Find Us on the Web! http://www.apnet.com

This book is printed on acid-free paper. ∞

The cover design is an inverted part of a time series of weekly amounts of rye sold in Cologne, 1542–1648, showing a long-term trend line. The time series is reminiscent of the Cologne cathedral.

Academic Press
525 B St. Suite 1900, San Diego, CA 92101-4495, USA
1300 Boylston Street, Chestnut Hill, MA 02167, USA

Academic Press Limited
24-28 Oval Road, London NW1 7DX, UK
http://www.hbuk.co.uk/ap/

Library of Congress Cataloging-in-Publication Data

Visualization of categorical data / [edited by] Jörg Blasius, Michael Greenacre.
p. cm.
Includes bibliographical references and index.
ISBN 0-12-299045-5 (alk. paper)
1. Multivariate analysis--Graphic methods. I. Blasius, Jörg, 1957- . II. Greenacre, Michael L.
QA278.V57 1997
001.4'226--dc21

97-35782
CIP

Printed in the United States of America
97 98 99 00 EB 9 8 7 6 5 4 3 2 1

Visualization of Categorical Data

Edited by

JÖRG BLASIUS
Zentralarchiv für Empirische Sozialforschung
University of Cologne
Cologne, Germany

MICHAEL GREENACRE
Departament d'Economia i Empresa
Pompeu Fabra University
Barcelona, Spain

ACADEMIC PRESS
San Diego London Boston
New York Sydney Tokyo Toronto

Contents

Chapter 33

Graphical Display of Latent Budget Analysis and Latent Class Analysis, with Special Reference to Correspondence Analysis

L. Andries van der Ark and Peter G. M. van der Heijden

1 Introduction

Latent budget analysis (LBA) and latent class analysis (LCA) are methods for the analysis of contingency tables. They are equivalent techniques that lead to an identical visualization of the results of the data analyses. It is not widely known that LBA and LCA results can be visualized. Aided by two clarifying examples, we will illustrate these visualizations, and we will also show the relation between the graphical representation of LBA and LCA and that of correspondence analysis (CA), another method for the analysis of contingency tables.

The first set of data was originally published and analyzed by Guttman (1971). It is a two-way contingency table about the principal worries of Israeli adults (Table 1). The row variable is a combination of residence and father's residence, denoted by "residence," with $I = 5$ categories indexed by i. The column variable is the principal worry of the respondents, denoted by "worry," with $J = 8$ categories indexed by j.

The frequency of the cell corresponding to the ith row category and the jth column category is denoted by n_{ij}. The marginal row and column frequencies are denoted by $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$ respectively. The total number of respondents is denoted by $n = \sum_i \sum_j n_{ij}$ (= 1554).

2 The Latent Budget Model

From the data matrix of Table 1 we can construct the matrix of proportions $\mathbf{P}$, with elements p_{ij}, by dividing each element of the data by n: $p_{ij} = n_{ij}/n$. The marginal proportions of the rows and columns are denoted by $r_i \equiv p_{i\cdot} = \sum_j p_{ij}$ and $c_j \equiv p_{\cdot j} = \sum_i p_{ij}$, respectively. Since "residence" is an explanatory variable and "worry" is a response variable, we investigate the conditional proportions of "worry," given "residence," denoted by $p_{j|i} \equiv p_{ij}/p_{i\cdot} = n_{ij}/n_{i\cdot}$, rather than the unconditional proportions p_{ij}. This allows us to compare the categories of the variable "worry" between residence groups. If we collect the $p_{i\cdot}$ as entries of the $I \times I$ diagonal matrix $\mathbf{D}_r$ then the conditional proportions $p_{j|i}$ are found in the matrix $\mathbf{D}_r^{-1}\mathbf{P}$, which is presented in Table 2.

The rows of $\mathbf{D}_r^{-1}\mathbf{P}$ are vectors that contain only nonnegative elements and add up to one. We call such vectors *budgets*, in general, and the rows of $\mathbf{D}_r^{-1}\mathbf{P}$ *observed budgets* (in correspondence analysis these rows are referred to as *row profiles*). Normally, $\mathbf{D}_r^{-1}\mathbf{P}$ is of full rank, that is, $\text{rank}(\mathbf{D}_r^{-1}\mathbf{P}) = \min(I, J)$, equal to 5 in this example. In the latent budget model $\mathbf{D}_r^{-1}\mathbf{P}$ is approximated by $\mathbf{D}_r^{-1}\mathbf{\Pi}$, a matrix of conditional probabilities $\pi_{j|i}$, of rank K [$K \le \min(I, J)$], such that $\pi_{j|i}$ is a mixture of K conditional probabilities $\pi_{j|k}$ ($k = 1, \ldots, K$). The mixing parameters are denoted by $\pi_{k|i}$. The latent budget model can be written as

$$\pi_{j|i} = \sum_{k=1}^{K} \pi_{k|i}\pi_{j|k} \tag{1}$$

The parameters in (1) are subject to the equality constraints

$$\sum_{j=1}^{J} \pi_{j|i} = \sum_{k=1}^{K} \pi_{k|i} = \sum_{j=1}^{J} \pi_{j|k} = 1 \tag{2}$$

and inequality constraints

$$0 \le \pi_{j|i} \le 1,\ 0 \le \pi_{k|i} \le 1,\ 0 \le \pi_{j|k} \le 1 \tag{3}$$

The idea for the latent budget model was introduced by Goodman (1974) and elaborated by Clogg (1981), de Leeuw *et al.* (1990), van der Heijden *et al.* (1992), and Siciliano and van der Heijden (1994). There are two ways to interpret the parameters of the latent budget model, which we will call the *mixture model interpretation* and the *MIMIC-model interpretation* (Multiple Indicator Multiple Cause model; Goodman, 1974). The mixture model interpretation is as follows. If we collect $\pi_{j|i}$ in an $I \times J$ matrix, then the rows of this matrix, denoted by $\boldsymbol{\pi}_i^{\mathrm{T}} = [\pi_{1|i} \ldots \pi_{j|i} \ldots \pi_{J|i}]$, are vectors with nonnegative elements that add up to one. These vectors are called *expected budgets*. The latent budget model writes these expected budgets as a mixture of the vectors $\tilde{\boldsymbol{\pi}}_k^{\mathrm{T}} = (\pi_{1|k} \ldots \pi_{j|k} \ldots \pi_{J|k})$ ($k = 1, \ldots, K$), which are typical or *latent budgets.* We can write (1) as

$$\boldsymbol{\pi}_i^{\mathrm{T}} = \pi_{1|i}\tilde{\boldsymbol{\pi}}_1^{\mathrm{T}} + \cdots + \pi_{k|i}\tilde{\boldsymbol{\pi}}_k^{\mathrm{T}} + \cdots + \pi_{K|i}\tilde{\boldsymbol{\pi}}_K^{\mathrm{T}} \tag{4}$$

Table 1: Principal worries of Israeli adults (Guttman, 1971)*

Residence/ father's residence	Principal worries[a] ENL	SAB	MIL	POL	ECO	OTH	MTO	PER	Total
Asia/Africa	61	70	97	32	4	81	20	104	469
Europe/America	104	117	218	118	11	128	42	48	786
Israel; father Asia/Africa	8	9	12	6	1	14	2	14	66
Israel; father Europe/America	22	24	28	28	2	52	6	16	178
Israel; father Israel	5	7	14	7	1	12	0	9	55
Total	200	227	369	191	19	287	70	191	1554

[a]ENL, enlisted relative; SAB, sabotage; MIL, military situation; POL, political situation; ECO, economic situation; OTH, other; MTO, more than one worry; PER, personal economics.
*Reprinted by permission of the Psychometric Society.

Table 2: Observed budgets

Residence/ father's residence	Principal worries[a] ENL	SAB	MIL	POL	ECO	OTH	MTO	PER	Total
Asia/Africa	.130	.149	.207	.068	.009	.173	.043	.222	1.000
Europe/America	.132	.149	.277	.150	.014	.163	.053	.061	1.000
Israel; father Asia/Africa	.121	.136	.182	.091	.015	.212	.030	.212	1.000
Israel; father Europe/America	.124	.135	.157	.157	.011	.292	.034	.090	1.000
Israel; father Israel	.091	.127	.255	.127	.018	.218	.000	.164	1.000

[a]ENL, enlisted relative; SAB, sabotage; MIL, military situation; POL, political situation; ECO, economic situation; OTH, other; MTO, more than one worry; PER, personal economics.

Hence, each expected budget is built up out of the K latent budgets, and the mixing parameters determine to what extent. If we interpret the latent budget model as a MIMIC model, then $\pi_{k|i}$ denote what proportion of row category i belongs to some latent class k, and $\pi_{j|k}$ denote how the subjects in each latent class k respond to the column categories j.

A schematic representation of a mixture model and a MIMIC model is given in Figure 1. For the mixture model the squares represent the expected budgets π_i and the circles the latent budgets $\tilde{\pi}_k$. The arrows represent $\pi_{k|i}$ and determine how each expected budget is built up in terms of the latent budgets. For the MIMIC model the squares on the left and right represent the row and column categories, repectively. The arrows on the left-hand side represent $\pi_{k|i}$ and the arrows on the right-hand side represent $\pi_{j|k}$. Hence the MIMIC model shows what proportion of each row category falls into each latent category and what proportion of each latent category responds to each column category.

In general, the latent budget model is not identifiable if $K > 1$ and no constraints other than (2) and (3) are imposed on the model. Therefore different sets of parameter estimates may be obtained for different starting values, but they provide the same estimates of the expected budgets. For a discussion of identifiability in the latent budget model we refer to de Leeuw *et al.* (1990) and van der Ark and van der Heijden (1996).

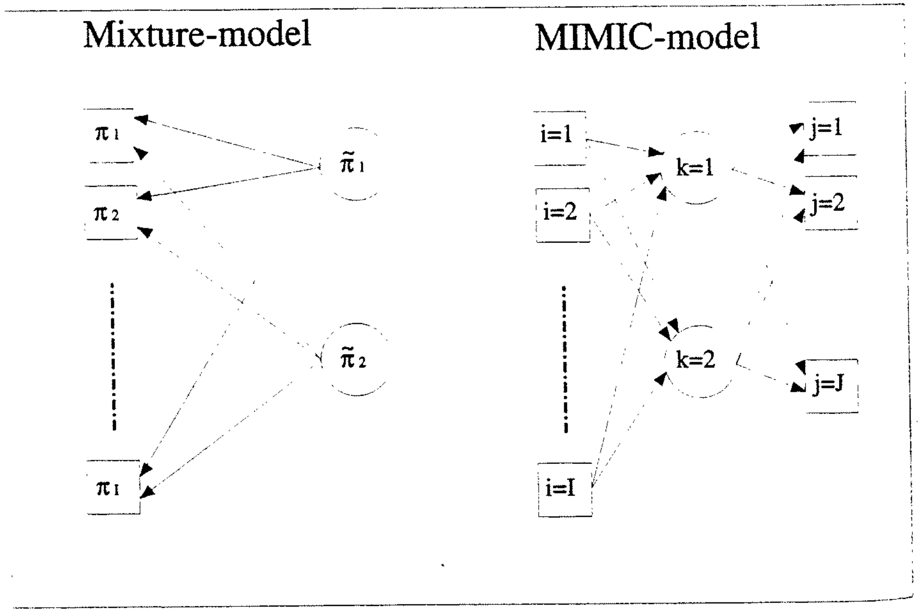

Figure 1: Graphical display of a mixture model and MIMIC model.

Table 3: $K = 1, K = 2$, and $K = 3$ latent budget solutions for the data of Table 1

	$K = 1$ latent budget solution	$K = 2$ latent budget solution		$K = 3$ latent budget solution		
	$k = 1$	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 3$
Mixing parameters						
Asia/Africa	1.000	.383	.617	.000	.477	.523
Europe/America	1.000	.832	.168	.235	.633	.133
Israel; father Asia/Africa	1.000	.424	.576	.116	.402	.482
Israel; father Europe/America	1.000	.721	.279	.436	.353	.210
Israel; father Israel	1.000	.576	.424	.205	.447	.348
Latent budgets						
Enlisted relative (ENL)	.129	.132	.123	.100	.149	.109
Sabotage (SAB)	146	.147	.145	.105	.170	.128
Military situation (MIL)	.238	.286	.145	.021	.429	.011
Political situation (POL)	.123	.187	.000	.250	.145	.000
Economic situation (ECO)	.012	.106	.006	.016	.015	.004
Other (OTH)	.185	.180	.194	.508	.000	.329
More than one worry (MTO)	.045	.054	.028	.000	.084	.000
Personal economics (PER)	.123	.000	.359	.000	.009	.420
Likelihood ratio χ^2	121.5	29.23		6.490		
Degrees of freedom	28	18		10		
Probability	.000	.050		.846		

The matrix $\mathbf{D}_r^{-1}\mathbf{P}$ in Table 2 was analyzed using the maximum likelihood estimation procedure of de Leeuw *et al.* (1990). The results of the latent budget analysis with $K = 1$, $K = 2$, and $K = 3$ latent budgets are presented in Table 3. We can see that the model with $K = 1$ latent budgets does not fit the data. In the model with $K = 2$ latent budgets, the goodness of fit has improved and now $100(121.5 - 29.2)/121.5 = 75.9\%$ of the dependence is modeled, but the fit is still not satisfactory. The model with $K = 3$ latent budgets fits the data very well, with 94.7% of the dependence modeled. We have transformed the parameter estimates such that $\hat{\pi}_{j=8|k=1} = \hat{\pi}_{j=4|k=2} = 0$ in the $K = 2$ latent budget model and $\hat{\pi}_{j=7|k=1} = \hat{\pi}_{j=8|k=1} = \hat{\pi}_{j=6|k=2} = \hat{\pi}_{j=7|k=3} = \hat{\pi}_{j=4|k=3} = \hat{\pi}_{k=1|i=1} = 0$ in the $K = 3$ latent budget model. These transformations were chosen so that as many parameter estimates as possible equal zero without altering the goodness of fit (see van der Ark and van der Heijden 1996). This facilitates the interpretation of the parameter estimates.

We will now interpret the parameter estimates for the model with $K = 3$ latent budgets to get insight into the data. One can characterize the latent budgets by the values of their categories presented in Table 3, but it is more appropriate to characterize these relative to the average. Therefore we interpret the latent budgets by comparing the estimates $\hat{\pi}_{j|k}$ $(k = 1, 2, 3)$ with the column marginals, $p_{.j}$, which are also the elements of the latent budget in the $K = 1$ latent budget model, and attach a label to them. For example, the marginal proportion of MIL is 0.238. In the

$K = 3$ latent budget solution we can see that the estimated proportions of MIL are $0.021, 0.429$, and 0.011, respectively, for the first, second, and third latent budgets. Hence the second latent budget is characterized more than the other two budgets and more than average by people who feel the military situation is their principal worry. When we attach a label to the second latent budget, this feature should be considered. Besides MIL, the second latent budget is characterized by ENL and SAB, also larger than their respective marginals, which also deal with the endangerment of daily life by war and the undetermined category MTO ("more than one worry"). Hence this latent budget can be labeled "concerns for safety." In a similar way, we find that the first latent budget is characterized by POL and OTH, while personal and military concerns (PER, ENL, SAB, MIL) have very low existence or are absent. Hence we can label this latent budget "political and other worries." The third latent budget is dominated by PER, worries about personal economics, with OTH also present. Categories that denote nonegocentric concerns (MIL, POL, ECO) are almost absent in the third latent budget, hence this latent budget can be labeled "personal worries."

After the latent budgets are interpreted by the column categories we examine how the categories of the explanatory (row) variable are composed out of these latent budgets. For example, in the $K = 3$ latent budget solution, the category "residents from Europe and America" (EA) contributes 63.3% to the second latent budget. Hence EA can be described as a group whose principal worries are determined for the larger part by "concerns for safety."

3 Latent Class Model

The latent budget model is equivalent to the latent class model for two variables (see Clogg, 1981; van der Heijden *et al.*, 1992). The latent class model can be written as

$$\pi_{ij} = \sum_{k=1}^{K} \pi_k \pi_{i|k} \pi_{j|k} \tag{5}$$

For (5) we can write

$$\pi_{ij} = \sum_{k=1}^{K} \pi_k \frac{\pi_{ik}}{\pi_k} \frac{\pi_{jk}}{\pi_k} = \sum_{k=1}^{K} \pi_{ik} \frac{\pi_{jk}}{\pi_k} \Leftrightarrow \frac{\pi_{ij}}{\pi_i} = \sum_{k=1}^{K} \frac{\pi_{ik}}{\pi_i} \frac{\pi_{jk}}{\pi_k} \tag{6}$$

where the last expression is the equation of the latent budget model [see (1)]. Note that the latent budget model and the latent class model for two variables have the parameters $\pi_{j|k}$ in common. Equation (6) implies that, in the case of two variables, for each latent budget solution there is one corresponding latent class solution and vice versa. Therefore the estimation procedures and the unidentifiability of the model, mentioned in the previous section on LBA, apply to LCA as well (see van der Ark and van der Heijden, 1996). However, if we have an identified latent budget solution, such as presented in Table 3, we can get the corresponding latent class parameters $\pi_{i|k}$ and π_k by using Bayes' theorem and the law of total probability

$$\pi_{i|k} = \frac{\pi_i \pi_{k|i}}{\sum_{k=1}^{K} \pi_i \pi_{k|i}} \quad \text{and} \quad \pi_k = \sum_{i=1}^{I} \pi_i \pi_{k|i} \tag{7}$$

The latent class parameter estimates for Table 1, corresponding to the mixing parameter estimates from Table 3 and reparameterized through (7), are presented in Table 4.

The reason for using either the latent class model or the latent budget model depends on the types of manifest variables. Since the latent class model studies the joint probabilities π_{ij}, the model is more appropriate if the row variable and the column variable are both response variables. The response variables are then independent given the latent class. The latent budget model is more appropriate if one of the variables is an explanatory variable and the other a response variable. Only if we regard "residence" as a response variable, then it is appropriate to interpret the latent class solution of Table 4. This might be considered if one accepts that a person can choose the country in which he or she lives. In this case we have a latent variable with three classes that determines the "principal worries" and "residence" of the respondents. We did not find, however, an appropriate way to label the classes in this way.

Since the latent class model comprises only response variables, the model can be extended easily to more than two variables. The latent class model with four variables, for example, is then

$$\pi_{ghij} = \sum_{k=1}^{K} \pi_k \pi_{g|k} \pi_{h|k} \pi_{i|k} \pi_{j|k} \tag{8}$$

From (8) we can see that the general latent class model is equivalent to the law of total probability where the response variables are independent conditional on the latent classes.

Table 4: $K = 1, K = 2$, and $K = 3$ latent class solutions for the data of Table 1

	$K = 1$ latent class solution	$K = 2$ latent class solution		$K = 3$ latent class solution		
	$k = 1$	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 3$
Residence						
Asia/Africa	.302	.176	.544	.000	.268	.560
Europe/America	.506	.640	.248	.656	.596	.238
Israel; father Asia/Africa	.042	.027	.071	.027	.032	.073
Israel; father Europe/America	.115	.126	.094	.276	.075	.086
Israel; father Israel	.035	.031	.043	.040	.029	.044
Worry		See latent budgets Table 3				
Latent class probabilities	1.000	.658	.324	.181	.537	.282

Table 5: Cross-classification of four manifest variables (McCutcheon, 1987)*

			Cooperation		
Purpose	**Accuracy**	**Understanding**	**Interested**	**Cooperative**	**Impatient hostile**
Good	Mostly true	Good	419	35	2
		Poor/fair	71	25	5
	Not true	Good	270	25	4
		Poor/fair	42	16	5
Depends	Mostly true	Good	23	4	1
		Poor/fair	6	2	0
	Not true	Good	43	9	2
		Poor/fair	9	3	2
Waste	Mostly true	Good	26	3	0
		Poor/fair	1	2	0
	Not true	Good	85	23	6
		Poor/fair	13	12	8

*Reprinted by permission of Sage Publications, Inc.

Table 5 contains the cross-classification of four response variables collected in the 1982 General Social Survey (see McCutcheon, 1987b, p. 31). The data comprise the evaluation of 1202 respondents in terms of the respondent's attitude toward the purpose of surveys and the accuracy of surveys in general and the respondent's cooperation and understanding of the survey. In Table 6 a latent class solution with three latent classes published by McCutcheon (1987b, p. 43) is presented. After performing latent class analysis, McCutcheon characterized the three latent classes by the type of respondent who belongs to each of them. Some of the parameters have been restricted post hoc to facilitate interpretation (see Table 6). The three respondent types (classes) are "Ideal," those who have a positive attitude toward surveys and understand the questions well; "Believers," those who have a positive attitude toward surveys but do not really grasp their content; and "Skeptics," those who mistrust surveys although they understand the questions rather well. For further discussion of LCA see McCutcheon (Chapter 32).

4 Visualization of the Latent Budget Model

Latent budgets are K vectors in the J-dimensional space of the response variable. For example, the $K = 3$ latent budgets of the latent budget model in Table 3 can be viewed as three vectors in an eight-dimensional space. The heads of these K vectors span a $(K - 1)$ dimensional subspace; that is, if $K = 1$ then the head of the latent budget is a point, if $K = 2$ the heads of the latent budgets can be connected by a one-dimensional line segment, and if $K = 3$ the heads of the latent budgets are

Table 6: $K = 3$ latent class solution of the data of Table 5 (McCutcheon, 1987)*

	Respondent types		
Manifest Variables	$k = 1$ **(ideal)**	$k = 2$ **(believers)**	$k = 3$ **(skeptics)**
Purpose			
Good	.887[a]	.887[a]	.110
Depends	.060[a]	.060[a]	.228
Waste	.053	.053	.661
Accuracy			
Mostly true	.617[a]	.617[a]	.000[b]
Not true	.383	.383	1.000
Cooperation			
Interested	.943	.683	.649
Cooperative	.057	.260	.248
Impatient/hostile	.000[b]	.058	.103
Understanding			
Good	1.000[b]	.338	.765
Poor/fair	.000	.662	.235
Latent class probabilities (π_k)	.619	.223	.158

[a] Equality constraints imposed.
[b] Exact indicator restriction imposed.
* Reprinted by permission of Sage Publications, Inc.

the vertices of a triangle (Figure 2). Because the expected budgets, π_i, are mixtures of the latent budgets, $\tilde{\pi}_k$ [see (4)], the expected budgets can be viewed as vectors whose heads lie in the space (point, line segment, triangle, ...) spanned by the latent budgets. The precise position of the expected budgets in this space is the weighted average of the latent budgets, where the weights are the mixing parameters, $\pi_{k|i}$. The mixing parameters serve as coordinates in a so-called barycentric coordinate system, which in the $K = 3$ case is also known as the "triangular coordinate system" (see e.g., Greenacre, 1993, p. 15). The models with $K = 2$ and $K = 3$ latent budgets can be visualized by depicting the space spanned by the latent budgets and plotting the expected budgets onto this space by means of their mixing parameters.

In Figure 3 we show the graphical display of the $K = 2$ latent budget model (for the parameter estimates, see Table 3). Here the line segment spanned by the two latent budgets is presented, with the head of the first latent budget on the right-hand side and the head of the second latent budget on the left-hand side. Now the first expected budget (AA), with mixing parameter estimates 0.383 and 0.617, is made up 38.3% by the first latent budget and 61.7% by the second latent budget. If we scale the line segment from 0 (the second latent budget) to 1 (the first latent budget), then the position of (AA) is .383, hence closer to the second latent budget than to the first latent budget. If the mixing parameter estimates were (1.000, .000), then the expected budget would be equal to the first latent budget and be positioned on the right end of

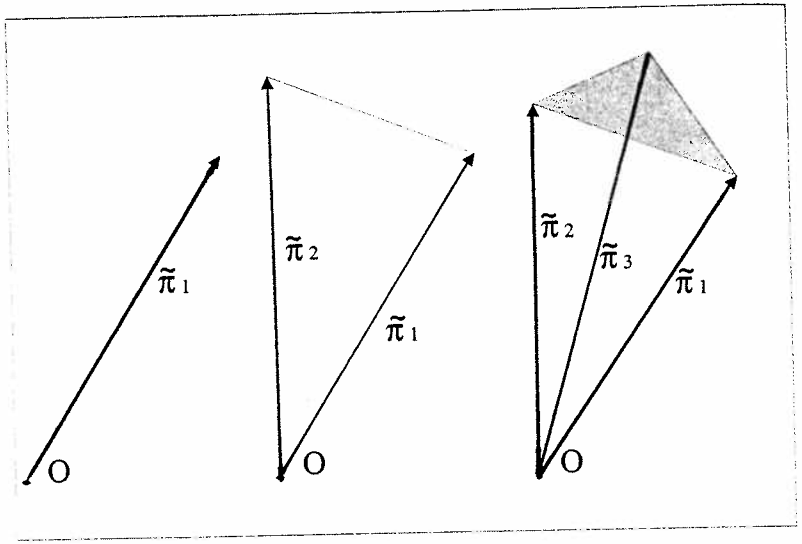

Figure 2: Visualization of $K = 1$, $K = 2$, and $K = 3$ latent budget model.

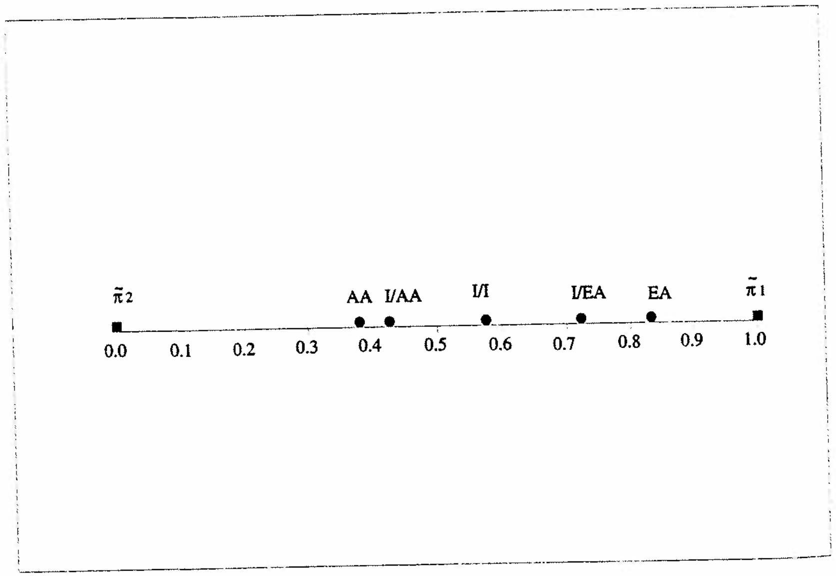

Figure 3: Graphical display of the $K = 2$ latent budget model.

the line segment. By depicting the $K = 2$ latent budget model in this way, we can see immediately that the expected budgets are composed of the latent budgets as their weighted average, where $\hat{\pi}_{1|i}$ and $\hat{\pi}_{2|i}$ $(i = 1, \ldots, 5)$ denote the weights.

In Figure 4 a graphical representation is given of the $K = 3$ latent budget model (for the parameter estimates, see Table 3). By convention, the vertices are at equal distance and the upper vertex of the triangle represents the head of the first latent budget, the right-hand vertex represents the head of the second latent budget, and the left-hand vertex represents the head of the third latent budget. The side opposite a vertex is the area where the corresponding mixing parameters are zero; for example, the first expected budget (AA) with mixing parameter estimates (0.000, 0.477, 0.523) lies on the bottom side of the triangle, because the first parameter estimate is zero. The scales of this triangular coordinate system are drawn as dotted lines parallel to the three sides of the triangle. The second mixing parameter estimate $\hat{\pi}_{2|1} = 0.477$ of AA positions the point between the fourth and the fifth dotted line that parallels the left side of the triangle. The third mixing parameter estimate $\hat{\pi}_{3|1} = 0.523$ positions AA between the fifth and the sixth dotted line parallel to the right side of the triangle.

Figures 3 and 4 can be interpreted as a mixture model as well as a MIMIC model. The mixture model interpretation is a tool for understanding the composition of the expected budgets in terms of the latent budgets. The closer the expected budgets lie to the latent budgets, the greater the probability that a residence group will resemble the

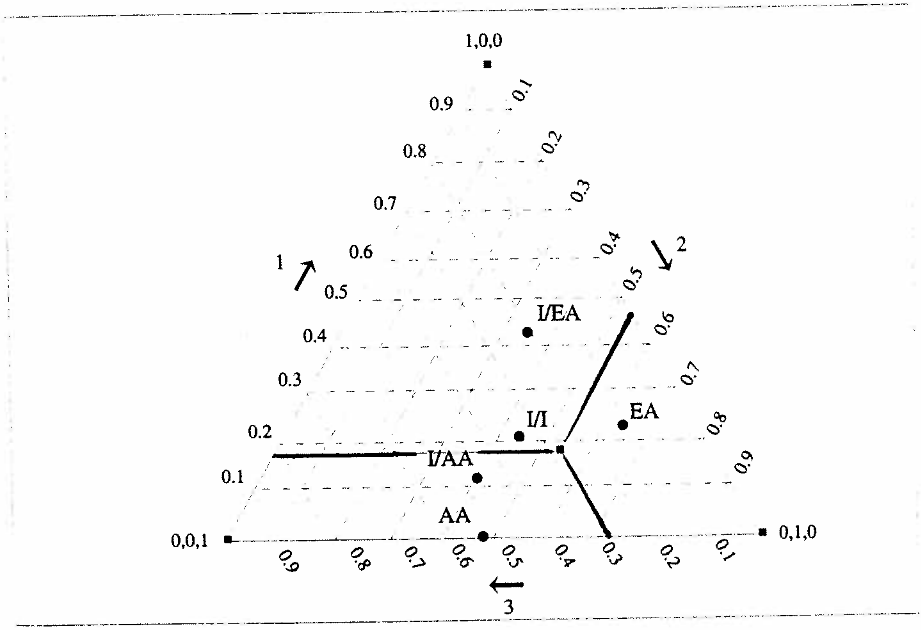

Figure 4: Graphical display of the $K = 3$ latent budget model.

latent budget, and the distance between two expected budgets determines the relative similarity among them. The vector of marginal column proportions with elements $\pi_{\cdot j} = \sum_i \pi_{ij}$ ($j = 1,\ldots,J$) has also been plotted in Figure 4 as a solid square. This vector is the latent budget in the independence model and therefore represents the average budget. The vector of marginal proportions can serve as a reference vector for the expected budgets. Hence, if an expected budget is closer to a latent budget than the vector of marginal column proportions, then the expected budget resembles the particular latent budget more closely than average. The coordinates of the vector of marginal proportions are $\pi_k = \sum_i \pi_i \pi_{k|i}$, that is, 0.181, 0.537, 0.282. For example, from Table 3 we can see that I/I has mixing parameter estimates (0.205, 0.447, 0.348) and is closer to the vector of marginal proportions than any other row category. Hence the Israeli residents whose fathers also live in Israel display the most average pattern of worries.

The MIMIC model interpretation is a guide to an additional characterization of the latent budgets. We can consider Figure 4 such that the triangle displays the probability to enter the latent budgets; that is, the vertices denote the probability 1 that the subjects of a row category i belong to the corresponding latent budget ($\pi_{k|i} = 1$) and a probability 0 that they belong to the other latent budgets. In this interpretation the picture shows how the marginal row probabilities are distributed over the latent budgets, and we can label the budgets by this distribution. The point that denoted the vector of marginal column proportions now represents the average distribution of all subjects in the contingency table. If a row category is closer to a latent budget than the point representing the overall average, then the latent budget is characterized more than average by that row category. If the distance between two points in the figure is large, then the distribution of those two categories over the latent budgets is not similar; if the distance is small, then the two categories are distributed over the latent budgets in more or less the same way. We can see that the first latent budget (1, 0, 0) is represented more than average by I/EA, EA, and I/I, the second latent budget (0, 1, 0) can be interpreted as a budget typical for those who live in Europe or America, and the third latent budget (0, 0, 1) is represented more than average by AA, I/AA, and I/I.

The categories of the column variable can also be represented graphically. This can be done if we rescale the elements of the latent budgets from $\pi_{j|k}$ into $\pi_{k|j}$ by

$$\pi_{k|j} = \frac{\pi_{j|k} \sum_{i=1}^{I} \pi_i \pi_{k|i}}{\pi_j} \tag{9}$$

[see (7)]. In Figure 5 a graphical representation of $\pi_{k|j}$ in the $K = 3$ latent budget solution is given for the data of Table 1 and the rescaled latent budget parameters [see (9)] are given in Table 7.

Figure 5 cannot be interpreted in terms of the mixture model for the rows, but must be interpreted according to the MIMIC model; that is, the vertices denote $\pi_{j|k} = 1$ ($k = 1, 2, 3$) and the squares in Figure 5 are the column categories. Their position in the triangle determined by $\pi_{k|j}$ denotes how the marginal probability of a particular observed category j is distributed over the three latent budgets. If one of the categories were positioned on a vertex, this category would be present only in the particular latent budget. Hence, MTO is present only in the second latent budget. If a category were positioned in the center of the figure [i.e., coordinates are ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)], then the responses to that category would be equally distributed over the latent

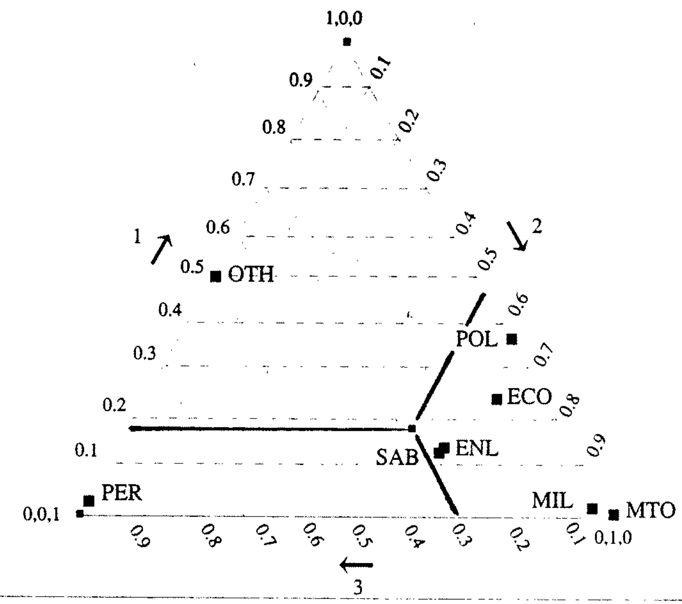

Figure 5: Graphical representation of the rescaled latent budgets in the $K = 3$ latent budget solution.

Table 7: Rescaled latent budget elements of Table 3

Worries	$k = 1$	$k = 2$	$k = 3$
Enlisted relative (ENL)	.141	.620	.238
Sabotage (SAB)	.129	.625	.247
Military situation (MIL)	.017	.969	.014
Political situation (POL)	.368	.632	.000
Economic situation (ECO)	.239	.669	.092
Other (OTH)	.498	.000	.502
More than one worry (MTO)	.000	1.000	.000
Personal economics (PER)	.000	.038	.962
Latent class probabilities (π_k)	.181	.537	.282

budgets. If two points were plotted close together, for example ENL and SAB, then these categories have a similar distribution over the latent budgets. In this way we visualize the characterization of the latent budgets that has been given in Section 2. In Figure 5 the point that denoted the average distribution of all subjects over the latent budgets, with coordinates π_k ($k = 1, 2, 3$), is also plotted and now serves as a reference point for the column categories.

Notice that if we examine $\pi_{k|j}$ instead of $\pi_{j|k}$ the marginal column effects have disappeared. This means that if a marginal column proportion is very small, for example, the marginal proportion of ECO (.012), and we examine the actual proportions of the latent budgets with elements $\pi_{j|k}$, then ECO hardly plays a role in the interpretation of the latent budgets, because of its low marginal frequency. Categories with large marginal column proportions, on the other hand, tend to dominate, for example, MIL, which has a marginal proportion of 0.238. These differences disappear if we examine $\pi_{k|j}$, where we see how each category is distributed over the latent budgets.

Figures 4 and 5 can be overlaid. In this case the figure has to be interpreted according to the MIMIC model. Thus, the plot indicates to what extent the categories of the row and the column variables appear in a certain latent budget. This may help to interpret the latent budgets not only by means of the column categories but also by means of the row categories.

5 Visualization of the Latent Class Model

The idea of rescaling the parameters $\pi_{j|k}$ into $\pi_{k|j}$ can also be used to visualize the latent class parameters. If we have two response variables, we can depict $\pi_{k|i}$ ($i = 1,\ldots,I$) and $\pi_{k|j}$ ($j = 1,\ldots,J$) simultaneously. If we assume that the variables "residence" and "worry" from Table 1 are both response variables, visualization of the latent class model with three latent classes would be equivalent to Figures 4 and 5. The plot must be interpreted as a MIMIC model, however; that is, the picture reveals how the categories of the variables are distributed over the latent classes. In this way we can easily characterize the latent classes by the closeness of the category points to the corners of the triangle. If we overlay Figures 4 and 5, then we have a simultaneous representation of the row variable and the column variable.

As mentioned in Section 3, the latent class model can easily be extended to more than two variables. If we have more than two variables, say four, with corresponding latent parameters $\pi_{g|k}$ ($g = 1 \ldots G$), $\pi_{h|k}$ ($h = 1 \ldots H$), $\pi_{i|k}$ ($i = 1 \ldots I$), and $\pi_{j|k}$ ($j = 1 \ldots J$), [see (8)], they can be transformed into $\pi_{k|g}$ ($g = 1 \ldots G$), $\pi_{k|h}$ ($h = 1 \ldots H$), $\pi_{k|i}$ ($i = 1 \ldots I$), and $\pi_{k|j}$ ($j = 1 \ldots J$) by

$$\pi_{k|g} = \frac{\pi_k \pi_{g|k}}{\pi_g}; \; \pi_{k|h} = \frac{\pi_k \pi_{h|k}}{\pi_h}; \; \pi_{k|i} = \frac{\pi_k \pi_{i|k}}{\pi_i}; \; \pi_{k|j} = \frac{\pi_k \pi_{j|k}}{\pi_j} \qquad (10)$$

A graphical display of Table 6 is given in Figure 6. The manifest variables are depicted simultaneously. The rescaled parameter estimates obtained with (10) are given in Table 8.

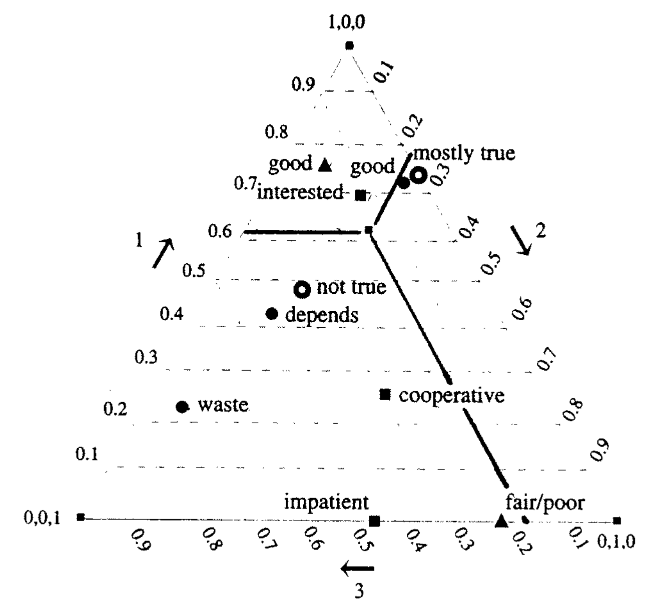

Figure 6: Display of McCutcheon solution.

Table 8: Reparameterized latent class solution of Table 6

	Respondent types		
Manifest Variables	$k = 1$ **(ideal)**	$k = 2$ **(believers)**	$k = 3$ **(skeptics)**
Purpose			
Good	.718	.259	.023
Depends	.429	.155	.416
Waste	.220	.079	.700
Accuracy			
Mostly true	.735	.265	.000
Not true	.493	.178	.329
Cooperation			
Interested	.696	.182	.122
Cooperative	.266	.438	.296
Impatient/hostile	.000	.443	.557
Understanding			
Good	.759	.093	.148
Poor/fair	.000	.799	.201
Latent class probabilities (π_k)	.619	.223	.158

In Figure 6 the solid circles denote the variable "purpose," the open circles denote "accuracy," the solid squares denote "cooperation," and the solid triangles denote "understanding." Figure 6 displays the characterization of the latent classes according to the results in Table 4. The first class (ideal respondents) is characterized more than average by all most positive categories of the variables. The second class (believers) is mostly characterized by a fair to poor understanding of surveys, but they are more cooperative than average. The third class (skeptics) is characterized by negative categories of all variables.

6 Relation of LBA and LCA to Correspondence Analysis

A problem with the display of LCA and LBA is that only the relative distances are visualized; Figures 4, 5, and 6 are equilateral triangles, while there are always two latent classes (budgets) that are more similar to each other than to the third one. We are able to solve this problem by using correspondence analysis (CA), and the solution follows from the relation of LBA and LCA to CA. Because LBA and LCA are equivalent, we will refer only to LCA in this section.

The relation between LCA and CA is rather close and has been studied before by, among others, Gilula (1979, 1983, 1984), Goodman (1987), de Leeuw and van der Heijden (1991), and van der Ark and van der Heijden (1996). Visualization of both models will give more insight into this relation. We recapitulate here first the analytic results of de Leeuw and van der Heijden (1991) and then illustrate the results by visualizing them.

Consider a two-way matrix with observed proportion p_{ij} of rank M. We define CA as

$$p_{ij} = p_{i\cdot}p_{\cdot j}\left(1 + \sum_{m=1}^{M} \alpha_m r_{im} c_{jm}\right) \tag{11}$$

where the scores r_{im} and c_{jm} are centered: $\sum_i p_{i\cdot} r_{im} = \sum_j p_{\cdot j} c_{jm} = 0$, and standardized: $\sum_i p_{i\cdot} r_{im}^2 = \sum_j p_{\cdot j} c_{jm}^2 = 1$. The parameters α_m are the singular values obtained from a singular value decomposition of the matrix with elements $(p_{ij} - p_{i\cdot}p_{\cdot j})/\sqrt{(p_{i\cdot}p_{\cdot j})}$. When the matrix of proportions has full rank, then $M = \min(I-1, J-1)$. Decomposition (11) is also known as the canonical analysis of a contingency table (Gilula, 1984; Gilula and Haberman, 1986). In this context α_m is the mth canonical correlation between the quantified row and column variable, where the scores r_{im} are used as quantifications for the rows, and the scores c_{jm} are used as quantifications for the columns. These scores are often called the "standard coordinates" of a CA solution.

Suppose that we use only M^* ($1 \leq M^* < M$) dimensions of decomposition (11) to derive elements

$$p_{ij}^* = p_{i\cdot}p_{\cdot j}\left(1 + \sum_{m=1}^{M^*} \alpha_m r_{im} c_{jm}\right) \tag{12}$$

and we collect these approximations p_{ij}^* in a matrix $\mathbf{P}^*$. The matrix $\mathbf{P}^*$ is a reduced rank matrix of rank $M^* + 1$, and it provides an optimal approximation of the observed matrix in a least-squares sense (see, for example, Greenacre, 1984). Notice that $\mathbf{P}^*$ need not be a probability matrix, that is, a matrix with nonnegative elements adding up to one. Although it can be shown that $\sum_{ij} p_{ij}^* = 1$, some elements may be negative.

Both CA and LCA are reduced rank models. If a matrix can be decomposed by a K-class LCA, then it can also be decomposed by a $(K-1)$-dimensional CA. However, contrary to what was stated by van der Heijden *et al.* (1989), the reverse does not hold in general. This can be seen from the fact that the factorization provided by LCA consists of nonnegative parameters only, whereas the parameters of CA may be negative. There is one special case, though. De Leeuw and van der Heijden (1991) prove that LCA and CA are equivalent in the two-class, one-dimensional case and then provide a counterexample to illustrate that this is not true in general for higher dimensions.

Let us now discuss the implications of these results for data analysis. Observed contingency tables that are of reduced rank seldom occur. If a matrix does not have a reduced rank, then we can still calculate the decomposition provided by (12). If we then consider only $M^* < \min(I-1, J-1)$ dimensions, then $\mathbf{P}^*$ need not be a probability matrix (see earlier). Therefore for CA estimated by least squares the above has limited practical relevance. It is relevant, however, for CA estimated by maximum likelihood, as proposed by Goodman (1985) and Gilula and Haberman (1986) (see also Siciliano *et al.*, 1993). Their model is

$$\pi_{ij}^* = a_i b_j\left(1 + \sum_{m=1}^{M^*} f_m u_{im} v_{jm}\right) \tag{13}$$

where the parameters have identification restrictions identical to those in (11) and (12): $\sum_i a_i u_{im} = \sum_j b_j v_{jm} = 0$ and $\sum_i a_i u_{im}^2 = \sum_j b_j v_{jm}^2 = 1$. A choice of M^* determines the rank of the matrix with elements π_{ij}^* and because (13) is estimated by maximum likelihood, this yields a probability matrix of reduced rank when $M^* < \min(I-1, J-1)$. This shows that for $K = 2$ the estimates of expected probabilities of both models will be equal, and therefore the fit of both models will be equal as well. For $K = 3$ it turns out that often, but not always, LCA and CA have identical estimates of expected probabilities (see van der Ark and van der Heijden, 1996, for more details). This is relevant for the visualization of LBA and LCA.

7 Simultaneous Visualization of the Correspondence Model and the Latent Budget Model

CA is usually employed to make graphical representations. The categories of the variables are plotted onto an M^*-dimensional space with M^* orthogonal axes.

An important concept in the visualization of CA is the chi-squared distance. The chi-squared distance $\delta^2_{i,i'}$ between rows i and i' of $\mathbf{\Pi}^*$ is defined as

$$\delta^2_{i,i'} = \sum_{j=1}^{J} \frac{\left((\pi^*_{ij}/\pi^*_i) - (\pi^*_{i'j}/\pi^*_{i'})\right)^2}{\pi^*_j} \tag{14}$$

From equation (14) we can see that the chi-squared distance is the squared difference between two expected row budgets π^*_{ij}/π^*_i weighted by the marginal proportion of the column π^*_j. Now we can plot each category of the row variable using $r_{im}\alpha_m$ ($m = 1 \ldots M^*$) as coordinates. Then the Euclidean distances between the rows in the plot are equal to chi-squared distances between the rows of $\mathbf{\Pi}^*$. We can also plot each category of the column variable using $c_{jm}\alpha_m$ ($m = 1 \ldots M^*$) as coordinates. Here the Euclidean distances between the columns in the plot are equal to chi-squared distances between the columns of $\mathbf{\Pi}^*$. These two graphical displays that are often produced for CA estimated by maximum likelihood could be enriched by supplementing them with points for latent budgets, when CA and LCA yield identical estimates of expected probabilities, and this would lead to an interpretation of the CA solution from a different perspective.

An example for the data in Table 1 is given in Figure 7. Let us denote the maximum likelihood estimates of the model by $\hat{\mathbf{\Pi}}^*$. Now in Figure 7 the row profiles of $\hat{\mathbf{\Pi}}^*$ are plotted onto the first two principal axes of the correspondence model, using $r_{im}\alpha_m$ ($i = 1\ldots5$; $m = 1,2$) as coordinates. The columns of $\hat{\mathbf{\Pi}}^*$ have been plotted in the picture as well using standard coordinates c_{jm} ($j = 1\ldots8$; $m = 1,2$). We can project the latent budgets onto Figure 7 to illustrate the relation between LBA and CA. We find the coordinates of the latent budgets by projecting them as supplementary points in the CA space, that is, $\sum_j \pi_{j|k} c_{jm}$ ($k = 1\ldots3$, $m = 1,2$). The coordinates are given in Table 9. The horizontal axis differentiates basically on the origins of the respondents with (AA) and (I/AA) on the left-hand side, (EA) and (I/EA) on the right-hand side, and (I/I) in between, while the vertical axis differentiates on the actual residence of the respondents, residents of Israel on the upper side and citizens living abroad on the lower side. Notice that the triangle in Figure 7 is not the same as the triangle in Figure 4, because in Figure 4 by convention the distances between the latent budgets are unity, whereas in Figure 7 they are measured in the chi-squared metric. Also by convention, in Figure 4 the first latent budget is placed on top, whereas in Figure 7 the position of the latent budgets depends on the axes. Thus Figure 7 can

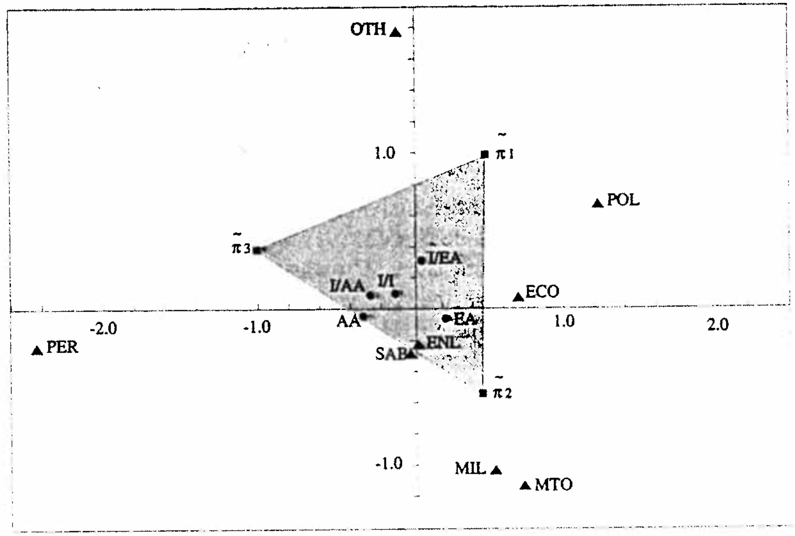

Figure 7: Latent budgets plotted in CA space.

Table 9: Coordinates of the expected budgets and latent budgets in Figure 7

	Dim. 1	Dim. 2
Row profiles (black dots)		
Asia/Africa	−.330	−.051
Europe/America	.212	−.058
Israel; father Asia/Africa	−.285	.089
Israel; father Europe/America	.069	.326
Israel; father Israel	−.101	.100
Column profiles (black triangles)		
Enlisted relative (ENL)	.057	−.238
Sabotage (SAB)	.101	−.288
Military situation (MIL)	.512	−1.069
Political situation (POL)	1.227	.669
Economic situation (ECO)	.687	.081
Other (OTH)	−.069	1.779
More than one worry (MTO)	.520	−1.160
Personal economics (PER)	−2.443	−.268
Latent budgets (black squares)		
First latent budget	.300	.997
Second latent budget	.440	−.545
Third latent budget	−1.032	.399

be viewed as a plot of the latent budget solution scaled in chi-squared distances, which allows a visual comparison of the similarities of the three latent budgets.

8 Discussion

We have shown how to visualize the results of LBA and LCA and how these visualizations are related to the visualizations of CA. A K-budget LBA, that is, a K-class LCA, is equivalent to a $(K - 1)$-dimensional CA when they yield the same estimated expected frequencies. In such a case, the visualization of the results of LBA and LCA can be plotted onto the CA map and vice versa.

LBA is a technique that can be used best when we have one explanatory and one response variable, and the question of interest is how the expected budgets can be composed of a smaller amount of typical or latent budgets. LCA can be used best when we want to study the relation between two or more discrete response variables. The question of interest is whether we can split up the sample into K latent classes such that the relation among the variables is satisfactorily explained by the classes.

On the other hand, CA visualizes how row profiles can be explained by continuous axes, which can be interpreted as latent traits. If the row profiles are equivalent to expected budgets, then the difference between LCA/LBA and CA could be summarized as the choice between a trait or state explanation of the latent budgets.

When the models have the same expected frequencies, plotting the latent budget solution or the latent class solution onto the CA map gives us the benefits of both models. On the one hand, we can see at a glance how the expected budgets are built up of prototypes, and on the other hand, we can assign latent trait scores to the expected budgets. An extra advantage is that the map allows a valid distance interpretation.

References

Clogg, C.C. (1981). Latent structure models of mobility. *American Journal of Sociology,* **86,** 836-68.

de Leeuw, J., and van der Heijden, P.G.M. (1991). Reduced rank models for contingency tables. *Biometrika,* **78,** 1, p229-32.

de Leeuw,J., van der Heijden, P.G.M., and Verboon, P. (1990). A latent time budget model. *Statistica Neerlandica,* **44,** 1, p1-21.

Gilula, Z. (1979). Singular value decomposition of probability matrices: probability aspects of latent dichotomous variables. *Biometrics,* **66,** 339-44.

Gilula, Z. (1983). Latent conditional independence in two-way contingency tables: a diagnostic approach. *British Journal of Mathematical and Statistical Psychology,* **36,** 114-22.

Gilula, Z. (1984). On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika,* **71,** 523-9.

Gilula, Z., and Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association,* **81,** 780-8.

Greenacre, M.J. (1993). *Correspondence Analysis in Practice.* London: Academic Press.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. I. A modified latent structure approach. *American Journal of Sociology,* **79,** 1179-259.

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. *Annals of Statistics,* **13,** 10-69.

Goodman, L.A. (1987). New methods for analyzing the intrinsic character of qualitative variables using cross-classified data. *American Journal of Sociology,* **93,** 3, 529-83.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika,* **36,** 329-47.

McCutcheon, A.L. (1987). *Latent Class Analysis.* Newsbury Park : Sage.

Mooijaart, A., van der Heijden, P.G.M., and van der Ark, L.A. (1996). A least squares algorithm for a mixture model for compositional data. (submitted).

Siciliano, R., Mooijaart, A., and van der Heijden P.G.M. (1993). A probabilistic model for nonsymmetric correspondence analysis and prediction in contingency tables. *Journal of the Italian Statistical Society,* **1,** 85-106.

Siciliano, R., and van der Heijden, P.G.M. (1994). Simultaneous latent budget analysis of a set of two-way tables with constant-row-sum data. *Metron,* **52,** 155-80.

van der Ark, L.A. and van der Heijden, P.G.M. (1996). *Geometry and Identifiability in the Latent Budget Model.* Methods Series MS-96-3. Utrecht: ISOR.

van der Heijden, P.G.M., Mooijaart, A., and De Leeuw, J. (1989). Latent budget analysis. In Decarli, A., Francis, B.J., Gilchrist, R., and Seeber, G.U.H. (Eds.). *Statistical Modelling.Proceedings of GLIM 89 and the 4th International Workshop on Statistical Modelling, Lecture Notes in Statistics 57,* p301-13. Berlin: Springer-Verlag.

van der Heijden, P.G.M., Mooijaart, A., and De Leeuw, J. (1992). Constrained latent budget analysis. In Marsden, P.V., *Sociological Methodology* 1992, Vol 22, Chapter 9 279-320. Cambridge: Blackwell Publishers.