can be used. The latent budgets are typical distributions of goods.

If we want to know whether there are significant sex differences in this characterization, or or significant age differences, or significant sex-age interactions, we investigate this using the fact that LBA can be considered as a loglinear model for an unobserved contingency table. In the algorithm the values p_{i+k} derived from (3) can be coded into a matrix of order 2x9xK. We can fit loglinear models to this matrix as an intermediate step between (3) and (4a) in order to constrain α_{ik}. For this purpose redefine $p_{ijk} \equiv p_{asjk}$ and $\alpha_{ik} \equiv \alpha_{ask}$, where a (a=1,2) and s (s=1,...,9) index age and sex respectively. In the unrestricted case LBA finds estimates under loglinear model [ASB][GB] for the unobserved table. Let p_{as+k}* be elements of a matrix to which a loglinear model is fitted inbetween steps (3) and (4a). We use the fact that fitting loglinear models comes to the same as fitting margins of p_{as+k} to p_{as+k}*. Firstly, since we are interested in the decomposition of budgets $\pi_{asj}/\pi_{as+} = p_{asj+}/p_{as++}$, margins p_{as++} have to be fitted to p_{as+k}*. Fitting margins p_{as++} and p_{+s+k} to p_{as+k}* constrains $\alpha_{ask}=\alpha_{a'sk}$. The model for the unobserved four-way table is [AS][SB][GB], showing that age is not directly related to the budgets. Similarly, fitting p_{as++} and p_{a++k} constrains $\alpha_{ask}=\alpha_{as'k}$. The model for the unobserved table is [AS][AB][GB], showing that sex is not directly related to the budgets. Notice that these models can also be fit using equality constraints. Last, fitting p_{as++}, p_{a++k} and p_{+s+k} can be interpreted as a sex and an age effect on the budgets, but no interaction effect. This corresponds with the model [AS][AB][SB][GB], showing that there is no joint effect of age

Table 2: Fit measures for shoplifting data, from K=2 to K=5 (columns), and for distinct forms of restrictions on the α_{ik}. To each cell .1 is added. The baseline model (K=1) has fit G^2=19111 and df=204. The proportion of amelioration compared to the fit of the baseline model is given by Pr, for example, for K=2 unrestricted Pr=(19111-7170)/19111=.625. The sex-effect model is fitted only for K=2: because there are two sex categories, the β_k do not change for K>2.

	K=2	K=3	K=4	K=5
Unrestricted	7170 (df=176) Pr=.625	3263 (150) Pr=.829	1672 (126) Pr=.912	713 (104) Pr=.963
No interaction age and sex	7184 (df=183) Pr=.624	3351 (164) Pr=.825	1797 (147) Pr=.906	791 (132) Pr=.959
Age effect only	7713 (df=185) Pr=.595	6445 (168) Pr=.663	5632 (153) Pr=.705	5560 (140) Pr=.709
Sex effect only	13358 (df=192) Pr=.301	XX	XX	XX

Table 3: Shoplifting data, parameters α_{ik} and β_{jk} for solution K=3 latent budgets

Age			
Males	k=1	k=2	k=3
<12	.777	.027	.196
12-14	.602	.000	.398
15-17	.323	.097	.580
18-20	.052	.384	.564
21-29	.000	.481	.519
30-39	.008	.418	.574
40-49	.004	.361	.635
50-64	.000	.262	.738
65+	.072	.175	.753
Females			
<12	.800	.200	.000
12-14	.677	.323	.000
15-17	.269	.720	.011
18-20	.050	.882	.068
20-29	.023	.916	.061
30-39	.038	.903	.060
40-49	.033	.869	.098
50-64	.044	.794	.162
65+	.070	.636	.294

Goods	k=1	k=2	k=3	p_{+j}
Clothing	.000	.522	.056	.216
Clothing ass.	.012	.106	.068	.066
Provisions/tobacco	.026	.091	.133	.086
Writing material	.317	.000	.057	.112
Books	.016	.031	.100	.049
Records	.015	.003	.096	.037
Household goods	.004	.044	.051	.035
Sweets	.182	.003	.019	.061
Toys	.187	.002	.006	.058
Jewelry	.167	.061	.000	.073
Perfume	.024	.079	.005	.039
Hobbies, Tools	.023	.000	.205	.074
Other	.025	.056	.205	.096
Mean	.294	.379	.327	

and sex on the budgets. Fitting p_{as+k} leaves the α_{ask} unconstrained: the loglinear model corresponds to [ASB][GB], showing that there can also be interaction between age and sex on the budgets.

The above procedure can be generalized straightforwardly to the situation of more than two row variables, or more than one column variable. If there is more than one column variable, then we can also generalize our model to SLCA (compare section 3). This will be useful if we can assume that the observed column variables are related due to an unobserved categorical latent variable. If this is not the case, then the generalization discussed for the rows will be more natural.

We will now discuss an analysis of the shoplifting data. We have fitted various models up to K=5 latent budgets (see table 2). Due to the large sample size (n=33101) none of the models fits in a strict sense. However, when we compare the fit of models with the fit for K=1 (i.e. baseline loglinear model [AS][G]), then several models provide an adequate fit. The unrestricted solution for K=3 is shown in table 3. Although this solution is unidentified we may interpret it because the range of unidentification is very small (see below). We can calculate the importance of the budgets using the LCA parameters $\pi_k^X=\Sigma_{as}p_{as+}\alpha_{ask}$, where π_k^X denotes the proportions of observations falling into the three latent budgets. Starting with the interpretation of the latent budgets β_k we find, by concentrating on those parameter values for which $\beta_{jk} > p_{+j}$, that in β_1 writing materials, sweets, toys and jewelry is stolen more than average; in β_2 clothing, clothing accessories, provisions/tobacco, household goods and perfume is stolen more than average; and in β_3 provisions/tobacco, books, records, household goods, hobbies/tools and other is stolen more than average. We can interpret the α_{ask} most easily by concentrating on those α_{ask} for which $\alpha_{ask}>\pi_k^X$. This shows that β_1 is used more than average by children; β_2 is used more than average by males with age 18-39, but most notably by females older than 14; and β_3 is used more than average by males older than 12.

We can visualize these results in the following way. The α_{ask} can be represented in a two-dimensional subspace of a three-dimensional space since $\Sigma_k\alpha_{ask}=0$. This two-dimensional subspace is shown in figure 4. The rows are represented by points, and these points lie in a triangle with corner points the extremes (1,0,0), (0,1,0) and (0,0,1). MEAN has coordinates π_k^X. Each of the points in figure 4 corresponds with a budget: the 18 row points with π_i, the three corner points with β_k, and MEAN with the mean budget with values p_{+j}.

We can make a similar representation for the columns by rescaling the β_{jk}. This is most simply done by using the p_{ijk} found in (3) and calculating $\beta_{jk}^*=p_{+jk}/p_{++k}$ instead of $\beta_{jk}=p_{+jk}/p_{+j+}$ (compare (7)). Values β_{jk}^* indicate the proportion of observations of activity j found in latent budget k. Since $\Sigma_k\beta_{jk}^*=1$, the J activities can be represented in a two-dimensional subspace of a three-dimensional space (see figure 5).

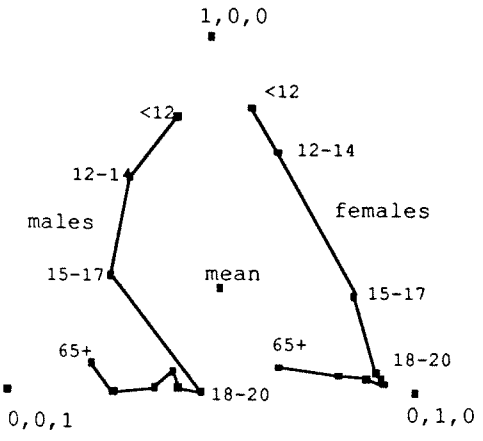

Figure 4: plot of groups

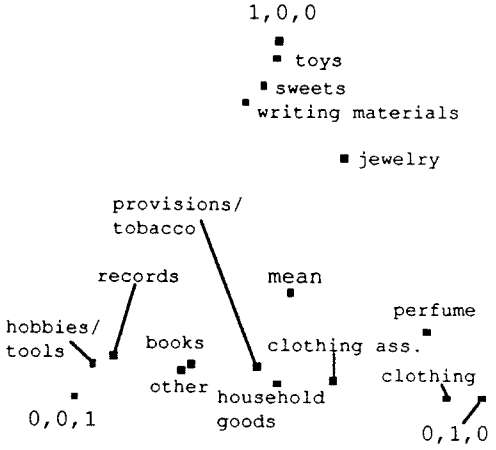

Figure 5: plot of goods

Figures 4 and 5 are very similar to the two-dimensional correspondence analysis representations where the categories represent also budgets (see van der Heijden et al., 1989). Gilula (1984) proposed to circumvent the identification problem in LCA of two-way tables by using correspondence analysis, where the identification problem is solved by using principal axes. Figures 4 and 5 also shed more light on the identification problem in LBA (see also de Leeuw et al, 1989). The identification problem implies for figure 4 that the budgets π_i for the 18 row points remain the same, while the β_k (three corner points) can become different. We can place the three-corner points closer to the 18 points, making the β_k less extreme. However, restrictions (2c) imply that points may not fall outside the triangle. Hence we can identify the solution by placing the most extreme rows on the edges of the triangle. For figure 4 we can restrict $\alpha_{4,1}$, $\alpha_{8,1}$, $\alpha_{2,2}$, $\alpha_{10,3}$, $\alpha_{11,3}$ to zero without loss of fit. For figure 5 we can restrict $\beta_{1,1}$, $\beta_{7,1}$, $\beta_{4,2}$, $\beta_{12,2}$, $\beta_{9,3}$ and $\beta_{10,3}$ to zero: thus we make the β_k more extreme. Because many of these parameters are nearly zero in the unrestricted solution, we conclude that the range of unidentification is small.

Returning to our analysis of the data, table 2 shows that a considerable gain in fit can be obtained in going from K=3 to K=5. Compared to the baseline model (K=1) the gain in going from K=5 to K=6 is less important (for K=6 G^2=407). Compared with K=3, for K=5 one extra budget is found for males older than 30 and females older than 40, who mainly steal more than average provisions/tobacco, and to a lesser extent books, household goods, perfume and hobby/tools. The second extra budget is found for girls younger than 20, who steal more than average sweets, jewelry and perfume. The three budgets already found for K=3 change somewhat: jewelry is not anymore among the goods stolen more than average by young children in general, provisions/tobacco is not anymore among the goods stolen by males and females (budget 2 and 3 in table 3).

Table 2 shows that the effects of age and sex are both important. Also the interaction cannot be neglected.

5. Conclusion

Latent budget analysis is introduced as a tool for the analysis of ordinary contingency tables. It was used before by Clogg (1981) for the analysis of social mobility tables, and by de Leeuw and van der Heijden (1988) for the analysis of time-budgets. We have given attention to tests for collapsibility and tests for indifference. Considerable attention is given to the relation of LBA with LCA, simultaneous LCA, and correspondence analysis. In the last section a possible way to generalize the model to higher-way contingency tables is introduced.

References

C.C.Clogg (1981). Latent structure models of mobility. *American Journal of Sociology*, *86*, 836-868.

C.C.Clogg & L.A.Goodman (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*, 762-771.

de Leeuw, J. & van der Heijden, P.G.M. (1988). The analysis of time-budgets with a latent time-budget model. In: E. Diday et al. (Eds.), *Data analysis and informatics 5*, Amsterdam: North-Holland.

de Leeuw, J., van der Heijden, P.G.M., & Verboon, P. (1989). *A latent time budget model.* Leiden: Department of Psychology (PRM 01-89).
Fienberg, S.E. (1980). *The analysis of cross-classified categorical data (2nd ed.).* Cambridge: MIT-Press.
Gilula, Z. (1979). Singular value decomposition of probability matrices: probabilistic aspects of latent dichotomous variables. *Biometrics, 66,* 339-344.
Gilula, Z. (1983). Latent conditional independene in two-way contingency tables: a diagnostic approach. *The British Journal of Mathematical and Statistical Psychology, 36,* 114-122.
Gilula, Z. (1984). On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika, 71,* 523-529.
Gilula, Z. (1986). Grouping and association in contingency tables: an exploratory canonical correlation approach. *Journal of the American Statistical Association, 81,* 773-779.
Gilula, Z. & Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association, 81,* 780-788.
Gilula, Z. & Haberman, S.J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association, 83,* 760-771.
Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with discussion). *International Statistical Review, 54,* 243-309.
Goodman, L.A. (1987). New methods for analyzing the intrinsic character of qualitative variables using cross-classified data. *American Journal of Sociology, 93,* 529-583.
Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* New York: Academic Press.
Haberman, S.J. (1979). *Analysis of qualitative data (Vol 2).* New York: Academic Press.
Lauro, N.C. & d'Ambra, L. (1984). L'analyse non symmetrique des correspondances. In E. Diday et al. (Eds.), *Data analysis and informatics 3.* Amsterdam: North Holland.
Lauro, N.C. & Siciliano, R. (1988). *Correspondence analysis and modelling for contingency tables: symmetrical and non-symmetrical approaches.* Napoli: Universita di Napoli.
Luijkx, R. (1987). Loglinear modelling with latent variables: the case of mobility tables. In: W. Saris & I. Gallhofer (Eds.), *Sociometrics Research: Vol.2.* London: MacMillan.
Mooijaart, A. (1982). Latent structure analysis for categorical variables. In: K.G.Joreskog and H. Wold, *Systems under indirect observation.* Amsterdam: North Holland.
Srole, L. et al. (1962). *Mental health in the metropolis: the Midtown Manhattan study.* New York: McGraw-Hill.
Takane, Y. (1987). Analaysis of contingency tables by ideal point discriminant analysis. *Psychometrika, 52,* 493-514.
van der Heijden, P.G.M. (1987). Correspondence analysis of longitudinal categorical data. Leiden: D.S.W.O.-Press
van der Heijden, P.G.M., de Falguerolles, A., & de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and loglinear analysis (with discussion). *Applied Statistics*

Acknowledgements: writing this paper is partly made possible by a grant of the Netherlands Organisation for the Advancement of Pure Research (N.W.O.) for the first author.