Constrained latent budget analysis

Peter G.M. van der Heijden*, Ab Mooijaart** and Jan de Leeuw***

Summary

Latent budget analysis is a model for the decomposition of vectors of conditional probabilities (observed budgets) into a prespecified number of underlying, or latent, vectors of conditional probabilities (latent budgets). The model was originally proposed by Clogg in the context of square social mobility tables. In this paper we discuss the model in the context of two–way contingency tables and elaborate upon possible extensions of the latent budget model by considering constraints to be imposed upon the parameters.

Keywords: latent budget analysis, latent class analysis, EM–algorithm, contingency table analysis, data analysis.

*University of Utrecht, Department V.O.S., section methods, The Netherlands.
**University of Leiden, Department of Psychometrics and Research Methods, The Netherlands.
***University of California Los Angeles, Departments of Psychology and Statistics, California, U.S.A.

1. Introduction

In two–way contingency tables interest often goes out to the relation between an explanatory variable and a response variable. One way to study this asymmetric relation is by comparing the proportions conditional on each level of the explanatory variable.

We introduce some notation. We denote the proportions as p_{ij} where i (i=1,...,I) indexes the levels of the explanatory (row) variable and j (j=1,...,J) the levels of the response (column) variable. If we add up over an index we will replace the index by '+': $p_{i+} = \Sigma_j p_{ij}$. The proportions p_{ij} are derived from frequencies n_{ij} as $p_{ij}=n_{ij}/N$, where $N=n_{++}$.

The conditional proportions we are interested in are p_{ij}/p_{i+}. Since *independence* in the contingency table implies that for the theoretical probabilities π_{ij} we have $\pi_{ij}/\pi_{i+}=\pi_{+j}$, we can have an impression of the *relation* between the explanatory and the response variable by comparing values p_{ij}/p_{i+} for different i.

We denote a vector with the conditional proportions of row i by $\mathbf{p_i}$, and we coin it 'budget'. The model that we discuss in this paper is a model for the theoretical row budget $\boldsymbol{\pi}_i$ in the sense that the I theoretical budgets are a linear combination (a mixture) of K *latent* budgets $\boldsymbol{\beta}_k$, having elements β_{jk}, where k (k=1,...,K) indexes the latent budgets. All budgets add up to one, i.e. $\mathbf{u'p_i} = \mathbf{u'}\boldsymbol{\pi}_i = \mathbf{u'}\boldsymbol{\beta}_k = 1$, where $\mathbf{u}$ is a unit column vector of appropriate length. The model is

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

with parameter restrictions

$$0 \leq \alpha_{ik} \leq 1\,,\ \alpha_{i+} = 1 \qquad (2a)$$

$$0 \leq \beta_{jk} \leq 1,\ \beta_{+k} = 1 \qquad (2b)$$

The parameters α_{ik} are the coefficients in the linear combination and can be interpreted as proportions due to (2a).

The model has (I–K)(J–K) degrees of freedom. For K=1 (1) is equivalent to the independence model. If K = max (I,J) the model is equivalent to a saturated model: in this case it is not restrictive. Under the assumption that the observations in each age group are multinomially distributed, the model is estimated with the EM–algorithm (see section 3). The model can be tested against the unrestricted alternative

using chi–squared tests such as the Pearson chi–square test and the likelihood ratio chi–square test. If the model is true, then the test statistic is asymptotically chi–squared distributed. The conditional test of the model with K=n latent budgets given that the model with K=n+m latent budgets is true (n,m ≥1, n+m≤max(I,J)) is *not* asymptotically chi–squared distributed since we are working in the domain of mixture distributions (see also Aitkin, Anderson and Hinde, 1981, and Everitt, 1988, who discuss this problem in the context of latent class analysis).

The model is not identified, since (by writing (1) in matrix notation)

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

Here $\mathbf{D}_r$ is diagonal with marginal probabilities π_{i+} as elements, Π is the matrix with probabilities π_{ij}, and $\mathbf{A}$, $\mathbf{A}^*$, $\mathbf{B}$ and $\mathbf{B}^*$ are matrices with row parameters α_{ik} and α_{ik}^* and column parameters β_{jk} and β_{jk}^* respectively. $\mathbf{T}$ is a square KxK matrix with $\mathbf{Tu} = \mathbf{1}$, $\mathbf{u}$ being a unit column vector, the length of which depends on the context. Since $\mathbf{Tu} = \mathbf{1}$, it follows that $\alpha_{i+}^* = 1$ and $\beta_{+k}^* = 1$. $\mathbf{T}$ also has to be chosen in such a way that $0 \le \alpha_{ik}^* \le 1$ and $0 \le \beta_{jk}^* \le 1$ (compare the restrictions (2a) and (2b)). Usually we choose the matrix $\mathbf{T}$ that results in as many zeroes in $\mathbf{A}^*$ or $\mathbf{B}^*$ as possible, under the restriction that $\mathbf{B}^* \ge 0$ and $\mathbf{A}^* \ge 0$ repectively. This number of zeroes is maximally K(K–1), being the number of free elements of $\mathbf{T}$. If we choose as many zeroes in $\mathbf{A}^*$ as possible, this results in a simplified interpretation because, if row i has parameter estimates α_{ik}^* equal to zero, its estimate of expected budget π_i is derived from less than K latent budgets β_k^*. Many details on the identification problem of the latent budget model can be found in de Leeuw, van der Heijden and Verboon (1989).

Example

We now give a small example to further motivate our model. Caussinus (1986) presents a matrix of 5 age groups by 4 types of cancer. See table 1. We want to study how the type of cancer depends on age. The latent budget model gives K typical distributions of cancer types. These typical distributions are the latent budgets β_k. For age group i the K row parameters α_{ik} show how the expected budget π_i of group i is a mixture of these K typical distributions.

We first determine how many typical (latent) distributions we need to give an adequate description of table 1. For K=1 (independence) we find a likelihood ratio statistic G^2=110.67 (df is (I–K)(J–K)=12), for K=2 G^2=45.02 (df is 6), and for K=3 G^2=.31 (df is 2). The model with three latent budgets has a very good fit. In table 2 the solutions for K=1,2,3 are given. For K=1 α_{i1}=1 and $\beta_{j1}=\pi_{+j}$. Notice that we have chosen T in such a way that for K=2 there is one α_{ik}=0 in each column k, and for K=3 there are two α_{ik}=0 in each k. We will only interpret the solution with

Table 1: Number of cancers of types A, B, C, D in age groups 4 to 8

Age group	Cancer type A	B	C	D	Total
4	55	108	15	19	197
5	175	138	18	41	372
6	230	191	76	67	564
7	381	334	194	117	1026
8	174	262	80	55	571
Total	1015	1033	383	299	2730

4, age <50; 5, age 50–60, 6, age 60–70; 7, 70–80; 8, age ≥80

K=3. The distribution π_1 of age group 1 is built up for 77.3% from the second latent budget β_2 and for 22.7% from the third latent budget β_3. This is departing much from the average distribution of probabilities π_{+j} that is built up for 60% from β_1, for 16.7% from β_2 and for 23.3% from β_3. Latent budgets can be easily interpreted by comparing values β_{jk} with their corresponding marginal probabilities π_{+j}. This shows that β_1 has a slightly higher probability of having cancer type B and a slightly lower probability of having cancer type C. Latent budget β_2 has a much higher probability of having cancer type B, at the expense of types A, C and D. Latent budget type β_3 has a markedly higher probability of cancer type A at the expense of C. We conclude that age group 1 has higher probabilities for A and B. Interpretations for the other age groups can be given in a similar way.

Earlier work

The model was first introduced in the context of (square) social mobility tables by Clogg (1981) as a reparametrization of latent class analysis for two–way contingency tables. *Latent class analysis* for two–way contingency tables can be written as

$$\pi_{ij} = \sum_{k=1}^{K} \pi_k \pi_{ik} \pi_{jk} \qquad (3)$$

with restrictions

Table 2: Latent budget analysis parameter estimates for data in table 1. Solutions for K=1 (independence), K=2 and K=3 latent budgets. $G^2(K=1) = 110.67$, $G^2(K=2) = 45.02$, $G^2(K=3) = .31$

Row parameters
Age group

	K=1	K=2		K=3		
4.	1.000	.055	.945	.0*	.773	.227
5.	1.000	.0*	1.0	.0*	.0*	1.0
6.	1.000	.667	.333	.610	.0*	.390
7.	1.000	1.0	.0*	.978	.022	.0*
8.	1.000	.490	.510	.510	.490	.0*
Mean	1.000	.620	.380	.600	.167	.233

Column parameters
Cancer type

	K=1	K=2		K=3		
A.	.372	.365	.383	.374	.228	.468
B.	.378	.330	.458	.319	.603	.370
C.	.140	.191	.057	.191	.085	.048
D.	.110	.114	.102	.115	.084	.114

$$\sum_{k=1}^{K} \pi_k = 1 \tag{4a}$$

$$\sum_{i=1}^{I} \pi_{ik} = 1 \tag{4b}$$

$$\sum_{j=1}^{J} \pi_{jk} = 1 \tag{4c}$$

and latent budget analysis is derived from latent class analysis as

$$\alpha_{ik} = \frac{\pi_k \pi_{ik}}{\sum_{k=1}^{K} \pi_k \pi_{ik}} \tag{5a}$$

$$\beta_{ik} = \pi_{ik} \tag{5b}$$

Latent class analysis is most often used in the context of contingency tables of more than two variables, where it aims to identify a latent variable that 'explains' the relations between the observed variables (see, for example, Goodman, 1974). Attention for latent class analyses of two–way contingency tables is relatively limited, with exceptions the theoretical contributions of Good (1969), Gilula (1979, 1983, 1984), Clogg (1981), Goodman (1987) and the work in social mobility research and marketing of Marsden (1985), Grover (1987), Grover and Srinivasan (1987) and Luijkx (1987). As far as we know the reparametrization (1) only appears in Clogg (1981) and references given below.

Van der Heijden et al. (1989) showed that latent budget analysis can also be understood as a restricted form of simultaneous latent class analysis (Clogg and Goodman, 1984) for one manifest variable. Consider a three–way table with one variable with groups as categories, indexed by i, and two dependent variables indexed by j and l. Simultaneous latent class analysis states that the dependent manifest variables are independent in each level of a latent variable, indexed by k. The model is

$$\frac{\pi_{ijl}}{\pi_{i++}} = \sum_{k=1}^{K} \pi_{ik}\pi_{jik}\pi_{lik} \tag{6}$$

with restrictions

$$\sum_{k=1}^{K} \pi_{ik} = 1 \tag{7a}$$

$$\sum_{j=1}^{J} \pi_{jik} = 1 \tag{7b}$$

$$\sum_{l=1}^{L} \pi_{lik} = 1 \tag{7c}$$

In the simple form of latent budget analysis we have only one row variable, indexed by i, and one dependent manifest variable, indexed by j. So we can ommit the parameters π_{lik}. If we further restrict π_{jik} so that $\pi_{j1k} = \ldots = \pi_{jik} = \ldots \pi_{jIk}$, so that the index i becomes irrelevant in this parameter, the model reduces to

$$\frac{\pi_{ij}}{\pi_{i+}} = \sum_{k=1}^{K} \pi_{ik}\pi_{jk} \tag{8}$$

with π_{ik} restricted as the latent budget analysis parameters α_{ik}, and π_{jk} restricted as

β_{jk}.

Unaware of earlier work De Leeuw and van der Heijden (1988) independently found decomposition (1) for the analysis of so–called time–budget data. Time budget data are a specific type of constant row sum data (also called 'compositional data') where the matrix to be analyzed has (groups of) individuals in the rows, activities in the columns, and proportions of time spent by individual i in row i. Also in the context of time budget analysis, De Leeuw, van der Heijden and Verboon (1990) discussed the identification of the model in more detail, and showed the relation of latent budget analysis with logcontrast principal component analysis (Aitchison, 1986), which is another method for the analysis of compositional data.

Preliminary results of latent budget analysis in the general context of contingency table can be found in Van der Heijden, Mooijaart and de Leeuw (1989) and de Leeuw and van der Heijden (1989), where the relation of latent budget analysis with (simultaneous) latent class analysis and (a maximum likelihood version of) correspondence analysis (Goodman, 1985, 1986; Gilula and Haberman, 1986, 1988) is discussed. Van der Heijden et al. (1989) indicate how latent budget analysis can be used for the analysis of higher way contingency tables.

This paper
In this paper we discuss possible extensions of latent budget analysis by showing how additional constraints can be imposed upon the row and column parameters. If the explanatory variable or the response variable is a composite variable derived from other variables, this information can also be used by imposing appropriate constraints.

In order to be able to discuss the constraints and their estimation properly, we first discuss estimation in the unconstraint case. In section 3 we discuss three types of constraints: fixed value constraints, equality constraints, and constraints using additional information about the rows and columns. Subsequently we discuss identifiability of the model when constraints are used, and the number of degrees of freedom. We end with an example.

2. Estimation in the unconstrained latent budget model

We estimate the model with the EM–algorithm (Dempster, Laird and Rubin, 1977). This is also the algorithm most often employed for the estimation of latent class analysis, see Goodman (1974) for a description. An alternative algorithm used for the estimation of latent class analysis is the Newton Raphson algorithm, see Haberman (1979) and Formann (1978). Here we will point out exactly how our version of Goodman's algorithm is related to the EM–algorithm. We need to do this so that we can show later how constraints on the parameters affect the algorithm.

The EM–algorithm is an algorithm for the estimation of missing values. In this context the observations on the latent variable are missing, and only the two–way margins n_{ij} of the three–way table with elements n_{ijk} are known. For the three–way table we have a model for π_{ijk}/π_{i++}, namely

$$\frac{\pi_{ijk}}{\pi_{i++}} = \alpha_{ik}\beta_{jk} \tag{8}$$

The loglikelihood for the unobserved matrix is

$$L = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk} \ln \frac{\pi_{ijk}}{\pi_{i++}} \tag{9}$$

We do not know the unobserved n_{ijk}, and neither do we know the parameter estimates for α_{ik} and β_{jk}. The EM–algorithm consists of two steps: the expectation step and the maximization step. In the *E–step* the expectation of the loglikelihood for the unobserved data n_{ijk} is found conditional on the observed frequencies n_{ij} and the model parameters. So the expectation of the sufficient statistics of the complete data matrix have to be expressed in terms of the model parameters, so we need an expression for n_{ijk}. For this step the current best estimates of α_{ik} and β_{jk} are taken. Thus we find new unobserved frequencies $\underline{n}_{ijk}$ as

$$\underline{n}_{ijk} = n_{ij}\frac{\pi_{ijk}}{\pi_{ij+}} = n_{ij}\frac{\pi_{ijk}/\pi_{i++}}{\pi_{ij}/\pi_{i+}} = n_{ij}\left(\frac{\alpha_{ik}\beta_{jk}}{\sum_{k=1}^{K}\alpha_{ik}\beta_{jk}}\right) \tag{10}$$

In the *M–step* the likelihood for the unobserved data is maximized as a function of the model parameters. So the following function is to be maximized over α_{ik} and β_{jk}:

$$\begin{aligned} f(\alpha_{ik},\beta_{jk}) = &\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\ln(\alpha_{ik}\beta_{jk}) - \\ &\sum_{i=1}^{I}\gamma_i\Big(\Big(\sum_{k=1}^{K}\alpha_{ik}\Big)-1\Big) - \sum_{k=1}^{K}\delta_k\Big(\Big(\sum_{j=1}^{J}\beta_{jk}\Big)-1\Big) \end{aligned} \tag{11}$$

where γ_i and δ_k are so–called Lagrange multipliers. For finding the $\{\alpha_{ik}\}$ and the $\{\beta_{jk}\}$ that maximize $f(\alpha_{ik}, \beta_{jk})$ under the restrictions given above, we might as well rewrite (11) as both (12a) and (12b) and maximize (12a) over $\{\alpha_{ik}\}$ and (12b) over

$\{\beta_{jk}\}$ separately:

$$f(\alpha_{ik},\beta_{jk}) = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{i+k}\ln \alpha_{ik} - \sum_{i=1}^{I} \gamma_i((\sum_{k=1}^{K} \alpha_{ik})-1) + \text{constant} \qquad (12a)$$

$$f(\alpha_{ik},\beta_{jk}) = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{+jk}\ln \beta_{jk} - \sum_{k=1}^{K} \delta_k((\sum_{j=1}^{J} \beta_{jk})-1) + \text{constant} \qquad (12b)$$

Maximizing (12a) and (12b) gives as solutions

$$\underline{\alpha}_{ik} = \frac{n_{i+k}}{n_{i++}} \qquad (13a)$$

$$\underline{\beta}_{jk} = \frac{n_{+jk}}{n_{++k}} \qquad (13b)$$

These new estimates are used in (10) as current best estimates in the next step of the algorithm.

De Leeuw et al. (1990) prove that in this application of the EM–algorithm the likelihood is increased in each step of the algorithm. Therefore the algorithm converges to a maximum, though not necessarily a global maximum.

Estimates of expected frequencies m_{ij} are

$$m_{ij} = n_{i+}\frac{\pi_{ij}}{\pi_{i+}} \qquad (14)$$

Notice that the model restricts the two observed variables in the unobserved frequencies of the three–way table to be conditionally independent given the level of the latent variable:

$$\pi_{ijk} = \pi_{i++}\alpha_{ik}\beta_{jk} = \pi_{i++}\frac{n_{i+k}}{n_{i++}}\frac{n_{+jk}}{n_{++k}} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \qquad (15a)$$

or, denoted as a hierarchical loglinear model,

$$\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)} \qquad (15b)$$

This should not come as a surprise since we already found that latent budget analysis is equivalent to latent class analysis, and latent class analysis is sometimes presented as a loglinear model with latent variables (see, for example, Haberman, 1979, or Hagenaars, 1986, 1988).

3. Constraining the parameters

Since latent budget analysis is closely related to latent class analysis, constraining the parameters in latent budget analysis is also closely related to those in latent class analysis. Langeheine (1989) gives an insightful overview of constraints in latent class analysis, and relates the types of constraints to the different ways latent class analysis is presented, namely by Goodman (1974) as a product of conditional probabilities and by Haberman (1979) as a loglinear model with a latent variable. We make use of both representations to discuss constraints in latent budget analysis, and also introduce a set of constraints that is very similar to the linear logistic constraints on the conditional probabilities introduced by Formann (1982, 1985, 1989).

3.1 Fixed value constraints for row and column parameters

We first introduce some notation. We denote parameters constraint to fixed values by adding a '*' as superscript, i.e. α_{ik}^* and β_{jk}^*. In the presence of fixed value constraints we denote parameters to be estimated by adding a '+' as superscript, i.e. parameters to be estimated are α_{ik}^+ and β_{jk}^+. From the sets of parameters $\{\alpha_{ik}^*\}$, $\{\beta_{jk}^*\}$, $\{\alpha_{ik}^+\}$ and $\{\beta_{jk}^+\}$ we derive the elements α_{ik} and β_{jk} of the matrices A and B. For this purpose the sets $\{\alpha_{ik}^+\}$ and $\{\beta_{jk}^+\}$ are sometimes transformed to find their corresponding $\{\alpha_{ik}\}$ and $\{\beta_{jk}\}$ that have the property that $\alpha_{i+}=1$ and $\beta_{+k}=1$.

Fixed value constraints for specific row and column parameters can be built in easily. We will show this for the situation with fixed row parameters. The results for the column parameters are identical, if we replace α by β, and indices i by k and k by j.

Say we restrict the parameter for row i in budget m to be fixed to some constant c, i.e. $\alpha_{im}^* = c$, $0 \le c \le 1$. Since the elements α_{ik} of the matrix A have to add up rowwise to 1, we have

$$\alpha_{ik} = (1 - \sum_{m=1} \alpha_{im}^*)\, \alpha_{ik}^+, \text{ with } \sum_{k=1} \alpha_{ik}^+ = 1 \qquad (16)$$

So in order to find the elements α_{ik} of the matrix A, the loglikelihood to be maximized in the M–step over the parameters α_{ik}^+ is:

$$f(\alpha_{ik}^{+}, \beta_{jk}) = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{i+k} \ln \left((1 - \sum_{m=1} \alpha_{im}^{*}) \alpha_{ik}^{+} - \right.$$

$$\left. \sum_{i=1}^{I} \gamma_i ((\sum_{k=1}^{K} \alpha_{ik}^{+}) - 1) \right) + \text{constant} \qquad (17)$$

where in (17) the sum is taken over these index pairs (i,k) that correspond with the index pairs of free parameters α_{ik}^{+}. Maximizing this function we find

$$\alpha_{ik}^{+} = \frac{n_{i+k}}{n_{i++}} \qquad (18)$$

where n_{i++} is derived from the n_{i+k} with index pairs (i,k) that correspond with free parameters α_{ik}^{+}. Having (18) we find new estimates $\underline{\alpha}_{ik}$ using (16). Since we can rewrite the loglikelihood in a part with row parameters and a part with column parameters, and maximize the loglikelihood over these parts separately (see (11), (12a) and (12b)), the estimation of the column parameters β_{jk} is unaffected by the presence of fixed value constraints for α_{ik}–parameters, and therefore remains as in (13b).

In the situation of fixed value constraints for β_{jk}–parameters, we can define β_{jk} in a similar way as α_{ik} in (16), and specify a loglikelihood as in (17). When we maximize this loglikelihood over β_{jk}^{+} we find

$$\beta_{jk}^{+} = \frac{n_{+jk}}{n_{++k}} \qquad (19)$$

where n_{++k} is derived from the n_{+jk} that correspond with the index pairs (j,k) of free parameters β_{jk}^{+}. The estimates β_{jk}^{+} can be used to obtain the elements β_{jk} of the matrix B, similar to the way the α_{ik}^{+} are used to obtain the elements α_{ik} in (16).

If there are both row and column parameters constrained to fixed values, the loglikelihood can be rewritten in two parts that can be maximized separately. If we do this the parameters α_{ik} and β_{jk} can be derived using α_{ik}^{+} in (18) and β_{jk}^{+} in (19).

Example

In the introduction we analyzed the data given in table 1. The results were displayed in table 2. It was concluded that the solution with K=3 latent budgets gave an adequate fit. Six α_{ik} parameters were already constrained to zero to identify the model. Now we will constrain some additional α_{ik} parameters to simplify the interpretation. The parameters that we will constrain are α_{12}^{*}=.75, α_{31}^{*}=2/3,

$\alpha_{42}^*=0$ and $\alpha_{51}^*=.5$. With these round values the interpretation simplifies. Notice that due to the restriction $\alpha_{i+}=1$ there are no free α_{ik} parameters left. If we fit the model with these constraints we find a fit of $G^2 = .58$ for df = 15–9 = 6, where the number of free β_{jk} parameters is nine. We still have an extremely good fit. The parameter estimates are given in table 3a.

Test for indifference of columns

Consider the case that $\beta_{j1} = ... = \beta_{jk} = ... \beta_{jK} = \pi_{+j}$,. i.e. the j'th element of each latent budget is equal to the average probability to fall into category j. This constraint can be tested using a fixed value constraint since the maximum likelihood estimate of π_{+j} is p_{+j}. The consequence of this constraint is that $\pi_{ij}/\pi_{i+} = \pi_{+j}$, so by constraining $\beta_{j1} = ... = \beta_{jk} = ... \beta_{jK} = \pi_{+j}$ we can test whether the difference between the expected budgets π_i are due to differences in other column categories than column category j. If such a test cannot be rejected, the interpretation can simplify considerably: it is not necessary to characterize the groups (rows) in terms of differences in their use of column j.

Table 3: Latent budget analysis parameter estimates for data in table 1.
Table 3a: Fixed value constraints for the row parameters. G^2=.583, df=6.
Table 3b: Fixed value constraints for the row parameters and an equality constraint for the column parameters. G^2=.614, df=7.

Row parameters	*Table 3a*			*Table 3b*		
Age group						
		K=3			K=3	
4.	.0*	.75*	.25*	.0*	.75*	.25*
5.	.0*	.0*	1.0*	.0*	.0*	1.0*
6.	.67*	.0*	.33*	.67*	.0*	.33*
7.	1.0*	.0*	.0*	1.0*	.0*	.0*
8.	.5*	.5*	.0*	.5*	.5*	.0*
Mean	.618	.159	.223	.618	.159	.223
Column parameters						
Cancer type						
		K=3			K=3	
A.	.374	.228	.468	.374	.227	.223
B.	.319	.603	.370	.325	.600	.370
C.	.191	.085	.048	.187	.086*	.047
D.	.115	.084	.114	.114	.086*	.113

3.2 Equality constraints

We start immediately with the situation of both equality constraints and fixed value constraints. Parameters with superscript '*' are again fixed to some specific value. The parameters with superscript '+' are free or constrained to be equal to other parameters. These two types of parameters constitute the matrix A with elements α_{ik} and B with elements β_{jk}. We only consider equality constraints between row parameters, and equality constraints between column parameters, but do not consider equality constraints between row and column parameters. Due to this, the estimation problem in the M–step can be simplified since we can rewrite the loglikelihood in a part with parameters for A and a part with parameters for B, and maximize these parts separately.

We first consider constraints on row parameters α_{ik}. We found in (17) in the case with only fixed value constraints the loglikelihood function to be maximized over the parameters α_{ik}^+ is

$$f(\alpha_{ik}^+, \beta_{jk}) = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{i+k} \ln\left((1 - \sum_{m=1} \alpha_{im}^*)\,\alpha_{ik}^+ - \sum_{i=1}^{I} \gamma_i\left(\left(\sum_{k=1}^{K} \alpha_{ik}^+\right) - 1\right)\right) + \text{constant} \qquad (17)$$

and the parameters α_{ik}^+ can be transformed into their corresponding α_{ik} through (16).

Let some parameters α_{ik} be restricted to be equal. We introduce some extra notation. Let A_{ik} be the set $\{(i,k)\}$ for which equality constraints are imposed, where for (i,k) the first combination is taken (the index i running faster than the index k), the others do not exist, i.e. if $\alpha_{12}=\alpha_{21}$, then the set $\{(1,2),(2,1)\}$ is A_{12}, and A_{21} does not exist; if no equality constraint is imposed for some α_{ik}, then A_{ik} has only element (i,k). Call W the set of existing index pairs (i,k) of the sets A_{ik}. Let the number of parameters in each A_{ik} be equal to c_{ik}, and let $f_{ik} = \Sigma_{\text{index pairs in A(ik)}}\, n_{i+k}$. Instead of (17) we now find

$$f(\alpha_{ik}^+, \beta_{jk}) = \sum_{i=1}^{I}\sum_{k=1}^{K} f_{ik} \ln\left((1 - \sum_{m=1} \alpha_{im}^*)\,\alpha_{ik}^+ - \sum_{i=1}^{I} \gamma_i\left(\left(\sum_{k=1}^{K} c_{ik}\alpha_{ik}^+\right) - 1\right)\right) + \text{constant} \qquad (20)$$

where the sum is taken over the index pairs in W, so that each different parameter α_{ik}^{+} is used only once in (20) and combinations (i,k) for fixed parameters are not used.

As for (17) the idea is now to maximize the loglikelihood in (20) over α_{ik}^{+} so that the α_{ik}^{+} that maximize.(20) can be used (16) to find the parameters α_{ik}. However, only in some instances the maximization of (20) over α_{ik+} yields direct estimates for α_{ik+}. We will give some examples.

Example 1: equalities within rows only. In this case we can maximize $f(\alpha_{ik}^{+},\beta_{jk})$ in (20) in each row separately. As a result we find

$$\alpha_{ik}^{+} = \frac{f_{ik}}{c_{ik} f_{i+}} \tag{21}$$

The parameters α_{ik}^{+} can be used in (16) to find the parameters α_{ik}. If there are no elements in row i constraint to some fixed value, then $\alpha_{ik}^{+} = \alpha_{ik}$ (and $f_{i+} = n_{i++}$). In this case the parameter α_{ik} for the elements in the equality constraint defined by A_{ik} is the mean of the parameters if these would have been unconstrained to be equal. If some elements in row i are fixed, then $f_{ik}/c_{ik}f_{i+}$ is the mean of the parameters α_{ik}^{+} if these would not have been constraint to be equal; f_{i+} is only derived from the non–fixed α_{ik}. We conclude that, if there are only equality constraints in each row separately, then we can first estimate all α_{ik}^{+} as if they are completely free (also the parameters constraint to be equal); second, in each row we take the mean of the α_{ik}^{+} constrained to be equal; and, third, use (16) to find the elements of A.

Example 2: $\alpha_{ik} = \alpha_{i'k}$ If there is an equality constraint in two rows but in the same column of A, for example, $\alpha_{11} = \alpha_{21}$, then the estimate found for α_{11} is

$$\alpha_{11} = \frac{n_{1+1}+n_{2+1}}{n_{1++}+n_{2++}} \tag{22}$$

and the other parameters in row 1 and 2 are to be derived from n_{i+k}/n_{i++}, and then adjusted in a way similar to (16) so that $\alpha_{i+} = 1$. This procedure can be generalized to more elements α_{i1} as long as these elements are in the same column.

Example 3: $\alpha_{ik} = \alpha_{ik'} = \alpha_{i'k}$. This equality constraint is an example of an equality constraint that can not be solved by direct estimation. The result is that an iterative procedure has to be applied. We were not able to find iterative proportional fitting

procedures that could be used to fit this equality constraint.

Our conclusion is that, if other equality constraints than the examples above are to be fitted, it should be checked whether direct estimates exist. If not, a Newton Raphson or similar procedure should be applied in each step of the EM–algorithm. Generally it is not advisable to use such equality constriants when using the EM algorithm, since it will make the algorithm very time consuming.

Example
In section 3.1 we constrained the α_{ik} parameters using specific value constraints. The results were displayed in table 3a. As an example we will now constrain two β_{jk}–parameters to be equal, namely $\beta_{32}=\beta_{42}$. If we impose this constraint additional to the constraints we already imposed to the α_{ik} parameters, we find a fit of $G^2=.614$ (df = 7). The parameter estimates are given in table 3b. Other equality constraints such as $\beta_{11}=\beta_{21}$ lead to a significant reduction of fit.

A test for collapsibility of rows
If $\alpha_{ik} = \alpha_{i'k}$ for all k in row i and i', then the theoretical budgets $\boldsymbol{\pi}_i$ for i and $\boldsymbol{\pi}_{i'}$ for i' are built up in the same way from the latent budgets $\boldsymbol{\beta}_k$, i.e. $\boldsymbol{\pi}_i = \mathbf{B}'\boldsymbol{\alpha}_i = \mathbf{B}'\boldsymbol{\alpha}_{i'} = \boldsymbol{\pi}_{i'}$, where $\boldsymbol{\alpha}_i$ is a kx1 column vector with values α_{ik}. We can test whether $\alpha_{ik} = \alpha_{i'k}$, i.e. we can test whether the expected row budgets are equal.

The equality of expected row budgets is used earlier by Breiger (1981), Goodman (1981), Gilula (1986), and Gilula and Krieger (1989) as a criterium for collapsibility of rows. In a similar way we can also interpret the test that $\alpha_{ik} = \alpha_{i'k}$ for all k as a test for collapsibility of rows i and i'. This idea is closely related to the work of Gilula (1986) and Gilula and Krieger (1989) who applied similar tests for equality of the parameters of two or more rows (or two or more columns) in a maximum likelihood version of correspondence analysis. The closeness of the work of Gilula and the procedure described here follows immediately in these cases where correspondence analysis and latent budget analysis are equivalent (see de Leeuw and van der Heijden, 1989). In the cases that correspondence analysis and latent budget analysis are not equivalent, the tests can give different results.

3.3 Constraints using additional information about the row and column categories

If we have additional information about the rows and/or columns of the two–way table, we can constrain the corresponding row and column parameters to be a function of this additional information. This has to be done in the M–step of the EM–algorithm.

Our sets of parameters α_{ik} and β_{jk} are conditional probabilities, and therefore we use

a the multinomial logit model for conditional probabilities discussed extensively by Bock (1975). We first discuss this model, and then apply it to our parameter estimates.

The multinomial logit model for the conditional probabilities π_{ij}/π_{i+} is

$$\frac{\pi_{ij}}{\pi_{i+}} = \frac{\exp z_{ij}}{\sum_{n=1}^{J} \exp z_{in}} \tag{23}$$

In (23) the values z_{ij} are unconstrained. Bock (1975) constrains the π_{ij} by constraining the elements of $\mathbf{Z}$ to

$$\mathbf{Z} = \mathbf{X\Omega Y} \tag{24}$$

where $\mathbf{X}$ is a model matrix for the row parameters, $\mathbf{Y}$ is contrast matrix describing structure in the column categories, and $\mathbf{\Omega}$ is a parameter matrix to be estimated. We will use the multinomial logit model both for constraining the row parameters as well as for constraining the column parameters. For the row parameters we leave away $\mathbf{Y}$, and for the column parameters we assume that $\mathbf{K} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

This approach to constraining parameters is also used in other contexts. Formann (1982, 1985, 1989) constrains latent class analysis in a way very similar to what we do, but concentrates on latent class analysis on dichotomous variables (see also Langeheine, 1989). Shigemasu and Sugiyama (1989) use this parameterization for latent class analysis of choice behavior. Takane (1988) uses the multinomial logit model to restrict the conditional probabilities in ideal point discriminant analysis.

Constrained row parameters

Let the additional information for the rows of our matrix Π be collected in a matrix $\mathbf{X}$, having I rows index by i and M columns indexed by m (m=1,...,M). We use this information by defining the following version of the multinomial logit model for the (conditional) row parameters α_{ik}:

$$\alpha_{ik} = \frac{\exp\left[\sum_{m=1}^{M} x_{im}\gamma_{km}\right]}{\sum_{n=1}^{K} \exp\left(\sum_{m=1}^{M} x_{im}\gamma_{km}\right)} \tag{25}$$

where the parameters γ_{km} can be collected in a parameter matrix Γ of order KxM.

Model (25) can be identified by constraining $\gamma_{1m} = 1$.

We can specify the loglikelihood with α_{ik} constrained as in (25) as

$$f(\gamma_{km},\beta_{jk}) = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{i+k} \ln \left(\frac{\exp\left(\sum_{m=1}^{M} x_{im}\gamma_{km}\right)}{\sum_{n=1}^{K} \exp\left(\sum_{m=1}^{M} x_{im}\gamma_{km}\right)} \right) + \text{constant} \qquad (26)$$

In each M–step of the EM–algorithm we have to maximize $f(\gamma_{km},\beta_{jk})$ over γ_{km} in order to α_{ik}. This can be done by fitting the multinomial logit model to the values n_{i+k}, using these as if they were observed frequencies. The way in which we find the parameter estimates for β_{jk} is unaffected: it is done in the usual way as in (18b).

An important special case of model (1) is when the matrix **X** is a design matrix describing the factorial structure in the rows. In this case each multinomial logit model is equivalent to a loglinear model for unconditional probabilities: two models are equivalent if the loglinear model has the same parameters as the multinomial logit model (these describe the relations between the explanatory and the response variables), and additionally the parameters that describe the relations between the explanatory variables (see Bock, 1975). This is evident if K=2, since then the multinomial logit model simplifies to the ordinary logit model (see Fienberg, 1980, ch.6, who shows the equivalence of the loglinear model and the logit model).

The equivalence of the multinomial response model and the hierarchical loglinear model gives us further insight into the constrained latent budget model. A result of Goodman (1971) shows that, if a hierarchical loglinear model is fitted to n_{i+k} that includes parameters for k (which correspond to a column of '1's in **X**), that then the estimates for the latent values n_{ijk} are still estimates of expected frequencies under a loglinear model for the latent matrix. So if the constraints imposed using (25) define a multinomial logit model that is equivalent to a hierarchical loglinear model including terms for the latent variable, then on a stationary point the latent probabilities still follow a hierarchical loglinear model. Thus constraining the latent budget model does not affect the relation with loglinear analysis that we discussed in section 2 (equation (15b)).

As an example, let the matrix Π have rows that can be stratified by two variables, with categories of the first indexed by i, and of the second indexed by s. Consider that the matrix **X** does not constrain the row parameters α_{isk}, i.e. it specifies a saturated multinomial logit model. If we generalize equation (15a) in a straigthforward way to the latent four–way table, we find that the joint row variable indexed by i and s is independent from the column variable indexed by j, conditional

on the level of the latent variable indexed by k:

$$\pi_{isjk} = \pi_{is++}\alpha_{isk}\beta_{jk} = \pi_{is++}\frac{n_{is+k}}{n_{is++}}\frac{n_{++jk}}{n_{+++k}} = \frac{\pi_{is+k}\pi_{++jk}}{\pi_{+++k}} \qquad (27a)$$

or in terms of loglinear models,

$$\begin{aligned}\log \pi_{isjk} = u &+ u_{1(i)} + u_{2(s)} + u_{3(j)} + u_{4(k)} + u_{12(is)} + \\ &u_{14(ik)} + u_{24(sk)} + u_{124(isk)} + \\ &u_{34(jk)}\end{aligned} \qquad (27b)$$

where the u–terms add up to zero over each index.

If we constrain (27b) by using design matrices $\mathbf{X}$ that lead to hierarchically constrained multinomial logit models, we can constrain the sets of parameters $\{u_{14(ik)}\}$, $\{u_{24(sk)}\}$, and $\{u_{124(isk)}\}$ to be zero. An example is to constrain $u_{124(isk)} = 0$ for all (i,s,k) by deleting the columns of $\mathbf{X}$ that describe interaction between i and s: if a model with this constraint fits adequately, we can conclude that the rows (i,s) can be described in terms of the latent budgets $\boldsymbol{\beta}_k$ by an effect of the variable indexed by i, and by an effect of the variable indexed by s, but that there is no interaction between these variables in their relation to the latent budgets.

Another example is to restrict interactions to be linear in the logarithm. Consider that we analyze a two way table with probabilities π_{ij}. Then the loglinear model for the latent probabilities π_{ijk} is given in (15b). Consider now that the categories of the row variable are ordered, and we assume that this order is reflected in the interaction parameters $\{u_{13(ik)}\}$ by the constraint $u_{13(ik)} = (v_i - \underline{v})u_{3(k)}^*$ where $\underline{v}$ is the average of a priori assigned scores $\{v_i\}$, and the set of parameters to be estimated is $\{u_{3(k)}^*\}$. Thus the latent budget model can still be understood as a loglinear model for the latent probabilities. Although these loglinear constraints are straightforward, they are not always revealed clearly by the conditional probability parametrization $\{\alpha_{ik}\}$ that we employ for latent budget analysis. For the first example we find (compare (1))

$$\begin{aligned}\alpha_{ik} &= \frac{\exp\left(u_{4(k)} + u_{14(ik)} + u_{24(sk)}\right)}{\sum_{n=1}^{K} \exp\left(u_{4(n)} + u_{14(in)} + u_{24(sn)}\right)} \\ &= \frac{\tau_{4(k)}\tau_{14(ik)}\tau_{24(sk)}}{\sum_{n}^{K} \tau_{4(n)}\tau_{14(in)}\tau_{24(sn)}}\end{aligned} \qquad (28)$$

where $\tau_{4(k)} = \exp(u_{4(k)})$, etcetera. We see that the constraint $u_{124(isk)} = 0$ is reflected by the numerator of (5), but not by the denominator: due to the denominator each combination of i and s is divided by a distinct constant. However, since this constant is a function of i and s separately, often the parameters α_{ik} reflect the restriction that $u_{124(isk)} = 0$ roughly.

Constrained column parameters

So far for constraints on the row parameters. Using the multinomial logit model described in (23) and (24) we are also able to define constraints on the column parameters. Assume that the information on the column categories of Π is collected in the matrix **Y**, having J rows and H columns, indexed by h (h=1,...,H). Now we can constrain the β_{jk} parameters as

$$\beta_{jk} = \frac{\exp\left(\sum_{h=1}^{H} y_{jh}\psi_{kh}\right)}{\sum_{n=1}^{J} \exp\left(\sum_{h=1}^{H} y_{nh}\psi_{kh}\right)} \tag{29}$$

Similar to (26), we can specify the loglikelihood as

$$f(\alpha_{ik}, \psi_{kh}) = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{+jk} \ln \frac{\exp\left(\sum_{h=1}^{H} y_{jh}\psi_{kh}\right)}{\sum_{n=1}^{J} \exp\left(\sum_{h=1}^{H} y_{nh}\psi_{kh}\right)} + \text{constant} \tag{30}$$

This shows that we have to fit the version of the multinomial logit model defined in (6) in each M–step of the EM–algorithm. The parameters estimates for α_{ik} are found in the usual way as in (18a).

When the column categories are classified by more than one variable, the matrix **Y** can be used to describe the factorial structure of the columns. As for the rows, in this case the multinomial logit model in (29) is equivalent to a loglinear model for the unconditional probabilities in the following sense. The matrix **Y** defines a model that is fitted to each k separately. So if the are, for example, two variables for the column categories, and **Y** constrains these to be independent, the loglinear model fitted is the conditional independence model where the two variables are independent given k (see, for more details, Bock, 1975).

Results in Goodman (1971) show that, if we thus constrain the β_{jk}–parameters using hierarchical loglinear models, the model for the latent probabilities is still a hierarchical loglinear model. In this way we can also derive a constrained form of

simultaneous latent class analysis (6) from latent budget analysis by constraining the column variables to be independent in each level k. The constrained version of simultaneous latent class analysis found is

$$\frac{\pi_{ijl}}{\pi_{i++}} = \sum_{k=1}^{K} \pi_{ik}\pi_{jk}\pi_{lk} \qquad (31)$$

and this corresponds with the hierarchical loglinear model for the latent probabilities π_{ijlk}

$$\log \pi_{ijlk} = u + u_{1(i)} + u_{2(j)} + u_{3(l)} + u_{4(k)} + u_{14(ik)} + u_{24(jk)} + u_{34(lk)} \qquad (32)$$

Constraining both row and column parameters using additional information
If we constrain both row parameters with (25) as well as column parameters with (29) the estimation procedure does not fundamentally change: in the loglikelihood there are two parts to be maximized independently in each M–step of the algorithm.

Estimation
Bock (1975) shows how to find the estimates for the general multinomial logit model (23)–(24) using the Newton Raphson algorithm. He also explains how to fit the multinomial logit model in case of structural zeroes, which can be used in our context when there are additional fixed value constraints for some of the α_{ik}. When using the Newton Raphson algorithm, the advantage of using the multinomial logit model instead of a corresponding loglinear model is that estimation of the multinomial logit is computationally more efficient since less parameters have to be estimated.

However, this approach is extremely time consuming. Another approach can be used when the multinomial logit model (23)–(24) is equivalent to a *hierarchical* loglinear model is to use iterative proportional fitting (see Bishop, Fienberg and Holland, 1975; Fienberg, 1980, ch. 3,4). Iterative proportional fitting will be more efficient than the Newton Raphson procedure when the number of parameters to be estimated becomes large, or when direct estimates for the expected frequencies exist.

Iterative proportional fitting is also more in line with the EM algorithm: in each step of the algorithm a matrix with latent estimates of expected probabilities π_{ijk} is generated that has the property that $\pi_{i+k} = n_{i+k}/n_{+++}$ and $\pi_{+jk} = n_{+jk}/n_{+++}$, showing that two margins are fitted. In this paper the constraints (25) and (29) are choosen in such a way that the multinomial ligit models are equivalent to hierachical loglinear models so that we can use iterative proportional fitting.

Test for collapsibility of rows revisited
Some of the constrained latent budget models with hierarchical constraints for the

row parameters as in (25) also yield the equality constraints discussed in the subsection on collapsibility of rows in section 3.2). To be more precise, if there are no columns in **X** that related some of the row variables to the latent variable, then the row parameters $(\alpha_{i1},\ldots,\alpha_{iK})$ will be equal for distinct index values for these variables. To give an example, if there are two variables that constitute the rows indexed by i and s (say sex), and the model fitted to α_{isk} has only columns relating i to k, then $(\alpha_{is1},\ldots,\alpha_{isK})$ will be equal to $(\alpha_{is'1},\ldots,\alpha_{is'K})$ (i.e. the values for men are equal to those for women). This does not mean that these constraints can always be fitted using the procedure for equality constraints, since the 'i' can be constrained as well by (25). If (25) specifies a multinomial logit model that is equivalent with a conditional independence model fitting all possible parameters for i, than this multinomial logit model can be fitted using the procedure for equality constraints.

Since the above equality constraints define a test for collapsibility of rows, it is also possible to find the parameter estimates under such constraints by adding up the rows in the observed data matrix and doing an unrestricted latent budget analysis. For the example using the multinomial logit model that could be fitted using the above equality constraints, parameter estimates could be found by doing an unconstrained analysis on a matrix collapsed over the variable indexed by 's'. Of course, the fit of the model has to be evaluated using the uncollapsed observed data, and the uncollapsed estimates of expected frequencies that can be derived from the parameter estimates.

Misscelaneous issues related with loglinear analysis

Consider latent budget analysis of a contingency table with elements π_{ij}. The latent budget model generates a *latent* matrix with probabilities π_{ijk} for which i and j are *independent* conditional upon k, the levels of the latent variable (see (15b)). The expected two–way table and the expected latent three–way table are related by $\pi_{ij} = \pi_{ij+}$, and in the *marginal* table π_{ij} i and j are usually dependent. This follows from the collapsibility theorem that states that in this situation the values of $\{u_{12(ij)}\}$ can change when we add up over the third variable (compare Bishop et al., 1975). In fact, latent budget analysis is a model that can be used to describe the *departure from independence* by assuming a latent variable having more than one level: for one level the latent budget model comes down to the independence model: $\pi_{ij} = \pi_{ij1}$.

The situation becomes more complicated when we work with latent budget analysis for contingency tables with elements π_{isj}, where i and s index two (explanatory) row variables and j indexes the (response) column variable. The loglinear model for the latent table π_{isjk} is given in (27b), which shows that the u–parameters $\{u_{13(ij)}\}$, $\{u_{23(sj)}\}$, $\{u_{123(isj)}\}$ and $\{u_{1234(isjk)}\}$ are all zero. It follows from the collapsibility theorem that in the marginal table $\pi_{isj} = \pi_{isj+}$ these u–parameters will generally be different from zero. Again, for K=1 we find an interesting *baseline model* in which

$\{u_{13(ij)}\}$, $\{u_{23(sj)}\}$, $\{u_{123(isj)}\}$ are all zero since for K=1 we find $\pi_{ij} = \pi_{ij1}$. So if this model does not fit adequately due to the fact that this interaction cannot be neglected, we try to model this interaction by assuming K>1.

An interesting special case arises when we constrain the loglinear model for the latent matrix with elements π_{isjk} further, for example we assume that loglinear model (27b) has as extra restrictions that $\{u_{23(sj)}\}$, $\{u_{123(isj)}\}$ are zero. In this particular case these interactions will also be zero in the collapsed table π_{isj}, as the collapsibility theorem indicates. The α_{isk} parameters reflect this: for model (27b) with $u_{23(sj)} = u_{123(isj)} = 0$ we find that $\alpha_{isk} = \alpha_{is'k} = \alpha_{is''k} = ...$ for all possible (i,k). Thus we can impose this special class of equality constraints by choosing an appropriate loglinear model for the latent probabilities.

When we fit the loglinear model (4b) with only extra restriction $u_{123(isj)} = 0$, these u–terms will not be zero in the collapsed table with elements π_{isj}, as the collapsibility theorem indicates. So in the collapsed table there can be interaction between $u_{123(isj)}$, but by controlling for the latent variable this interaction disappears. The study of the row parameters can be simplified by studying only the average parameters $\Sigma_i (\pi_{is++}/\pi_{i+++})\alpha_{isk}$ and $\Sigma_s (\pi_{is++}/\pi_{+s++})\alpha_{isk}$ that correspond with the main effects. An example is given in section 4.

3.4 Degrees of freedom and identifiability

As usual the number of degrees of freedom is to be calculated as

$$\#\,df = \#\text{independent cells} - \#\text{independent parameters}$$

However, it is sometimes not easy to derive the number of independent parameters in case of parameter constraints.
The number of independent cells is I(J–1), where I is the total number of rows, and J is the total number of columns of the matrix. In the unrestricted case with K latent budgets the number of row parameters is I(K–1), the number of column parameters is (J–1)K, and due to the fact that these parameters are not independent since $\mathbf{AB'} = \mathbf{ATT^{-1}B'}$, with $\mathbf{Tu} = \mathbf{1}$, we have to subtract the total number of free elements of T from the estimated parameters. This number is K(K–1). Hence in the unrestricted case we get

$$\#df = I(J-1) - [I(K-1) + (J-1)K - K(K-1)] = (I-K)(J-K)$$

The difficulty in the derivation of the number of degrees of freedom is that, by imposing constraints, we sometimes identify the model (partly), so that we should not subtract K(K–1) from the number of degrees of freedom.

First we discuss fixed value constraints. As discussed in section 1, in the

unrestricted case we usually choose such a **T** that we get as many zero parameters estimates in **A** or **B** as possible, thus simplifying the interpretation. This can be seen as constraining these estimates to be zero. If we first constrain all these K(K–1) values, than each subsequent fixed value constraint fitted to an independent parameter decreases the number of independent parameters fitted with 1.

Similar things can be said about equality constraints. First we should constrain K(K–1) parameters so that the model is identified. The decrease in the number of independent parameters due to imposing equality constraints can then easily be derived.

By imposing constraints on row parameters (25) and/or column parameters (29) we sometimes identify the model. If this happens, K(K–1) values should not be subtracted from the number of parameters fitted. So the question is: when do these estimates identify the model? We are not yet able to give a general rule for this. If the model is identified by imposing constraints like (25) or (29), then the matrix T in $\mathbf{AB'} = \mathbf{ATT^{-1}B'}$ is restricted to be the identity matrix (compare Mooijaart, 1982), and we do not subtract K(K–1) from the number of estimated parameters. This seems to happen in all constrained models, except in the class of models where mulinomial logit constraints can be fitted by adding up the matrix over some of the row variables and performing an unconstraint analysis of the collapsed matrix (see section 3.3).

If one is not sure about the number of degrees of freedom, this number can be derived using the standard methods, such as determining the rank of the matrix with partial derivatives of the probabilities with respect to the parameters.

4. Example: Social Milieu and Secundary Education

In the Netherlands children go at the age of 11–12 from primary school to secundary school. Distinct types of secundary education can be chosen, with two main types: vocational types of education and general types of education. A choice will depend on aspects such as, among others, capacities of children, interests, advice of the primary school teacher, advice of parents. In educational research much interest goes out to in which way the social milieu of a child influences this choice.

In 1977 and 1981 information was collected from more than 37,000 children about their social milieu and aspects regarding their secundary education. Distinct variables were collected, see for a description CBS (1982) and Meester and de Leeuw (1983). We reanalyze part of the data that were published in Meester and de Leeuw (1983).

The variables we will use in our analysis are the scores on an intelligence test, social milieu (profession of father), sex and the level of education attained in 1981, i.e. after four years of secundary education. The intelligence test used was the (Dutch) Test for Intellectual Capacity (TIC), a figure exclusion test that consist of 33 items. The TIC scores were recoded as 1 for 1 to 14 items right, 2 for 15 to 17 right, 3 for 18 to 20, 4 for 21 to 23, 5 for 24 to 26, 6 for 27 to 29, and 7 for 30 to 33 items right. The social milieu of the family is measured by the profession of the father, in six categories: 1 is skilled and unskilled laborers, 2 is farmers and farm laborers, 3 is shopkeepers, 4 is lower employees, 5 is middle employees and 6 is higher employees, scientific and free professions. The last explanatory variable is the variable sex, with 2 categories. The response variable is the Level of education attained after 4 years, and these levels are 1. dropped out, 2. Junior vocational education (LBO), 3. General education, medium level (MAVO), 4. General education, high level (HAVO), 5. General education, preparing for university (VWO) and 6. Senior vocational training ((M)BO).

Meester and de Leeuw (1983) have eliminated all children having no TIC score (16,433 children), a missing value on level of education attained (38) or on an education type called extraordinary lower education (646) from the sample. Children having a father who is unemployed, or medically unfit for work are also eliminated (6,190). After these selections we have a sample of 16,236 children.

We will analyze the data with latent budget analysis by coding the levels of sex, social milieu and TIC as 2x6x7=84 rows and the level of education attained as 6 columns. The latent budget model with K=1 (independence) is equivalent to the loglinear model for the observed variables where the variables sex, social milieu and TIC are dependent, and independent from level of education attained. This model can be considered as our baseline model. It has a fit of $G^2 = 4612$, with df = 415.

The latent budget model with K>1 can be given a nice interpretation in this context, namely that of a MIMIC model (see Clogg, 1981, who used this interpretation for the analysis of social mobility tables). See figure 1 for K=2. The interpretation is

that we assume K latent classes inbetween the observed explanatory variables and the observed response variable. Each child has K probabilities to come into the K classes. These K probabilities add up to one and are determined by the explanatory variables, namely the child's sex, his/her TIC score, and the social milieu: these probabilities are given by the α–parameters. Once a child is in one of the K latent classes, he/she has J probabilities to attain each of the J levels of education. These J probabilities add up to one and are given by the β–parameters.

A sensible approach to the analysis is first to determine the number of latent classes K that is needed to give an adequate description of the data. See table 4a for the G_2 values for unrestricted models from K=2 to K=5. All of the models have to be rejected at p=.05 level. To check whether this could be due to the specific form of our models, we studied the residuals of the least restricted model, i.e. the model with K=5. We found no understandable patterns in the residuals, or specific outlier cells, so we assume that the misfit of the models is due to large sample size.

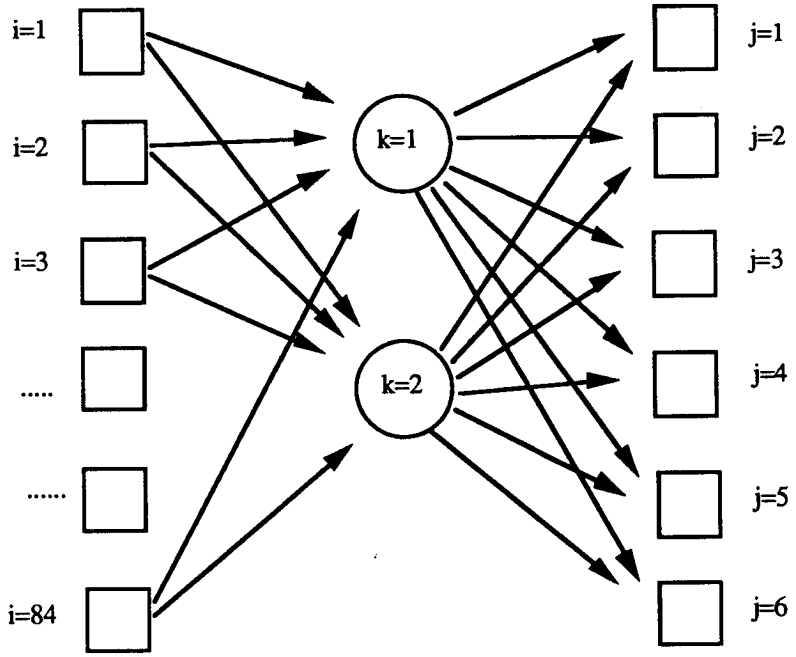


Figure 1: LBA as MIMIC model: children go with some probability from one out of 84 rows to two latent states, and from these states they go with some probability to the end levels in school

Table 4. Chi-squared tests for latent budget models

Table 4a. Tests to determine the number of latent budgets

	G^2	df	% of departure from K=1
K=1	4612	415	0.0
K=2	1113	328	75.8
K=3	441	243	90.4
K=4	226	160	95.1
K=5	116	79	97.5

Table 5: Latent budgets for K = 2, 3, 4 and 5 for educational level after four years of secundary school

	K=2		K=3			K=4			
	k=1	k=2	k=1	k=2	k=3	k=1	k=2	k=3	k=4
1. Drop out	.011	.100	.160	.014	.011	.184	.000	.008	.042
2. LBO	.000	.392	.658	.000	.000	.692	.188	.000	.000
3. MAVO	.171	.208	.121	.090	.325	.124	.198	.059	.357
4. HAVO	.346	.073	.000	.367	.232	.000	.000	.307	.420
5. VWO	.335	.000	.000	.530	.000	.000	.000	.625	.000
6. (M)BO	.137	.228	.061	.000	.432	.000	.614	.000	.181
	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Budget size	.423	.577	.343	.267	.389	.266	.225	.227	.283

	K=5					Independence
	k=1	k=2	k=3	k=4	k=5	K=1
1. Drop out	.190	.009	.000	.063	.025	.063
2. LBO	.809	.000	.222	.000	.000	.226
3. MAVO	.000	.099	.093	.725	.000	.192
4. HAVO	.000	.223	.000	.000	.878	.188
5. VWO	.000	.670	.000	.000	.000	.142
6. (M)BO	.000	.000	.685	.211	.097	.189
	1.000	1.000	1.000	1.000	1.000	1.000
	.228	.211	.188	.212	.161	

The column parameters of the latent budget model, i.e. the latent budgets, are shown for each of the models in table 5. For K=2 the latent budget for k=1 is the budget where children have a very low probability to drop out, they do not go to the lower vocational training (LBO) and less than average to medium and high vocational training (M)BO, but go instead more than average (see K=1) to general education.

For k=2 they drop out more than average, go more than average to vocational training, and less than average to general education. For K=3 in k=1 children go more than average to lower vocational training (LBO) or drop out, and to a certain extent they go to medium general education (MAVO) and (M)BO. In budget k=2 children go more than average to higher general education (HAVO and VWO), and in budget k=3 they go more than average to medium and higher general education (MAVO and HAVO) and higher vocational training (MBO), but not to general education, preparing for university (VWO). In K=4 budget k=3 is similar to budget 2 of K=3. Compared with K=3 the budgets having large probabilities for vocational training change: a new budget, namely k=2, is found, derived from k=1 and k=3 in K=3. It shows a high probability to do (M)BO, and a relatively high probability to do LBO and MAVO. In K=5 the budget for VWO is still approximately the same (k=2), and again the budgets with non–zero probabilities for vocational training split up. Overall, we find, in going from K=3 to K=5, that there is always only one budget with a non–zero probability to go to VWO and zero probabilities for vocational training LBO and (M)BO. If the number of budgets increase, a more refined description is given of budgets having non–zero probabilities to go to vocational training.

Given the large sample size, we are satisfied with the description that latent budget analysis with K=3 offers. Although significant, the discrepancy between the G^2 = 441 and df = 243 is not enormous, the model describes 90.4% of the departure from independence between the explanatory variables and the response variables (i.e. .904 = (4612–441)/4612). The gain in percentage in going from K=3 to K=4 is relatively small. Therefore we will now study this model more carefully.

We start with a study of the row parameters. We have derived plots of the row parameters, separately for each TIC–score and each sex. This gives 7x2=14 plots, shown in figure 2. In each plot we have set out horizontally the six levels of social milieu, and vertically the probability to go to one of the latent budgets. Each plot has 18 points, namely children in each of the 6 levels of social milieu can go to each of the three latent budgets; points belonging to the same latent budgets are connected, so that each plot has three lines. We have chosen to display the parameters in this way for the following reasons.

Firstly, if sex would have no influence on the probability to go to latent budgets, then plots on the left (boys) would be identical to plots on the right (girls). This way of displaying the parameters clearly shows the influence of sex by looking in which way each pair of plots differs. Secondly, if the social milieu would have no influence on the probability to go to the latent budgets, then all lines would be horizontal, and departures from this are easily displayed in this way. It is clear that the probability to go to a latent budget will be strongly influenced by the TIC–score, since the levels attained do not only reveal differences between types of education, but also niveau. Therefore, in going from the plots on the top (TIC–

Figure 2. Plots of row parameters for unrestricted model with K=3 budgets Separate plots for each combination of sex and TIC-score. A line in pl connects probabilities to go to identical latent budgets. Horizontally the level of social milieu are set out (1=laborer, 2=farmer, 3=shopkeeper, 4=low employee, 5=medium employee, 6=higher employee), vertically th probabilities to go to each of the three latent budgets

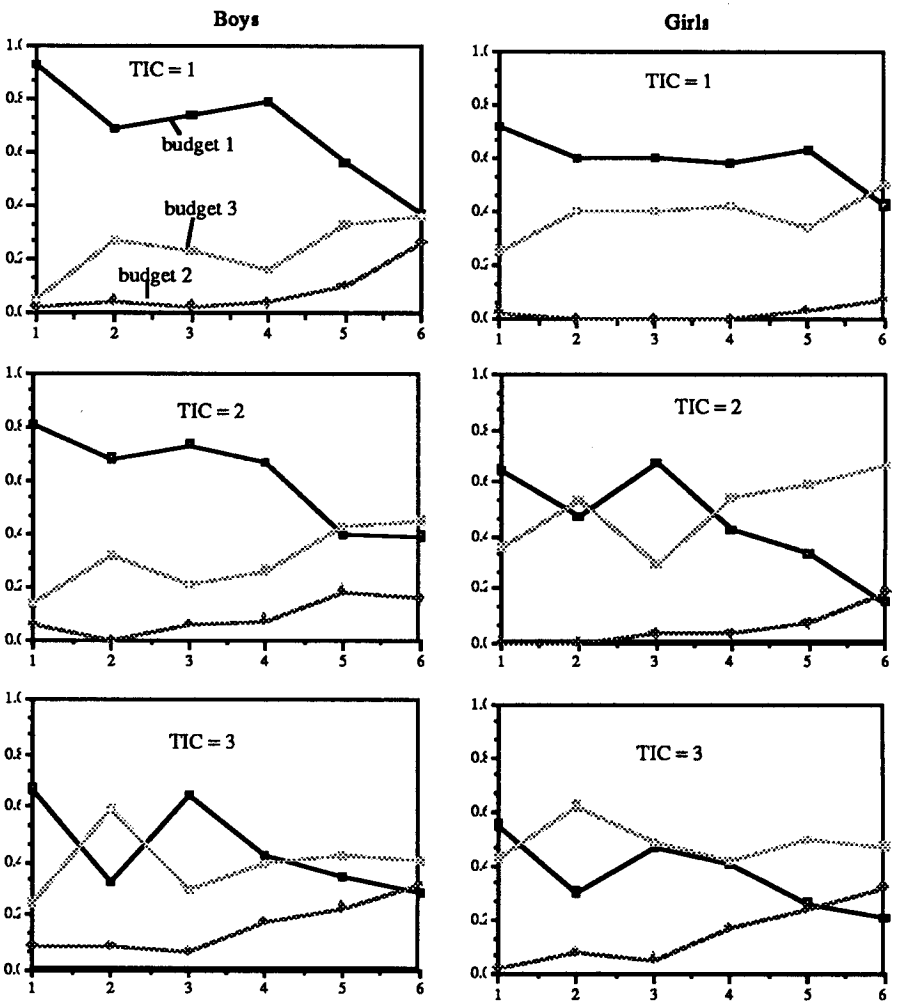
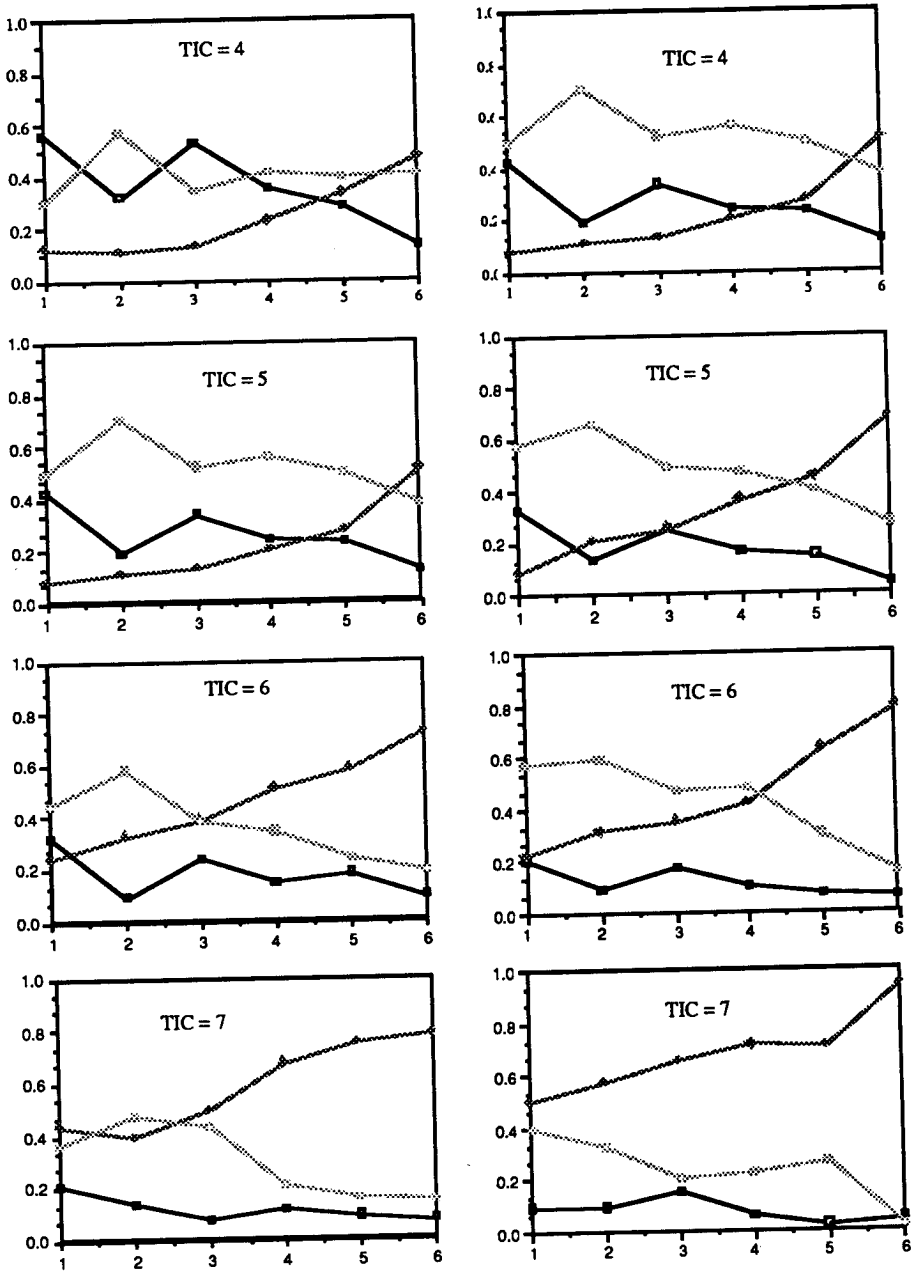


Figure 2 continued



score equals 1) to the bottom (TIC–score equals 7) the line with open blocks drops generally; this is not surprising since this line shows the probability to go to latent budget 1, which is the budget in which 65.8% of the children go to LBO and 16.0% drops out: children more often drop out or go to LBO when their TIC–score is lower.

Many striking aspects from these plots can be emphasized. To pick out a few: in all levels of TIC, children with fathers being medium or higher employees (5 and 6) have a probability much higher than average to go to latent budget 2, which is the budget for higher general education (HAVO, 36.7%) and prepatory for university (VWO, 53.0%). Their probability to go to budget 1 (drop out and LBO) is much lower. The reverse holds for children whose parent is skilled or unskilled laborer: given their TIC–score, their probability to go to budget 1 is in general the highest. On average, children whose parent is farmer (2) are more likely than average, given their TIC–score, to go to latent budget 3, where they have a high probability to follow medium vocational training ((M)BO). Notice that the latent budget parameters, being probabilities, allow for an easy interpretation: they do not only show that something is going on (for example, girls go on average less to budget 1 than boys), they also show how strong the effects are.

We are now going to test the effects in the plots of figure 2 by constraining the α_{ik}–parameters using the factorial structure given by the explanatory variables. Test results are shown in table 4b. First we constrain all the lines to be horizontal, i.e. in our design matrix **X** used in equation (24) we have a general term, a column for sex, 6 columns for TIC, and six columns for the interaction of sex and TIC. We restrict ourselves to tests using hierarchical models, so that we can use iterative proportional fitting in each step of the EM–algorithm, and we will denote the models for α_{ik} by placing the highest levels of the terms fitted between brackets. So this model can be denoted as ST, for the interaction between sex and TIC score. It has a very poor fit: $G^2 = 2101$. In the population the lines are obviously not horizontal. A similar model that constrains each boy plot to be identical to its corresponding girl plot fits much better, but still quite poor: this model PT (P for profession father – social milieu) has $G^2 = 727$. Notice that these two models could have been fitted by equality constraints on the parameters also (see section 3.3).

We will now impose systematically all possible constraints on the row parameters, with as most restrictive model the model with only main effects Sex, Profession (Social Milieu), and TIC–score (see table 4b). The model SP,ST,PT assumes that there is an interaction of Sex and Profession on the latent budgets, of Sex and TIC on the latent budgets, and of Profession and TIC on the latent budgets, but that there is no interaction between Sex, Profession and TIC jointly on the latent budgets. The fit is reasonably good ($G^2 = 508$, df = 297), and the drop in fit compared to the unconstrained model is not large: $G^2 = 508 - 441 = 67$, df = 297 – 243 = 54. In going down in figure in table 5b we find that the fit diminishes to G^2=627 for

Table 4b. Models constraining α–parameters for K=3. Hierarchical models, only terms including latent budgets are given. S is sex, T is TIC–score, P is profession father (social milieu). In blocks the models, their fit (G^2) and #df is given, next to lines terms constrained to be zero and conditional tests are given. For more details, see text

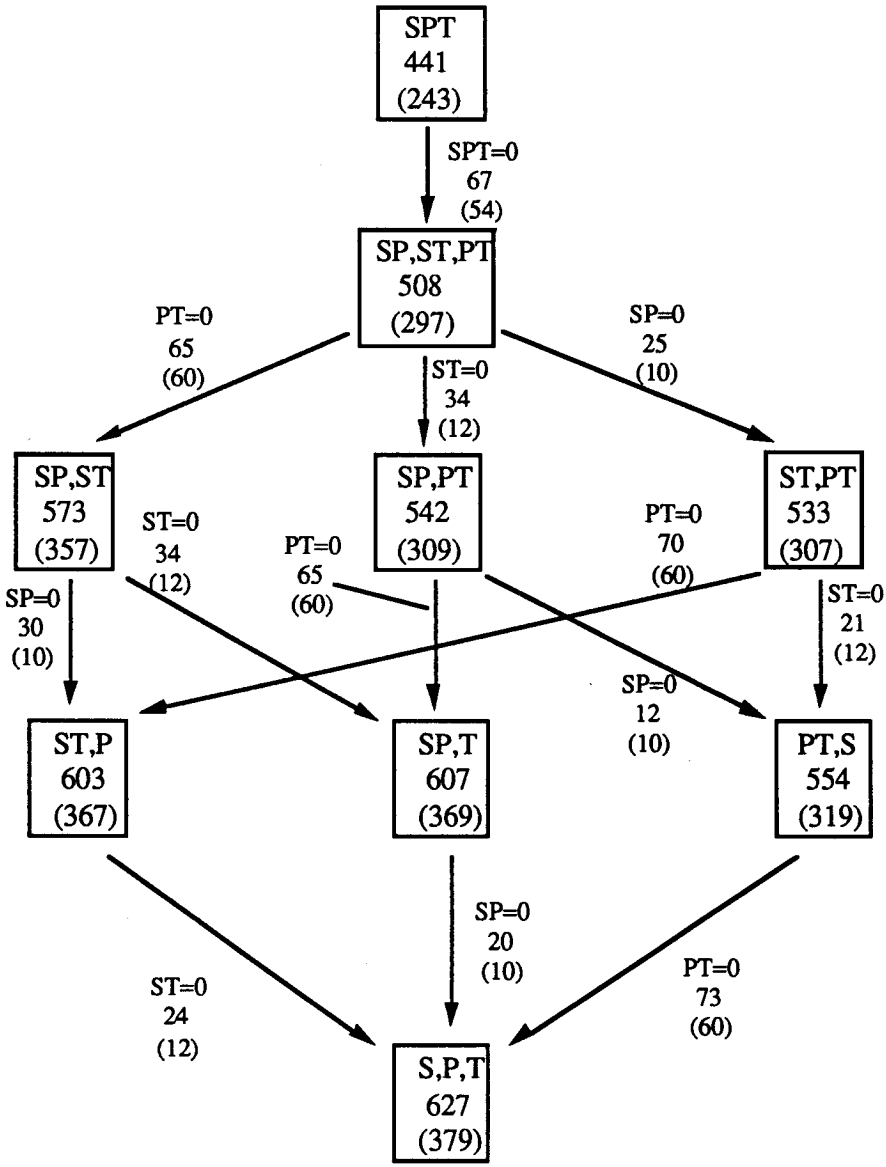


Table 6: latent budget for K = 3 for model S,P,T in table 4b

	K=3		
	k=1	k=2	k=3
1. Drop out	.177	.025	.005
2. LBO	.701	.038	.006
3. MAVO	.092	.090	.331
4. HAVO	.000	.337	.228
5. VWO	.015	.500	.000
6. (M)BO	.015	.011	.430
	1.000	1.000	1.000
Budget size	.304	.274	.422

df=379 Compared with the unconstrained model we win 379–243=136 degrees of freedom, whereas the loss in fit is 627–441=186. We certainly loose information, but interpretation becomes much easier, and the parameter estimates gain stability. Some information is lost by imposing PT=0 (conditional tests G^2 are between 65 and 73 for 60 df) and somewhat more by imposing SP=0 (conditional tests are between 12 and 30 for 10 df) and by imposing ST=0 (conditional tests are between 21 and 34 for 12 df). We now interpret the most restrictive model S,P,T.

The latent budget parameters are similar to the unrestricted model with K=3, compare table 6 with table 5. The plots with α_{ik}–parameters are sometimes quite different, see figure 3. These plots are perhaps most easily studied by deriving weighted averages of row parameters (see section 3.3), so that we obtain parameters for TIC only, for social milieu only, and for sex only, see figure 4.

In figure 4 we see in the plot for TIC–score that the probability to go to budget 1 (mainly LBO, drop out) decreases as TIC increases, the probability to go to budget 2 (mainly VWO, HAVO) increases as TIC increases, and the probability to go to budget 3 (MAVO, HAVO, MBO) increases from TIC 1 to 4, and then decreases smoothly. In the plot for Social Milieu the probability to go to budget 1 is lower for children of farmers (2) and medium and higher employees (5,6), the probability to go to budget 2 increases rapidly for children from lower to higher employees, and the probability to go to budget 3 is somewhat higher than average for children from farmers and somewhat lower for children from higher employees. In the plot for Sex we find that there is no difference for boys and girls in their probability to go to budget 1. There is a difference in their probability to go to budget 2 and 3: for boys these probabilities are approximately equal, whereas girls go less often to budget 3 and more often to budget 2. In the plots of figure 3 the effects shown in the three plots of figure 4 are brought together.

Figure 3: Plots of row parameters for restricted model 6 (see table 4
Separate plots for each combination of sex and TIC–score

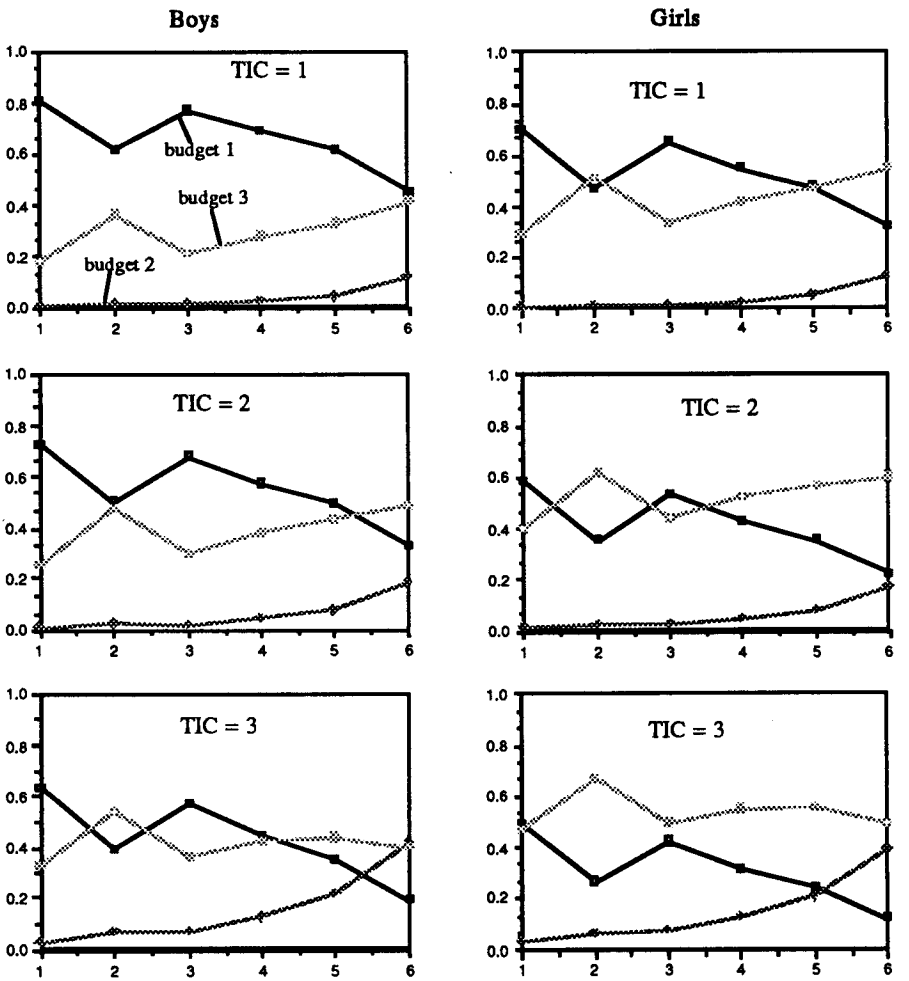


Figure 3 continued

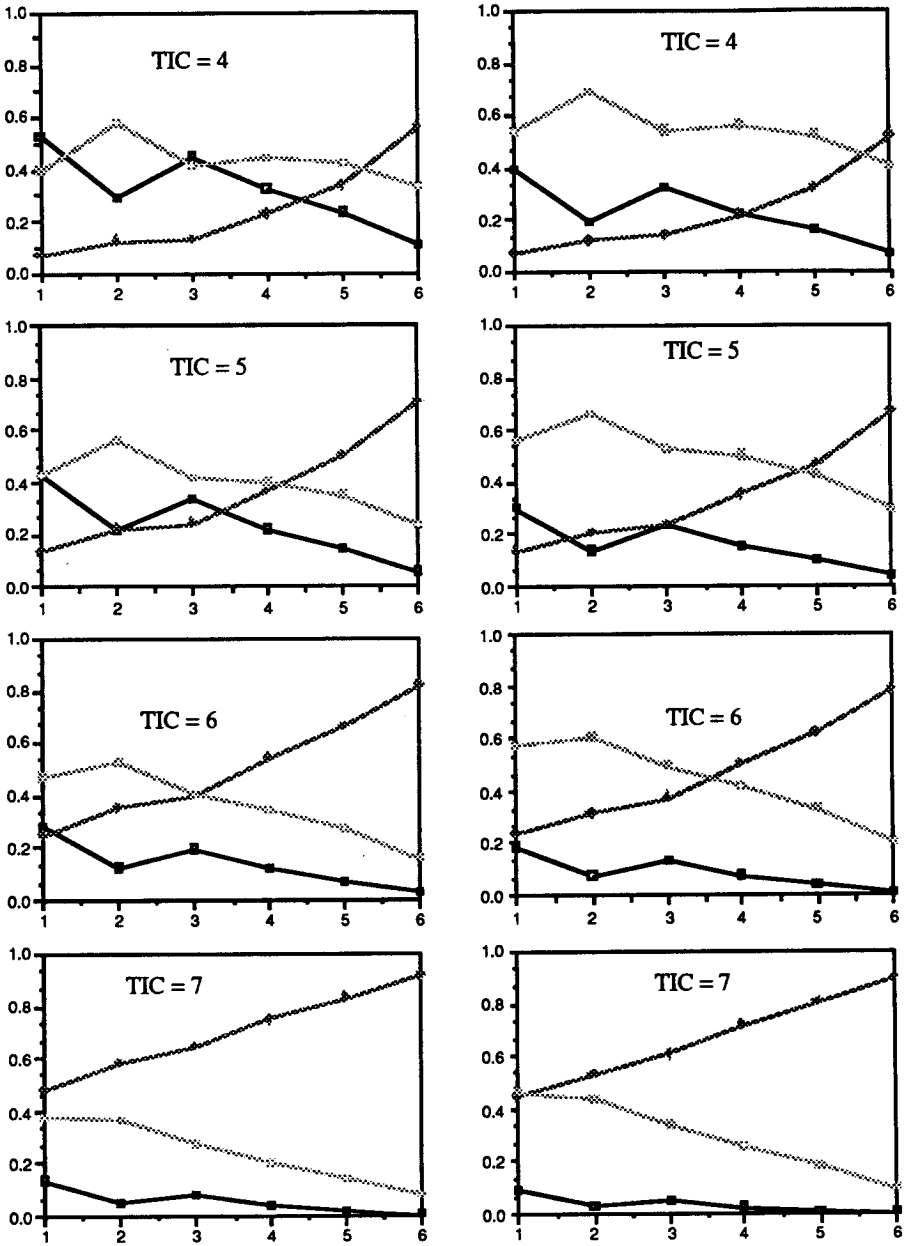
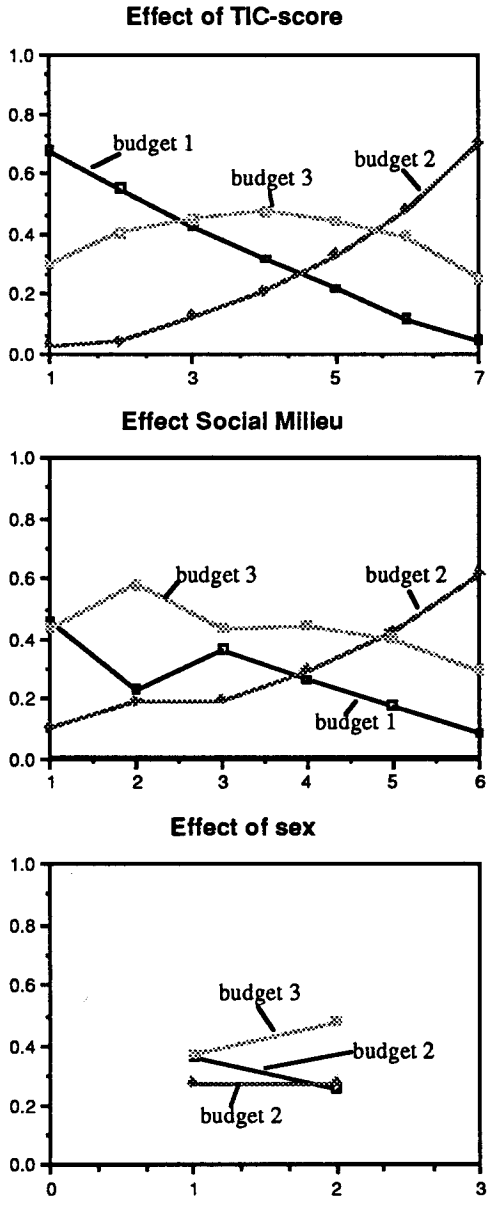


Figure 4, main effects for model 6 in table 4b.
Horizontally the levels of each of the explanatory variables can be found, vertically the probabilities to g to each of the latent budgets



Latent budget analysis gives much insight into these data. The MIMIC–model like interpretation that we use shows the probabilities with which children, having a specific background, go to specific latent budgets. These latent budgets specify the probabilities to reach specific ending levels of education. Also in this example the parameters are very easily interpreted, so that it is easy to indicate the processes that operate in the relation between explanatory variables such as TIC, sex and social milieu on the one hand and secundary education on the other.

5. Conclusion and discussion

We showed how to built in three types of constraints in latent budget analysis, making the model quite flexible for many situations. The constraints were fitted using the EM–algorithm. There are two problems with this approach, first, the amount of computing time, and, second, generality of the constraints used.

First, by constraining the model the number of iterations needed to reach convergence increases rapidly (although each iteration costs only little time). For example, for some of the constrained models fitted to the SMVO data the number of iterations was sometimes more than 2500, were the stopping criterion used was a difference in subsequent G^2 of $.2x10^{-6}$. This large number of iterations is also partly due to the largeness of the matrix analyzed and the extreme stopping criterion. There are two ways out of this problem. A first way is by using methods to speed up the EM algorithm. A second way to to use a different algorithm, for example Newton Raphson or Fletcher Powell.

By using a different algorithm like Fletcher Powell we will also be able to tackle the second problem, namely to impose more general constraints than the constraints imposed in this paper (see, for some examples, Langeheine, 1989). We are currently working on this problem.

References

Aitchison, J. (1986). *The statistical analysis of compositional data.* London: Chapman and Hall.

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A, 144,* 419–461.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis.* Cambridge: M.I.T.Press

Bock, R.D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw–Hill.

Breiger, R.L. (1981). The social class structure of occupational mobility. *American Journal of Sociology, 87,* 578–611.

C.B.S. (1982). *School career and origin of pupils in secondary education. Part 2: cohort 1977, choice of school type.* (in dutch). The Hague: Staatsuitgeverij.

Caussinus, H. (1986). In discussion of: L.A. Goodman, Some useful extensions of the usual correspondence analysis approach and the usual loglinear models approach in the analysis of contingency tables. *International statistical review, 54,* 243–309.

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

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

de Leeuw, J. and 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, p.159–166.

de Leeuw, J., and van der Heijden, P.G.M. (1989). *Reduced rank models for contingency tables.* Leiden: Department of Psychometrics and Research Methods, internal report PRM 89–04.

de Leeuw, J., van der Heijden, P.G.M., and Verboon, P. (1990). A latent time budget model. *Statistica Neerlandica.*

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

Everitt, B.S. (1988), *Multivariate Behavioral Research,* 23, 531–538

Fienberg, S.E. (1980). *The analysis of cross–classified categorical data. 2nd ed.* Cambridge: M.I.T.–Press.

Formann, A.K. (1978). A note on parameter estimation for Lazarsfeld's latent class analysis. *Psychometrika, 43,* 123–126.

Formann, A.K. (1982). Linear logistic latent class analysis. *Biometrical Journal, 24,* 171–190.

Formann, A.K. (1985). Constrained latent class models: theory and applications. *British Journal of Mathematical and Statistical Psychology, 38,* 87–111.

Formann, A.K. (1989). Constrained latent class models: some further applications. *British Journal of Mathematical and Statistical Psychology, 42,* 37–54.

Gilula, Z. (1979). Singular value decomposition of probability matrices: probabilistic aspects of latent dichotomous variables. *Biometrics,* **66**, 339–344.

Gilula, Z. (1983). Latent conditional independence in two–way contingency tables: a diagnostic approach. *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 eploratory canonical correlation approach. *Journal of the American Statistical Association, 81,* 773–779.

Gilula, Z. and Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*, **81**, 780–788.

Gilula, Z. and 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.

Gilula, Z. and Krieger, A.M. (1989). Collapsed two–way contingency tables and the chi–square reduction principle. *Journal of the Royal Statistical Society, Series B, 51*, 425–433.

Good, I.J. (1969). Some applications of the singular decomposition of a matrix. *Technometrics*, **11**, 823–831.

Goodman, L.A. (1971). Partitioning of chi–square, analysis of marginal contingency tables, and estimation of epected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association, 66*, 339–344.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Goodman, L.A. (1981). Criteria for determining whether certain categories in a cross–classification table should be combined with special reference to occupational categories in occupational mobility tables. *American Journal of Sociology, 87*, 612–650.

Goodman, L.A. (1985). The analysis of cross–classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**, 10–69.

Goodman, L.A. (1986). Some useful extensions to the usual correspondence analysis approach and the usual loglinear approach in the analysis of contingency tables (with comments). *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.

Grover, R. (1987). Estimation and use of standard errors of latent class model parameters. *Journal of marketing research, 24*, 298–304.

Grover, R. and Srinivasan, V. (1987). A simultaneous approach to market segmentation and to market structuring. *Journal of Marketing Research, 24*, 139–153.

Haberman, S.J. (1979). *Analysis of qualitative data (2 vols.)*. New York: Academic Press.

Hagenaars, J.A. (1986). Symmetry, quasi–symmetry, and marginal homogeneity on the latent level. *Social science research, 15*, 241–255.

Hagenaars, J.A. (1988). Latent structure models with direct effects between indicators. Local dependence models. *Sociological methods and research, 16*, 379–405.

Langeheine, R. (1989). New developments in latent class theory. In: R. Langeheine

and J. Rost. *Latent trait and latent class models.* New York: Plenum Press. pp. 77–108.

Luijk, R. (1987). Loglinear modelling with latent variables: the case of mobility tables. In: W.Saris and I.Gallhofer (Eds.) *Sociometrics Research: Vol.2.* London: MacMillan.

Marsden, P.V. (1985). Latent structure models for relationally defined social classes. *American Journal of Sociology, 90,* 1002–1021.

Meester, A. and de Leeuw, J. (1983). *Intelligence, social milieu and the school career* (in dutch). Leiden: Department of Data Theory.

Mooijaart, A. (1982). Latent structure analysis for categorical variables. In: K.G. Joreskog and H. Wold (Eds.) *Systems under indirect observation.* Amsterdam: North Holland.

Shigemasu, K., and Sugiyama, N. (1989). *Latent class analysis of choice behavior.* Presentation at the Annual Meeting of the Psychometric Society, July 1989, University of California Los Angeles.

Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika, 52,* 493–513.

van der Heijden, P.G.M., Mooijaart, A. and de Leeuw, J. (1989). Latent budget analysis. In: A. Decarli, B.J. Francis, R.Gilchrist and G.U.H.Seeber (Eds.). *Statistical Modelling. Proceedings, Trento, 1989.* Berlin, Springer Verlag.

Appendix: the SMVO data

School types: 1=drop out, 2=LBO, 3=MAVO, 4=HAVO, 5=VWO, 6=(M)BO. Social Milieu (SES) 1=skilled and unskilled laborers 2=farmers and farm laborers, 3=shopkeepers, 4=lower employees, 5=middle employees, 6=higher employees. TIC scores: number of items correct are 1=1–14, 2=15–17, 3=18–20, 4=21–23, 5=24–26, 6=27–29, 7=30–33

		Boys						Girls					
School type		1	2	3	4	5	6	1	2	3	4	5	6
SES 1	TIC 1	43	126	23	5	2	17	.28	87	24	13	3	35
	2	41	172	58	20	9	28	29	131	57	15	0	74
	3	50	271	83	58	24	87	67	209	128	59	6	141
	4	64	268	131	93	44	111	64	200	157	95	34	194
	5	43	202	121	113	47	109	35	163	177	105	39	201
	6	11	78	60	62	43	78	20	54	106	92	48	103
	7	4	15	20	23	27	19	2	10	22	40	38	28

SES 2	TIC 1	3	13	1	1	1	8	2	8	5	1	0	5
	2	3	18	9	0	0	10	2	14	10	4	0	12
	3	2	18	12	15	3	23	5	18	16	19	3	26
	4	8	25	15	14	9	47	0	18	23	21	8	46
	5	5	25	16	12	16	35	0	13	28	21	15	39
	6	2	4	7	20	11	22	5	6	19	37	15	30
	7	0	3	2	5	7	9	0	4	4	12	17	10
SES 3	TIC 1	11	17	6	1	1	10	7	12	11	2	0	8
	2	9	37	11	6	2	10	6	29	11	5	1	11
	3	23	59	26	12	6	29	16	43	30	19	4	38
	4	12	72	34	23	14	38	18	39	39	36	13	49
	5	11	40	26	37	25	36	16	32	54	54	25	39
	6	7	20	26	25	30	25	11	12	28	41	20	24
	7	3	1	7	9	12	9	2	3	3	16	7	3
SES 4	TIC 1	9	29	13	4	1	4	3	15	6	3	0	10
	2	9	38	21	5	4	13	10	24	26	7	2	29
	3	12	56	47	37	15	27	12	54	40	37	15	35
	4	11	62	52	54	26	43	15	39	64	56	27	61
	5	12	48	62	55	37	30	9	31	54	87	44	52
	6	6	15	33	40	45	24	7	11	35	49	39	39
	7	3	4	7	17	23	7	2	3	5	23	26	9
SES 5	TIC 1	5	25	14	9	3	9	6	20	8	3	1	12
	2	8	26	30	23	7	11	9	22	24	19	4	30
	3	13	60	65	39	35	50	10	42	50	44	33	59
	4	20	79	91	94	71	70	17	58	97	82	55	79
	5	11	58	70	95	95	63	11	44	89	103	101	70
	6	9	39	44	71	107	40	5	17	46	117	104	47
	7	4	7	9	28	57	12	2	3	28	49	70	21
SES 6	TIC 1	4	6	10	6	4	3	5	2	6	1	1	5
	2	7	14	15	11	5	12	4	3	6	18	2	11
	3	5	31	34	39	21	23	5	16	24	33	16	21
	4	10	16	45	54	52	36	9	16	44	83	46	29
	5	7	16	44	71	105	28	7	7	40	80	83	27
	6	3	12	24	40	85	19	8	7	32	66	100	15
	7	3	4	9	16	52	9	1	3	10	29	51	1