Correspondence Analysis in the Social Sciences

Recent Developments and Applications

edited by

Michael Greenacre
University of South Africa
Pretoria, South Africa

and

Jörg Blasius
University of Cologne
Germany

ACADEMIC PRESS

Harcourt Brace & Company, Publishers
London San Diego New York Boston Sydney Tokyo Toronto

The values in each column are as follows:

- MASS – mass [see sections 3.2.5 and 3.2.6].
- SQCOR – squared correlation (of profiles with subspace, in this case the solution subspace is two-dimensional), or qualities [see sections 3.2.24 and 3.2.30].
- INR – inertia of each profile point [see sections 3.2.22 and 3.2.28].
- LOC1 – principal coordinate on axis 1 [see sections 3.2.16 and 3.2.17].
- QCOR1 – squared correlation (of profile with axis 1) [see sections 3.2.23 and 3.2.29].
- INR1 – proportion of inertia on axis 1 explained by each profile [see sections 3.2.21 and 3.2.27].
- LOC2, QCOR2 and INR2 – corresponding quantities for principal axis 2.

The above output is in the same format as the SimCA and BMDP output, except in SimCA all quantities are multiplied by 1000 and written as integers to save space in the tables. Notice that the column SQCOR (quality) is the sum of the columns QCOR1 and QCOR2. If the solution was three-dimensional, then SQCOR would be the sum of QCOR1, QCOR2 and QCOR3, and so on. In the full solution, with K axes, SQCOR = 1.

4

Correspondence Analysis and Contingency Table Models

Peter G. M. van der Heijden, Ab Mooijaart and Yoshio Takane

4.1 INTRODUCTION

Sometimes correspondence analysis (CA) is presented as a model-free technique (see Benzecri *et al.* 1973). The idea is that CA is helpful when we want 'to let the data speak for itself'. The idea is that no assumptions are made about the distribution which yielded the values in the matrix to be studied. No prejudices of the researcher lead the analysis of the data. CA is simply a tool to make a graphical representation of the data. The generalized singular value decomposition performed in the computation of CA is helpful in projecting most information to the first few dimensions of the full-dimensional solution.

We have some difficulty with calling correspondence analysis 'model-free'. We rather adopt another definition of a model, namely that a model is a nonlinear projection of the data on a (usually low-dimensional) parameter space (see discussion of van der Heijden *et al.* 1989; see also de Leeuw 1988). This nonlinear projection can be optimal in terms of some criterion, and often used criteria are least squares, generalized least squares, or maximum likelihood (ML). In earlier chapters CA is estimated using a singular value decomposition that has optimality properties in a least squares sense.

We prefer this definition because we find that by choosing CA to represent the data graphically an explicit choice is made to emphasize certain aspects of the data. One aspect is to emphasize the association between the row and the column variable instead of the margins. A second choice is to study the

association by using certain metrics for the row space and the column space, and certain weights for the row points and the column points. A term like 'model-free' suggests that no explicit choices are made in studying the data, and we do not agree with this. For a related discussion we refer to the discussion of the paper of van der Heijden *et al.* (1989).

This leads us to a presentation of CA as a model that can be fitted using different criteria. In our view CA as presented in earlier chapters is optimized in terms of a least-squares criterion for a transformation of the observed proportions. A different approach is to optimize CA by ML. This will be discussed in section 4.2. In section 4.3 we will show the close relation that exists between CA and the latent class model, a model that assumes that the manifest variables are independent given the level of a latent variable. It turns out that latent class analysis and CA are models that are often equivalent. In section 4.4 we describe some relations between CA and log–linear models, models having log–bilinear terms, and ideal point discriminant analysis (IPDA). In this section it is shown that CA can be interpreted as a tool for residual analysis. It is also shown that if certain conditions are fulfilled, the estimates obtained with CA will be very similar to the estimates obtained with models having log–bilinear terms. We end with a discussion about the relative merits of CA and other statistical models.

Throughout this paper the same data matrix will be analyzed with different models. This matrix is displayed in Table 4.1. The matrix is concerned with results of an election held in The Netherlands in 1986 for the so-called Second Chamber (comparable with the House of Commons). The rows of the matrix are six types of city, namely rural, industrialized rural, commuter, small, middle large and large. The political parties are subdivided into six categories, namely PvdA (labour party), CDA (christian democrats), VVD (right-winged liberals), D'66 (left-winged liberals), small left-winged parties and small right-winged parties (mostly religious). We will consider this table to be a one in a

TABLE 4.1
Voters for the 1986 elections in The Netherlands. In the rows city types are found, and in the column political parties are given. For more details, see the text.

	Political party					
City type	PvdA	CDA	VVD	D'66	Left	Right
Rural	285	482	186	49	21	60
Rural, industrialized	620	914	308	102	42	97
Commuter	355	460	347	104	36	47
Small city	336	337	168	62	27	46
Middle large city	548	455	233	91	47	43
Large city	903	516	343	153	110	37

thousand random sample of voters (see Central Bureau for Statistics 1987, for more details).

Since this book focuses on correspondence analysis approximated by least squares, we present here the results of the correspondence analysis of our table (see also van de Geer 1989). In Figure 4.1(a) and 4.1(b) the graphical representation is given of the first two dimensions of the five-dimensional space.

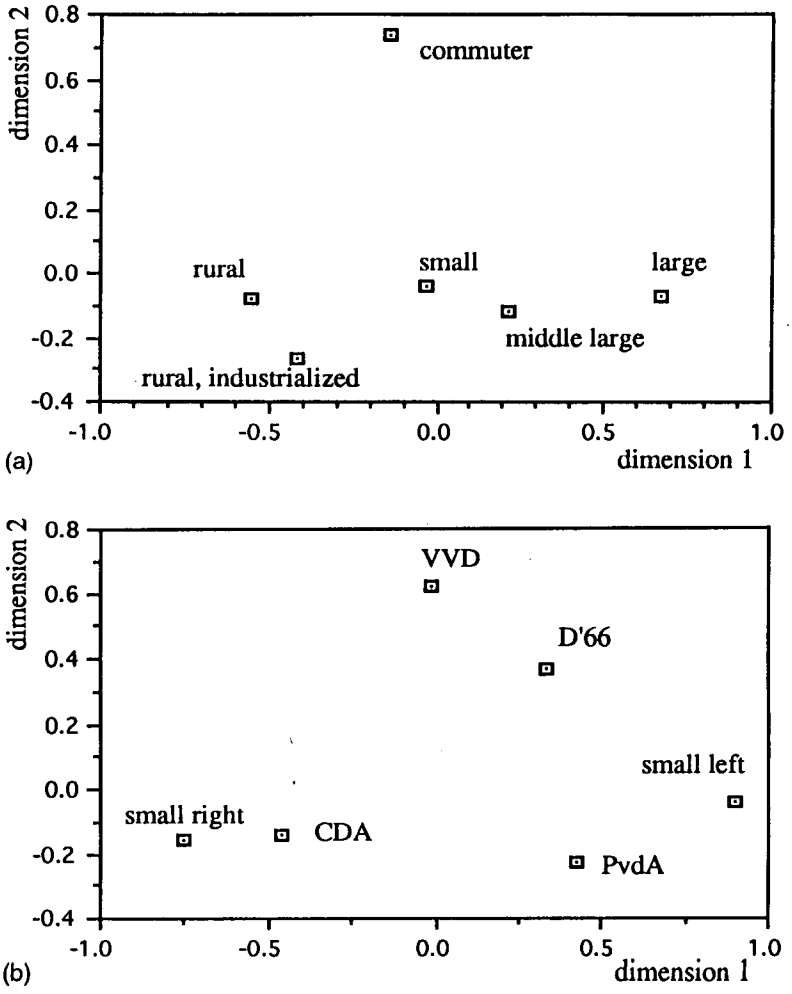

FIGURE 4.1 (a) Correspondence analysis solution in two dimensions, row points. See text for explanation. (b) Correspondence analysis solution in two dimensions, column points. See text for explanation.

The singular values are 0.19 and 0.10, displaying 0.77 and 0.22 of the total inertia. In both figures the points are represented using principal coordinates (see Greenacre, Chapter 1). Roughly, we find rural towns in the bottom left corner, larger cities in the bottom right corner, and commuters at the top. It turns out that in the rural towns people vote more than average for religious parties (small right and CDA), in large cities they vote more than average for left-winged parties, and in commuter towns they vote more than average for the VVD. Both in large cities and in commuter cities people vote more than average for D'66.

4.2 CORRESPONDENCE ANALYSIS

CA approximated by least squares

Our starting point in a comparison of CA with other contingency table models is to show how the observed proportions p_{ij} are decomposed in CA. The reason for this starting point is that statistical models are usually defined in terms of theoretical probabilities π_{ij} that are assumed to hold in some population. Thus by starting with the observed proportions, a later comparison with other statistical models becomes easier.

The reconstitution formula can be written as

$$p_{ij} = p_{i+}p_{+j}\left(1 + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk}\right)$$

where $\mu_k = \sqrt{\lambda_k}$ is the square root of the kth principal inertia and the standard coordinates r_{ik} and c_{jk} are normalized in the usual way. The graphical representations such as in Figure 4.1 are usually made with principal coordinates, which are given by $f_{ik} = r_{ik}\mu_k$ and $g_{jk} = c_{jk}\mu_k$, in which case the reconstitution formula can be written equivalently as:

$$p_{ij} = p_{i+}p_{+j}\left(1 + \sum_{k=1}^{K} \mu_k^{-1} f_{ik} g_{jk}\right)$$

In earlier chapters CA is estimated by means of the singular value decomposition of the matrix of standardized residuals with elements

$$s_{ij} = \frac{(p_{ij} - p_{i+}p_{+j})}{\sqrt{p_{i+}p_{+j}}}$$

This decomposition is optimal in terms of least squares: by using only K^* of the K dimensions, i.e. by using only the first K^* terms in the reconstitution formula we obtain a weighted least-squares approximation of the proportions p_{ij}.

It may be noted, in passing, that it is relatively straightforward to incorporate linear constraints in correspondence analysis (see Chapter 5). They

come in different guises, redundancy of qualitative data (Israëls, 1984), canonical correspondence analysis (ter Braak 1986), canonical analysis with linear constraints (Böckenholt and Böckenholt 1990). These techniques use external information to limit the space in which rows and columns of contingency tables are represented. Analogous developments in modelling approaches to contingency tables will be presented later in this chapter.

As we said before, the correspondence analysis solution discussed in section 4.1 for Table 4.1 is presented with principal coordinates f_{ik} and g_{jk}. So in Figure 4.1(a) for row point i on dimension k, f_{ik} is used as coordinate, and in Figure 4.1(b) for column point j on dimension k, g_{jk} is used as coordinate. In the tradition to see correspondence analysis as a tool to represent the matrix graphically (see section 4.1) this is quite useful, because thus the Euclidean distances in Figures 4.1(a) and 4.1(b) are approximations of the so-called chi-squared distances between the rows (columns) of the matrix (see the earlier chapters). In the modelling tradition it is more standard to show the estimates for the parameters r_{ik}, c_{jk} and μ_k. These estimates are given in panel A of Table 4.2. We will now compare these estimates with estimates of models approximated by ML.

CA approximated by ML

Correspondence analysis approximated using least squares has quite a long history. A much shorter tradition exists since 1985, when Goodman (1985, 1986) started studying CA as a model to be estimated by ML (see also Gilula and Haberman 1986, 1988). In this tradition there are population probabilities π_{ij} that follow a statistical model, in this case CA:

$$\pi_{ij} = \alpha_i \beta_j \left(1 + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk}\right)$$

Under the assumption that the frequencies follow a (product) multinomial or a Poisson distribution, a likelihood function is set up that is maximized over the unknown parameters α_i, β_j, μ_k, r_{ik} and c_{jk}. If the parameter estimates r_{ik} and c_{jk} are unconstrained, then it turns out that $\alpha_i = p_{i+}$ and $\beta_j = p_{+j}$ (see Goodman 1985, Siciliano *et al.* 1990).

The CA model defined above is not restrictive because $K = \min(I-1, J-1)$. Therefore the CA estimated by least squares and the CA estimated by ML are identical for $K = \min(I-1, J-1)$. The model becomes restrictive if either restrictions are imposed on the parameters r_{ik} and/or c_{jk}, or on the dimensionality of the solution. For example, in the model

$$\pi_{ij} = \alpha_i \beta_j \left(1 + \sum_{k=1}^{K^*} \mu_k r_{ik} c_{jk}\right)$$

where $0 \leqslant K^* < K$, for K^* dimensions $\mu_k \geqslant 0$. If this model fits adequately, then it is not necessary to study more than K^* dimensions. It is also possible

TABLE 4.2

Parameter estimates for the data in Table 4.1. In panel A we find the estimates for correspondence analysis approximated by least squares; in panel B estimates for correspondence analysis approximated by maximum likelihood; in panel C estimates for the RC(2) association model; in panel D estimates for the latent budget model with three latent budgets; in panel E estimates for ideal point discriminant analysis; in panel F rescaled estimates for ideal point discriminant analysis (see text).

	Panel A		Panel B		Panel C		Panel D			Panel E		Panel F	
	0.191	0.102	0.190	0.102	0.194	0.093	0.458	0.294	0.248			0.097	0.046
Rows													
1	−1.274	−0.241	−1.262	−0.248	−1.272	−0.094	0.659	0.103	0.238	−0.917	−0.055	−1.287	−0.082
2	−0.955	−0.825	−0.964	−0.824	−0.937	−0.883	0.627	0.163	0.210	−0.665	−0.593	−0.933	−0.881
3	−0.317	2.304	−0.321	2.305	−0.339	2.272	0.450	0.188	0.362	−0.242	1.529	−0.340	2.271
4	−0.067	−0.121	−0.066	−0.125	−0.076	−0.149	0.471	0.286	0.242	−0.049	−0.106	−0.069	−0.157
5	0.501	−0.375	0.514	−0.366	0.483	−0.458	0.388	0.383	0.229	0.350	−0.319	0.491	−0.474
6	1.528	−0.232	1.525	−0.234	1.541	−0.160	0.228	0.537	0.234	1.094	−0.103	1.535	−0.153
Columns													
1	0.973	−0.704	0.984	−0.703	0.957	−0.790	0.295	0.697	0.000	0.133	−0.047	0.980	−0.689
2	−1.059	−0.441	−1.058	−0.437	−1.059	−0.398	0.623	0.164	0.078	−0.143	−0.031	−1.053	−0.455
3	−0.037	1.945	−0.037	1.949	0.015	1.875	0.000	0.000	0.711	−0.007	0.133	−0.052	1.951
4	0.772	1.167	0.784	1.150	0.810	1.297	0.000	0.062	0.179	0.104	0.079	0.766	1.159
5	2.070	−0.111	1.959	−0.114	2.004	0.248	0.002	0.077	0.032	0.281	−0.001	2.070	−0.015
6	−1.732	−0.493	−1.778	−0.536	−1.848	−0.320	0.080	0.000	0.000	−0.235	−0.034	−1.731	−0.499

to test whether some parameters are fixed, or equal to other parameters. For example, if the parameters $r_{ik} = 0$ for $k = 1, \ldots, K^*$, then the point i falls into the origin, and the result is that $\pi_{ij}/\pi_{i+} = \pi_{+j}$ for point i. As another example, if $r_{ik} = r_{i'k}$ for $k = 1, \ldots, K^*$, then the points i and i' are located in the same place. A third possible restriction that can be tested is whether the scores are equidistant, for example, whether the scores r_{ik} are $(-3, -1, 1, 3)$ if $I = 4$, or, whether $r_{ik} = c_{ik}$ if we deal with the analysis of a square table. See Gilula and Haberman (1988) for other kinds of restrictions that may be imposed.

The correspondence analysis model with constraints on the dimensionality of the solution only, has $(I - K^* - 1)(J - K^* - 1)$ degrees of freedom. Thus we can test whether one or two dimensions are sufficient to describe the association in Table 4.1. For Table 4.1 we fit the model with $K^* = 0$, $K^* = 1$ and $K^* = 2$ dimensions. For $K^* = 0$ the correspondence analysis model comes down to the independence model: $\pi_{ij} = \pi_{i+}\pi_{+j}$. This model has a likelihood ratio chi-square statistic $G^2 = 420.6$ for $(I - 1)(J - 1) = 25$ degrees of freedom. The model with one dimension only ($K^* = 1$) has a $G^2 = 95.1$ for 16 degrees of freedom. The model with two dimensions ($K^* = 2$) has a $G^2 = 6.7$ for 9 degrees of freedom.

It turns out that the model with two dimensions is not significantly different from the saturated model where $K^* = K = 5$. The estimates for r_{ik}, c_{jk} and μ_k are given in panel B of Table 4.2. It can be seen that the estimates found in the ML solution are very similar to the estimates found in the least squares solution found in panel A. In fact, for this example there is hardly any difference.

Comparison

It is much easier to estimate CA by least squares than by ML, so one should have good reasons to do the latter instead of the former. Compared with CA approximated by least squares, it is a great advantage of the version of CA estimated by ML that restrictions on the model parameters can be tested. Very natural questions can be answered about a CA solution, such as 'how many dimensions should I interpret?'; 'is the distance between two row points or two column points in my graphical display significantly different from zero?'; 'is the distance of a point to the origin significantly different from zero?'; and 'are the scores for the rows and/or the columns equidistant?'.

It should be emphasized, however, that it is only useful to perform such tests when the assumption is fulfilled that the data stem from a (product)-multinomial or a Poisson distribution. This assumption can be violated if the observations leading to the frequencies are dependent, for example, if the answers of respondents in households are related. Another case in which the usefulness of this approach is doubtful is if the data are not a random sample

from the population one had in mind by setting up the study. This often happens in surveys. If the non-response is selective in the sense that it is larger in some cells than in other cells, then the test results only tell us how to generalize from a sample to some unclear population.

Often the results from CA approximated by least squares and those from CA approximated by ML will be very similar (see, for example, Goodman 1985, 1986). We also found this for the example discussed above (see panel A and B in Table 4.2). We already indicated that if $K^* = K$, then these approximations are identical because they do not impose restrictions on the data. However, there are also situations in which they may be rather different. One such situation is when in the CA approximated by least squares

$$\left(1 + \sum_{k=1}^{K^*} \mu_k r_{ik} c_{jk}\right)$$

is negative. This might occur when, for example, $K^* = 1$ and μ_1 is large. Then for a one-dimensional solution it is most likely that there will be reconstituted proportions that are smaller than zero. In the one-dimensional CA solution approximated by ML the probabilities π_{ij} are always positive, and therefore either the estimates for μ_1 or for some parameters r_{i1} or c_{j1} will be quite different from those found in the CA approximated by least squares.

A question that remains is how the ML estimates of the scores should be interpreted. One answer is that there is no difference in interpretation between the ML estimates and the least squares estimates, but that these estimates have other properties (see the statistical literature). One could also argue as follows: when CA is performed by least squares (as explained and illustrated in Chapters 1 to 3), we have a geometric interpretation of the results in the form of profiles, masses, projections of profiles onto optimal subspaces, etc. When CA is estimated by ML a similar but different interpretation is found. The key to the solution of the interpretation problem is that now the matrix with elements $\hat{\pi}_{ij}$ is the starting point for the interpretation, and not the matrix with observed proportions p_{ij}. For the ML solution the points in a representation like Figure 4.1(a) do not represent *observed* profiles with elements p_{ij}/p_{i+}, but estimates of *expected* profiles with elements $\hat{\pi}_{ij}/\hat{\pi}_{i+}$. Masses of the points are the same, and the metric of the row and column spaces are also the same. However, whereas for the least squares solution in Table 4.2 the scores represent an optimal least squares projection to a two-dimensional subspace, for the ML solution the matrix with elements $\hat{\pi}_{ij}$ is perfectly represented in a two-dimensional space. The fit of the model, $G^2 = 6.7$ for 9 degrees of freedom in this case, helps us to evaluate whether the matrix with elements p_{ij} may be represented by the matrix with elements $\hat{\pi}_{ij}$. The principle of ML tells us that, if there is a true two-dimensional graphical display for the population, the parameter estimates in panel B of Table 4.2 make the observed data most likely. So the scores in panel B of Table 4.2 are not only useful for plots by

analogy with CA, they can be understood in terms of profile points and the like in their own right.

4.3 SOME RELATIONS BETWEEN CA, LATENT BUDGET ANALYSIS AND LATENT CLASS ANALYSIS

4.3.1 Latent budget analysis

Consider the matrix in Table 4.3. The elements of this matrix are derived from the matrix in Table 4.1 as $p_{j|i} = n_{ij}/n_{i+}$. The proportion $p_{j|i}$ is a conditional proportion: it is the proportion of voters that have voted for political party j given that they live in city type i. It is easy to study a matrix as in Table 4.3 by comparing the conditional proportions $p_{j|i}$ with the marginal proportions p_{+j}, that are given in the last line of Table 4.3. It shows that PvdA is voted more than average (i.e. 0.340) in large cities (i.e. 0.387), CDA is voted more than average in rural cities (0.445, 0.439 versus 0.353), VVD is voted more than average in commuter towns (0.257 versus 0.177), and so on. All these results were already clear from Figures 4.1(a) and 4.1(b), because CA is based on an analysis of a matrix with elements $p_{j|i}$ and a matrix with elements $p_{i|j}$. The rows of conditional proportions $p_{j|i}$ in Table 4.3 are the row profiles. The average profile is given in the last line of the table. One might ask whether there exists a small number of *typical* profiles, say three, that have *generated* the observed profiles. Thus the observed profiles would be approximated by a weighted average of three unknown profiles. The question then is: what are these unknown profiles, and how do they approximate the observed profiles.

TABLE 4.3
Voters for the 1986 elections in The Netherlands. In the rows city types are found, and in the column political parties are given. Conditional proportions: given the type of city, what is the distribution of voters for the political parties?

	Political party					
City type	PvdA	CDA	VVD	D'66	Left	Right
Rural	0.263	0.445	0.172	0.045	0.019	0.055
Rural, industrialized	0.298	0.439	0.148	0.049	0.020	0.047
Commuter	0.263	0.341	0.257	0.077	0.027	0.035
Small city	0.344	0.345	0.172	0.064	0.028	0.047
Middle large city	0.387	0.321	0.164	0.064	0.033	0.030
Large city	0.438	0.250	0.166	0.074	0.053	0.018
Total sample	0.340	0.353	0.177	0.063	0.032	0.037

This solution is displayed in the lower part of panel D of Table 4.2. Here we find as a first typical profile 0.295, 0.623, 0.000, 0.000, 0.002, 0.080, where the CDA and small right-winged parties have much higher proportions than their average (0.623 versus 0.340; 0.080 versus 0.037). We call this the religious profile. The second typical profile is 0.697, 0.164, 0.000, 0.062, 0.077 and 0.000, showing that here the PvdA and small left-winged parties are much higher than their average. We call this the left-winged profile. The third typical profile is 0.000, 0.078, 0.711, 0.179, 0.032, 0.000, showing that here the liberal parties VVD and D'66 are higher than their average. We call this the liberal profile. Notice that this is very similar to the three clusters of parties that we found in the correspondence analysis plots in Figure 4.1(b).

In the top part of panel D of Table 4.2 we find how the observed profiles are approximated by the typical profiles. On average, the proportions of voters making use of the three profiles are 0.458, 0.294 and 0.248, so 0.458 of the voters are in the religious profile, 0.294 in the left-winged profile and 0.248 in the liberal profile. The top part of panel D shows that, for example, the rural city profile is built up for 0.659 of the religious profile, for 0.103 of the left-winged profile, and for 0.238 of the liberal profile. We see that, on the whole, the profiles of rural and industrialized rural are built up more than average by the typical religious profile (0.659, 0.627 versus the average 0.458); the profiles of middle large and large more than average by the typical left-winged profile; and the profile of commuter cities by the typical liberal profile.

After this informal introduction of the model, we now introduce some notation. Let us denote theoretical profile elements as $\pi_{j|i} \equiv \pi_{ij}/\pi_{i+}$. Let us index the typical profiles to be estimated by $x (x = 1, \ldots, X)$, so that we can denote the element j of typical profile x as $\pi_{j|x}$, since $\Sigma_j \pi_{j|x} = 1$ (see bottom part of panel D). Let the probabilities that relate the rows to the typical profile be denoted by $\pi_{x|i}$, since $\Sigma_x \pi_{x|i} = 1$ (see top part of panel D). Then the model is

$$\pi_{j|i} = \sum_{x=1}^{X} \pi_{x|i} \pi_{j|x}$$

This model is called the latent budget model, and for many more details we refer to de Leeuw *et al.* (1990) and van der Heijden *et al.* (1992). Here the term 'budget' is synonymous with 'profile', and the term 'latent' corresponds to the fact that unknown, typical profiles ('budgets') are sought.

The model is usually approximated by ML. In its unconstrained form it has $(I - X)(J - X)$ degrees of freedom (for constraints we refer to van der Heijden *et al.* 1992). For one typical budget the model comes down to the independence model, having only the 'typical' budget with elements p_{+j}. We saw earlier that for this model our example gives a test statistic $G^2 = 420.6$ (df = 25). For two latent budgets (typical profiles) the model has a chi-square of $G^2 = 95.1$ (df = 16) and for three latent budgets the model has a chi-square of $G^2 = 6.7$

(df = 9). Thus the solution that we discussed above corresponds with a model that fits the data.

The latent budget model is not identified, but it can be identified by imposing fixed value constraints upon the parameters. The identification problem is similar to that in factor analysis. For the example discussed above we have imposed six constraints to the elements of the latent budgets, namely that elements 3 and 4 of latent budget 1, element 3 and 6 of latent budget 2 and elements 1 and 6 of budget 3 are all constrained to zero. For more details we refer to de Leeuw *et al.* (1990). Van der Heijden *et al.* (1992) indicate how general classes of constraints can be imposed on the parameters $\pi_{x|i}$ and $\pi_{j|x}$. These constraints are particularly useful in the analysis of higher-way tables.

4.3.2 *Equivalence of the latent class model and the latent budget model*

The latent class model is a closely related model for the analysis of contingency tables. For example, if there are three variables A, B and C, then the basic latent class model assumes the existence of a categorical latent variable X that explains the relation between the observed variables. Given the level of X the observed variables A, B and C are independent (for more details, see Goodman 1974).

Less attention is given to the latent class model for a two-way table

$$\pi_{ij} = \sum_{x=1}^{X} \pi_x \pi_{i|x} \pi_{j|x}$$

(but see Clogg 1981). One of the reasons might be that this model is not identified, whereas the latent class model for three or more variables is identified (see Mooijaart 1982, Goodman 1987, de Leeuw and van der Heijden 1991).

This model is equivalent to the latent budget model $\pi_{j|i} = \Sigma_x \pi_{x|i} \pi_{j|x}$: we can derive $\pi_{x|i}$ from the latent class parameters π_x and $\pi_{i|x}$ as

$$\pi_{x|i} = \frac{\pi_x \pi_{i|x}}{\sum_{z}^{X} \pi_z \pi_{i|z}}$$

For the practice of data analysis this implies that the latent class model with X latent classes provides the same estimates of expected frequencies as the latent budget model with X latent budgets. The parameter estimates from the latent class model can be used to obtain the estimates for the latent budget model, and vice versa. For a discussion of relative advantages of one model over the other we refer to van der Heijden *et al.* (1992).

It turns out that the latent budget model (and the latent class model) is also closely related to CA, because all models are reduced rank models for contingency tables. We follow here the presentation of de Leeuw and van der Heijden (1991), who describe this relation in terms of theoretical probabilities.

4.3.3 CA and LCA as reduced rank models

Let π_{ij} be the elements of a matrix Π with theoretical probabilities, where $\Sigma_i \Sigma_j \pi_{ij} = 1$. Then Π has rank R ($R \leqslant \min(I, J)$) if it is possible to decompose Π by $\Pi = \mathbf{X}\mathbf{Y}'$, where $\mathbf{X}$ is some matrix of order $I \times R$ and $\mathbf{Y}$ is some matrix of order $J \times R$. If $R = \min(I, J)$, then Π has full rank, and if $R < \min(I, J)$, then Π has a reduced rank R. Notice that in practice it is unlikely to encounter an observed matrix $\mathbf{P}$ having a reduced rank, that is a rank smaller than $\min(I, J)$.

We will now show that CA defining a rank R model is equivalent to the more general 'model' that some matrix has rank R. This is done by showing that both models imply each other.

If the matrix Π has rank R, then the matrix can always be decomposed by a CA with $R - 1$ dimensions (see Lancaster 1958; see also de Leeuw and van der Heijden 1991). The reverse is evident: if the matrix Π is decomposed by a CA with $R - 1$ dimensions, then the matrix Π has rank R. This can be easily seen by rewriting the reconstitution formula as

$$p_{ij} = p_{i+}p_{+j}\left(1 + \sum_{k=1}^{R-1} \mu_k r_{ik} c_{jk}\right) = p_{i+}p_{+j} + \sum_{k=1}^{R-1} (\mu_k^{1/2} p_{i+} r_{ik})(\mu_k^{1/2} p_{+j} c_{jk})$$

$$= p_{i+}p_{+j} + \sum_{k=1}^{R-1} r_{ik}^* c_{jk}^*$$

where $r_{ik}^* \equiv \mu_k^{1/2} p_{i+} r_{ik}$ and $c_{jk}^* \equiv \mu_k^{1/2} p_{+j} c_{jk}$. This shows that $p_{i+}p_{+j}$ is a rank 1 matrix, and the $R - 1$ terms defined by $r_{ik}^* c_{jk}^*$ constitute a rank $R - 1$ matrix. If we denote the matrix Π of rank R as Π_R, and CA decomposing a rank R matrix as CA_R, then we can say that Π_R and CA_R are equivalent.

From the equation for latent class analysis of two-way tables we immediately see that the latent class model is a reduced rank model. If the latent variable has R levels, then the latent class model defines a matrix of rank R. We denote this model as LCA_R.

It is evident that if LCA_R is true for some matrix Π, that for this matrix Π_R is true, and therefore CA_R is true. The reverse does not hold. The reason is that the parameters in LCA are all non-negative, whereas in CA the parameters may be both positive or negative. Therefore LCA is more restrictive than CA, and therefore if CA_R is true for some matrix Π, then it might be that LCA_R is not true. In other words, there are matrices Π of rank R for which LCA_R will not be true. However, it is obvious that for rank 1 (independence), both models are equivalent. Second, for full rank, i.e. $R = \min(I, J)$, both models are also equivalent. Third, in contrast with statements of Gilula (see, for example, Gilula and Haberman 1986), de Leeuw and van der Heijden

(1991) prove that for a matrix Π of rank 2 the latent class model is equivalent to CA.

4.3.4 Conclusions

Several conclusions follow from these results. First, although the parametrization from CA and LCA are very different, they extract similar information from the data.

Second, if both models are fit by ML (which is the usual criterion for latent class analysis but not for CA, but see section 4.3), then they will have the same fit for $R = 2$. For $R > 2$, the models CA_R and LCA_R may have the same fit (this is almost always the case in situations where we tried this out), but it may also be that CA_R has a better fit than LCA_R, because the latter is more restrictive than the former. For the example discussed in this section and in section 4.2 the both models yield equivalent results. This can be seen from the fact that, if we compare the latent budget model with the correspondence analysis approximated by ML (so that both models are optimized using the same criterion), both for rank 2 and for rank 3 both models have the same likelihood ratio chi-square. For rank 2 this is true for every possible matrix (since CA with one dimension and the latent budget model with two latent budgets are equivalent) but this need not be true for rank 3.

Third, since for rank 2 CA_R and LCA_R are equivalent, they yield the same estimates of expected frequencies. This result may be used in the ML estimation of CA: it is possible to estimate the latent class model as a first step, and then to decompose the estimates of expected frequencies with ordinary CA in order to obtain the ML estimates of the CA model with one dimension.

In situations where both models are fit by ML, and they yield the same fit, it is not possible to prefer one model over the other based on statistical considerations. However, the parameters of both models have different interpretations: the parameters of CA can be interpreted as optimal scores, whereas the parameters of LCA are conditional probabilities, and therefore a choice for either CA or the latent class model will depend on the substantive research question about the data. LCA has the problem that its solution is not identified. However, this identification problem seems to be well understood (see de Leeuw *et al.* 1990).

Goodman (1987) compares both models, and shows an empirical example. In van der Heijden *et al.* (1990) it is shown that CA and the latent class model can provide graphical representations that are closely related. We refer to van der Heijden (1992) for a comparison of CA and the latent class model in the context of square tables where interest goes out to the off-diagonal frequencies. It is shown in van der Heijden *et al.* (1990) that multiple correspondence analysis (discussed in part two of this book) and the latent class model are also closely related.

4.4 CORRESPONDENCE ANALYSIS AND LOG-LINEAR MODELS

In earlier sections we have discussed a ML version of CA, and the relation of (the ML version of) CA with the latent class model. In this paragraph we will discuss a specific way to view the relation of CA with log-linear models. We refer to van der Heijden *et al.* (1989) and the papers cited there for many more details concerning this approach. An approach that is slightly different, but basically the same, is discussed by Novak and Hoffman (1990).

4.4.1 CA as a tool for residual analysis

The starting point of the relationship of CA with log-linear models is the relation of CA with the independence model. This relation can easily be derived from both the reconstitution formula as well as from the relation between the Pearson chi-square with the sum of the squared singular values. The reconstitution formula given in section 2 can be rewritten as

$$p_{ij} = p_{i+}p_{+j}\left(1 + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk}\right) = p_{i+}p_{+j} + p_{i+}p_{+j} \sum_{k=1}^{K} \mu_k r_{ik} c_{jk}$$

which shows that in CA the (scaled) residuals $(p_{ij} - p_{i+}p_{+j})/p_{i+}p_{+j}$ are decomposed into K dimensions, where $p_{i+}p_{+j}$ are estimates of expected probabilities under the statistical model defining independence of the categories of the row from the column variable. In $p_{i+}p_{+j}\Sigma_k \mu_k r_{ik} c_{jk}$ the marginal proportions p_{i+} and p_{+j} correct for the fact that some rows and columns have higher marginal frequencies. Another property showing the relation of CA with the independence model is

$$\frac{\chi^2}{n} = \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{k=1}^{K} \lambda_k$$

where χ^2 is the Pearson chi-square statistic used for testing independence between the row and the column variable, and n is the sample size. This equation shows that the departure from independence, as measured by the Pearson chi-square, is decomposed over the K dimensions of the CA solution.

So there is a close connection between independence and CA: CA can be considered as a tool for residual analysis. By decomposing the residuals from the independence model a graphical representation is obtained of the interaction between the row and column variables. However, this tool is not very flexible, because the model for which the residuals are decomposed is always the independence model. In many applications this model is very often naïve: often we are interested in a model less naïve and less restrictive than independence.

The reader should notice that we are discussing a rather limited use of CA here: the above interpretation is only useful if the data stem from a (product)-

multinomial or a Poisson distribution. If this is not the case, then the interpretation of CA in terms of residual analysis will not be very fruitful (but see, for an application of the tools discussed here in the context of multiple correspondence analysis, van der Heijden and Meijerink 1989).

4.4.2 *A generalization of correspondence analysis*

When we want to study the residuals from less restricted models we can use a generalization of CA proposed by Escofier (1983, 1984). This generalization is

$$p_{ij} = q_{ij} + s_i t_j \sum_{k=1}^{K} \mu_k r_{ik} c_{jk}$$

Here the row scores r_{ik} and column scores c_{jk} are standardized such that

$$\sum_{i=1}^{I} s_i r_{ik} = 0 \qquad \sum_{i=1}^{I} s_i r_{ik}^2 = 1 \qquad \sum_{j=1}^{J} t_j c_{jk} = 0 \qquad \sum_{j=1}^{J} t_j c_{jk}^2 = 1$$

The terms s_i are used as weights for the row points, and as the metric defined by s_i^{-1} for the space of the column points, and the terms t_j are used as weights for the column points, and provide the metric defined by t_j^{-1} for the space of the row points (compare Chapter 1). These terms can be chosen by the user. For the elements q_{ij} we can choose estimates of expected probabilities under some contingency table model. If we choose $s_i = p_{i+}$, $t_j = p_{+j}$, and $q_{ij} = p_{i+}p_{+j}$, then we find 'ordinary' CA. Many different types of applications can be worked out using this generalization. We will discuss the two applications that have received most attention thus far.

4.4.3 *Correspondence analysis of incomplete contingency tables*

In the standard approach of CA an analysis is performed of a complete two-way contingency table. However, sometimes we would like to analyze only part of the cells in a contingency table. As a first example, consider a two-way table of regions by years, with the number of property crimes in the cells. It might be that in some of the regions no data are available for all years. So some of the frequencies are simply missing. A standard CA is not useful here, because it starts from the assumptions that there are no missing frequencies. Yet we would like to study the interaction in the available frequencies with a tool like CA.

As a second example, some of the cells can be structurally zero, i.e. they refer to combinations that cannot logically occur, such as diagonal cells in import–export tables. In such tables countries are the categories of the row variable as well as of the column variable. In the cells amounts of goods or money are provided that countries import from other countries (and that the

latter countries export to the former countries). The values in the diagonal cells are undefined: it is not clear what should be filled in here, because countries do not export and/or import to themselves. This may cause a problem in a standard analysis of the data. A solution is to focus attention to the off-diagonal cells only.

Thirdly, for substantive reasons we may decide that we want to apply a model to some cells but not to others. For example, in a social mobility table it might be that we are only interested in the off-diagonal cells, that reflect changes. In the diagonal cells we find the number of father–son couples that have the same profession. Processes that lead to having the same profession are usually considered to be different from processes that lead to profession-differences for fathers and sons. Therefore it might be useful to model only the off-diagonal cells, so that only one process is studied. A similar example will be discussed in the next section.

In standard CA the independence model is the baseline model. In log-linear analysis the independence model is adjusted to a so-called quasi-independence model to deal with the three examples given above, where estimates under the independence model are $p_{i+}p_{+j}$, under the quasi-independence model they are $\hat{\alpha}_i\hat{\beta}_j$ for the cells for which we would like to fit (quasi-)independence. For the other cells the estimated probabilities are simply equal to the observed proportions.

We can analyze the residuals from quasi-independence by the generalization of CA in the following way: we choose $q_{ij} = \hat{\alpha}_i\hat{\beta}_j$. If we now choose $s_i = \hat{\alpha}_i$ and $t_j = \hat{\beta}_j$, then the generalization of CA has again the property that the Pearson chi-square (now for testing quasi-independence) is closely related to the sum of the squared singular values. This property was lost for the generalization in its most general form. So, like in ordinary CA, the information as measured by Pearson's chi-square (which is the criterion used to decide whether the residuals contain interesting information) is decomposed into a number of dimensions.

As de Leeuw and van der Heijden (1988) show, it turns out that this application is equivalent to a procedure to deal with missing data in CA for a long time (see, for example, Greenacre, 1984, Ch. 8). This procedure adjusts the table to be analyzed, *before* an actual analysis is performed. Specifically, for the cell values in which one is *not* interested, 'independent' values are filled in. These independent values are calculated iteratively as $p_{ij}^{(m+1)} = p_{i+}^{(m)}p_{+j}^{(m)}$ for iteration m, and iterating ends after convergence. As it turns out, for the converged values p_{ij}, $p_{ij} = \hat{\alpha}_i\hat{\beta}_j$.

This shows, first, that a well-known procedure in the CA tradition can be understood in terms of a residuals approach of log-linear models. It also shows that this application is easily carried out by ordinary CA programs.

De Leeuw and van der Heijden (1988) give more details about this procedure, and in van der Heijden *et al.* (1989) it is shown how this procedure can

TABLE 4.4

Number of students, subdivided by faculties and years. First six columns: first-year students; last six columns: graduate students.

Faculty	Year						Year					
	1935	1946	1957	1968	1979	1983	1935	1946	1957	1968	1979	1983
ARTS	194	384	654	1129	4224	4101	144	46	188	527	1114	2075
THEOLOGY	167	236	123	317	246	248	124	37	56	53	152	100
MEDICINE	549	1504	868	2438	1940	2418	391	507	650	1320	2021	2174
MATHPHYS	257	560	796	2046	2537	2496	171	237	316	924	1218	1635
TECHNICS	296	1355	1251	2756	3140	3957	212	272	457	1050	1403	1727
AGRICULT	58	176	184	510	1155	959	71	106	71	138	343	649
LAW	464	768	365	2079	3417	4993	420	493	358	790	1385	2348
ECONOMIC	65	671	633	1584	2018	2513	81	314	226	571	807	868
SOC-CULT	0	0	313	1258	1194	1340	0	0	68	291	498	561
PSYCHOL	0	0	256	947	1247	1228	0	0	72	190	744	797
PEDAGOGY	0	0	59	542	1278	1112	0	0	10	71	670	867
GEOGRAPH	24	236	157	358	592	626	24	15	35	103	365	473
OTHER	0	0	0	0	224	719	0	0	0	0	112	432

be used fruitfully in the context of square contingency tables. They also show how it can be generalized to deal with the departure from other models in this context, such as the symmetry model and the quasi-symmetry model.

4.4.4 *Correspondence analysis as a tool for the analysis of residuals from conditional independence in a higher-way table*

In order not to make the discussion too abstract, we present this application by means of an example, omitting technical details (for these the references at the end of this section can be consulted).

Consider Table 4.4. It is a table with faculties in the rows, and types of student in specific years in the columns. In the cells we find the number of first-year or graduate students in different faculties in The Netherlands at successive time points. The faculties are the arts (including philosophy), theology, medicine (including veterinary and dentistry), mathematics and physics (together), technical science, agriculture, laws, economic sciences, socio-cultural sciences, psychology, pedagogy, geography, and others. The years chosen are 1935, 1946, 1957, 1968, 1979 and 1983, which are evenly spaced, except for the last year (1983). Some of the cells in the matrix have zero frequencies, due to the fact that some faculties did not exist at that time. So in fact, we could interpret Table 4.4 as an incomplete contingency table, and handle these zero cells appropriately (see above).

However, we do a standard CA first. The Pearson chi-square is 12 883 (df is 25). This chi-square reflects the fit of a (log-linear) model where the row variable is independent of the column variables, i.e. the faculties have no relation with the types of student and the years. The estimates of expected frequencies under this log-linear model have the following property: the relative proportions of first-year students and graduates are equal for each faculty, and both the number of first-year students and the number of graduate students develop for each faculty proportionally in the same way. In the terminology of CA we would say that under this log-linear model the row profiles are equal. But since this model is rejected (this is revealed by the Pearson chi-square, which is significant), it makes sense to study the departure from the log-linear model.

The standard CA shows how Table 4.4 departs from this situation. The first four singular values (with percentages of chi-square) are 0.245 (54.4%), 0.141 (18.0%), 0.104 (9.9%) and 0.083 (6.3%) so we can restrict attention to the first two dimensions (the singular values for dimension 3 and 4 do not differ enough to warrant attention to any of these two). These two dimensions display 72% of the total dependence between the row and the column variables (see Figure 4.2). Care should be taken that the points geography, economics and agriculture are not very well represented in the first two dimensions: 2.6%, 22.4% and 42.9%, respectively, are displayed in the first two dimensions;

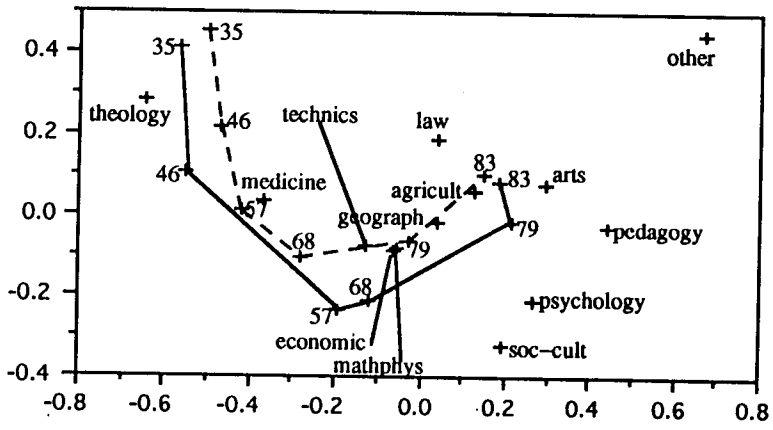

FIGURE 4.2 Correspondence analysis of Table 4.4. Joint plot, distances between rows and distances between column approximate chi-square distances. Solid line: first year students; dotted line: graduate students. See text for explanation.

similarly, only 8.6% and 50.5% are displayed from the graduate years 1979 and 1983.

The time points are roughly ordered in the first dimension from 1935 to 1983. We can conclude that, compared to the average faculty profile given by p_{+j}, the values p_{ij}/p_{i+} of many of the faculty profiles are either regularly growing or diminishing from 1935 until 1983, or have a peak in the middle period. Compared to the average faculty profile having elements p_{+j}, the faculties on the left of the origin have more students in the earlier years, whereas the faculties on the right of the origin have more students (than the average) in the later years. In the early years relatively many students were in theology and medicine. In recent years this has changed towards pedagogy, psychology, arts and 'other'. The peak for mathematics and physics (mathphys) and economy is found in the middle. All faculties are ordered along dimension 1 according to their peak compared with the average row profile.

Interpretation of the second dimension is more difficult. It seems that this dimension is a quadratic transformation of the first dimension. This effect is often found in CA when the first dimension is dominant. It is usually called the horseshoe effect, or Guttman effect. Different interpretations exist for this phenomenon (compare van Rijckevorsel 1986). For example, in Gifi (1990) an interpretation in terms of Hermite–Chebychev polynomials is given. Greenacre (1984, Ch. 8) presents an interpretation in terms of chi-square distances. Schriever (1983, 1986) presents an interpretation in terms of order dependence in categorical variables. It is clear from these interpretations that the second and higher dimensions need no substantive interpretations.

Let us now consider the solution in Figure 4.2 in terms of log-linear models again. In log-linear models the interaction in the three-way contingency table is attributed to different sources. There are three two-factor interactions, namely the two-factor interactions between faculties and years (do some faculties have more students in some years, relative to the average number of students over years?), between faculties and types of student (do some faculties have more graduate students, relative to the average proportion of graduate students over all the faculties?), and between years and types of student (are there more graduate students in some years, relative to the average proportion of graduate students?). Then there is also three-factor interaction: the way in which the relation between faculties and years differs for different types of student.

When we consider the standard CA of Table 4.4, it should first be noticed that the two-factor interaction between years and types of student is not displayed, because it is contained in the column margins of Table 4.4. This information could be used explicitly in choosing an appropriate way to code and analyze the three-way table: now the two-factor interaction between years and type of students is not displayed, but (two) different choices could have been made, namely either coding years and faculties interactively (then the two-factor interaction between years and faculties would not have been displayed) or coding faculties and students interactively (then the two-factor interaction between faculties and students would not have been displayed).

What *is* displayed in Figure 4.2 is the two-factor interaction between faculties and years, the two-factor interaction between faculties and types of student, and the three-factor interaction. So standard CA displays the departure from the log-linear model where faculty is independent from year and type of student jointly. The two-factor interaction between faculties and years is clearly overwhelming, and in fact it overshadows the other interactions. One reason is that this two-factor interaction consists for a large part of zeros in Table 4.4, namely that some faculties were founded in a later year. Therefore one might wonder whether it is possible to eliminate this two-factor interaction, in order to have a better view of the two-factor interaction between faculty and type of student, and the way in which this interaction differs in different years.

This can be accomplished by using the generalization of CA to study the departure from a different log-linear model. In this different log-linear model there is interaction between years and types of student (as before), but there is also interaction between faculties and years. This log-linear model is a conditional independence model: under this model faculties and types of student are independent for each year. So, if we study the *departure* from this model, we get a display of the two-factor interaction between faculties and types of student, and of the three factor interaction (i.e. how this interaction differs for different years). The Pearson chi-square for this new log-linear

model is 2973, with 62 df. This model is still rejected, so in principle (generalized) CA can be useful to study the structure in the residuals. The generalized CA solution has first four singular values 0.118 (55.7%), 0.069 (19.3%), 0.056 (12.4%) and 0.050 (10.0%). Following the elbow criterion, we study only two dimensions. This solution in Figure 4.3 is closely related to the solution in

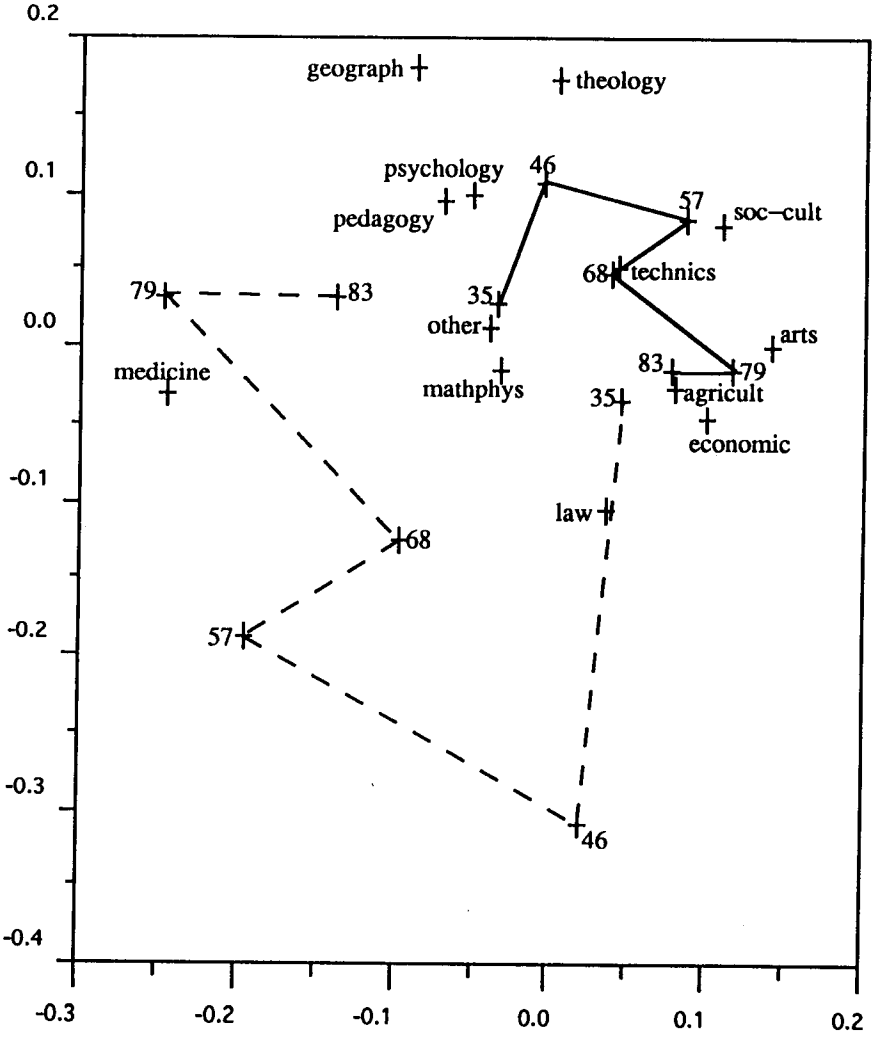

FIGURE 4.3 Generalized correspondence analysis of Table 4.4, where the two-factor interaction between faculties and years is eliminated. See text for explanation.

Figure 4.2: in the full-dimensional spaces corresponding to Figures 4.2 and 4.3 the distances between corresponding year points are identical. This holds, for example, for the distance between 1946 first-year students and 1946 graduates. However, whereas the (weighted) average of the two points in Figure 4.2 is clearly far from the origin (displaying the two-factor interaction between faculty and year), in Figure 4.3 it is located near the origin, because the interaction between faculties and years is eliminated from the (generalized) CA solution. Therefore we find in Figure 4.3 that each pair of year points has a weighted average of zero (when the marginal frequencies are used as weights). The first-order interaction between faculties and types of student is displayed on the first two dimensions: the faculties with relatively more first-year students (i.e. socio-cultural sciences (soc-cult), technical sciences (technics), arts and theology) are located in the direction of the first quadrant (solid line), whereas the faculties with relatively more graduates (medicine, law, mathematics and physics) are plotted in the direction of the third quadrant. Three-factor interaction can be derived from the plot as follows: medicine for example (but also pedagogics, psychology and geography) seems to have relatively more graduates in later years, and in these same years relatively less first-year students (as compared to the average). On the other hand, the law faculty (and also economy, agriculture) has relatively many first-year students in later years, and many graduates in early years. Inspection of the contributions shows that we have to be careful with the interpretation of the position of the points for mathematics/physics, technical sciences, psychology and pedagogy, because, respectively, only 29.9%, 47.2%, 51.0% and 44.7% of the total squared distance to the origin are projected on the first two dimensions.

We conclude that the approach to use log-linear models to eliminate information from standard CA solutions can be usefully applied when specific interactions in a higher-way table are so strong that they overwhelm the other interactions. For more applications, see van der Heijden *et al.* (1989). For more (technical) details concerning this application we refer to Escofier (1983, 1984), van der Heijden *et al.* (1989), and Takane *et al.* (1991b). Escofier (1983) calls the solution in Figure 4.3 a solution of an 'intra analysis', and she proposes to do this 'intra analysis' jointly with an 'inter analysis', which is the analysis of the marginal table of faculty by year. Van der Heijden and de Leeuw (1985) and van der Heijden *et al.* (1989) show the connection of 'intra analysis' with log-linear models, they show how this application can be extended to tables of more than three variables, and to more complicated log-linear models than 'simple' conditional independence models. They also propose to use CA and log-linear analysis complementary to each other: log-linear analysis could be used to answer the question of which variables are related by detecting the important interactions; in the second step CA can be used, for example, to explore the residuals when a model in which we are

interested does not fit. Takane *et al.* (1991) show that this approach is closely related to other approaches with similar aims.

4.4.5 *Conclusion*

We think that by relating CA to log-linear models many insights are obtained. To mention a few:

- A new way to use CA is indicated, namely to see it as a tool for residual analysis. This might be particularly fruitful for large tables.
- It can also be emphasized that, given that the CA procedures study the residuals from log-linear models, these residuals should be meaningful, i.e. the models under study should not fit. This holds for the applications making use of the generalization of CA, but also for ordinary CA. In the applications that we see in the literature this is seldomly checked.
- For the analysis of higher-way tables by means of stacking categories of variables more insight is obtained into how to make an appropriate choice between different ways to stack variables.
- For incomplete contingency tables it is shown that a CA procedure already known can be understood in terms of the quasi-independence model.
- It is emphasized that for many applications and research questions the independence model is simply not the appropriate baseline model. Very often less restrictive models are needed to study these aspects of the data one is interested in. In the CA approach, the aspects we are *not* interested in are incorporated in the model, so that the interesting aspects are in the residuals.

For an extensive discussion and appreciation of this approach we refer to the discussion following the paper of van der Heijden *et al.* (1989).

4.5 CORRESPONDENCE ANALYSIS AND THE RC-ASSOCIATION MODELS

In this section we will discuss relations between CA and the RC-association model. This relation is mainly that, if certain conditions are fulfilled, parameter estimates found in CA are very similar to parameter estimates found in the RC-association model. Consider our example in Table 4.1. The results of the analysis with the so-called RC(2)-association model are displayed in panel C of Table 4.2. The similarity between the estimates in panel C and panel B is considerable. This is not a peculiarity of our example, but a systematic result, that will be worked out below.

In order to appreciate the form of the RC-association model, we first have to define the log-linear model for two-way tables. The saturated (i.e. unrestricted) log-linear model is

$$\log \pi_{ij} = u + a_i + b_j + c_{ij}$$

where the parameters add up to zero over each subscript. If $c_{ij} = 0$ we find the independence model.

The interaction parameters $u_{12(ij)}$ are often related to quantities called the 'log-odds ratio' $\theta_{ii'jj'}$, which is defined for four cells (i,j), (i,j'), (i',j) and (i',j'). For the saturated model the log-odds ratio is defined as

$$\theta_{ii'jj'} = \log [(\pi_{ij}\pi_{i'j'})/(\pi_{ij'}\pi_{i'j})] = u_{12(ij)} + u_{12(i'j')} - u_{12(ij')} - u_{12(i'j)}$$

For the whole table $(I-1)(J-1)$ log-odds ratios describe the full pattern of associations. The definition of $\theta_{ii'jj'}$ shows that the log-odds ratio is independent of the margins, which is an attractive property for measures of interaction, because it allows one to compare the interaction strength in different subtables or tables. The definition also shows that under the independence model $\theta_{ii'jj'} = 0$.

There obviously is a need for models that are more restrictive than the saturated model, yet less restrictive than the independence model. Many models are proposed for this purpose, but recently models with bilinear terms (multiplicative terms) in the logarithm have received much attention.

4.5.1 *The RC-association model*

The basic model is the RC-association model (Goodman 1979, 1985, 1986, Andersen 1980), defined as

$$\log \pi_{ij} = u + a_i + b_j + \phi v_i w_j$$

where the v_i and w_j have to be constrained in order to identify the model. One way to identify them is by

$$\sum_{i=1}^{I} p_{i+} v_i = 0 \qquad \sum_{i=1}^{I} p_{i+} v_i^2 = 1 \qquad \sum_{j=1}^{J} p_{+j} w_j = 0 \qquad \sum_{j=1}^{J} p_{+j} w_j^2 = 1$$

These identifying constraints are chosen in such a way that they are very similar to those used for CA. The term $\phi v_i w_j$ is a bilinear term, and therefore the model is not log-linear any more. This term can be seen as a rank one constraint to the matrix of interaction parameters $u_{12(ij)}$. If I and J are large, the number of interaction parameters can be reduced considerably.

In terms of log-odds ratios, the model defines

$$\theta_{ii'jj'} = \log [(\pi_{ij}\pi_{i'j'})/(\pi_{ij'}\pi_{i'j})] = \phi(v_i - v_{i'})(w_j - w_{j'})$$

This shows that $\theta_{ii'jj'}$ is a function of the association strength ϕ, of the difference between the scores for i and i', and of the difference between the scores for j and j'. The smaller any of these three quantities is, the more the quantity $\theta_{ii'jj'}$ approaches zero, which implies that in the subtable of cells (i,j), (i',j'), (i',j) and (i,j') the association approaches independence.

4.5.2 *The RC(K)-association model*

More recently Goodman (1985, 1986) proposed a natural extension of this model, namely the RC(K)-association model. This model is defined as

$$\log \pi_{ij} = u + a_i + b_j + \sum_{k=1}^{K} \phi_k v_{ik} w_{jk}$$

with constraints

$$\sum_{i=1}^{I} p_{i+} v_{ik} = 0 \qquad \sum_{i=1}^{I} p_{i+} v_{ik} v_{ik'} = \delta^{kk'} \qquad \sum_{j=1}^{J} p_{+j} w_{jk} = 0$$

$$\sum_{j=1}^{J} p_{+} w_{jk} w_{jk'} = \delta^{kk'}$$

where $\delta^{kk'} = 1$ if $k = k'$, and $\delta^{kk'} = 0$ if $k \neq k'$, that is, similar to CA, the scores are uncorrelated for different factors. If $K = \min(I-1, J-1)$, then $\Sigma_k \phi_k v_{ik} w_{jk} = u_{12(ij)}$, but if $K < \min(I-1, J-1)$, then the RC(K) association model is more restrictive than the saturated model.

In terms of log-odds ratios, the model is

$$\theta_{ii'jj'} = \log[(\pi_{ij}\pi_{i'j'})/(\pi_{ij'}\pi_{i'j})] = \Sigma_k \phi_k (v_{ik} - v_{i'k})(w_{jk} - w_{j'k})$$

The insight into the log-odds ratio $\theta_{ii'jj'}$ is getting more difficult now, because it is a function of parameters ϕ_k, differences of score i and i' on K dimensions, and differences of scores j and j' on K dimensions.

Consider again Table 4.2, panel C. In the first line the estimates for ϕ_k are found, then a 6×2 matrix with estimates for v_{ik} follows, and then a 6×2 matrix with estimates for w_{jk} follows. This model has $(I-K-1)(J-K-1)$ degrees of freedom, so for this example, where $K = 2$, the number of degrees of freedom is 9. The chi-square statistic is 5.828. The RC-association model, for which $K = 1$, the number of degrees of freedom is 16, and the chi-square is $G^2 = 90.2$.

Although the parameter estimates in Table 4.2, panel C are very close to these in panel A and B (reasons for this closeness are explained below), the interpretation of these parameters is different. Interpretation is simplified by means of a graphical display. A useful graphical display is obtained by using the coordinates $\hat{v}^*_{ik} = \hat{\phi}_k^{1/2} \hat{v}_{ik}$ for row point i on dimension k, and coordinates $\hat{w}^*_{jk} = \hat{\phi}_k^{1/2} \hat{w}_{jk}$ for column point j on dimension k (see Figure 4.4); thus

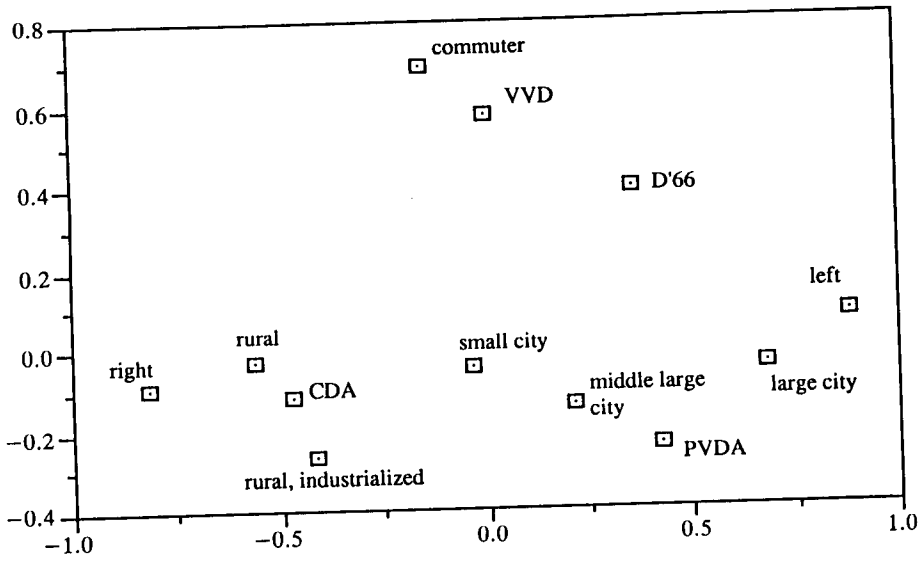

FIGURE 4.4 Graphical display of RC(2)-association model parameter estimates. See text for explanation.

$\Sigma_k \hat{v}^*_{ik}\hat{w}^*_{jk} = \Sigma_k \hat{\phi}_k \hat{v}_{ik}\hat{w}_{jk}$. Notice that, compared to Figure 4.1(a) and 4.1(b), the configuration of both the row points and the column is very similar, except for a stretching of the second dimension: on this dimension the distances are larger now. The reason is that in the CA representation distances between the row points, and distances between the column points, approximate chi-square distances, because $\hat{f}_{ik} \equiv \hat{\mu}_k \hat{r}_{ik}$ and $\hat{g}_{jk} \equiv \hat{\mu}_k \hat{c}_{jk}$ are used as coordinates. Here we do not use $\hat{\phi}_k$ but $\hat{\phi}_k^{1/2}$ to scale the row and column scores. Thus we may make a simultaneous display of row and column scores.

In Figure 4.4 estimates for $\Sigma_k \hat{v}^*_{ik}\hat{w}^*_{jk}$ could be derived to find the association for cell (i,j). This shows that the association for cell (i,j) is zero when row point i and column point j make a right angle with the origin. The association is positive when the angle is larger than 90°, and negative when the angle is smaller than 90°. The more the row point and the column point are away from the origin, and the smaller the angle between the row and column point, the more extreme becomes the association. Similarly, $\hat{\theta}_{ii'jj'} = \Sigma_k (\hat{v}^*_{ik} - \hat{v}^*_{i'k})(\hat{w}^*_{jk} - \hat{w}^*_{ik})$ should be used to get the estimated log-odds ratio for the four cells (i,j), (i',j'), (i,j') and (i',j). So, if row points i and i' are close, and column points j and j' are close, then the log-odds ratio for the corresponding subtable approaches zero, and under the model the subtable of expected frequencies approaches independence. For many details concerning the

interpretation of these graphical displays, and a discussion of some alternatives, we refer to Goodman (1991).

4.5.3 *Estimation and restrictions*

This model is usually fitted by ML. Apart from restrictions on the number of factors M, various additional restrictions can be imposed on the parameter estimates, such as fixed value constraints and equality constraints. If the row parameters are fixed, for example, to equidistant scores $v_i = \{-1, 0, 1\}$ if $I = 3$, the model becomes log-linear, and only w_j (and the scaling factor ϕ) has to be estimated. This model is called the C-association model. Similarly, if the column scores are fixed, and the row scores v_i have to be estimated, we find the R-association model. If both the row scores and the column scores are fixed to equidistant scores, then only ϕ has to be estimated, and this model is called the U-association model (see Goodman 1985 for details). Also, see Gilula and Haberman (1988). For square tables, it is possible to restrict $v_i = w_i$, and to eliminate the influence of the diagonal cells by including one parameter for each diagonal cell. The RC-association model is quite flexible in the sense that it can be adjusted to deal with many situations.

4.5.4 *Relation to CA*

For the example discussed above we found that the estimates for the RC(2)-association model were very similar to CA. This is not a coincidence. It is well known that CA, the RC-association model and the RC(M)-association model are closely related. We will now discuss the reasons for this.

First, it is shown by Goodman (1981) that if $k = 1$ and the proportions come from a discretized bivariate normal distribution (or a distribution that is bivariate normal after a suitable transformation of the rows and columns), CA in one dimension is closely related to the RC-association model with one dimension: it turns out that $\mu_1 \approx \phi, r_{i1} \approx v_i$ and $c_{j1} \approx w_j$. Second, the CA representation in k dimensions can be written in an adapted version of the reconstitution formula as

$$m_{ij} = p_{i+}p_{+j}\left(1 + \sum_{k=1}^{k^*} \mu_k r_{ik} c_{jk}\right)$$

where m_{ij} is the reconstituted value for cell (i, j). Escoufier (1982) noted that if

$$x = \sum_{k=1}^{k^*} \mu_k r_{ik} c_{jk}$$

is small compared to one (so that $\log(1+x) \approx x$) we can rewrite the reconstitution formula as

$$\log m_{ij} \approx u + a_i + b_j + \sum_{k=1}^{K^*} \mu_k r_{ik} c_{jk}$$

where $u = 0$, $a_i = \log p_{i+}$ and $b_j = \log p_{+j}$. Since r_{ik} and c_{jk} are normalized in a similar way to v_i and w_j, Escoufier's condition roughly reduces to the situation that the departure from independence is not too large.

The conclusion is that in CA the interaction is decomposed approximately in a log–multiplicative way: the graphical displays show approximations of log–bilinear parameters. For empirical examples of this relation we refer, for example, to Goodman (1985). This close relation between CA and models with log–bilinear terms also holds for more complicated models. Examples are given in van der Heijden and Worsley (1988) and Green (1989) for a three-way table, and in van der Heijden (1992) for square tables where the CA approach to incomplete tables (see section 4.4.1) is compared with adjusted versions of the RC(M)-association model. Van der Heijden and Mooijaart (1991) show that the scores found in examples using the CA approach of section 4.4.1 where the departure from quasi-symmetry is studied, are very similar to the parameter estimates obtained in comparable models with log–bilinear terms.

Our experience is that the condition that $\log(1+x) \approx x$ should be small, is not very restrictive. Even if for some of the cells x is rather large compared to one, then generally the interpretation will not change drastically. A reason might be that, if for some cells (i,j), x is large compared to one, this will not necessarily mean that the factorization of the matrix of cells (i,j) in the association model context will be very different from the factorization in the CA context. More research is needed in this area.

4.6 CORRESPONDENCE ANALYSIS AND IDEAL POINT DISCRIMINANT ANALYSIS

4.6.1 Ideal point discriminant analysis

Ideal point discriminant analysis (IPDA; Takane 1987, 1989, Takane *et al.* 1987) is a model that is inspired by the unfolding interpretation of CA. In unfolding (Coombs 1964) it is aimed to represent the rows and columns of a two-way table in one common spatial representation. Various specification can be given, but the general idea of unfolding is that, if row i has certain dissimilarities with the J columns, these dissimilarities are reflected by the distances of row i to the J columns. In correspondence analysis (approximated by least squares) a distance interpretation can be found if asymmetric scaling of rows and columns is used. Such an asymmetric scaling is obtained if for the

row points of correspondence analysis coordinates r_{im} are used as coordinates, and for the column points g_{jk} (see section 2). Then the column points are in the centroids of the row points:

$$g_{jk} = \sum_{i=1}^{I} \frac{p_{ij}}{p_{+j}} r_{ik}$$

Thus column point j is in the weighted average of the row points, where the conditional proportions p_{ij}/p_{+j} are used as weights. It is allowed to calculate distances between row points i and column point j, and the order of these distances reflects the order of the dissimilarities. (Notice that this property does not hold for CA approximated by ML, unless instead of the observed conditional proportions p_{ij}/p_{+j} the conditional probabilities π_{ij}/π_{+j} are used as weight.) For more details concerning an unfolding interpretation of CA, see Takane (1980), Heiser (1981), Ihm and van Groenewoud (1984), ter Braak (1986) and Greenacre (1989).

Using this as a starting point, IPDA is defined as

$$\pi_{j|i} = \frac{w_j \exp(-d_{ij}^2)}{\sum_{j'}^{J} w_{j'} \exp(-d_{ij}^2)}$$

where d_{ij}^2 is defined as the Euclidean distance between row point i and column point j, column point j being derived as the weighted average of the row points. So, let x_{ik} be the IDPA parameter for the coordinate of row point i on dimension m, y_{jk} the IPDA column coordinate, then the column point j is derived from the row points i as

$$y_{jk} = \sum_{i=1}^{I} \frac{p_{ij}}{p_{+j}} x_{ik}$$

and the squared Euclidean distance d_{ij}^2 is

$$d_{ij}^2 = \sum_{k=1}^{K^*} (x_{ik} - y_{jk})^2$$

Notice that the coordinates y_{jk} are not parameters: the parameters that have to be estimated are the row parameters x_{ik} and the weights w_j (which will often approximate the marginal column proportions p_{+j}).

The denominator of the model for $\pi_{j|i}$ is only included in the model to ensure that $\Sigma_j \pi_{j|i} = 1$. It follows that $\pi_{j|i}$ is linearly related to w_j when $\exp(-d_{ij}^2)$ is fixed, or to $\exp(-d_{ij}^2)$ when w_j is fixed. This latter statement implies that, for fixed w_j, the conditional probability $\pi_{j|i}$ becomes larger if the distance d_{ij} becomes smaller, i.e. if row i is closer to column j.

Consider now the example in Table 4.1. The fit of the ideal point discriminant model with $K^* = 2$ dimensions is $G^2 = 7.02$ (df is 16). For $K^* = 1$ the fit is $G^2 = 90.48$ (df is 20), and for $K^* = 3$ the fit is $G^2 = 2.64$ (df = 13). The

coordinates for $K^* = 2$ are presented in panel E of Table 4.2. A graphical display is found in Figure 4.5. In this display the distances between a row point and the column points are well defined. But now the distances between the rows, and the distances between the columns are not well defined. When we compare this with CA, we find that in a similar asymmetric display (where the column points are in the weighted average from the row points) the distances from each row to the columns are also well defined, but the distances between the column points are equal to chi-squared distances.

4.6.2 *Relation to correspondence analysis and to RC-association models*

It was indicated that IPDA was inspired by the unfolding interpretation of

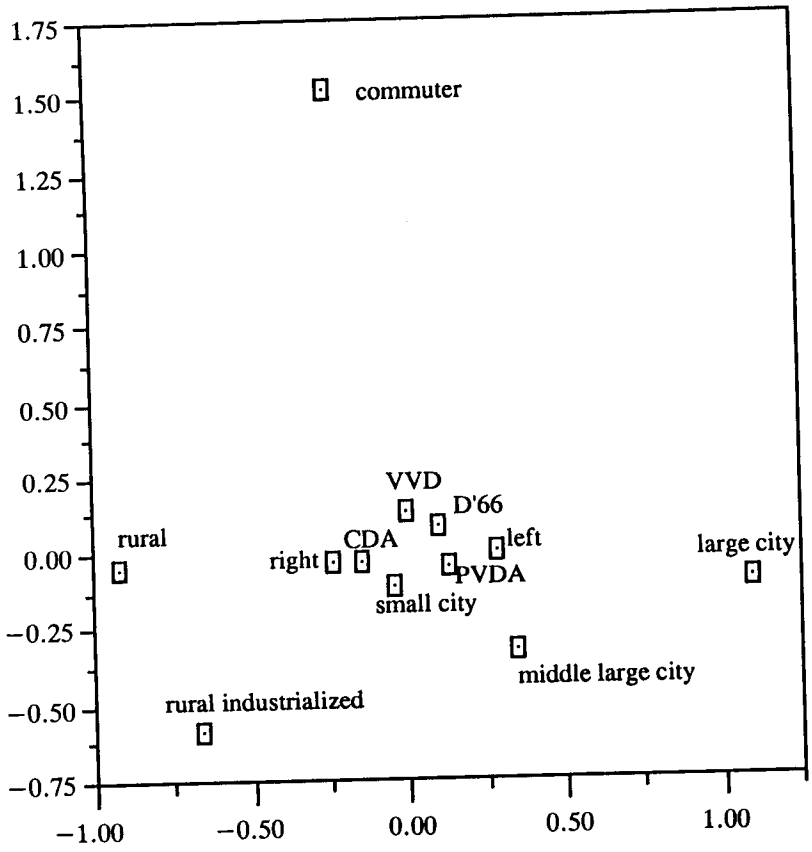

FIGURE 4.5 Graphical display of ideal point discriminant analysis parameter estimates. See text for explanation.

CA, because the IDPA column coordinates y_{jk} are derived as the weighted averages of the IPDA row parameters x_{ik}. This property holds for CA approximated by LS, but not by CA approximated by ML. Compared with CA approximated by ML and to the RC(M)-association model, in IPDA less parameters are estimated, because the column coordinates y_{jk} are derived from the row parameters x_{ik}. This usually does not lead to a great loss of fit, but, on the contrary, it does lead to a gain in number of degrees of freedom. It often leads to an increased stability of the parameters that *are* estimated in IPDA, in the sense that, if in some models one or more parameters are fixed, then the standard errors of the remaining parameters often become smaller.

Takane (1987) has shown that IPDA can be written as a constrained conditional version of the RC-association model:

$$\pi_{j|i} = \frac{b_j \exp\left(\sum_{k=1}^{K^*} \phi_k v_{ik} w_{jk}\right)}{\sum_{j'=1}^{J} b_{j'} \exp\left(\sum_{k'=1}^{K^*} \phi_{k'} v_{ik'} w_{jk'}\right)}$$

in the sense that w_j corresponds to w_j, $\phi_k v_{ik}$ corresponds to x_{ik}, and w_{jk} corresponds to the constrained y_{jk}. Takane (1987) calls this latter constraint critical in the comparison of IPDA and the RC-association model. The correspondence of IPDA with the RC-association model can be seen by working out $-d_{ij}^2$ as $-\Sigma_k(x_{ik}-y_{jk})^2 = -\Sigma_k x_{ik}^2 + 2\Sigma_k x_{ik}y_{jk} - \Sigma_k y_{jk}^2$, it follows that

$$\exp(-d_{ij}^2) = c_i c_j^* \exp\left(2 \sum_{k=1}^{K^*} (x_{ik}y_{jk})\right)$$

The parameter c_i cancels out in the numerator and denominator of the conditional model, and c_j^* is absorbed in w_j. This shows that, in IPDA, it is allowed to set $\Sigma_i p_{i+} x_{ik} = 0 = \Sigma_j p_{+j} y_{jk}$ without influencing the fit of the model. In order to simplify the comparison of IPDA with the RC-association model we now rescale x_{ik} to x_{ik}^* with the property that $\Sigma_i p_{i+} x_{ik}^* = 1$ and y_{jk} to y_{jk}^* with the property that $\Sigma_j p_{+j} y_{jk}^* = 1$. So we are rescaling $x_{ik}y_{jk} = \omega_k x_{ik}^* y_{jk}^*$. The corresponding parameter estimates are shown in panel F of Table 4.2. The similarity of the rescaled IPDA parameter estimates with the RC-association parameter estimates is striking, especially for x_{ik}^*. Notice that twice the parameter estimates for ω_k correspond to the parameter estimates for ϕ_k, since $-\Sigma_k(x_{ik}-y_{jk})^2$ corresponds to $2\Sigma_k x_{ik}y_{jk}$.

For more details concerning the relation of IPDA with CA and the RC(M)-association model, see Takane (1987) and Goodman (1991).

4.7 DISCUSSION

In this chapter we have compared the usual CA approach (which we called CA

by least squares) with different sorts of models fitted by ML. As we see it, the most important advantage of CA by least squares is that the computations are straightforward, using the singular value decomposition, eigenvalue decomposition or reciprocal averaging algorithm. On the other hand, if we know that the assumptions of a (product)multinomial or Poisson distribution are fulfilled, then we should use this information in our model fitting, and thus ML estimation is a more natural candidate for the CA model.

When we compare CA estimated by ML with association models with log–bilinear terms (see section 4.4), a draw back of the correspondence analysis models is that the parameters for the interaction depend on the marginal distributions. This makes association models preferable. Another reason to prefer association models is that they are very easily extended to deal with higher-way tables, to deal with structural zeros, and to deal with specific patterns in the association. These are probably the more important reasons that there is a tendency in the Anglo-Saxon literature to prefer the RC association model over the CA model. In fact, we found only one exception to this, namely Wasserman and Faust (1989), who favour the use of the CA model over the RC association model for social network analysis. For other comparisons between the models we refer to Goodman (1985, 1986, with discussion; 1987, 1991, with discussion).

IPDA is closely related to both CA and to the RC-association model. CA estimated by least squares and IPDA have the unfolding interpretation in common (see section 4.4.2). The RC-association model and IPDA have the model formulation in common, where the IPDA can be seen as a version of the RC-association model parametrized differently and having additional constraints. These additional constraints lead to an increased stability of the model parameters, while they usually do not lead to a substantial reduction in fit.

APPENDIX: SOFTWARE

Most of the analyses were performed with special purpose software that we wrote ourselves. This special purpose software can be obtained from the authors upon request. However, there is much software available from other sources. In general, we refer to Andersen (1990), who wrote a program library called CATANA to support the analyses in his book, and Agresti (1990), who also discusses software.

CA estimated by least squares can be performed in software packages such as SPSS, BMDP and SAS. CA estimated by ML was fitted with a computer program written in APL by the second author (see also Siciliano *et al.* 1990). We have limited experience with a computer program sent to us by Haberman

(see Gilula and Haberman 1986). Recently a program has been written in GLIM3.77 by de Falguerolles and Francis (in press).

The CA procedures to decompose residuals from log-linear models can usually be fitted with ordinary CA programs, if the input matrix is adjusted (see van der Heijden *et al.* 1989). These adjustments are done by us in special APL-programs.

There is more software available to fit association models with log-bilinear terms. See for references Agresti (1990). There are also programs written in GLIM3.77 by Dessens *et al.* (1985) and Becker (1989).

There is a computer program to perform IPDA that can be obtained from Takane.