A COMBINED APPROACH OF CONTINGENCY TABLE ANALYSIS USING CORRESPONDENCE ANALYSIS AND LOGLINEAR ANALYSIS

Peter G.M. van der Heijden*, Antoine de Falguerolles & Jan de Leeuw*****

Summary

Correspondence analysis and one of its generalizations are presented as tools that can be used for the analysis of residuals from loglinear models. By recognizing relations between correspondence analysis and loglinear analysis a better understanding of correspondence analysis is obtained. Furthermore, it is shown how these relations can be used to arrive at a combined approach to contingency table analysis using both loglinear analysis and correspondence analysis.

Keywords: CORRESPONDENCE ANALYSIS, LOGLINEAR ANALYSIS, ASSOCIATION MODELS, EXPLORATORY DATA ANALYSIS, ASYMMETRY

* Department of Psychometrics and Research Methods, University of Leiden, Hooigracht 15, 2312KM Leiden, The Netherlands
** Laboratoire de Statistique et Probabilité, Université Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse Cedex, France
*** Departments of Psychology and Mathematics, University of California Los Angeles, Los Angeles, CA 90024, U.S.A.

1. Analysis of categorical data

Correspondence analysis (CA) as a method for the analysis of categorical data has quite a long history. The starting point of this history is usually set in 1935, and since then, it has been reinvented several times. Hence at present CA is known under different names, such as reciprocal averaging, dual (or optimal) scaling, canonical correlation analysis and simultaneous linear regressions. Greenacre (1984) discusses the differences in approach between CA and these methods (compare also Nishisato, 1980, ch. 3). The term CA originates from France, where it is quite popular. This is probably due to Benzécri and associates. They developed CA in the sixties, and their work culminated in standard references as Benzécri et al. (1973), and the series Pratique de l'Analyse des Données (Benzécri et al., 1980, 1986). Furthermore, since 1976 they have a special journal dedicated almost exclusively to (the applications of) CA, called Les Cahiers de l'Analyse des Données. CA places a heavy emphasis on geometrical representations, and probably this is one of the reasons this approach became so popular, at least in France. One reason for its developmental lag outside France is perhaps the language problem. However, recent publications in English have done much to alleviate this situation (for example Nishisato, 1980; Gifi, 1981; Jambu & Lebeaux, 1983; Greenacre, 1984; Lebart, Morineau & Warwick, 1984; compare also the papers of Hill, 1974; Healy & Goldstein, 1976; Deville & Malivaud, 1983), and it is our impression that CA is becoming rapidly popular outside France (compare the bibliography of Nishisato, 1986). Another reason might be that CA is often introduced without any reference to other methods of statistical treatment of categorical data which have proven their usefulness and flexibility.

A major difference between CA and most other techniques for categorical data analysis lies in the use of models. In loglinear analysis (LLA), for example, a distribution is assumed under which the data are collected, then a model for the data is hypothesized and estimations are made under the assumption that this model is true, and finally these estimates are compared to the observed frequencies in order to evaluate the model. In this way it is possible to make inferences about the population on the basis of sample data.

In CA it is claimed that no underlying distribution has to be assumed, no model has to be hypothesized, but a decomposition of the data is obtained in order to study the "structure" in the data. The primary interest is in the presentation of the structure in the observed data in some optimal way. A rationale for this approach is that the observed data matrix is approximated in a least squares sense by some data matrix of lower rank. Strictly speaking, conclusions about the data then may not be generalized to the population (although in practice it is difficult to prevent people from doing this).

A first approach to bridge this gap between CA and model-based approaches was to show under which conditions CA results (i.e. row and column scores, singular values) are similar to those of the logmultiplicative RC model (see Goodman, 1979, 1981a,b; in Goodman, 1985, 1986 this model is discussed under the name RC association model). Recently, the gap between CA and model-based approaches is bridged further by Goodman (1985, 1986), and Gilula & Haberman (1986), who introduced a version of CA as a model.

In this paper we summarize work that departs from a different angle, namely that of CA as a tool for residual analysis. In this approach CA is viewed as a technique that can be used for the exploration of residuals from loglinear models. The usage of a generalization of CA plays an important role in this approach. It will be shown that CA and LLA are closely related. This will lead us to two general conclusions:

1. A better understanding of CA procedures is obtained by recognizing these relations between CA and LLA. This facilitates the understanding of CA for those who are better accustomed to LLA.
2. It is possible to combine CA and LLA in one combined approach of contingency table analysis. To put it roughly, in this approach LLA is used for the assessment of the importance of interaction effects, and CA is used for the further exploration of these interactions. In this context CA may be particularly useful when the number of interaction parameters to be interpreted becomes large.

The distinction between point 1 and 2 may seem a bit artificial. However, our impression is that experienced users of loglinear and related models often do not seem to feel a need for CA, and therefore the second point does not seem relevant to them. But for them, however, the first point may still be of interest.

In the sequel we will first give a short introduction to (generalized) CA, and LLA. In section 3 we will present some results indicating relations between CA and LLA. In section 4 we illustrate a combined approach for the analysis of a square contingency table.

2. Tools of analysis: correspondence analysis, generalized correspondence analysis, and loglinear analysis

2.1 Correspondence analysis

We will discuss CA here with an emphasis on the geometrical aspects. For details and proofs, we refer to Benzécri et al. (1973, 1980, 1986), Nishisato (1980), Gifi (1981), Greenacre (1984) and Lebart, Morineau & Warwick (1984), and the references given there.

CA is a technique by which it is possible to find a multi-dimensional representation of the dependence between the rows and columns of a two-way contingency table. This representation is found by allocating scores to the row and column categories, and displaying the categories as points, where the scores are used as coordinates of these points. These scores can be normalized in such a way that distances between row points and/or between column points in Euclidean space are equal to so-called chi-square distances.

Consider a two-way contingency table P with proportions p_{ij}, having I rows ($i=1,\dots,\underline{I}$) and J columns ($j=1,\dots,\underline{J}$). An index is replaced by '+' when summed over the corresponding variable, e.g. $\Sigma_j p_{ij} = p_{i+}$. Chi-square distances can be computed between rows as well as between columns. We will proceed by considering chi-square distances between rows. These distances are computed on the profiles of the rows of a matrix, where the profile of row i is the vector of conditional proportions p_{ij}/p_{i+}. The chi-square distance between rows i and i' is defined as

$$(1)\ \delta^2(i,i') = \Sigma_j \frac{(p_{ij}/p_{i+} - p_{i'j}/p_{i'+})^2}{p_{+j}} .$$

Formula (1) shows that $\delta^2(i,i')$ is a measure for the difference between the profiles of row i and i'. When i and i' have the same profile, $\delta^2(i,i') = 0$. The configuration of I row points is located in a Euclidean space of dimension (I-1). The profile of column proportions p_{+j}, being the mean row profile, is in the centroid of this space.

Up to now we have only discussed the situation for the row categories. CA is symmetric in the sense that similar results hold for the columns: it is also possible to construct a (J-1)-dimensional space in which the mean column profile is placed in the centroid.

The CA solution can be found as follows: let P be the matrix to be analyzed; D_r and D_c diagonal matrices with respectively marginal row proportions p_{i+} and column proportions p_{+j}; $E = D_r tt' D_c$, where t is a unit vector, the length of which depends on the context. Elements of E have the form

$$(2)\ e_{ij} = p_{i+}p_{+j}.$$

We then compute the singular value decomposition (SVD) of the matrix $D_r^{-\frac{1}{2}}(P-E)D_c^{-\frac{1}{2}}$. Elements of this matrix have the value $(p_{ij}-e_{ij})/e_{ij}^{\frac{1}{2}}$; they are proportional to standardized residuals. These residuals are decomposed with (3):

$$(3)\ D_r^{-\frac{1}{2}}(P-E)D_c^{-\frac{1}{2}} = U\Lambda V'$$

where $U'U = I = V'V$, and Λ is a diagonal matrix with singular values λ_α in descending order.

The row and column scores are normalized as follows:

$$(4a)\ R = D_r^{-\frac{1}{2}}U$$

(4b) $C = D_c^{-\frac{1}{2}}V$

So $R'D_rR = I = C'D_cC$. Furthermore $t'D_rR = 0 = t'D_cC$: for each dimension row scores and column scores have weighted average 0 and therefore weighted variance 1.

The relation between the row and column points is specified by the 'transition formulas'

(5a) $\tilde{R} = D_r^{-1}PC$

(5b) $\tilde{C} = D_c^{-1}P'R$

where $\tilde{R} = R\Lambda$ and $\tilde{C} = C\Lambda$. Equations (5) show that row points $\tilde{R}$ are the weighted averages of the column points C while at the same time column points $\tilde{C}$ are the weighted averages of the row points R.

By substituting (4) in (3), one finds, after some manipulating,

(6) $P = E + D_rR\Lambda C'D_c = D_r(1 + R\Lambda C')D_c$

which is known under the name 'reconstitution formula'. Equation (6) shows that CA decomposes the departure from independence in matrix P. It follows that CA only makes sense when these residuals are not merely a result of random variation from independence (in practice this is very often not checked for). Whether this is the case or not, can be tested by using the Pearson $\mathbf{X}^2$ statistic:

(7) $\mathbf{X}^2 = n\, \Sigma_i \Sigma_j\, (p_{ij}-e_{ij})^2/e_{ij}$

where n is the sample size. The relation between $\mathbf{X}^2$ and the squared singular values in Λ^2 follows from (3) and (7):

(8) trace $\Lambda^2 = \mathbf{X}^2/n$

In the French literature trace Λ^2 is often called the 'total inertia'. In the English literature $\mathbf{X}^2/n$ is often referred to as Pearson's coefficient of contingency. Equation (8) shows that CA decomposes the chi-square value for testing independence in the matrix.

Concluding, we see that we have a better understanding of what CA does. A naive model, the independence model, is chosen, and the residuals of this model are decomposed. In this way a picture of the interaction between the row and column variable is obtained.

However, sometimes the independence model is too naive. In order to have the possibility to decompose residuals from models different from the independence model, while retaining CA properties, we can make use of a generalization of CA.

2.2 A generalization of correspondence analysis

In this section we discuss a generalization of CA, proposed by Escofier (1983, 1984). It is easiest to discuss this generalization using its reconstitution formula

$$(9)\ P = Q + S_r R \Lambda C' S_c$$

where Q is an arbitrary matrix, having the same order as P; S_r and S_c are two diagonal matrices of arbitrary weights for row and column points, respectively, under the restriction that trace $S_r = 1$ and trace $S_c = 1$. A comparison of (9) with (6) reveals that CA is generalized in two ways. Firstly, it is possible to decompose the departure from other matrices than the independent matrix E; secondly, the weights S_r and S_c of the row and column points are not necessarily defined by the margins of P. We will not discuss any properties of this generalization here, but refer to references in the sections to follow.

Escoufier (1985) also discusses other alternatives of generalizing CA, but we will use only the particular generalization given in (9), because it is a convenient one for the decomposition of residuals from various loglinear models.

2.3 Loglinear analysis

LLA is a well known method to study structural relations between variables in a contingency table. We introduce here only some notation, for the two-variable case. With a two-way matrix, the unrestricted loglinear model has the form

$$(10)\ \log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

where π_{ij} denotes the theoretical proportion for cell (i,j), and the u-parameters can be constraint in different ways. The generalization to notation for higher-way tables is straightforward. We follow the convention to denote loglinear models by placing the variables that constitute the highest fitted margins between square brackets, i.e., for example, [1][2] for the independence model.

It is not common practice to interpret individual u-parameters. One reason may be their number, which often becomes very large, especially when there are (higher-order) interactions, or when the number of categories is large. Nevertheless, it is sometimes possible to attempt an interpretation by representing the parameters graphically. An alternative, to be discussed in section 3.2, is to use CA to assess which restrictions can be placed on the parameters.

3. CA and LLA as complementary tools of analysis

In this section we will show how CA and LLA can be seen to be related. This leads to a better understanding of CA. We will also indicate how this can lead to a combined use of both methods. For additional results we refer to Van der Heijden & De Leeuw (1985), de Falguerolles & Van der Heijden (1987), Van der Heijden (1987), Van der Heijden & Worsley (in press), De Leeuw & Van der Heijden (in press).

3.1 CA as a technique decomposing residuals from loglinear models for higher way tables

In section 2.1 it was emphasized that simple CA decomposes residuals from the independence model. Recently it is shown that CA is also related to LLA in case of higher way contingency tables (Van der Heijden & De Leeuw, 1985). In this section we will summarize these results, going from two-way tables to three- and higher-way tables respectively.

For a two-way table with proportions p_{ij} it is clear that the p_{ij} can be interpreted as ML-estimates of expected proportions under the saturated model [12]. On the other hand, model [1][2] implies that $\hat{\pi}_{ij} = p_{i+}p_{+j} = e_{ij}$ (see (2)). So, for a two-way table CA can be interpreted in terms of the difference between models [12] and [1][2].

Results are more interesting for analysis of higher-way tables by means of so-called "multiple tables", i.e. two-way tables constructed by concatenating "slices" of the higher-way tables, thus creating "interactive" variables. First consider the three-way table P with proportions p_{ijk}. In order to do CA, the two-way multiple table $P^{i[jk]}$ is constructed by coding the categories of variables 2 and 3 interactively. We denote the elements of $P^{i[jk]}$ and $E^{i[jk]}$ as $p_{i[jk]}$ and $e_{i[jk]}$. Values $e_{i[jk]}$ are equal to ML-estimates of expected proportions under model [1][23]: $\hat{\pi}_{ijk} = p_{i++}p_{+jk} = p_{i[++]}p_{+[jk]} = e_{i[jk]}$. So when the three-way matrix P is flattened to the multiple table $P^{i[jk]}$, CA can be interpreted in terms of the difference between the loglinear models [123] and [1][23]. When multiple tables $P^{j[ik]}$ or $P^{k[ij]}$ would have been constructed, values in E would follow model [2][13] or [3][12], respectively.

When we have a higher-way table P, we can construct a multiple table $P^{[a][b]}$, where a and b index the elements of two groups of variables coded interactively. A and B together constitute the higher way table, where $A \cap B = \emptyset$. In this case CA can be interpreted in terms of the difference between loglinear models [AB] and [A][B]: elements e_{ab} of E are equal to $e_{ab} = p_{a+}p_{+b} = \hat{\pi}_{ab}$.

Thus in the analysis of multiple tables, ordinary CA can be interpreted as a tool to analyze residuals from specific loglinear models, namely, for the three-way table, [1][23], [2][13] or [3][12]. The generalization of CA discussed in section 2.2, makes it possible to analyze the difference between less restricted loglinear models and the saturated model. For this purpose estimated proportions π_{ijk} are taken as elements of Q (see (9)), and the margins of multiple tables of P (being equal to those for Q) are taken as diagonal elements of S_r and S_c. Escofier (1983) and Van der Heijden (1987) discuss geometrical implications of using less restricted loglinear models.

In Van der Heijden & De Leeuw (1985) and Van der Heijden (1987) this approach is illustrated with some examples. In this application generalized CA solutions can be obtained using ordinary CA programs, namely by feeding the program with the matrix (P-Q+E). Comparing formula (6) with (9), this is easily seen. Therefore, we can interpret the distances between points in the generalized CA solution as ordinary chi-square distances (1) derived on the matrix (P-Q+E).

We conclude the following. Firstly, we know more about CA now. For higher-way tables we conclude that CA of multiple tables results in a solution showing the difference between two loglinear models. The results above also show that CA does not cover all dependence in the data matrix. For example, for multiple table $P^{i[jk]}$ the first-order interaction between variables 2 and 3 does not affect the solution.

This leads to our second point, namely that we can use these results to come to a combined approach using both CA and LLA. For example, when the most suitable multiple table must be chosen in order to analyze a higher-way table, these findings can be used in two ways. Let us consider again the three-way table. Firstly, when some first-order interaction does not seem particularly interesting, the corresponding two variables should be coded interactively. Secondly, when one is especially interested in the relation of some variable with the other two, this variable should not be coded interactively with another variable. Ordinary CA shows the two first-order interactions of this variable with the other two, and the second order interaction. The generalization of CA can be used to suppress other first-order interactions. Of course, it is important to first check whether the residuals are statistically significant and therefore contain interesting information.

Another implication is that in some cases it is possible to use CA and LLA complementary to each other. In Van der Heijden & De Leeuw (1985) it is proposed to use the two techniques to find answers to different questions: in a first step LLA can answer the question as to which variables are related by detecting the important interactions; in a second step CA can answer the question as to how these variables are related, i.e., which categories occur more often together than expected, or less often. This is done by displaying the interactions graphically. Some of the computable X^2-measures can be decomposed with ordinary CA, not only for two-way tables but also for higher-way tables. This circumvents problems with the interpretation of parameters (compare section 2.3). We will add a third step using results discussed in the next section.

3.2 Using correspondence analysis for finding loglinear models with restrictions on the interaction

In the section above CA was used to represent interactions graphically, thus circumventing the problem of interpreting a large amount of parameters. Another way to solve this problem is to stay within the LLA approach and to restrict the interaction parameters in some form or another, for example, to have a product form. This is done in the RC association model (Andersen, 1980; Goodman, 1979, 1981a,b, 1985, 1986). In such a model the number of parameters to be interpreted is considerably reduced, especially when the number of categories of the interacting variables is large.

In the two-variable case, CA and the RC association model are related in the following way. The CA representation in k dimensions can be written in an adapted version of (6) as

$$(11)\quad m_{ij} = p_{i+}p_{+j}\,(1 + \Sigma_{\alpha=1}^{k}\lambda_{\alpha}r_{i\alpha}c_{j\alpha})$$

where m_{ij} is the reconstituted proportion for cell (i,j). Escoufier (1982) noted that if $x=\Sigma_{\alpha=1}^{k}\lambda_{\alpha}r_{i\alpha}c_{j\alpha}$ is small compared to 1 (so that $\log(1+x) \approx x$) we can rewrite (11) as

$$(12)\ \log m_{ij} \approx u + u_{1(i)} + u_{2(j)} + \Sigma_{\alpha=1}^{k}\lambda_{\alpha}r_{i\alpha}c_{j\alpha}.$$

where $u=0$, $u_{1(i)}=\log(p_{i+})$, $u_{2(j)}=\log(p_{+j})$.

Another result is that if k=1, and the proportions stem from a discretized bivariate normal distribution (or a distribution that is bivariate normal after a proper transformation of the rows and columns), then no matter how large the singular value λ_1, (12) is closely related to the RC association model

$$(13)\ \log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + \phi u^{*}_{1(i)}u^{*}_{2(j)}$$

(Goodman, 1981a). In this expression $\phi = \lambda_1$, $r_{i1} = u^{*}_{1(i)}$ and $r_{j1} = u^{*}_{2(j)}$, where $u^{*}_{1(i)}$ and $u^{*}_{2(j)}$ are normalized in the same way as $r_{i\alpha}$ and $c_{j\alpha}$. The difference between (12) and (13) is that (12) is an approximation in k factors, whereas (13) has only one factor. Compared to saturated model (10), model (13) can be interpreted as a model allowing for interaction, where

$$(14)\ u_{12(ij)} = \phi u^{*}_{1(i)}u^{*}_{2(j)}.$$

A multi factor RC association model is introduced in Goodman (1985). For more details and related material we refer to Goodman (1985, 1986). For loglinear models with restrictive interactions in higher way tables we refer to Clogg (1982), De Leeuw (1983), Agresti (1984), and Pannekoek (1985). All loglinear models, with or without restrictions on the interactions, are easily estimated by means of the program GLIM. In Breen (1984, 1985) a procedure is described for estimating interactions restricted as (14) using GLIM.

Concluding, this gives us a better understanding of CA again. From the above we can conclude that in CA the interaction is decomposed approximately in a log-multiplicative way: the graphical CA displays show approximations of log-multiplicative parameters. Secondly, we can use these results to come to a further integration of CA and LLA, by recognizing that CA can give indications for restricting the (unrestricted) loglinear interaction parameters. This adds a third step to the two steps discussed in section 3.1: after the second step, in which CA was used to explore the interactions, we can fit models with restrictions on interaction parameters that are easy to interpret. Such models have as an obvious advantage above CA that they can be tested - possibly providing justification for the interpretation of the CA solution. In Worsley (1987), Van der Heijden (1987) and Van der Heijden & Worsley (in press) examples are given for three-way tables. Overfitting the data might be checked by using cross-validation procedures, such as described in Bonett & Bentler (1983).

3.3 Correspondence analysis of incomplete tables

Now we will use the generalization of CA to study the departure from independence in incomplete tables (compare Van der Heijden, 1985, 1987; De Leeuw & Van der Heijden, in press). The loglinear models used in this context are quasi-independence models. We will consider incomplete two-way tables only, although our results generalize to independence models for higher way tables discussed in section 3.2 such as [1][23], [2][13] and [3][12], or more generally, models [A][B] for groups of variables A and B. Quasi-independence models are useful when some entries in the table are missing (e.g. not known for some years), structurally zero, or uninteresting for some other reason.

The mathematical form of the model is very simple. Compared to saturated model (10), this model defines independence for only a limited number of cells, collected in S, while the values for the cells (i,j) not in S are not restricted:

(15) $u_{12(ij)} = 0$ when (i,j) is in S,

$\pi_{ij} = p_{ij}$ when (i,j) is not in S.

When the departure from model (15) is significant, it makes sense to study the residuals from (15) using the generalization of CA.

However, starting from the generalization of CA as defined in equation (9), we still have several possibilities. It will be clear that for Q we take the matrix of quasi-independent proportions. For S_r and S_c, the diagonal matrices of weights, we first reformulate π_{ij} in (15) as $\pi_{ij} = a_i b_j$ if cell (i,j) is in S. If we take a_i and b_j as diagonal elements of S_r and S_c, respectively, then the sum of squares of the singular values becomes

$$(16)\ \mathrm{tr}\ \Lambda^2 = \Sigma_i \Sigma_j \frac{(p_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}} = \mathbf{X}^2/n$$

This property also holds for ordinary CA (compare (7) and (8)), but was lost for the generalization in its most general form. Here each dimension can again be interpreted as a decomposition of part of the $\mathbf{X}^2$-statistic, which is the criterion used to decide whether the residuals contain interesting information.

Further insight in this choice of S_r and S_c can be derived from the fact that it leads to a CA procedure for dealing with missing values. This procedure is proposed by Nora (1975; see also Greenacre, 1984, pp. 236-244). First choose the dimensionality h. Then so-called "reconstitution of order h" is the iterative process

$$(17)\ p_{ij}^{(m+1)} = p_{i+}^{(m)} p_{+j}^{(m)} (1 + \Sigma_{\alpha=1}^{h} \lambda_\alpha^{(m)} r_{i\alpha}^{(m)} c_{j\alpha}^{(m)})$$

which is applied to cells (i,j) not in S (compare formula (6)). For (i,j) in S we simply set $p_{ij}^{(m)} = p_{ij}$ for all m. The solution will, in general, depend upon the choice of the dimensionality h. Iterative reconstitution of order zero simplifies to

$$(18)\ p_{ij}^{(m+1)} = p_{i+}^{(m)} p_{+j}^{(m)}$$

for all (i,j) not in S. This is exactly identical to our form of CA of incomplete tables. Iterating (18) converges to "independent" values

$a_i b_j$, computed for the cells not in S. For these cells, fitting the independence model on $p_{ij}^{(m)}$ gives residuals equal to zero.

We can also derive from the above that CA of incomplete tables can be computed using classical CA programs. We just have to fill in values computed with (18). This makes this usage of the generalization easily interpretable: it comes down to ordinary CA of the matrix $P^{(m)}$. An example will be given in section 4.2.

We conclude that by introducing the quasi-independence model in the CA approach we have broadened the scope of data analysis for CA. Furthermore a better understanding is obtained of the CA approach to do reconstitution of order zero.

3.4 Conclusions

In this chapter we have introduced a residual analysis approach of CA. First of all, we have seen that CA of higher-way tables by means of multiple tables can be interpreted in terms of loglinear models. Secondly, the scores obtained with CA are often approximately the same as parameters from the RC association model. Thirdly, the CA procedure for treating missing cells is closely related to the quasi-independence model.

These findings can be used, of course, by performing LLA and CA in a complementary way: LLA is used to answer the question of which relations (parameters) in a contingency table are important, while CA can show the precise form of these relations. In this way one benefits from the strong aspects of each of the two techniques: an advantage of LLA is that it gives a precise definition of what is meant by the presence or absence of interaction between variables, and an advantage of CA is that its properties can be used for an exploration of how categories of the interacting variables are related. We also have shown how CA can be used as a step in finding parsimonious loglinear models. We come back to these points in the discussion (section 5).

Below we will illustrate a combined approach of data analysis for square tables, using both LLA and CA. Furthermore, some results will be presented regarding the analysis of asymmetry using CA.

4. Example for a square matrix, with some new results

Usually, square matrices are analyzed by means of loglinear models. Here we will discuss some of the more important models, and see what CA has to offer for the analysis of residuals of these models. The models are the independence model, the quasi-independence model, the symmetry model, and the quasi-symmetry model (see Bishop et al., 1975). For all of these, the complementary CA's have special properties, either in the computations, in the solutions, or in both (see also Van der Heijden, 1987).

4.1 Decomposing residuals from independence

A starting point in the analysis of square matrices is often the comparison of observed proportions and proportions expected under the assumption that the independence model holds. Compared to the saturated model (10), the independence model restricts

$$(19) \quad u_{12(ij)}=0$$

and estimates of expected proportions are computed as $\hat{\pi}_{ij} = p_{i+}p_{+j}$. The difference between p_{ij} and $\hat{\pi}_{ij}$ can be tested with (I-1)(J-1) df. When (19) fits badly (which is usually the case), either a less restrictive model can be fitted to all cells, or residuals for individual cells can be studied for significant departure from (19). We will now study the residuals of this model using CA.

<u>Example</u>

The example is taken from Harshman et al. (1982), and deals with car switching. In 1979 recent new car buyers were surveyed to collect information on their old and new car. The cars are categorized in 16

Table 1: 1979 Car-switching data. Rows are cars disposed of, columns are new cars. Labels are described in table 2.

	SUBD	SUBC	SUBI	SMAD	SMAC	SMAI	COML	COMM	COMI	MIDD	MIDI	MIDS	STDL	STDM	LUXD	LUXI	Totals
SUBD	23272	1487	10501	18994	49	2319	12349	4061	545	12622	481	16329	4253	2370	949	127	110708
SUBC	3254	1114	3014	2656	23	551	959	894	223	1672	223	2012	926	540	246	37	18344
SUBI	11344	1214	25986	9803	47	5400	3262	1353	2257	5195	1307	8347	2308	1611	1071	288	80793
SMAD	11740	1192	11149	38434	69	4880	6047	2335	931	8503	1177	23898	3238	4422	4114	410	122539
SMAC	47	6	0	117	4	0	0	49	0	110	0	10	0	0	0	0	343
SMAI	1772	217	3622	3453	16	5249	1113	313	738	1631	1070	4937	338	901	1310	459	27139
COML	18441	1866	12154	15237	65	1626	27137	6182	835	20909	566	15342	9728	3610	910	170	134778
COMM	10359	693	5841	6368	40	610	6223	7469	564	9620	435	9731	3601	5498	764	85	67919
COMI	2613	481	6981	1853	10	1023	1305	632	1536	2738	1005	990	454	991	543	127	23282
MIDD	33012	2323	22029	29623	110	4193	20997	12155	2533	53002	2140	61350	28006	33913	9808	706	315900
MIDI	1293	114	2844	1242	5	772	1507	452	565	3820	3059	2357	589	1052	871	595	21137
MIDS	12981	981	8271	18908	97	3444	3693	1748	935	11551	1314	56025	10959	18688	12541	578	162714
STDL	27816	1890	12980	15993	34	1323	18928	5836	1182	28342	938	37380	67964	28881	6585	300	256354
STDM	17293	1291	11243	11457	41	1862	7731	6178	1288	20942	1048	30189	15318	81808	21974	548	230211
LUXD	3733	430	4647	5913	6	622	1652	1044	476	3068	829	8571	2964	9187	63509	1585	108236
LUXI	105	40	997	603	0	341	75	55	176	151	589	758	158	756	1234	3124	9162
Total	179075	15339	142259	180654	616	34215	112978	50756	14784	183858	16199	278226	150804	194228	126429	9139	1689559

Table 2: Description labels in tabel 2 (a more elaborated description can be found in Harshman et. al., 1982).

1. SUBD: Subcompact/Domestic
2. SUBC: Subcompact/Captive Imports
3. SUBI: Subcompact/Imports
4. SMAD: Small Specialty/Domestic
5. SMAC: Small Specialty/Captive Imports
6. SMAI: Small Specialty/Imports
7. COML: Low Price Compacts
8. COMM: Medium Price Compacts
9. COMI: Import Compact
10. MIDD: Midsize Domestic
11. MIDI: Midsize Imports
12. MIDS: Midsize Specialty
13. STDL: Low Price Standard
14. STDM: Medium Price Standard
15. LUXD: Luxury Domestic
16. LUXI: Luxury Import

segments, yielding a transition matrix of 16 x 16 (see table 1). The abbreviation of the segments is explained in table 2. For more details we refer to Harshman et al. (1982). Note that the row and column margins of the same cars are sometimes very different, indicating a loss or gain in share of the market.

Testing for independence, we find X^2 = 1,357,827, with 225 df. This exorbitant value is due to the extreme sample size. It shows us that CA can be done in order to find structure within the residuals from independence.

The CA solution has as first five singular values (with proportions of chi-square) .542 (.366), .390 (.189), .332 (.137), .273 (.093) and .232 (.067). A plot of the first two dimensions is shown in figure 1. The cars disposed of have large labels, the new cars small labels. Clearly, the CA is heavily dominated by the diagonal elements of the table (corresponding large and small labels are often very near). A special position of the luxury cars can be noticed: people having a LUXD buy much more than average again a LUXD (contributions to the

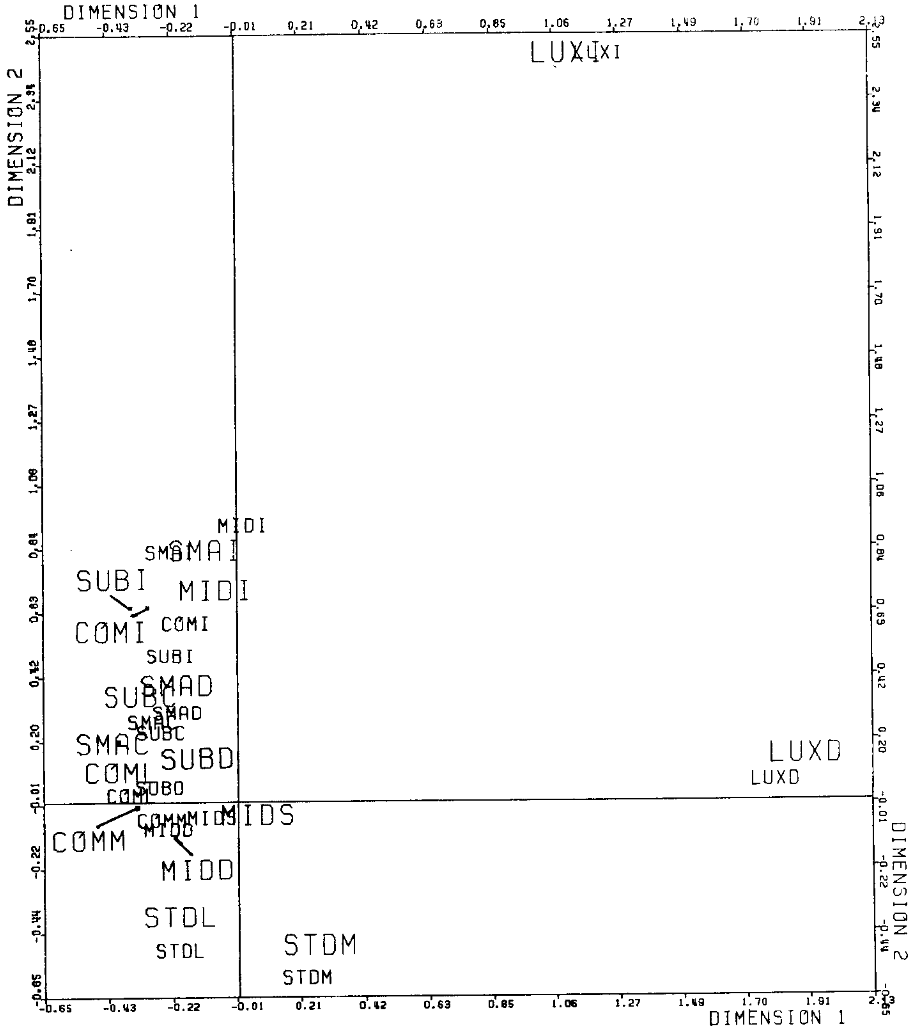

Figure 1: Ordinary correspondence analysis of car-switching data. Large labels are cars disposed of, small labels are new cars. Labels are explained in table 2.

first singular value about .8), and the same holds for LUXI. This tendency to stick to the type of luxury car one already has, overshadows information on the relations between other cars. We can see, however, on the second dimension a line of points, going from standard cars (STDL and STDM), via domestic midsize cars, domestic compact cars, domestic small specialties and domestic subcompacts to the import cars. This shows that import cars are interchanged more often than average with subcompacts and small specialties, which in turn are interchanged more often than average with compacts and midsize cars. These are interchanged more often than average with standards. The average is the profile of the total sample (the margin of the table). The first two dimensions show the most important aspects of the residuals. In the next section these data will be discussed again, with the attention restricted to the off-diagonal values.

4.2 Decomposing residuals from quasi-independence

Another independence model with which transition matrices are often analyzed, is the quasi-independence model (see section 3.3). In the context of square matrices this model can be useful when substantive interest is for some reason restricted to the off-diagonal cells. Because $\hat{\pi}_{ii}=p_{ii}$ in this context, the residuals are zero for these cells. Therefore, diagonal cells do not contribute to the total inertia $\Sigma_\alpha \lambda_\alpha^2$ (see (16)). This is an elegant way to remove the usual dominating influence of the diagonal elements in classical CA. Section 3.3 shows that we can interpret generalized CA in this context as an ordinary CA of the matrix, where independent values are imputed for the diagonal cells.

<u>Example</u>

In section 4.1 we found that the luxury cars dominated the solution because of their high diagonal proportions. Using CA of incomplete tables as specified above, the diagonal proportions do not influence the solution anymore. The quasi-independence model fits much better than the independence model: $\mathbf{X^2}$ = 235,914 compared with $\mathbf{X^2}$ = 1,357,827.

The first two dimensions of CA are shown in figure 2. Larger labels denote the cars disposed of. The first five singular values are .250 (.361), .206 (.245), .139 (.112), .126 (.091) and .090 (.047). We have drawn clusters around the corresponding car types, in order to make the main structure better discernible. Note that figure 2 may only be interpreted in terms of car-type changes.

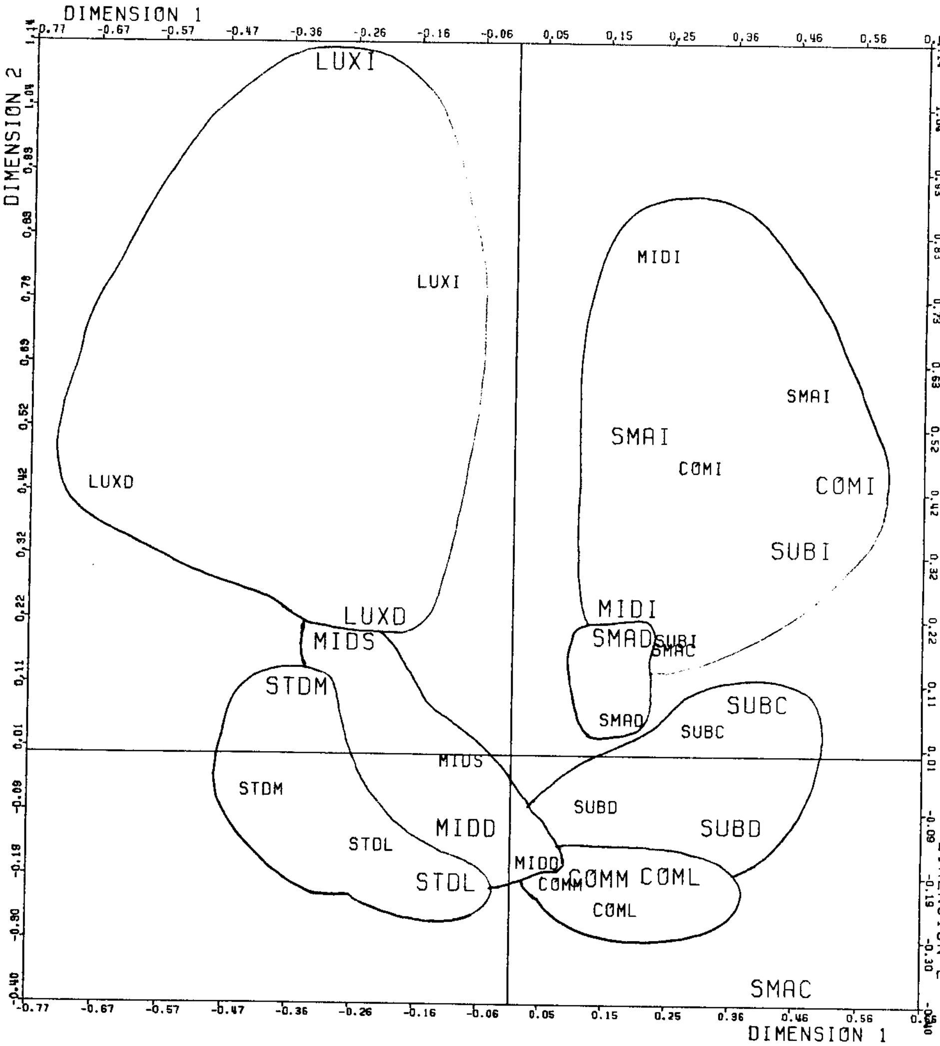

Figure 2: Correspondence analysis of incomplete table (diagonal cells ignored by decomposing residuals from quasi-independence). Large labels are cars disposed of, small labels are new cars.

Going from left to right we find a rough order from expensive to cheap: luxury, standard, midsize, compact/small, subcompact. The second dimension discriminates the import cars from the domestic cars. The location of the row point for SMAC, an import type, is somewhat peculiar in this plot, namely at the bottom right. The explanation is that, when we restrict attention to SMAC owners who change car type, SMAC is followed more than average by SUBD, SMAD, COMM and MIDD (small labels), which all are in the lower part of the plot. On the whole, it seems that, when people buy a new car of a different type, they remain more than average in the same cluster, or go to a neighboring cluster. Although the plot reconstitutes only 61% of the scaled residuals, it suggests to try to fit models incorporating symmetries in the data.

4.3 Decomposing residuals from symmetry

Recently much attention has been given to asymmetry in square matrices. An important line of contributions restricts attention to the skew-symmetric part of the asymmetric matrix. This skew-symmetric part can be found as

$$(20)\quad P = M + N$$

where P is the asymmetric square matrix of observed proportions, M is a symmetric matrix with values

$$(21)\quad m_{ij} = m_{ji} = (p_{ij} + p_{ji})/2,$$

and N has elements $n_{ij}=-n_{ji}=(p_{ij}-p_{ji})/2$, and is a skew-symmetric matrix. A justification for decomposition (20) is that values m_{ij} are equal to ML-estimates for the symmetry model

$$(22)\quad \log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

where $u_{1(i)} = u_{2(i)}$ and $u_{12(ij)} = u_{12(ji)}$ (compare the saturated model (10)). So N is a matrix of residuals. The difference between p_{ij} and π_{ij} can be tested with I(I-1)/2 df.

The attention for skew-symmetry is due to Gower (1977) and Constantine & Gower (1978), who show that the SVD of N has the form:

$$(23)\quad N = U\Lambda V' = U\Lambda JU',$$

where J is a block diagonal matrix with blocks $\begin{matrix} 0 & 1 \\ -1 & 0 \end{matrix}$, and the singular values are ordered in pairs: λ_1, λ_1, λ_2, λ_2, ... If the number of rows I of N is odd, $\lambda_I=0$ and $J_I=0$. Thus, when we make a plot of singular vectors U and UJ for paired dimensions, the configuration of row points is the same as the configuration of column points, up to a rotation of 90^o. A parsimonious plot can be made by showing only the row points, and giving the direction of the rotation.

When rotations from row to column points are made clockwise, Gower (1977) and Constantine & Gower (1978) show that in the plot of row points elements n_{ij} and n_{ji} are approximated by twice the area of triangle of row i, row j, and the origin (this triangle will further be noted as OIJ). Element n_{ij} is positive if the rotation from i to j is counterclockwise, n_{ij} is negative when the rotation is clockwise. (These relations hold approximately when a paired dimension gives only an approximation of N.) Points on a line through the origin have no asymmetric relation because area OIJ is zero. Points near the origin have small asymmetries. Points on a line not going through the origin form an additive scale because the areas of the triangles add up: for example when the order on the line is I, J, K, we find OIJ + OJK = OIK. Points on different sides from the line OI have opposed signs in fitted skew-symmetries. Finally, if all points are at one side of the paired dimension (compared with the origin), circular triads are absent in the data.

The decomposition of skew-symmetry has the following relation with CA. Decomposition (23) is a special form of the generalization of CA (compare equation (9)): when we take Q = M, then N=(P-Q), and $S_r=S_c=I$. We must note, though, that it is rather unusual not to weight individual rows and columns in the CA approach - the latter happens when we multiply N with identity matrices. So in fact weights for the residuals are sought for. Furthermore, in order to keep the properties of the SVD of skewsymmetry, the weights for the rows have to be equal to those for the columns: $S_r=S_c=S$. This can be done by taking $\pi_{i+}=\pi_{+i}$ as the elements of S. Matrix $S^{-\frac{1}{2}}NS^{-\frac{1}{2}}$ is skew-symmetric, and therefore the SVD of this matrix shows the properties discussed above. The CA decomposition becomes

$$(24)\quad P = Q + SR\Lambda JR'S$$

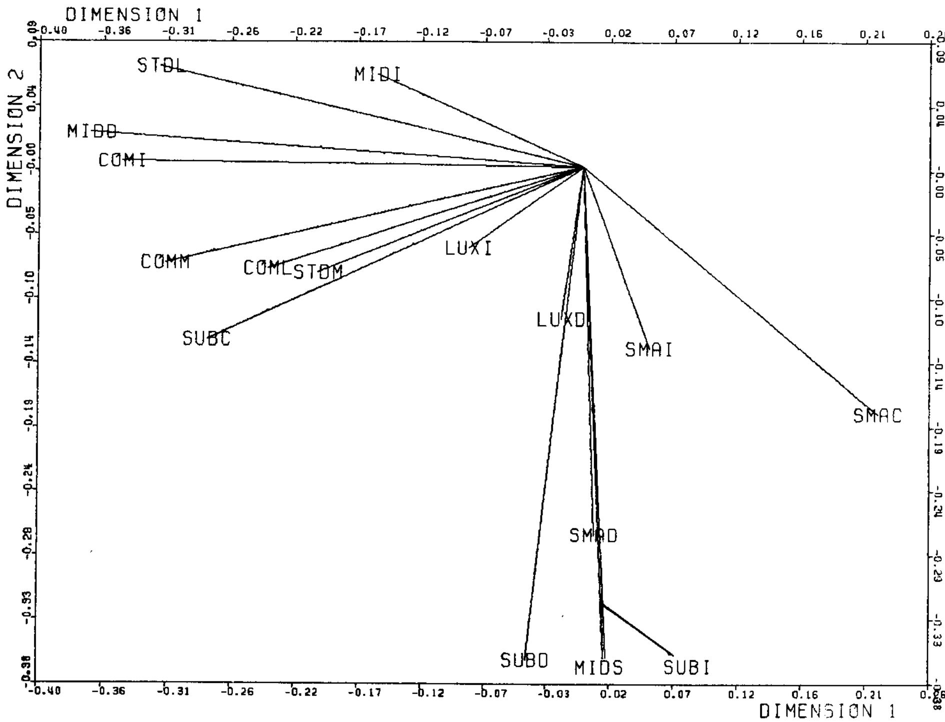

Figure 3: Generalized correspondence analysis decomposing residuals from symmetry. Going counter-clockwise, the (approximated) residuals are positive.

However, since the margins of P are not equal to those of Q, we may find the situation that all points are on one side of the space, showing that circular triads are absent.

Example

Testing for symmetry in the car data gives X^2=213,837, with 120 df. It follows that a study of the residuals from symmetry might reveal interesting information. The first two dimensions have singular values .217, and display 84% of the total inertia. The restrictions for model (22) show that in principle two types of asymmetry can be found, namely asymmetry due to the margins, and asymmetry not due to the margins. Figure 3 shows mainly asymmetry due to the margins. All car types are placed at one side of the origin, showing that (in the first two dimensions) circular triads are absent. This implies that it is possible to reorder the rows and columns of N in such a way that the upper-triangular matrix contains only positive elements, and the lower-triangular matrix only negative. The order to be taken is the order going clockwise from SMAC to MIDI in figure 3.

Going counter-clockwise from MIDI to SMAC we find cars that lost relatively much of their market share, as contrasted to cars that won (this can be easily checked in table 1 by inspecting the margins). So people having a MIDI buy more often another car than the other way around. People having a SUBD more often buy a small specialty, a SUBI or a MIDS, than the other way around. They buy less often a luxury car, a standard car (STDM and STDL) than the other way around, etcetera. People having another car than SMAC, more often buy a SMAC than vice versa.

4.4 Decomposing residuals from quasi-symmetry

In the residuals from the symmetry model the asymmetry is partly due to unequal margins of P, and partly due to other aspects than unequal margins. The quasi-symmetry model (Caussinus, 1965; Bishop et al., 1975) is a model with which it is possible to study whether the latter asymmetry, not due to different margins, can be neglected. This model can be formulated as

$$(25)\ \log \pi_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)},$$

where $u_{12(ij)}=u_{12(ji)}$ (compare saturated model (10)). The model has (I-1)(I-2)/2 df. Compared with symmetry model (22) the restriction $u_{1(i)}=u_{2(i)}$ is dropped. If $p_{i+}=p_{+i}$, the quasi-symmetry model simplifies to the symmetry model.

It will be clear that the matrix Π, following model (25), is not necessarily symmetric. However, the residual matrix (P-M)=N is skew-symmetric. We want to study the residual matrix with CA, while also retaining the properties of an SVD of skew-symmetry. As for the residuals of symmetry, this can be accomplished by taking the same set of weights for the rows and the columns, such as taking again $(p_{i+}+p_{+i}/2)$ as elements of S. Now we have the same situation as in the

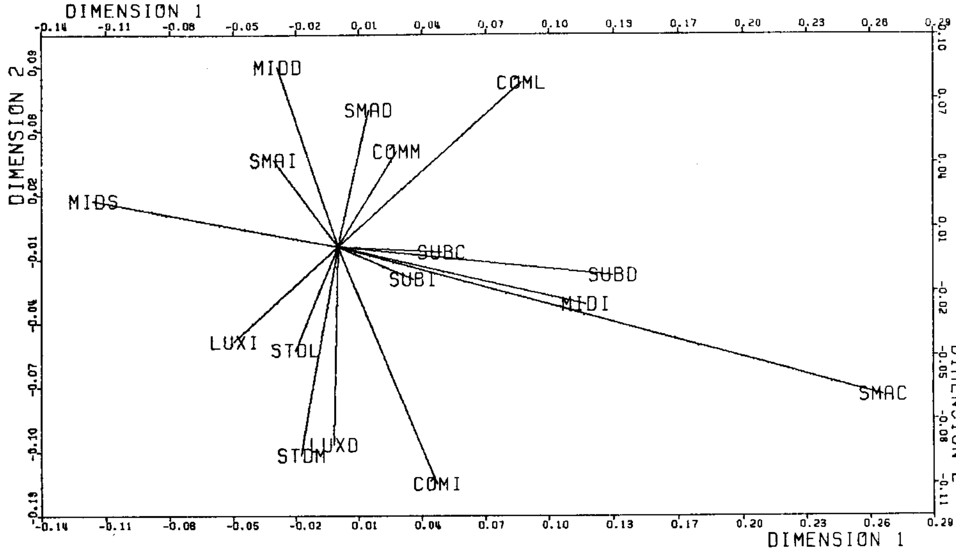

Figure 4: Generalized correspondence analysis decomposing residuals from quasi-symmetry. Going counter-clockwise, the (approximated) residuals are positive.

foregoing section. The reconstitution formula can be written as in (24), with the difference that M=Q is not symmetric.

Example

Here we will discuss asymmetry in the car switching data, to the extent that this asymmetry is not a result of a higher or lower market share. Compared with figure 3, asymmetry due to the margin is suppressed from the solution. Compared with figure 2 symmetric aspects are suppressed from the solution. In order to investigate whether the part of the asymmetry that is not due to the margins is still significant, we fitted the quasi-symmetry model to the data. This model had X^2=35,048, with 105 df. This departure is still considerable.

The first pair of dimensions has singular values .068 (.548), the second pair .048 (.270), and the third pair .029 (.102). The first pair of dimensions is shown in figure 4. Restricting attention to points with a contribution to the singular values of more than 5%, only domestic types of cars are left. Because the residuals are positive when we go counter-clockwise, we see that, corrected for marginal frequencies, SUBD is followed more often by COML and MIDD than the other way around, and the same for COML to MIDD and MIDS, MIDD to MIDS, MIDS to STDM/LUXD and STDM/LUXD to SUBD. Considering the price of these types, there seems to be a tendency of 'upward mobility' (buyers going from SUBD, COML, MIDD, MIDS to STDM/LUXD), with a sudden downward mobility from STDM/LUXD to SUBD. The order STDM/LUXD to SUBC might be interpreted as 'upward mobility' to the extent that it reflects that people buy a cheap second car. Note that the SMAC type of cars has very large areas with other cars, but because of its lower margins, it does not have a large influence to the determination of the factorial axes. The second pair of dimensions is determined by asymmetries between the import types of cars ordered as SUBI via SMAI to MIDI. In conclusion, we may say that there are asymmetric relations mainly in two groups: between domestic cars and between import cars.

4.5 Conclusion

We have shown how CA can be used for the analysis of square matrices. It can help to find relations between cells for which the residuals differ significantly from the selected model. CA can also provide a clear view of the important asymmetries in the square matrix.

Implicit in our discussion above we have introduced the notion to suppress information from a CA solution. Starting with the decomposition of residuals from independence (figure 1), we did first suppress the information about the diagonal by using the quasi-independence model, thus concentrating attention on the off-diagonal cells (figure 2). The information in the off-diagonal cells was already contained in the CA decomposing residuals from independence, but not optimally shown on the first dimensions. Going from quasi-independence to quasi-symmetry, we suppress the symmetric information from the solution, thus having an optimal view of the asymmetry (that is not due to the margins; figure 4). So we have the line of nested models independence - quasi-independence -

quasi-symmetry. Another line is the line symmetry - quasi-symmetry. Decomposing residuals from the symmetry model displays both asymmetry due to the margins and asymmetry not due to the margins (figure 3). By decomposing residuals from quasi-symmetry we have suppressed the asymmetry due to the margins from the solution, thus having an optimal view of the asymmetry not due to the margins (figure 4).

5. General discussion and conclusions

Our approach was to sketch a residual analysis interpretation of CA. Of course, it is important to use CA only in those cases that the residuals contain meaningful information: the loglinear model under study should not fit too well.

The main advantage of a residual analysis interpretation is two-fold. First of all, we obtain a better understanding of CA by relating it to LLA. This holds for the results displayed in section 3 mainly. Secondly, we have shown how CA can be used for the analysis of residuals of various loglinear models. In this respect CA might replace the cumbersome interpretation of a large amount of parameters. LLA is used for fitting and testing of specific models; CA is used for constructing meaningful geometrical representations of specific aspects of the data. So CA can be used both as in intermediate step for finding better fitting loglinear models, and as an end-point of the analysis.

We think that in the context of CA, loglinear models can play the following roles (compare de Falguerolles & van der Heijden, 1987). Firstly, some LLA model may have a central place in the analysis, and CA is used to supplement LLA by an analysis of the residuals of this model. Thus CA will only be useful if the departure from the model is significant, and if it makes sense to try to study structure in the residuals. In general, this makes sense only when there are many cells for which the observed values depart from the expected ones. CA will not be very useful if there is only one outlying cell. It will neither be useful if the total number of cells is small: in this case the structure in the departure might as well be studied directly (i.e. without CA). Secondly, CA makes it possible to suppress the dominant "uninteresting" interaction. This means we are not interested in parameters of the model, but only in residuals. Thirdly, CA can be used as an exploratory tool in order to find a loglinear model that is sparse and easily interpretable.

As was discussed in the introduction, Goodman's results indicate in what way ordinary CA is related to model-based approaches, by introducing the RC association and RC correlation models (see Goodman, 1985, 1986). Relations between the generalized CA decompositions and models are not yet known, but we think that this is a subject worth investigating.

REFERENCES

Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley.

Andersen, E.B. (1980). Discrete statistical models with social science applications. Amsterdam: North Holland.

Benzécri, J.P. et collaborateurs (1973). Analyse des données (Vols. 1-2). Paris: Dunod.

Benzécri, J.-P., et collaborateurs (1980). Pratique de l'analyse des données (Vols. 1-3). Paris: Dunod.

Benzécri, J.-P., et collaborateurs (1986). Pratique de l'analyse des données en économie (Vol. 5). Paris: Dunod.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). Discrete Multivariate Analysis. Cambridge: M.I.T. Press

Bonett, D.G. & Bentler, P.M. (1983). Goodness-of-fit procedures for the evaluation and selection of log-linear models. Psychological Bulletin, 93, 149-166.

Breen, R. (1984). Fitting non-hierarchical association loglinear models using GLIM. Sociological Methods and Research, 13, 77-107.

Breen, R. (1985). Log-multiplicative models for contingency tables using GLIM. GLIM Newsletter, 10.

Caussinus, H. (1965). Contribution a l'analyse statistique des tableaux de corrélation. Annales de la Faculté des Sciences de l'Université de Toulouse, 29, 77-182.

Clogg, C.C. (1982). Some models for the analysis of association in multi-way cross-classifications having ordered categories. Journal of the American Statistical Association, 77, 803-815.

Constantine, A.G. & Gower, J.C. (1978). Graphical representation of asymmetric matrices. Applied Statistics, 27, 297-304.

Escofier, B. (1983). Analyse de la difference entre deux mesures sur le produit de deux mêmes ensembles. Cahiers de l'Analyse des Données, 8, 325-329.

Escofier, B. (1984). Analyse factorielle en reference a un modèle: application a l'analyse de tableaux d'échanges. Revue de Statistique Appliquée, 32 (25-36).

Escoufier, Y. (1982). L'analyse des tableaux de contingence simples et multiples. Rivista Internazionale di Statistica (Metron), 40, 53-77.

Escoufier, Y. (1985). L'analyse des correspondances: ses propriétés et ses extensions. Bulletin of the International Statistical Institute, Proceedings of the 45th session, Amsterdam

Falguerolles, A. de & Heijden, P.G.M. van der, (1987). Sur l'analyse factorielle des correspondances et quelques unes de ses variantes. Revue de Statistique Appliquée.

Gifi, A. (1981). Non-linear multivariate analysis. Leiden: University of Leiden, Department of Data Theory.

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

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. Journal of the American Statistical Association, 74, 537-552.

Goodman, L.A. (1981a). 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. (1981b). Association models and the bivariate normal for contingency tables with ordered categories. Biometrika, 68, 347-355.

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. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. International Statistical Review, 54, 243-309.

Gower, J.C. (1977). The analysis of asymmetry and orthogonality. In: J. Barra et al. (Eds.), Recent developments in statistics. Amsterdam: North Holland.

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

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

Healy, M.J.R. & Goldstein, H. (1976). An approach of the scaling of categorized attributes. Biometrika, 63, 219-229.

Heijden, P.G.M. van der, (1986). Transition matrices, model fitting, and correspondence analysis. In: E. Diday (Ed.), Data analysis and informatics IV. Amsterdam: North Holland.

Heijden, P.G.M. van der (1987). Correspondence analysis of longitudinal categorical data. Leiden: D.S.W.O.-Press.

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

Heijden, P.G.M. van der, & Worsley, K.J. (in press). Comment on "Correspondence analysis used complementary to loglinear analysis". Psychometrika.

Hill, M.O. (1974). Correspondence analysis: a neglected multivariate method. Applied Statistics, 23, 340-354.

Jambu, M. & Lebeaux, M.-O. (1983). Cluster analysis and data analysis. Amsterdam: North-Holland.

Lebart, L., Morineau, A. & Warwick, K.M. (1984). Multivariate descriptive statistical analysis. Correspondence analysis and related techniques for large matrices. New York: Wiley.

Leeuw, J. de (1983). Models and methods for the analysis of correlation coefficients. Journal of Econometrics, 22, 113-137.

Leeuw, J. de, & Heijden, P.G.M. van der (in press). Correspondence analysis of incomplete tables. Psychometrika.

Luce, R.D. (1963). Detection and recognition. In: R.D. Luce et al. (Eds.), Handbook of Mathematical Psychology (Vol. 1). New York: Wiley.

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

Nishisato, S. (1986). Quantification of categorical data: a bibliography 1975-1986. Toronto: Microstats.

Nora, C. (1975). Une méthode de reconstitution et d'analyse de données incomplètes. Unpublished doctoral dissertation, Université P. et M. Curie, Paris.

Pannekoek, J. (1985). Log-multiplicative models for multiway tables. Sociological Methods and Research, 14, 137-153.