Statistica Applicata Vol. 4, n. 3, 1992

THREE APPROACHES TO STUDY THE DEPARTURE FROM QUASI–INDEPENDENCE

Peter G.M. van der Heijden*

University of Utrecht, Faculty of Social Sciences, Department of Methods, The Netherlands.

Correspondence analysis, models with logbilinear terms for the association, and the latent budget model are used to study the departure from the loglinear quasi–independence model in such cases where some of the cell frequencies are missing or of no interest. Approaches are chosen in which these cells do not influence the fit of the model or technique. It is indicated how to make graphic displays of the scores and parameters for the association. As an example, a square contingency table is analyzed where interest is focused on association in off–diagonal cells. For this example it is shown that these graphic displays are remarkably similar. Since this paper will appear in a special issue of Statistica Applicata devoted to the Italian and Dutch schools of data analysis, some attention will be given to the Dutch contributions in the field of contingency table analysis. For an overview of the Italian contributions in this field we refer to the paper of Siciliano (this issue).

Key words: correspondence analysis, association models, logbilinear models, latent class analysis, latent budget analysis, quasi–independence, structural zeros, data analysis

1. INTRODUCTION

Many approaches to the analysis of contingency table can be adopted. Three important ones (with a bias to approaches with Dutch contributions) are the following. First, correspondence analysis and extensions such as multiple correspondence analysis and optimal scaling techniques such as PRINCALS, OVERALS are worked out in detail by the Department of Data Theory, Leiden, the Netherlands (Gifi, 1990). In this approach categories are quantified in such a way that correlations or eigenvalues are maximized. Second, loglinear analysis and recent extensions are worked out mostly by authors in the U.S.A.. In loglinear analysis the logarithm of cell frequencies is decomposed by a linear model. Extensions come down to restricting the interaction parameters (for references, see below). Third, in latent class analysis associations between observed variables are explained by categorical latent variables. In this approach important contributions

* *I gratefully acknowledge the helpful comments of Antoine de Falguerolles*

were made by Goodman and Clogg (see below). A recent book is written by the Dutch author Hagenaars (1990).

There has been much interest over the last decade in relations between statistical models, such as loglinear analysis and latent class analysis, and techniques for the analysis of contingency tables, such as correspondence analysis. Important contributions include those of Goodman (1981) and Escoufier (1982), who study the relation between correspondence analysis and the RC–association model (see also Goodman, 1986, for an overview). In the Netherlands, another approach was adopted by van der Heijden and de Leeuw (1985), who showed how correspondence analysis can be used to study the residuals from loglinear models (see also van der Heijden et al., 1989a). Goodman (1987) also paid attention to the close relation between correspondence analysis, association models and latent class analysis, while de Leeuw and van der Heijden (1991) showed that specific cases of correspondence analysis and latent class analysis are equivalent.

In the above papers, attention was given to the standard applications of these models and techniques. In this paper we will focus on a more specific application of these models, namely on the way in which these approaches deal with the departure from the quasi–independence model. We will show that using correspondence analysis to deal with this departure is very similar to an approach using association models in which bilinear terms describe this association, and also similar to the approach of latent budget analysis, which is a reparametrization of latent class analysis.

The quasi–independence model is a model used to ignore certain cells in a contingency table where interest is focused on independence in the other cells. The need to ignore certain cells can result from the fact that these values are not known, for example, in a secondary analysis where some of the frequencies are missing for certain years or for certain areas. Another occasion for ignoring cell values is where they are structurally zero: certain combinations of levels of distinct variables cannot logically occur, for example, males (a level of the variable sex) cannot have menstruation problems (a level of medical problems). A third example is where we do not want to model some of the cells, because substantive interest is restricted to the remaining cells, for example, in a social mobility table interest might focus on off–diagonal frequencies that corresponding to people changing their social mobility status. A last example is where, a posteriori, we decide that we can ignore some of the cells in order to see what goes on the remaining cells, as, for example, when certain cells dominate the initial solution.

In the paper of van der Heijden et al. (1989a) it was argued that correspondence analysis could be seen as a tool to decompose the residuals from the loglinear independence model, and that a generalization of correspondence analysis could be used to decompose the residuals from other loglinear models, such as the quasi–independence model. In the discussion of this paper, Gower (1989) asked why correspondence analysis should be used to decompose the departure from models rather than modelling the departure from these models from the start. This paper can be seen as a contribution to this discussion: it is shown how modelling approaches can be used to model the departure from quasi–independence. The paper also explores a topic raised by Caussinus (1986), who was interested in the relation between correspondence analysis used to decompose the residuals from quasi–independence and association models.

We will illustrate the similarities of the different approaches by an example taken from Harshman et al. (1982) which deals with changing cars. In 1979, recent car buyers were surveyed to collect information on their old and new cars. The cars are categorized in 16 segments yielding a transition matrix of 16x16 (see table I). In the analysis of the example interest will go out to the association in the off–diagonal cells, i.e. in the cells of people changing car type. The diagonal cells are considered uninteresting.

Table I: 16 rows are cars disposed off, 16 columns are new cars. The segments are 1. subcompact/domestic 2. subcompact/captive imports 3. subcompact/imports 4. small specialty/domestic 5. small specialty/captive imports 6. small specialty/imports 7. low price compacts 8. medium price compacts 9. import compact 10. midsize domestic 11. midsize imports 12. midsize specialty 13. low price standard 14. medium price standard 15. luxury domestic 16. luxury import.

23272	14871	05011	8994	49	2319	1234	9406	1	5451	2622	4811	6329	4253	2370	949	127
3254	1114	3014	2656	23	551	959	894	223	1672	223	2012	926	540	246	37	
11344	1214	25986	9803	47	5400	3262	1353	2257	5195	1307	8347	2308	1611	1071	288	
11740	1192	11149	38434	69	4880	6047	2335	931	8503	1177	23898	3238	4422	4114	410	
47	6	0	117	4	0	0	49	0	110	0	10	0	0	0	0	
1772	217	3622	3453	16	5249	1113	313	738	1631	1070	4937	338	901	1310	459	
18441	1866	12154	15237	65	1626	27137	6182	835	20909	566	15342	9728	3610	910	170	
10359	693	5841	6368	40	610	6223	7469	564	9620	453	9731	3601	5498	764	85	
2613	481	6981	1853	10	1023	1305	632	1536	2738	1005	990	454	991	543	127	
33012	2323	22029	29623	110	4193	20997	12155	2533	53002	2140	61350	2800	6339	13980	706	
1293	114	2844	1242	5	772	1507	452	565	3820	3059	2357	589	1052	871	595	
12981	981	8271	18908	97	3444	3693	1748	935	11551	1314	56025	10959	18688	12541	578	
27816	1890	12980	15993	34	1323	18928	5836	1182	28324	938	37380	67964	28881	6585	300	
17293	1291	11243	11457	41	1862	7731	6178	1288	20942	1048	30189	15318	81808	21974	548	
3733	430	4647	5913	6	622	1652	1044	476	3068	829	8571	2964	9187	63509	1585	
105	40	997	603	0	341	75	55	176	151	589	758	158	756	1234	3124	

2. CORRESPONDENCE ANALYSIS

Correspondence analysis decomposes a matrix with observed proportions p_{ij} (i=1,...,I; j=1,...J) by

$$P_{ij} = P_{i+}P_{j+}\left(1+\sum_{m}^{M}\lambda_m r_{im}c_{jm}\right) \qquad (1)$$

where M = min (I–1,J–1), $p_{i+} = \Sigma_j p_{ij}$, $\Sigma_i p_{i+} r_{im} = \Sigma_j p_{+j} c_{jm} = 0$ and $\Sigma_i p_{i+} r_{im} r_{in} = \Sigma_j p_{+j} c_{jm} c_{jn} = \delta^{mn}$, where $\delta^{mn} = 1$ if m=n, and $\delta^{mn} = 0$ if $m \neq n$. Different motivations can be given for this decomposition (see Gifi, 1990; Greenacre, 1984; Nishisato, 1980). One is that the row categories and the column categories are quantified in such a way that the correlation between the row variable and the column variable is maximized. The scores r_{i1} and c_{j1} produce this maximal correlation, that is equal to λ_1. Scores r_{i2} and c_{j2} produce the second maximal correlation λ_2 under the restriction that this second quantification is orthogonal to the first quantification, and so on for subsequent quantifications m. Another objective of the decomposition is to yield scores that can be used as coordinates in graphic displays which then show how the row variable is related to the column variable. By using $r_{im}\lambda_m^{1/2}$ ($c_{jm}\lambda_m^{1/2}$) as coordinates of row category i (column j) on dimension m, biplots of the interaction are obtained.

Quasi–independence, and the departure from quasi–independence

In van der Heijden et al. (1989a) a residual approach to CA is proposed. This approach starts from the observation that for ordinary CA $X^2/n = \Sigma_m \lambda_m^2$, where X^2 is the Pearson chi–square statistic for testing independence. The decomposition (1) is rewritten as $p_{ij} = p_{i+}p_{+j} + p_{i+}p_{+j}\Sigma_m \lambda_m r_{im} c_{jm}$, showing that CA is a tool for the decomposition of the departure from independence into M dimensions. A generalization of CA by Escofier (1983) is used to extend this idea to models which differ from the independence model. This generalization is

$$p_{ij} = q_{ij} + s_i t_j \sum_{m}^{M}\lambda_m r_{im} c_{jm} \qquad (2)$$

where q_{ij} are values that are not necessarily equal to $p_{i+}p_{+j}$ (i.e. independence), and the weights to scale the residuals s_i and t_j are not necessarily equal to the marginal proportions p_{i+} and p_{+j}. This generalization is used for study of the residuals from loglinear models. Examples are conditional independence models for tables of more than two variables, and models for square contingency tables such as quasi–independence, symmetry and quasi–symmetry. For each of these applications q_{ij} consists of estimates of expected frequencies under the model studied, and

suitable choices are made for s_i and t_j. This extension of CA to other models than the independence model is discussed in some detail in van der Heijden et al. (1989a).

In this paper we concentrate on a study of the departure from quasi–independence. Quasi–independence can be defined as $\pi_{ij} = \alpha_i\beta_j$ for $\{(i,j)\} \in S$, and $\pi_{ij} = p_{ij}$ for $\{(i,j)\}$ not in S (see Goodman, 1968, for details). Thus the cells in S are the cells we are interested in. If we choose in (2) the estimates of expected frequencies $\hat{\pi}_{ij}$ for q_{ij}, and $\hat{\alpha}_i$ and $\hat{\beta}_j$ for the weights s_i and t_j, then (2) has the property that $\Sigma_m\lambda_m^2 = X^2 / n = \Sigma_i\Sigma_j\left(p_{ij} - \hat{\pi}_{ij}\right)^2 / \hat{\pi}_{ij}$ for this application. It can also be proven that this leads to a CA procedure for imputing missing values, called 'reconstitution of order zero' (see Greenacre, 1984; de Leeuw and van der Heijden, 1988). Reconstitution of order zero comes down to iteratively computing $p_{ij}^{(v+1)} = p_{i+}^{(v)}p_{+j}^{(v)}$ for the cells not in S, where v indexes the iteration. Thus 'independent' values are iteratively imputed for these cells. After convergence we find estimates p_{ij}^* with the property that $p_{ij}^* = p_{i+}^*p_{+j}^*$, and therefore the residuals for the cells to be ignored are zero. These cells do not contribute to the so–called inertia $\Sigma_m\lambda_m^2$ (see de Leeuw and van der Heijden, 1988, for details).

The general form of this procedure is called 'reconstitution of order h', i.e.

$$p_{ij}^{(v+1)} = p_{i+}^{(v)}p_{+j}^{(v)}\left(1 + \sum_m^h \lambda_m^{(v)}r_{im}^{(v)}c_{jm}^{(v)}\right) \qquad (2)$$

where $p_{ij}^{(0)} = p_{ij}$ for $\{(i,j)\} \in \Sigma$, and an arbitrary value for $\{(i,j)\}$ not in S. Thus, for an h–dimensional solution the (imputed) values in S are equal to the values derived from reconstitution of order h. The solution will, in general, depend on the choice of the dimensionality h. This procedure is similar to the way the EM–algorithm operates: estimates for the scores are derived, then missing proportions are estimated using these new scores, then updates for the scores are derived, and so on, until convergence is reached.

Fruitful as reconstitution of order zero may seem as a way to ignore cells in a CA of a two–way contingency table, this approach has its drawback. If the value in cell (i,j) is to be ignored, then $p_{ij}^* - p_{i+}^*p_{+j}^* = 0$ (or, in terms of quasi–independence, $p_{ij} - \hat{\pi}_{ij} = 0$). Therefore in a full–dimensional biplot row vector i will be orthogonal to column vector j (see Gabriel, 1971). This is an artefact of the procedure, as a result of which these cells influence the full–dimensional solution, contrary to what was intended.

We therefore propose to analyze a matrix in which we intend to ignore certain

cell values with reconstitution of order h. Unlike reconstitution order zero, reconstitution of orders larger than zero no longer decomposes the residuals from quasi–independence. In spite of the fact that we loose this advantage, we nevertheless choose for reconstitution of order h because it avoids the artefact. The resulting approach will also turn out to be very similar to modelling approaches of sections 3 and 4 where the association is best approximated without any interference from cells to be ignored.

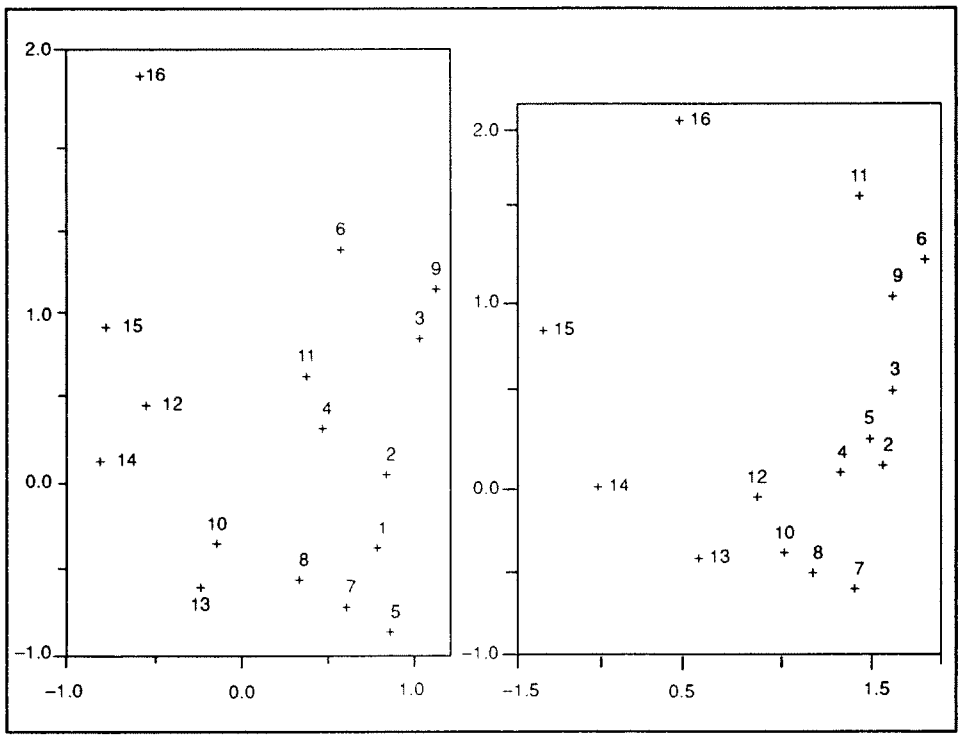
Example

For our example the choice for the quasi–independence model is such that S consists of the off–diagonal cells. Thus the quasi–independence model can be used to test whether the row and column variable are independent given that they have different levels. We will use reconstitution of order 2 for the diagonal elements of the car data. Thus we remove the dominating influence of the diagonal elements. The solution for reconstitution of order zero has been discussed in van der Heijden et al. (1989a). We use order 2 so that the results can be easily studied from a two–dimensional graphic display.

The quasi–independence model has a fit of $X^2 = 235{,}914$. The first two singular values are $\lambda_1 = .295$ (.469), $\lambda_2 = .235$ (.297). Thus 76.6% of the inertia $\Sigma_m \lambda_m^2$ for the adjusted table is displayed in these two dimensions (for the diagonal cells the inertia is displayed in full). A graphic display is given in figure 1a and 1b. In this display the scores $r_{im}\lambda_m^{1/2}$ and $c_{jm}\lambda_m^{1/2}$ are used as coordinates for the category points, so that the display can be interpreted as a biplot (see Gabriel, 1971). The display shows that the first dimension is a cheap–expensive dimension, with cheap car types on the right and expensive car–types on the left, while the second dimension is an import–domestic dimension, with import cars at the top, and domestic cars at the bottom. For a detailed discussion we refer to van der Heijden et al. (1989a), where the display for reconstitution of order zero is given. This display, however, explains only 60.5% of the inertia (which is for reconstitution of order zero the departure from quasi–independence as measured by Pearson's chi–square).

3. MODELS WITH LOGBILINEAR ASSOCIATION TERMS

In the Anglo–Saxon literature, the dominant tradition in contingency table analysis is loglinear modelling. The quasi–independence model is part of this tradition. In the last decade a good deal of attention has been given to models with logbilinear terms for the association between variables. The models discussed below are usually fit by maximizing the likelihood (see, for example, Goodman, 1979).



Figg.1a (left) and 1b (right): Left is the plot for row points, right is the plot for columns. Solution for reconstitution of order 2.

Let the saturated loglinear model for a two–way contingency table be

$$\log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{3}$$

with identifying restrictions, for example, $\Sigma_i u_{1(i)} = \Sigma_j u_{2(j)} = \Sigma_i u_{12(ij)} = \Sigma_j u_{12(ij)} = 0$. Under independence, $u_{12(ij)} = 0$.

The RC–association model (see Goodman, 1979) is

$$\log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + \phi u_i v_j \tag{4}$$

with identifying restrictions, for example, $\Sigma_i p_{i+} u_i = \Sigma_j p_{+j} v_j = 0$ and $\Sigma_i p_{i+} u_i^2 = \Sigma_j p_{+j} v_j^2 = 1$. Due to the logbilinear term $\phi u_i v_j$ the matrix of interaction parameters $u_{12(ij)} = \phi u_i v_j$ has rank 1.

The RC(M)–association model (Goodman, 1986) assumes that the matrix of interaction parameters has rank M ($0 \leq M \leq \min(I-1, J-1)$):

$$\log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + \sum_{m=1}^{M} \phi_m u_{im} v_{jm} \tag{5}$$

with identifying restrictions, for example, $\Sigma_i p_{i+} u_{im} = \Sigma_j p_{+j} v_{jm} = 0$ and $\Sigma_i p_{i+} u_{im} u_{in} = \Sigma_j p_{+j} v_{jm} v_{jn} = \delta^{mn}$, where $\delta^{mn} = 1$ if m=n, and $\delta^{mn} = 0$ if $m \neq n$ (compare the restrictions for (1)).

Relation with CA

In the study of the relation between CA and models with logbilinear terms attention has been given mainly to the relation between (1) and (4). First, if the data in the contingency table derive from a discretized bivariate normal distribution (or a distribution that is bivariate normal after a suitable transformation of the rows and columns), then the RC–association model and CA are closely related in the sense that $r_{i1} \approx u_i$, $c_{j1} \approx v_j$ and $\lambda_1 \approx \phi$ (see Goodman, 1981). Second, if $x = \approx {}_1 r_{i1} c_{j1}$ is small compared to unity, so that $\log (1+x) \lambda x$, it follows that $\lambda_1 r_{i1} c_{j1} \approx \phi u_i v_j$ (see Escoufier, 1982). This approximation will be closer when x is smaller. However, even when x is relatively large for some cells (i,j), the approximation can still be relatively close because it is not the values x that are studied but a factorization of these values in terms of ϕ, u_i and v_j. The result of Escoufier can be extended in a straightforward way to the RC(M)–association model (5) in the sense that then $x = \Sigma_m \rho_m r_{im} c_{jm}$, with the result that $\Sigma_m \lambda_m r_{im} c_{jm} \approx \Sigma_m \phi_m u_{im} v_{jm}$ when x is small.

Quasi–independence, and the departure from quasi–independence

Quasi–independence is defined as $\log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + \delta^{ij} u_{12(ij)}$ where $\delta^{ij} = 1$ if (i,j) is not in S, and $\delta^{ij} = 0$ if (i,j) " S. The result is that for the cells not in S $p_{ij} = \hat{\pi}_{ij}$, and for the other cells a form of independence will hold.

Starting from quasi–independence, we now want to model the association for the cells in S in a restrictive way. One such a model is (4) with an additional term $\delta^{ij} u_{12(ij)}$. Thus the cells not in S do not affect the fit of the model. The logbilinear term $\phi u_i v_j$ can be used to study the association in the cells in S. A similar adjustment can be made for model (5). Given the relation of the RC–association model to ordinary CA, we can expect a close similarity between such an adjusted RC–association model and the CA procedure in section 2 if the appropriate condition is fullfilled. For example, for the adjusted form of (5) this condition is that for the cells in S, $x = \Sigma_m^h \lambda_m^{(v)} r_{im}^{(v)} c_{jm}^{(v)}$ should be small compared to one.

Example

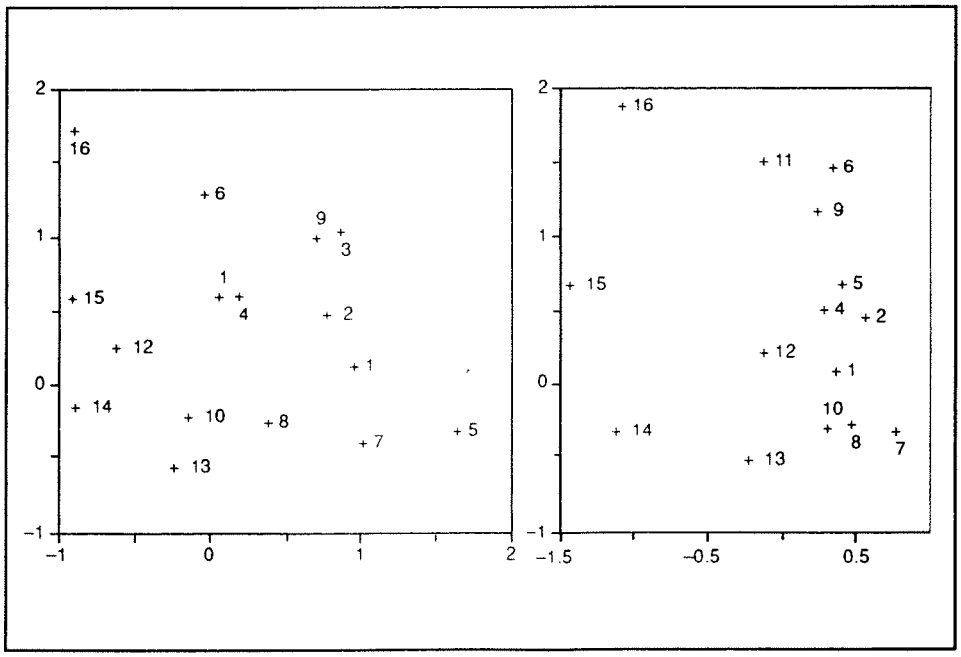
The model we choose for the car switching data is

$$\log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + \delta^{ij} u_{12(ii)} + \sum_{m=1}^{2} \lambda_m u_{im} v_{jm} \quad (6)$$

where $\delta^{ij} = 1$ if i=j, and $\delta^{ij} = 0$ if $i \neq j$. Thus the diagonal cells are ignored and they do

not affect the fit of the model. The fit is Pearson X^2 = 39,416 (likelihood ratio G^2 = 38,624) (df = 153). Compared with the quasi–independence model that has a fit of X^2 = 235,914 we modelled 83.3% of the departure from quasi–independence by including two logbilinear terms for the association. This is clearly better than the percentage of 76.6% found for the CA procedure.

A biplot can be made for the association $\Sigma_m \phi_m u_{im} v_{jm}$ by using scores $u_{im}\lambda_m^{1/2}$ and $v_{jm}\lambda_m^{1/2}$ as coordinates in a M–dimensional space. Such a biplot will facilitate the interpretation. A biplot is found when figures 2a and 2b are superimposed. The similarity with figures 1a and 1b is striking. The graphic representations obtained with the CA procedure to ignore specific cells, are very similar to the graphic representations that are obtained with the adjusted RC–association model. This illustrates that the relation that exists between CA and the RC–association model also holds for more elaborate models like model (6) and an adjusted CA procedure. In general the similarity will be less for those categories that have many cells in S for which $x = \Sigma_m^{\ h} \lambda_m^{(v)} r_{im}^{(v)} c_{jm}^{(v)}$ is large compared to one. For our example, the most extreme values for x are for row 16: x is 3 (rounded) for cell (16,6), 3 for cell (16,15) and 4 for (16,11). This explains the difference in the positions of point 16 in figures 1a and 2a.



Figures 2a (left) and 2b (right). Left is the plot for row points, right is the plot for columns. Solution for model with two terms for log–bilinear association.

Another example of this relation, not shown here, is that strikingly similar results obtain between CA used to decompose residuals from quasi-symmetry and a model with logbilinear terms to deal with asymmetry (see van der Heijden and Mooijaart, 1991). Our (limited) experience is that graphic representations obtained with (a generalization of) CA can be approximated to such an extent by models with logbilinear terms that the interpretation of the representations is basically the same (see also van der Heijden and Worsley, 1988; Green, 1989).

The RC-association modelling approach is far more flexible than the CA approach in that constraints on the parameters are easily imposed. It is not difficult to alter model (6) by imposing further constraints such as $u_{im} = v_{im}$, by adding parameters for the asymmetry, dropping parameters for the diagonal, and so on. The constraint $u_{im} = v_{im}$ assumes that the association is symmetric, and a graphic display will show only one point for corresponding row and column categories (see Becker, 1990, and van der Heijden and Mooijaart, 1991, for a more elaborate discussion of this model).

4. THE LATENT BUDGET MODEL

Latent budget analysis is a reparametrization of the latent class model (LCA), and it is closely related to CA. For an introduction to latent budget analysis and for more details, see van der Heijden et al. (1989b, 1990, in press), de Leeuw et al. (1990), de Leeuw and van der Heijden (1991).

Let A be the row variable, indexed by i, and B be the column variable, indexed by j. LBA is a model for conditional probabilities. The row vector with conditional proportions p_{ij}/p_{i+} is termed the *observed* budget for row i. Thus, a row budget is the conditional distribution of column categories for that row. LBA describes the *theoretical* row budgets having elements π_{ij}/π_{i+} as a mixture of T *latent* budgets, indexed by t (t=1,...,T), having elements $\pi_{jt}^{\bar{B}X}$ with $\pi_{jt}^{\bar{B}X} \geq 0$ and the bar over B indicating that $\Sigma_j \pi_{jt}^{\bar{B}X} = 1$. Let the mixture be defined by the parameters $\pi_{it}^{A\bar{X}}$, with $\pi_{it}^{A\bar{X}} \geq 0$ and $\Sigma_t \pi_{it}^{A\bar{X}} = 1$. The model is defined as

$$\frac{\pi_{ij}}{\pi_{i+}} = \sum_{t=1}^{T} \pi_{it}^{A\bar{X}} \pi_{jt}^{\bar{B}X} \qquad (7)$$

If the number of latent budgets T is 1, then (7) is equivalent to the independence model in which case $\pi_{it}^{A\bar{X}} = 11$ and $\pi_{jt}^{\bar{B}X} = \pi_{+j}$. If T = min (I,J), then the model is saturated. Under the assumption that the frequencies are derived from a product-

multinomial distribution, maximum likelihood estimates can be derived, and the model can be tested against the unconstrained alternative using chi–square tests. There are (I–T)(J–T) degrees of freedom.

Clogg (1981) presented LBA as a reparametrization of latent class analysis for two–way contingency tables (LCA). As a result of this equivalence many properties of LCA also hold for LBA. Let π^X_t be the probability to fall into latent class t, let $\pi^{\bar{A}X}_{it}$ be the conditional probability of level i in latent class t, and let $\pi^{\bar{B}X}_{jt}$ be the conditional probability of level j in latent class t. Then LCA is defined as a model for the latent probabilities π_{ijt} of falling into level i of variable A, level j of variable B, and level t of the latent variable X:

$$\pi_{ij} = \sum_t^T \pi_{ijt} = \sum_t^T \pi^X_t \pi^{\bar{A}X}_{it} \pi^{\bar{B}X}_{jt} \qquad (8)$$

LBA and LCA have in common the parameters $\pi^{\bar{B}X}_{jt}$. LBA parameters $\pi^{\bar{A}X}_{it}$ are derived from LCA parameters πs^X_t and $\pi^{\bar{A}X}_{it}$ by using Bayes rule: $\pi^{A\bar{X}}_{it} = \pi^X_t \pi^{\bar{A}X}_{it} / \Sigma_t \pi^X_t \pi^{\bar{A}X}_{it}$. LBA and LCA are also easily compared in terms of the latent probabilities π_{ijt}. For both models $\pi^{\bar{B}X}_{jt}$ is related to π_{ijt} as $\pi^{\bar{B}X}_{jt} = \pi_{+jt} / \pi_{++t}$;; for LCA $\pi^{\bar{A}X}_{it}$ is related to π_{ijt} as $\pi^{\bar{A}X}_{it} = \pi_{i+t} / \pi_{++t}$, whereas for LBA $\pi^{\bar{A}X}_{it}$ is related to π_{ijt} as $\pi^{A\bar{X}}_{it} = \pi_{i+t} / \pi_{i++}$.

Relations with correspondence analysis

De Leeuw and van der Heijden (1991) discuss LBA and CA as two reduced rank models for contingency tables, and show that they are closely related in terms of theoretical probabilities π_{ij}. From this perspective the main difference between the two models is that the parameters of LBA are constrained to be non–negative.

Consider a rank P matrix with probabilities π_{ij} denoted as $\prod(P)$. Let LCA(P) be a rank P matrix yielded by LBA with T=P latent classes, and let CA(P) be a rank P matrix yielded by CA with M+1=P dimensions (compare (1) and (8)). The first result is that $\prod(P)$ and CA(P) are equivalent: every rank P matrix can be parametrized by CA(P). Secondly, in general, if LCA(P) is true, then CA(P) is true, but the reverse does not hold since LCA(P) is more restrictive than CA(P). Third, it is clear that LCA(1) is equivalent with CA(1), both yielding the independence model. Fourth, if P=min(I,J), then both models are saturated. Fifth, de Leeuw and van der Heijden (1991) prove that LCA(2) and CA(2) are always equivalent. They give an example of a matrix for which CA(3) is true, but LBA(3) is not true.

This shows that LCA and LBA are closely related to CA. The above results imply that LCA(2) will have the same fit as CA(2) if they are fit with the same criterion (for example, maximum likelihood). For $2 < P < \min(I,J)$ the fit of CA(P) can be better than the fit of LCA(P).

Quasi–independence, and departure from quasi–independence

LBA with T=1 describes independence. One way to define an LBA–model that is equivalent to the quasi–independence model is by defining a specific budget for each cell to be ignored . For square tables where interest is focused on the off–diagonal cells it is possible to define an LBA model with I+1 latent budgets, i.e. I latent budgets for the diagonal cells and one general latent budget (see, for example, Clogg, 1981; Hagenaars, 1990). For example, if latent budget 4 is the budget for diagonal cell (4,4), then the idea is to constrain $\pi_{4t}^{A\bar{X}} = 0$ for $t \neq 4$, $\pi_{4t}^{\bar{B}X} = 1$ for t=4, and $\pi_{4t}^{\bar{B}X} = 0$ for $t \neq 4$. The result is that in the latent matrix the only nonnegative probability π_{ijt} in level t=4 is π_{444}, the other probabilities π_{ij4} being zero. Thus only one free parameter $\pi_{it}^{\bar{A}X}$ is included for each diagonal cell. Let such a free parameter be denoted by $\tilde{\pi}_{it}^{A\bar{X}}$. This approach to defining quasi–independence with LBA only works if all these free parameters $\tilde{\pi}_{it}^{A\bar{X}}$ turn out to be positive, which will be the case if there is clustering for all cells on the diagonal of the matrix. These cells do not then affect the fit of the model. In the context of transition matrices LBA can then be given a mover–stayer interpretation: the parameters $\tilde{\pi}_{it}^{A\bar{X}}$ specify the proportion of stayers, whereas the other parameters $\pi_{it}^{\bar{A}X}$ specify the movers (for more details, see Clogg, 1981).

In this approach LBA with one general latent budget approximates the off–diagonal probabilities by a rank 1 matrix. Similarly, LBA with two (or more) general latent budgets approximate the off–diagonal probabilities by a matrix of rank 2 (or more), where the parameters $\tilde{\pi}_{it}^{A\bar{X}}$ have the effect that $\hat{\pi}_{ii} = p_{ii}$ (see, for example, Clogg, 1981; Hagenaars, 1990).

Example

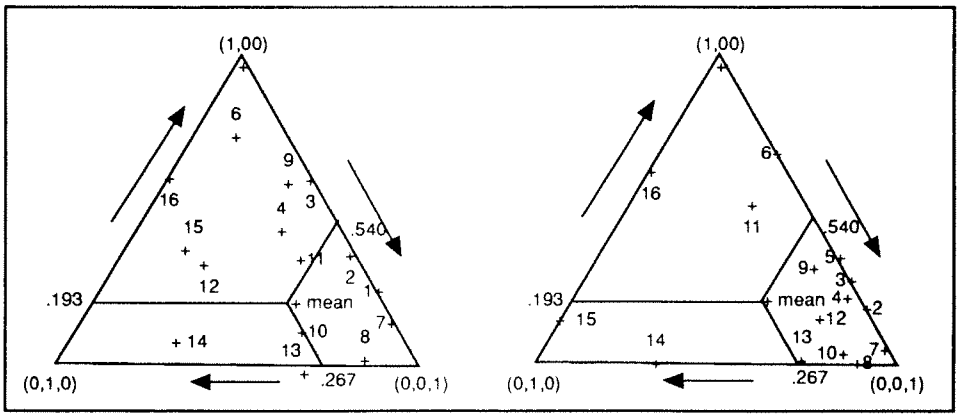
The aim is to fit a model for the departure from quasi–independence that is similar to the CA procedure described in section 2. For the car data such a model has 19 latent budgets, namely 16 for the diagonal cells and three general latent budgets. Thus the off–diagonal proportions are approximated by a rank 3 matrix, which is similar to what happens in reconstitution of order 2. This LBA model has

a fit of $X^2 = 56{,}209$ ($G^2 = 52{,}992$), df = 153. Since for quasi–independence $X^2 = 235{,}914$, 76.1 % is modelled by including two more general latent budgets into the model. The parameter estimates can be found in table II. The first column gives the parameter estimates for the stayers, ranging from .037 for car type 10 to .545 for car type 15. In columns 2, 3 and 4 we find the estimates $\hat{\pi}_{it}^{A\bar{X}}$, showing the probabilities of latent budgets 1 to 3. So, for people having a car type 1, .929 are movers, i.e. .203 of the people adopt latent budget 1 and .726 adopt latent budget 3. The latent budget estimates $\hat{\pi}_{jt}^{\bar{B}X}$ are given in columns 5, 6 and 7; they are most easily interpreted by comparing them with the unconditional probabilities $\hat{\pi}_{+j}$. For example, in latent budget 1 the probability of buying car type 6 is .088, which is much higher than the unconditional probability .016. Budget 1 is a budget with higher probabilities for buying an import car (3, 4, 6, 9, 16), the second latent budget is the budget for buying a more expensive domestic car (12, 15, 16), and the third budget is the budget for buying a cheaper domestic car (1, 7, 8, 10, 13).

Tab. II: parameter estimates for the latent budget solution. For car i=1, .071 are stayers, .203 adopt the first latent budget, .000 the second latent budget and .726 the third latent budget. In the first latent budget .115 buys car type 1, .017 car type 2, .200 car type 3, and so on. Estimates with a '*' are constrained in order to identify the solution (see de Leeuw et al., 1990).

FORM								
		t=1	t=2	t=3	t=1	t=2	t=3	
1	.071	.203	.000	.726	.115	.042	.160	.066
2	.047	.352	.013	.588	.017	.000	.013	.011
3	.198	.486	.017	.300	.200	.009	.089	.048
4	.184	.401	.094	.321	.226	.022	.115	.073
5	.011	.000	.000	.989	.001	.000	.000	.000
6	.134	.659	.098	.109	.088	*.000	.009	.016
7	.103	.063	.013	.822	.027	*.000	.118	.080
8	.070	.041	.108	.781	.006	.017	.049	.040
9	.048	.570	.047	.335	.028	.003	.007	.014
10	.037	.073	.259	.630	.037	.080	.170	.187
11	.132	.322	.115	.430	.033	.005	.003	.013
12	.197	.253	.372	.178	.165	.206	.161	.096
13	.196	.000	.237	.567	*.000	.117	.074	.152
14	.182	.065	.482	.271	*.000	.339	.034	.136
15	.545	.168	.228	.058	.042	.152	*.000	.064
16	.333	.403	.264	.000	.015	.005	*.000	.005
				1.000	1.000	1.000		

We will make the results of LBA comparable to those for CA and association models by making graphic representations. We can make these representations of



Figures 3a (left) and 3b (right). Left is the plot for row points, right is the plot for columns. Solution for latent budget model with three latent budgets. E.g., given row point 6 the observations fall mainly into the first latent budget; for row point 1 the observations fall into latent budget 1 and latent budget 3.

the latent budgets as follows (compare de Leeuw et al., 1990): let the elements $\hat{\pi}_{ijt}$ that correspond to the three general latent budgets be denoted by $\hat{\pi}_{ijt}{}^{*}$. By using only those elements $\hat{\pi}_{ijt}{}^{*}$, elements $\hat{\pi}_{it}^{A\bar{X}*} \equiv \hat{\pi}_{ijt}{}^{*} / \hat{\pi}_{i++}{}^{*}$ are the conditional probabilities of the movers adopting latent budget t given that one has car type i. Because T=3, the elements $\hat{\pi}_{it}^{A\bar{X}*}$ can be used to plot each i as a point in a three–dimensional space. Since $S_t\ \hat{\pi}_{it}^{A\bar{X}*} = 1$, these points lie in a two–dimensional subspace. This subspace is displayed in figure 3a. Similarly, we can derive rescaled column parameters $\hat{\pi}_{jt}^{\bar{B}X*} = \hat{\pi}_{ijt} / \hat{\pi}_{++t}$, which can be plotted into a two–dimensional subspace of a three–dimensional space. This two–dimensional subspace is displayed in figure 3b. The elements $\hat{\pi}_{jt}^{\bar{B}X*}$ specify the probabilities of latent budget t given that car type j is bought.

Figures 3a and 3b are, again, remarkably similar to figures 1a and 1b if we allow for an appropriate linear transformation of the coordinates in the figures. This also holds for the percentage of chi–square accounted for in both solutions, which is 76.1 % for LBA and 76.6 for CA. This similarity is the result of the fact that both approaches can be interpreted as reduced rank approximations for the cells in S, which are the off–diagonal cells in this example.

5. CONCLUSION

The starting point of this paper was an investigation of the similarity between CA and modelling approaches when CA is used to study the departure from quasi–independence. The similarity of CA with RC–association models and LCA/LBA was studied before in the context of what could be viewed as the departure from dependence. This was illustrated by means of an example: all three approaches yielded very similar graphic displays.

REFERENCES

Becker, M.P. (1990). Quasisymmetric models for the analysis of square contingency tables. *Journal of the Royal Statistical Society, Series B,, 52*, 369–378.

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.

de Leeuw, J. and van der Heijden, P.G.M. (1988b). Correspondence analysis of incomplete tables. *Psychometrika, 53*, 223–233.

de Leeuw, J., and van der Heijden, P.G.M. (1991). Reduced rank models for contingency tables. *Biometrika, 78*, 239–232

de Leeuw, J., van der Heijden, P.G.M., and Verboon, P. (1990). A latent time budget model. *Statistica Neerlandica*, 44, 1–22.

Escofier, B. (1983). Analyse de la difference entre deux mesures sur le produit de deux memes ensembles. *Cahiers de l'analyse des donnees, 8*, 325–329.

Escoufier, Y. (1982). L'analyse des tableaux de contingence simples et multiples. In: *Proc. International Meeting on the Analysis of Multidimensional Contingency Tables* (Rome 1981). Ed. R. Coppi. *Metron, 40*, 53–77.

Gabriel, K.R. (1971). The biplot–graphic display of matrices with applications to principal component analysis. *Biometrika, 58*, 453–467.

Gifi, A. (1990). *Non–linear multivariate analysis*. New York: Academic Press

Goodman, L.A. (1968). The analysis of cross–classified data: independence, quasi–independence, and interactions in contingency tables with and without missing entries. *Journal of the American Statistical Association, 63*, 1091–1131.

Goodman, L.A. (1979). Simple models for the analysis of association in cross–classifications having ordered categories. *Journal of the American Statistical Association, 70*, 755–768.

Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross–classifications having ordered categories. *Journal of the American Statistical Association, 76*, 320–334.

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.

Gower, J.C. (1989). Discussion of the paper by van der Heijden, P.G.M., de Falguerolles, A., and de Leeuw, J. (1989a). *Applied Statistics.*

Green, M. (1989). Discussion of the paper by van der Heijden, P.G.M., de Falguerolles, A., and de Leeuw, J. (1989a). *Applied Statistics.*

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* New York: Academic Press.

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

Harshman, R.A., Green, P.E., Wind, Y., and Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research. *Marketing Science, 1,* 205–242.

Nishisato, S. (1980). *Analysis of categorical data: dual scaling and its applications.* Toronto: University of Toronto Press.

van der Heijden, P.G.M., and de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika, 50,* 429–447.

van der Heijden, P.G.M., and Worsley, K. (1988). Comment on Correspondence analysis used complementary to loglinear analysis. *Psychometrika, 50,* 429–447.

van der Heijden, P.G.M., de Falguerolles, A., and de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and loglinear analysis. *Applied Statistics, 38,* 249–292.

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.

Van der Heijden, P.G.M., de Leeuw, J. and Mooijaart, A. (1990). On the relation between latent class analysis and correspondence analysis. Toulouse: *Proceedings of Fifth International Workshop on Statistical Modelling,* pp. 99–107.

van der Heijden, P.G.M., and Mooijaart, A. (1991). *A class of logbilinear models for the analysis of symmetric and skewsymmetric association.* Utrecht: ISOR, Methods Series MS– 91–1.

van der Heijden, P.G.M., Mooijaart, A., and de Leeuw, J. (in press). Constrained latent budget analysis. In: P. Marsden (Ed.) *Sociological Methodology 1992.*

RIASSUNTO

L'Analisi delle Corrispondenze, i modelli di associazione con termini logbilineari, i modelli a bilanci latenti sono usati per studiare le divergenze dal modello loglineare di quasi–indipendenza in quei casi in cui alcune delle celle di frequenza siano omesse o di non interesse. Sono stati scelti quegli approcci in cui queste celle non influenzano l'adattamento del modello o la tecnica. È indicato come effettuare una rappresentazione grafica dei parametri per l'associazione dei punteggi. Come esempio, si analizza una tabella di contingenza quadrata dove l'interesse è focalizzato sull'associazione nelle celle non diagonali. Per questo esempio si mostra come queste rappresentazioni grafiche siano molto simili. Poiché questo articolo appare su un numero speciale di Statistica Applicata dedicata alla scuola italiana e olandese di analisi dei dati, particolare attenzione sarà data ai contributi olandesi nel campo dell'analidi delle tabelle di contingenza. Per una rassrgna sui contributi italiani in questo campo rimandiamo all'articolo di Siciliano (in questo volume).