stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 85/82 AUGUSTUS
(DEPARTMENT OF MATHEMATICAL STATISTICS)

N.K. KESTER & B.F. SCHRIEVER

ANALYSIS OF ASSOCIATION OF CATEGORIAL VARIABLES BY NUMERICAL SCORES AND GRAPHICAL REPRESENTATION

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: Primary: 62H20, 62H17, 62H25
Secondary: 6207

Analysis of association of categorial variables by numerical scores and graphical representation *)

by

Nancy K. Kester **) & Bert Schriever

ABSTRACT

This paper is an expository treatment of correspondence analysis (CA) and homogeneity analysis (HA), two data analytic techniques which describe association among categorical variables in terms of numerical scores assigned to the categories. These scores are analogous to loadings of numerical variables on principal components. Thus CA and HA allow graphical representations of the data. We give rules for interpretation of such representations and apply them to a real data set.

KEY WORDS & PHRASES: *reciprocal averaging; principal component analysis; correspondence analysis; homogeneity analysis; biplot*

---

*) This report will be submitted for publication elsewhere.

**) Currently on leave from Bell Telephone Laboratories.

1. INTRODUCTION

Correspondence Analysis (CA) and Homogeneity Analysis (HA) are two relatives of Principal Component Analysis (PCA) which are especially suited to the analysis of nominal data. CA and HA describe the association among two or more nominal variables by constructing one or more sets of scores for the categories of the variables. These sets of scores are analogous to loadings on components of PCA. As in PCA, these scores may be used to produce graphical representations of the data. Moreover, in HA and CA, each set of scores for the categories defines derived numerical variables which imply an ordering of the nominal categories. Also, as in PCA, the derived numerical variables may be used in further analysis.

Early references to CA, for example, HIRSCHFELD (1935), are not well known; the major recent developments of CA and HA began with the work of Benzécri and his associates and is discussed in books by BENZÉCRI et. al. (1973) and by LEBART et. al. (1977). Further discussions on CA and HA appear in lecture notes, GIFI (1981), an article by HILL (1974), a book by NISHISATO (1980) and papers by SCHRIEVER (1982a,b).

This paper describes the three techniques as special cases of a general framework based on reciprocal averaging. General guidelines for interpretation of graphical representations are applied to each of these techniques. Also, alternate formulations are described.

Our discussion of these techniques is in the context of exploratory rather than confirmatory data analysis. This does not preclude the possibility of confirmatory analysis on a second sample; in fact, an example of such an analysis is described in a case study by MAAS-DE WAAL et. al. (1982).

Section 2 presents our general framework for these techniques; PCA, CA and HA are described as special cases in the subsequent three sections. Section 6 provides references on some available computer programs. An example of HA is given in section 7.

2. RECIPROCAL AVERAGING

2.1. Preliminaries

Matrices will be denoted by upper case letters. Usually the elements of such a matrix will be denoted by the corresponding lower case letter, doubly subscripted to indicate rows and columns. Diagonal elements of diagonal matrices, however, will be singly subscripted. All vectors, denoted by lower case letters, are column vectors. The transpose of a vector or matrix is indicated by the superscript ^{T}. The identity matrix of size $q \times q$ will be denoted I_q; a vector of unities will be denoted e. The notation $\mathrm{dist}^2(a,b)$ stands for the squared Euclidian distance between two column vectors a and b, i.e.,

$$\mathrm{dist}^2(a,b) = \sum_i (a_i - b_i)^2.$$

The singular value decomposition (svd) of real $n \times m$ matrix A of rank q is defined as

$$A = U \, \Psi \, W^T \tag{2.1}$$

where U is a real $n \times q$ matrix such that $U^T U = I_q$, W is a real $m \times q$ matrix such that $W^T W = I_q$, and Ψ is a $q \times q$ diagonal matrix with elements

$$\psi_1 \geq \psi_2 \ldots \geq \psi_q > 0.$$

Columns u_α of U (w_α of W) are called left (right) singular vectors of A. Left (right) singular vectors of A are exactly the eigenvectors of AA^T ($A^T A$) which correspond to non-zero eigenvalues. Furthermore, each non-zero eigenvalue is the square of a singular value of A. Thus from existence of the eigen decomposition of real, symmetric matrices AA^T and $A^T A$ the svd can be shown to exist (cf. RAO (1973), p.42). Moreover, if the singular values are distinct, then the singular vectors are unique up to a change of sign. (cf. WILKINSON (1965), p.5).

The svd of a matrix A was shown by HOUSEHOLDER & YOUNG (1938) to provide lower rank approximations to A that are best in the sense of least squares. For any integer $\tilde{q}$ such that $1 \leq \tilde{q} \leq q$,

$$\min_{\tilde{A}:\ \mathrm{rank}\,(\tilde{A}) = \tilde{q}} \ \sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij} - \tilde{a}_{ij})^2 \tag{2.2}$$

is achieved by

$$A^* = \sum_{\alpha=1}^{\tilde q} \psi_\alpha u_\alpha w_\alpha^T$$

where

$$A = \sum_{\alpha=1}^{q} \psi_\alpha u_\alpha w_\alpha^T$$

has svd (2.1). The minimal value of (2.2) is

$$\sum_{\alpha=\tilde q+1}^{q} \psi_\alpha^2 .$$

This suggests the following measure of goodness of fit of the rank $\tilde q$ approximation A^* to A

(2.3) $$\text{gof}\ (\tilde q) = \sum_{\alpha=1}^{\tilde q} \psi_\alpha^2 \Big/ \sum_{\alpha=1}^{q} \psi_\alpha^2 .$$

Since the ψ_α 's are ordered by magnitude, it is easy to see that

$$\tilde q/q \le \text{gof}(\tilde q) \le 1.$$

In practice it is not necessary to calculate all ψ_α to evaluate (2.3); we could use the identity

$$\sum_{\alpha=1}^{q} \psi_\alpha^2 = \text{trace}\ (AA^T).$$

2.2. Reciprocal averaging

Our general approach to the analysis of association between rows and columns of a real matrix A of size n×m and of rank q will be via reciprocal averaging. Reciprocal averaging constructs q sets of row scores, column scores and proportionality constants, denoted

$$x_\alpha = (x_{1\alpha}, x_{2\alpha}, \ldots, x_{n\alpha})^T,$$
$$y_\alpha = (y_{1\alpha}, y_{2\alpha}, \ldots, y_{m\alpha})^T,$$
$$\lambda_\alpha,$$

for $\alpha = 1,\ldots,q$.

The matrix A has elements a_{ij} which reflect the association of rows and columns. Associated with A are two diagonal matrices, R of size $n\times n$ and C of size $m\times m$, where the diagonal elements r_i and c_j are positive row and column weights. Thus R and C are non-singular.

DEFINITION 2.1. A solution of *reciprocal averaging* applied to A with respect to R and C consists of q sets of row scores x_α, column scores y_α, and proportionality constants λ_α which satisfy for $\alpha = 1,\ldots,q$

$$\begin{aligned} \lambda_\alpha x_\alpha &= R^{-1} A y_\alpha \\ \lambda_\alpha y_\alpha &= C^{-1} A^T x_\alpha \end{aligned} \tag{2.4}$$

where λ_α is maximal subject to

$$\begin{aligned} x_\alpha^T R x_\alpha &= 1,\ y_\alpha^T C y_\alpha = 1 \\ x_\alpha^T R x_\beta &= 0,\ y_\alpha^T C y_\beta = 0,\ \beta = 1,2,\ldots,\alpha-1. \end{aligned} \tag{2.5}$$

For a better understanding of equations (2.5), consider the following. Given a vector y_α of column scores $y_{j\alpha}$ and a proportionality constant λ_α, we compute a vector x_α of row scores as

$$x_\alpha = \frac{1}{\lambda_\alpha} R^{-1} A\, y_\alpha .$$

That is, row scores $x_{i\alpha}$ are proportional to $(R^{-1} Ay_\alpha)_i$, a weighted sum of column scores $y_{j\alpha}$ with weights a_{ij}/r_i. Similarly, column scores y_α are weighted sums of the x_α 's. Side conditions (2.5) simply require the q sets of scores to be mutually orthonormal with respect to the weights matrices.

The method of reciprocal averages was first named in HORST (1935); however, he and other authors (eg. HILL (1974) and NISHISATO (1980)) chose the weights matrices R and C as the row and column sums of A, respectively.

The existence of scores satisfying definition 2.1 is demonstrated by the following proposition.

PROPOSITION 2.1. *Suppose that the matrix* $R^{-\frac{1}{2}} AC^{-\frac{1}{2}}$ *has svd*

$$R^{-\frac{1}{2}} AC^{-\frac{1}{2}} = U \Psi W^T. \tag{2.6}$$

Then a solution of reciprocal averaging applied to A *with respect to* R *and* C *is given by*

$$x_\alpha = R^{-\frac{1}{2}} u_\alpha$$
$$y_\alpha = C^{-\frac{1}{2}} w_\alpha$$
$$\lambda_\alpha = \psi_\alpha$$

for $\alpha = 1,\ldots,q$.

PROOF. Multiply (2.6) on the right by w_α and on the left by $R^{-\frac{1}{2}}$ to show that

$$\lambda_\alpha x_\alpha = R^{-1} A y_\alpha.$$

Multiply (2.6) on the right by $C^{-\frac{1}{2}}$ and on the left by u_α to show that

$$\lambda_\alpha y_\alpha^T = x_\alpha^T AC^{-1}.$$

Side conditions (2.5) are easily verified. Since the singular values ψ_α in (2.6) are ordered by magnitude, it follows that λ_α is maximal. □

COROLLARY 2.2. *Reciprocal averaging scores* x_α *and* y_α *are right eigenvectors of the matries* $R^{-1}AC^{-1}A^T$ *and* $C^{-1}A^TR^{-1}A$, *respectively, corresponding to eigenvalue* λ_α^2, *for* $\alpha = 1,2,\ldots,q = \text{rank}(A)$.

Furthermore, the solution of reciprocal averaging is unique whenever the svd of $R^{-\frac{1}{2}}AC^{-\frac{1}{2}}$ is.

2.3. Interpretation of graphical representation

A graphical representation of the row scores and column scores produced by reciprocal averaging should provide insight into the structure of the matrix A. The graphical representation consists of n+m vectors in a q-dimensional space, where the i-th row's vector is

$$\xi_i = (\lambda_1^s x_{i1}, \lambda_2^s x_{i2}, \ldots, \lambda_q^s x_{iq})^T \tag{2.7}$$

and the j-th column's vector is

$$\eta_j = (\lambda_1^t y_{j1}, \lambda_2^t y_{j2}, \ldots, \lambda_q^t y_{jq})^T, \tag{2.7'}$$

where s and t are fixed constants.

The facts about geometrical interpretation listed below depend to a certain extent on the choices of s and t; these facts allow general guidelines for interpretation of ξ_i's and η_j's to be drawn. More specific applications of these facts to the special cases of PCA, CA and HA are discussed in sections 3.2, 4.3 and 5.3.

It follows from proposition 2.1 and corollary 2.2 that

$$R^{-1}AC^{-1} = X \Lambda Y^T,$$

$$R^{-1}AC^{-1}A^TR^{-1} = X \Lambda^2 X^T,$$

$$C^{-1}A^TR^{-1}AC^{-1} = Y \Lambda^2 Y^T.$$

These equations lead to the following facts.

FACT 1. *If* s+t = 1, *then*

$$\xi_i^T \eta_j = a_{ij}/r_i c_j \qquad \textit{for} \quad i = 1,\ldots,n;\ j = 1,\ldots,m.$$

FACT 2. *If* s = 1, *then*

$$\xi_i^T \xi_\ell = \sum_{j=1}^{m} \frac{a_{ij} a_{\ell j}}{r_i r_\ell c_j} \qquad \textit{for} \quad i,\ell = 1,\ldots,n.$$

FACT 2'. *If* t = 1, *then*

$$\eta_j^T \eta_h = \sum_{i=1}^{n} \frac{a_{ij} a_{ih}}{r_i c_j c_h} \qquad \textit{for} \quad h,j = 1,\ldots,m.$$

FACT 3. *If* s = 1, *then*

$$\mathrm{dist}^2(\xi_i,\xi_\ell) = \sum_{j=1}^{m} \frac{1}{c_j}\left\{\frac{a_{ij}^2}{r_i^2} + \frac{a_{\ell j}^2}{r_\ell^2} - 2\,\frac{a_{ij}a_{\ell j}}{r_i r_\ell}\right\} \text{ for } i,\ell = 1,\ldots,n.$$

FACT 3'. *If* t = 1, *then*

$$\mathrm{dist}^2(\eta_j,\eta_h) = \sum_{i=1}^{n} \frac{1}{r_i}\left\{\frac{a_{ij}^2}{c_j^2} + \frac{a_{ih}^2}{c_h^2} - 2\,\frac{a_{ij}a_{ih}}{c_j c_h}\right\} \text{ for } h,j = 1,\ldots,m.$$

From the side conditions (2.5) we formulate the following fact.

FACT 4. *The quantity*

$$r_i\, x_{i\alpha}^2$$

can be interpreted as the contribution of row i *to component* α, *for* $i = 1,\ldots,n;\ \alpha = 1,\ldots,q$.

FACT 4'. *The quantity*

$$c_j y_{j\alpha}^2$$

can be interpreted as the contribution of column j *to component* α , *for* $j = 1,\ldots,m;\ \alpha = 1,\ldots,q$.

2.4. Lower dimensional approximation

In most applications of reciprocal averaging, only the first $\tilde{q}$ out of a possible q components are calculated. Instead of an exact q-dimensional representation, we will consider an approximate $\tilde{q}$-dimensional representation in which the i-th row's vector is

$$\tilde{\xi}_i = (\lambda_1^s x_{i1}, \lambda_2^s x_{i2}, \ldots, \lambda_{\tilde{q}}^s x_{i\tilde{q}})^T$$

and in which the j-th column's vector is

$$\tilde{\eta}_j = (\lambda_1^t y_{j1}, \lambda_2^t y_{j2}, \ldots, \lambda_{\tilde{q}}^t y_{j\tilde{q}})^T.$$

Facts 1 through 3' of section 2.3 with $\tilde{\xi}_i$ replacing ξ_i and with $\tilde{\eta}_j$ replacing

η_j are approximations. The quality of these approximations may be assessed by the overall goodness of fit measure (2.3) and by the following measures for the i-th row

$$\text{(2.8)} \qquad \mathrm{gof}(\tilde{\xi}_i) = \tilde{\xi}_i^T \tilde{\xi}_i / \xi_i^T \xi_i,$$

and for the j-th column

$$\text{(2.8')} \qquad \mathrm{gof}(\tilde{\eta}_j) = \tilde{\eta}_j^T \tilde{\eta}_j / \eta_j^T \eta_j.$$

Clearly these measures range between 0 and 1; unlike the overall measure, for these measures no sharper lower bound can be given.

The application of facts 1 through 3' with $\tilde{\xi}_i$ or $\tilde{\eta}_j$ may be misleading for rows i or columns j for which the goodness of fit (2.8) or (2.8') is low.

3. PRINCIPAL COMPONENT ANALYSIS

3.1. Reciprocal averaging formulations of PCA

Association among m numerical variables is often expressed in an m×m covariance matrix, S, say of rank q. Reciprocal averaging applied to

$$A = S$$

(3.1) with respect to

$$R = C = I_m$$

yields column scores y_α and proportionality constants λ_α for $\alpha = 1,\ldots,q$ which are exactly the normalized eigenvectors (principal components) and the corresponding eigenvalues of S (by proposition 2.1). Since A is symmetric and since R = C, row scores x_α are identical to column scores y_α.

If the covariance matrix is based on n observations on m variables, these observations may be arranged in an n×m matrix B. Reciprocal averaging applied to

$$A = (I_n - \frac{1}{n} ee^T)\, B$$

(3.2) with respect to

$$R = n\, I_n,\ C = I_m$$

produces column scores y_α and proportionality constants λ_α which are eigenvectors and square roots of eigenvalues of

$$S = \frac{1}{n} A^T A$$

as above. Reciprocal averaging on (3.2) also produces scores x_α for the individuals (observations).

Alternatively, association among the m numerical variables can be expressed in a correlation matrix. We denote the average of the b_{ij}'s over i by $b_{\cdot j}$, and the variance of the j-th variable by s_j^2, where

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (b_{ij} - b_{\cdot j})^2 .$$

Reciprocal averaging applied to

$$A = (I_n - \frac{1}{n} ee^T)\, B\, \mathrm{diag}(\frac{1}{s_j})$$

(3.3) with respect to

$$R = n\, I_n,\ C = I_m$$

produces column scores y_α and proportionality constants λ_α that are respectively eigenvectors and square roots of eigenvalues of the $m \times m$ correlation matrix of the columns of B.

3.2. Graphical interpretation

For PCA in terms of reciprocal averaging on (3.2), the choices s = 0 and t = 1 in (2.7) and (2.7') allow especially convenient interpretation of the graphical representation. Our discussion will focus on formulation (3.2); the interpretation rules concerning variables are valid in the contex of formulation (3.1) if $s = t = \frac{1}{2}$.

Thus we interpret the $\tilde{q}$-dimensional graphical representation consisting of n vectors for the rows

$$\tilde{\xi}_i = (x_{i1}, x_{i2}, \ldots, x_{i\tilde{q}})^T \qquad \text{for } i = 1,\ldots,n$$

and m vectors for the columns

$$\tilde{\eta}_j = (\lambda_1 y_{j1}, \lambda_2 y_{j2}, \ldots, \lambda_{\tilde{q}} y_{j\tilde{q}})^T \quad \text{for } j = 1,\ldots,m.$$

The plot of $\tilde{\xi}_i$ and $\tilde{\eta}_j$ for $\tilde{q} = 2$ is precisely the principal component biplot described by GABRIEL (1971, section 3). We describe the interpretation rules as special cases of the facts of section 2.3. Many interpretation rules we give below are well known in the context of PCA and biplot.

One should take into account the goodness of fit measures when applying these rules. The goodness of fit measure (2.8') for the j-th variable is

$$\text{gof}(\tilde{\eta}_j) = \sum_{\alpha=1}^{\tilde{q}} \lambda_\alpha^2 y_{j\alpha}^2 \Big/ s_j^2.$$

The overall goodness of fit measure (2.3) is

$$\text{gof}(\tilde{q}) = \sum_{j=1}^{m} s_j^2 \,\text{gof}(\tilde{\eta}_j) \Big/ \sum_{j=1}^{m} s_j^2.$$

PCA-i.r. 1. *The inner product of* $\tilde{\eta}_j$ *with* $\tilde{\xi}_i$ *approximates* $\frac{1}{n}(b_{ij} - b_{.j})$. *In particular, vectors* $\tilde{\xi}_i$ *which are in the same direction as* $\tilde{\eta}_j$ *correspond to individuals who are above average on the* j-th *variable.* (fact 1).

PCA-i.r. 2. *The squared length of* $\tilde{\eta}_j$ *approximates the variance of the* j-th *variable.* (fact 2').

PCA-i.r. 3. *The cosine of the angle between* $\tilde{\eta}_j$ *and* $\tilde{\eta}_h$ *approximates the correlation of the* h-th *and* j-th *variable. Thus variable vectors which are in nearly the same direction (or nearly opposite directions, or nearly orthogonal) indicate variables with high positive (high negative, very weak) cor-*

relation. (fact 2').

PCA-i.r. 4. *The squared distance between* $\tilde{\eta}_j$ *and* $\tilde{\eta}_h$ *approximates the variance of the difference between the* j-th *and* h-th *variable.*. (fact 3).

PCA-i.r. 5. *The contribution of the* j-th *variable to the* α-th *component is* $y_{j\alpha}^2$, *the squared loading of the* j-th *variable on the* α-th *principal component.* (fact 4').

Direct application of corollary 2.2 yields

$$SY = Y \Lambda^2$$

which leads to an additional interpretation rule.

PCA-i.r. 6. *The covariance between the* j-th *variable and the* α-th *principal component is* $\lambda_\alpha^2 \, y_{j\alpha}$.

Furthermore, we have

$$XX^T = R^{-1}AY\Lambda^{-2}Y^TA^TR^{-1} = \frac{1}{n^2}(I_n - \frac{1}{n}ee^T)\, BS^{-1}B^T(I_n - \frac{1}{n}ee^T).$$

PCA-i.r. 7. *The squared distance between two individual vectors* $\tilde{\xi}_i$ *and* $\tilde{\xi}_\ell$ *approximates a standardized distance*

$$\frac{1}{n^2}(b_i - b_\ell)\, S^{-1}(b_i - b_\ell)^T,$$

where b_i *denotes the* i-th *row of* B.

4. CORRESPONDENCE ANALYSIS

4.1. Reciprocal averaging formulation of CA

Correspondence analysis describes the association among categories of two nominal variables, V_1 with m_1 categories and V_2 with m_2 categories. Information relevant to the association is summarized in an $m_1 \times m_2$ contingency table F of frequencies. We define diagonal matrices containing the marginals of F as

$$N_1 = \text{diag}\,(n_{1(h)})$$

where

$$n_{1(h)} = \sum_{j=1}^{m_2} f_{hj} \qquad \text{for } h = 1,\ldots,m_1,$$

and

$$N_2 = \text{diag}\,(n_{2(j)})$$

where

$$n_{2(j)} = \sum_{h=1}^{m_1} f_{hj} \qquad \text{for } j = 1,\ldots,m_2.$$

DEFINITION 4.1. *Correspondence analysis* applied to F is defined to be reciprocal averaging applied to

$$A = \frac{1}{n} F$$

(4.1) with respect to

$$R = \frac{1}{n} N_1, \; C = \frac{1}{n} N_2$$

where

$$n = \sum_{h=1}^{m_1} \sum_{j=1}^{m_2} f_{hj}.$$

From (4.1) it can be seen that CA constructs a score for the h-th row which is proportional to a weighted average of column scores, with weights

$$f_{hj}/n_{1(h)},$$

the conditional probability of column j given row h. The matrix F is of rank $q \le \min(m,n)$; the following proposition shows that CA on F yields at most q-1 components relevant to the association structure of F.

PROPOSITION 4.1. *The reciprocal averaging row scores* x_α, *column scores* y_α *and proportionality constants* λ_α *from* CA *satisfy*

$$0 \le \lambda_\alpha \le 1 \qquad \text{for } \alpha = 1,\ldots,q,$$

and can be chosen to satisfy

$$(4.2) \qquad \lambda_1 = 1,\ x_1 = e,\ y_1 = e.$$

PROOF. The row sums of $N_1^{-1}FN_2^{-1}F^T$ ($R^{-1}AC^{-1}A^T$ of corollary 2.2) are identically 1. An upper bound for eigenvalues of a non-negative matrix is the maximal row sum (cf. WILKINSON (1965), p.58). The eigenvalues are all real so the singular values must satisfy $0 \le \lambda_\alpha \le 1$. Furthermore, (4.2) satisfies (2.4) and (2.5) in the case of (4.1). □

In CA the first trivial component will be discarded; the second and higher components will be inspected to gain insight into the structure of F.

Furthermore, we note that

$$\frac{1}{n}\cdot\chi^2(F) = \sum_{\alpha=2}^{q} \lambda_\alpha^2$$

where

$$\chi^2(F) = n\left\{\sum_{h=1}^{m_1}\sum_{j=1}^{m_2} \frac{f_{hj}^2}{n_{1(h)}n_{2(j)}} - 1\right\},$$

the square contingency of F (cf. KENDALL & STUART (1979), p.587).

It can be shown (cf. HILL (1974)) that CA is algebraically equivalent to Fisher's contingency table analysis (cf. FISHER (1940)). Fisher's method, equivalently formulated by HIRSCHFELD (1935), was to assign scores to the categories of the nominal variables V_1 and V_2 such that the correlation between the derived numerical variables should be maximal. This approach was revived in the late 1960's by Benzécri and his associates. Thus the first non-trivial component from CA gives scores that yield maximal correlation, and further components give scores that yield maximal correlations subject to orthogonality to previous sets of scores.

4.2. Alternate formulations of CA

We now consider two alternate formulations of CA which yield the same scores for categories of V_1 and V_2 and closely related proportionality constants.

PROPOSITION 4.2. *The concatenated vectors* $(x_\alpha^T, y_\alpha^T)^T$, *where* x_α *and* y_α *are the row and column scores of* CA *on* F, *for* $\alpha = 1,\ldots,q$ *are identical to the first* q *sets of scores from reciprocal averaging applied to*

$$A = \frac{1}{4n}\begin{pmatrix} N_1 & F \\ F^T & N_2 \end{pmatrix}$$

(4.3) *with respect to*

$$R = C = \frac{1}{2n}\begin{pmatrix} N_1 & 0 \\ 0 & N_2 \end{pmatrix}.$$

Furthermore, proportionality constants λ_α *from* CA *are related to the first* q *proportionality constants of reciprocal averaging on* (4.3); *the latter are given by* $(1+\lambda_\alpha)/2$.

PROOF. Suppose that $m_1 \geq m_2$; otherwise apply CA to F^T. Define x_α for $\alpha = 1,\ldots,q,q+1,\ldots,m_2,m_2+1,\ldots,m_1$ to be eigenvectors of $N_1^{-1}FN_2^{-1}F^T$; define y_α for $\alpha = 1,\ldots,q,q+1,\ldots,m_2$ to be eigenvectors of $N_2^{-1}F^TN_1^{-1}F$; define λ_α^2 for $\alpha = 1,2,\ldots,m_1$ to be the corresponding eigenvalues (in decreasing order by magnitude). Clearly $\lambda_\alpha = 0$ for $\alpha = q+1,\ldots,m_1$.

Corollary 2.2 shows $x_\alpha, y_\alpha, \lambda_\alpha$ for $\alpha = 1,\ldots,q$ to be a solution of CA (i.e. reciprocal averaging on (4.1)). The corollary also shows that scores from reciprocal averaging on (4.3) are given by eigenvectors of

$$R^{-1}AC^{-1}A^T = \frac{1}{4}\begin{pmatrix} I_{m_1} + N_1^{-1}FN_2^{-1}F^T & 2N_1^{-1}F \\ 2N_2^{-1}F^T & I_{m_2} + N_2^{-1}F^TN_1^{-1}F \end{pmatrix}. \tag{4.4}$$

It is easily checked that the following are eigenvectors of (4.4):

$(x_\alpha^T,y_\alpha^T)^T$ for $\alpha = 1,\dots,q$, corresponding to eigenvalues $(1+\lambda_\alpha)^2/4$

$(x_\alpha^T,-y_\alpha^T)^T$ for $\alpha = 1,\dots,q$, corresponding to eigenvalues $(1-\lambda_\alpha)^2/4$

$(x_\alpha^T,y_\alpha^T)^T$ and $(x_\alpha^T,-y_\alpha^T)^T$ for $\alpha = q+1,\dots,m_2$,
corresponding to eigenvalue 1/4

$(x_\alpha^T,0^T)^T$ for $\alpha = m_2+1,\dots,m_1$,
corresponding to eigenvalue 1/4.

Therefore the first q sets of scores, corresponding to eigenvalues $(1+\lambda_\alpha)^2/4$, or to singular values $(1+\lambda_\alpha)/2$, are exactly the scores from CA (4.1). □

The contingency table F represents n observations on the variables V_1 and V_2; an alternate expression is by two indicator matrices, G_1 of size $n\times m_1$ and G_2 of size $n\times m_2$. A 1 in the i-th row and j-th column of indicator matrix G_k represents the selection of the j-th category of variable V_k by the i-th individual. All the other elements of the i-th row of G_k are 0. We may construct the contingency table F from the indicator matrices by

$$F = G_1^T G_2.$$

Reciprocal averaging on these indicator matrices is equivalent to CA on F, as shown below.

<u>PROPOSITION 4.3.</u> *The concatenated vectors* $(x_\alpha^T,y_\alpha^T)^T$, *where* x_α *and* y_α *are row and column scores of* CA *on* F, *i.e.* (4.1), *for* $\alpha = 1,\dots,q$ *are identical to the first* q *sets of column scores from reciprocal averaging applied to*

$$A = \frac{1}{2n} G$$

(4.5) *with respect to*

$$R = \frac{1}{n} I_n, \quad C = \frac{1}{2n}\begin{pmatrix} N_1 & 0 \\ 0 & N_2 \end{pmatrix},$$

where G *is the* $n\times(m_1+m_2)$ *matrix*

$$G = (G_1,G_2).$$

Furthermore the proportionality constants are derived from λ_α *of* CA *by*

$$\sqrt{(1+\lambda_\alpha)/2}.$$

PROOF. Analogous to the proof of proposition 4.2. □

Although the scores for the categories are the same in the CA formulations (4.1), (4.3) and (4.5), their graphical representations differ in a stretching or contracting of the axes. Moreover the goodness of fit measures derived from the CA formulations (4.3) and (4.5) will be lower than those of the formulation (4.1).

Another difference between formulations (4.1) and (4.5) is the creation of individual scores in (4.5). HILL (1974) calls (4.1) zero-order CA and (4.5) first-order CA. LEBART et.al. (1977) speak of CA in (4.1) as analysis of a "tableau de contingence", (4.3) as analysis of a "tableau de Burt", and (4.5) as analysis of a "tableau disjonctif".

4.3. Graphical interpretation

We consider graphical representation based on scores from (4.1), the traditional formulation of CA. Constants s and t in formulas (2.7) and (2.7') are both equal to 1. In a $\tilde{q}$-dimensional approximation, the h-th row of F is represented by

$$\tilde{\xi}_h = (\lambda_2 x_{h2}, \lambda_3 x_{h3}, \ldots, \lambda_{\tilde{q}+1} x_{h\tilde{q}+1})^T$$

and the j-th column of F by

$$\tilde{\eta}_j = (\lambda_2 y_{j2}, \lambda_3 y_{j3}, \ldots, \lambda_{\tilde{q}+1} y_{j\tilde{q}+1})^T$$

since the first component is trivial. The interpretation rules listed below are valid for these vectors, provided that their goodness of fit measures are high. These goodness of fit measures should be adjusted for the deletion of the first (trivial) component. Thus the overall measure (2.3) becomes

$$\mathrm{gof}(\tilde{q}) = \sum_{\alpha=2}^{\tilde{q}+1} \lambda_\alpha^2 \Big/ \sum_{\alpha=2}^{q} \lambda_\alpha^2 = n \sum_{\alpha=2}^{\tilde{q}+1} \lambda_\alpha^2 \Big/ \chi^2(F).$$

Similarly the row and column measures (2.8) and (2.8') become

$$\mathrm{gof}(\tilde{\xi}_h) = \sum_{\alpha=2}^{\tilde{q}+1} \lambda_\alpha^2 x_{h\alpha}^2 \Bigg/ \sum_{j=1}^{m_2} \frac{n}{n_{2(j)}} \left(\frac{f_{hj}}{n_{1(h)}} - \frac{n_{2(j)}}{n} \right)^2$$

and

$$\mathrm{gof}(\tilde{\eta}_j) = \sum_{\alpha=2}^{\tilde{q}+1} \lambda_\alpha^2 y_{j\alpha}^2 \Bigg/ \sum_{h=1}^{m_1} \frac{n}{n_{1(h)}} \left(\frac{f_{hj}}{n_{2(j)}} - \frac{n_{1(h)}}{n} \right)^2$$

(cf. CA-i.r. 2, below).

<u>CA-i.r. 1.</u> *The squared length of* $\tilde{\eta}_j$ *is approximately*

$$\sum_{h=1}^{m_1} \frac{n}{n_{1(h)}} \left(\frac{f_{hj}}{n_{2(j)}} - \frac{n_{1(h)}}{n} \right)^2 .$$

Thus when the conditional distribution of variable $V_1|V_2 = j$ *is similar to the marginal distribution of* V_1, *then* $\tilde{\eta}_j$ *is near the origin. In this case, in terms of the conditional distribution of* $V_1|V_2 = j$, *category* j *of* V_2 *is average.* (fact 2').

Analogously, the squared length of $\tilde{\xi}_h$ *indicates the extent to which the conditional distribution of* $V_2|V_1 = h$ *is similar to the marginal distribution of* V_2. (fact 2).

<u>CA-i.r. 2.</u> *The squared distance between* $\tilde{\eta}_j$ *and* $\tilde{\eta}_k$ *approximates*

$$\sum_{h=1}^{m_1} \frac{n}{n_{1(h)}} \left(\frac{f_{hj}}{n_{2(j)}} - \frac{f_{hk}}{n_{2(k)}} \right)^2 .$$

Thus whenever the conditional distributions of $V_1|V_2 = j$ *and* $V_1|V_2 = k$ *are similar,* $\tilde{\eta}_j$ *and* $\tilde{\eta}_k$ *will be close to each other.* (fact 3').

Analogously, the squared distance between $\tilde{\xi}_h$ *and* $\tilde{\xi}_\ell$ *indicates the similarity of the conditional distributions of* $V_2|V_1 = h$ *and* $V_2|V_1 = \ell$. (fact 3).

These distances are called χ^2-distances in the literature.

<u>CA-i.r. 3.</u> *The contribution of the* h-th *row* (j-th *column*) *of* F *to the* α-th *component is given by* $\frac{n_{1(h)}}{n} x_{h\alpha}^2$ (*by* $\frac{n_{2(j)}}{n} y_{j\alpha}^2$). (fact 4.4').

The interpretation of the graphical representation of CA formulation

(4.5) (and of (4.3)) is described in section 5.3 in the context of HA; however there are some differences (compare CA-i.r. 1 with HA-i.r. 2, and CA-i.r. 2 with HA-i.r. 4).

5. HOMOGENEITY ANALYSIS

5.1. Reciprocal averaging formulation of HA

Homogeneity analysis describes the association among categories of p nominal variables, say $V_1, V_2, \ldots, V_p$, where variable V_k has m_k categories. The p-dimensional contingency table F of size $m_1 \times m_2 \times \ldots \times m_p$ is constructed from n observations on variables $V_1, \ldots, V_p$. We denote by F_{jk} the $m_j \times m_k$ table of bivariate marginals of variables V_j and V_k. Notice that $F_{jk} = F_{kj}^T$. Also, we denote F_{jj}, the $m_j \times m_j$ diagonal matrix with univariate marginals for the variable V_j, by N_j. We define $m = \sum_{j=1}^{p} m_j$; the m×m super-diagonal matrix of N_j's will be denoted by N.

DEFINITION 5.1. *Homogeneity analysis applied to* F *is defined to be reciprocal averaging applied to*

$$A = \frac{1}{np^2}\begin{pmatrix} N_1 & F_{12} & \cdots & F_{1p} \\ F_{21} & N_2 & \cdots & F_{2p} \\ \vdots & \vdots & & \vdots \\ F_{p1} & F_{p2} & \cdots & N_p \end{pmatrix}$$

(5.1) with respect to

$$R = C = \frac{1}{np} N.$$

Clearly HA is a generalization of CA formulation (4.3) to the case of more than two variables. HA considers only bivariate associations; higher order associations may be studied by combining variables: replace V_j and V_k by one variable with $m_j \times m_k$ categories. Scores for the categories of the variables $V_1, \ldots, V_p$ are given by column scores from HA. As in CA, the first component is trivial.

PROPOSITION 5.1. *The column scores* y_α *and proportionality constant* λ_α *from* HA *satisfy*

$$0 \le \lambda_\alpha \le 1 \quad \textit{for} \quad \alpha = 1,..,q = \mathrm{rank}(A)$$

and can be chosen to satisfy

$$\lambda_1 = 1,\ y_1 = e.$$

PROOF. Analogous to the proof of proposition 4.1. □

Scores for the k-th category of variable V_j on the α-th component are denoted by $y_{j(k),\alpha}$; the $m_j \times 1$ vector of scores for the categories of V_j is denoted by $y_{j,\alpha}$; the marginal frequency of the k-th category of variable V_j is denoted $n_{j(k)}$.

Reciprocal averaging scores y_α, $\alpha \ge 2$, must have weighted average zero. In fact, these scores satisfy a stronger requirement.

PROPOSITION 5.2. *Column scores* y_α *from* HA *satisfy*

$$e^T N_j y_{j,\alpha} = 0 \quad \textit{for} \quad j = 1,..,p;\ \alpha = 2,..,q.$$

PROOF. $e^T N_j y_{j,\alpha} = \frac{1}{\lambda_\alpha}\frac{1}{p}\, e^T N_j N_j^{-1}(F_{j1},\ldots,F_{jp})y_\alpha = \frac{1}{\lambda_\alpha}\frac{1}{p}\, e^T N y_\alpha = 0.$ □

Since the weighted average of scores for each variable is zero, the scores for the two categories of a dichotomous variable must be of opposite sign.

The first non-trivial set of column scores y_α from HA are scores for nominal variables $V_1, V_2, \ldots, V_p$ such that the first principal component of the correlation matrix has maximal variance (see HILL(1974)). Further sets of reciprocal averaging scores are more difficult to interpret in this context.

5.2. An alternate formulation

An alternate representation of the n observations in the p-dimensional contingency table F is by p indicator matrices G_j of size $n \times m_j$ for

$j = 1,\dots,p$, or by $G = (G_1,G_2,\dots,G_p)$.

PROPOSITION 5.3. *The* q *sets of column scores from reciprocal averaging applied to*

$$A = \frac{1}{np} G$$

(5.2) *with respect to*

$$R = \frac{1}{n} I_n, \quad C = \frac{1}{np} N$$

are identical to scores from HA, *i.e.* (5.1). *Proportionality constants for* (5.2) *are the square roots of those derived from* (5.1).

PROOF. Straightforward. □

Proposition 5.3. shows that HA, formulated as (5.2), considers only bivariate marginals. Formulation (5.2) produces scores for each individual in addition to category scores and proportionality constants produced by formulation (5.1). Another difference between these two formulations is that (5.2) can be easily extended to handle missing observations (cf. GIFI (1981), p.116). This extension can be problematic, however, since the first component will no longer be trivial. This would affect the geometric interpretation rules as well as measures of goodness of fit. Other ways to handle missing observations, such that the first component remains trivial, are given in GIFI (1981), p.70. We restrict attention to the case of no missing data.

5.3. Graphical interpretation

HA is usually formulated as (5.2) and the graphical representation is constructed with $s = 0$ and $t = 1$. We discard the trivial solution from (5.2) with $\lambda_1 = 1$, $x_1 = e$, $y_1 = e$ and represent the i-th row of G in $\tilde{q}$ dimensions by

$$\tilde{\xi}_i = (x_{i2},x_{i3},\dots,x_{i\tilde{q}+1})^T;$$

similarly, the k-th category of the j-th variable is represented in $\tilde{q}$ dimensions by

$$\tilde{\eta}_{j(k)} = (\lambda_2 y_{j(k),2}, \ldots, \lambda_{\tilde{q}+1} y_{j(k),\tilde{q}+1})^T .$$

The interpretation rules below apply to vectors $\tilde{\xi}_i$ and $\tilde{\eta}_{j(k)}$ provided that goodness of fit measures are high. The overall goodness of fit of the $\tilde{q}$-dimensional representation is

$$\mathrm{gof}(\tilde{q}) = \sum_{\alpha=2}^{\tilde{q}+1} \lambda_\alpha^2 \Big/ (\frac{m}{p} - 1).$$

The goodness of fit measure for $\tilde{\eta}_{j(k)}$ is

$$\mathrm{gof}(\tilde{\eta}_{j(k)}) = \sum_{\alpha=2}^{\tilde{q}+1} \lambda_\alpha^2 \, n_{j(k)} y_{j(k),\alpha}^2 \Big/ (n - n_{j(k)})$$

(cf. HA-i.r. 2, below); the goodness of fit of variable V_j may be defined as the weighted average

$$\mathrm{gof}(\tilde{V}_j) = \sum_{k=1}^{m_j} \frac{n-n_{j(k)}}{n_{j(k)}} \mathrm{gof}(\tilde{\eta}_{j(k)}) \Big/ \sum_{k=1}^{m_j} \frac{n-n_{j(k)}}{n_{j(k)}} .$$

<u>HA-i.r. 1.</u> *The inner product of* $\tilde{\xi}_i$ *and* $\tilde{\eta}_{j(k)}$ *approximates* $n/n_{j(k)}$ *if individual* i *selected category* k *on variable* V_j, *and zero otherwise. Thus individual points are generally in the same direction as points representing the categories selected by the individual.* (fact 1).

<u>HA-i.r. 2.</u> *The squared length of* $\tilde{\eta}_{j(k)}$ *approximates*

$$(n-n_{j(k)})/n_{j(k)} .$$

Thus categories with large marginals frequently appear near the origin, while those with small marginal frequency are far from the origin. (fact 2').

<u>HA-i.r. 3.</u> *The inner product of* $\tilde{\eta}_{j(k)}$ *and* $\tilde{\eta}_{h(\ell)}$ *approximates*

$$\frac{n \, f_{j(k),h(\ell)}}{n_{j(k)} \; n_{h(\ell)}} - 1 .$$

Thus if j=h, *then the cosine of the angle between* $\tilde{\eta}_{j(k)}$ *and* $\tilde{\eta}_{h(\ell)}$ *approximates* *the correlation between two cells of a multinominal. For* $j \neq h$, *if* $\tilde{\eta}_{j(k)}$ *and* $\tilde{\eta}_{h(\ell)}$ *are nearly orthogonal (or in the same direction, or in opposite directions), then categoty* k *of* V_j *and category* ℓ *of* V_h *are weakly associated (or positively associated, or negatively associated).* (fact 2).

HA-i.r. 4. *The squared distance between* $\tilde{\eta}_{j(k)}$ *and* $\tilde{\eta}_{j(\ell)}$ *approximates*

$$n \frac{n_{j(k)} + n_{j(\ell)}}{n_{j(k)}\; n_{j(\ell)}}.$$

HA-i.r. 5. *The squared distance between* $\tilde{\eta}_{j(k)}$ *and* $\tilde{\eta}_{h(\ell)}$ *approximates*

$$n \frac{n_{j(k)} + n_{h(\ell)} - 2\, f_{j(k),h(\ell)}}{n_{j(k)}\; n_{h(\ell)}}.$$

Thus categories with high joint frequency (indicative of strong positive association) are plotted near each other. (fact 3').

HA-i.r. 6. *The contribution of the* k-th *category of* V_j *to the* α-th *component is*

$$\frac{n_{j(k)}}{n} \frac{1}{p}\, y^2_{j(k),\alpha}.$$

Furthermore we compute the total contribution for categories of one variable. This total contribution, often called the discrimination of V_j *on component* α, *is defined as*

$$\text{discr}\ (V_j,\alpha) = \sum_{k=1}^{m_j} \frac{n_{j(k)}}{n\ p}\, y^2_{j(k),\alpha}.$$

(fact 4').

The reciprocal averaging formulas (2.4) in HA lead to

$$\xi_{i\alpha} = \frac{1}{\lambda_\alpha^2} \frac{1}{p} \sum_{j=1}^{p} \sum_{k=1}^{m_j} g_{i,j(k)}\; \eta_{j(k),\alpha}$$

and

$$\eta_{j(k),\alpha} = \sum_{i=1}^{n} \frac{1}{n_{j(k)}}\, g_{i,j(k)}\; \xi_{i\alpha}.$$

This suggests another interpretation rule.

HA-i.r. 7. *A category point* $\tilde{\eta}_{j(k)}$ *is always exactly the center of gravity of the individual points* $\tilde{\xi}_i$ *for individuals* i *who selected the* k-th *category on* V_j.

Furthermore, we have

$$1 - \frac{1}{n} \sum_{k=1}^{m_j} \sum_{i=1}^{n} g_{i,j(k)} \, (\xi_{i\alpha} - \eta_{j(k),\alpha})^2 =$$

$$= \frac{2}{n} \sum_{k=1}^{m_j} \sum_{i=1}^{n} g_{i,j(k)} \, \xi_{i\alpha} \, \eta_{j(k),\alpha} - \frac{1}{n} \sum_{k=1}^{m_j} \sum_{i=1}^{n} g_{i,j(k)} \, \eta^2_{j(k),\alpha} =$$

$$= \frac{1}{n} \sum_{k=1}^{m_j} n_{j(k)} \, \eta^2_{j(k),\alpha} = p \, \lambda^2_\alpha . \mathrm{discr}(V_j,\alpha).$$

This leads to the following interpretation rule.

HA-i.r. 8. *When the discrimination of* V_j *on the* α-th *component is high, then on component* α, *for each* k, *scores* $\xi_{i\alpha}$ *for individuals who selected category* k *of variable* V_j *are near* $\eta_{j(k),\alpha}$.

6. AVAILABLE COMPUTER PROGRAMS

As shown in section 2.2, calculation for PCA, CA and HA require a singular value decomposition of $R^{-\frac{1}{2}}AC^{-\frac{1}{2}}$ and scaling of the resulting singular vectors by diagonal matrices. Thus these techniques can be performed with computer subroutines that are widely available. Further, various authors have developed special purpose programs for one or more of these techniques. These programs may offer various advantages, such as efficiency (for example, for HA, the required core storage is reduced by making use of the 0 – 1 nature of the indicator matrices) and convenience (output may include a line printer plot of pairs of components and additional information like goodness of fit and discrimination measures). However, various normalizations of the scores exist. We are aware of the following programs :

- biplot program of GABRIEL; PCA

- programs in LEBART et.al. (1977); also scheduled to appear as part of a package in Compstat '82; PCA, CA, HA
- programs in BENZÉCRI (1973), CA
- *canals* package of dept. of data theory, University of Leiden, contains PCA, CA, HA (c.f. DE LEEUW & VAN RIJCKEVORSEL (1980))
- *prinqual* program of TENENHAUS (1977)
- programs in NISHISATO (1980).

7. EXAMPLE OF HA

FIENBERG (1977, p.91) reproduces a four-dimensional contingency table, summarizing 4831 automobile accidents, taken from KIHLBERG et.al. (1964). The four variables are as follows: V_1 = *accident type (collision with vehicle, collision with object, rollover without collision, other rollover);* V_2 = *accident severity (not severe, moderately severe, severe);* V_3 = *driver ejected (not ejected, ejected)* and V_4 = *car type (small, compact, standard).* The main diagonal of table 7.1 gives the univariate marginals $n_{j(k)}$ for the variables $V_1,..,V_4$. The triangular portion of table 7.1 gives bivariate marginals $f_{j(k),h(\ell)}$. After division by np^2 = 77296, table 7.1 gives the upper portion of the symmetric matrix A of (5.1). Table 7.2 summarizes the results of HA formulation (5.2) with $\tilde{q}$ = 2.

V_1				V_2			V_3		V_4				
2526	0	0	0	1620	629	277	2325	201	151	234	2141	*Collision with vehicle*	V_1
	1195	0	0	745	315	135	1075	120	54	110	1031	*Collision with object*	
		454	0	128	252	74	293	161	79	57	318	*Rollover without collision*	
			656	121	333	202	431	225	66	69	521	*Other rollover*	
				2614	0	0	2436	178	186	269	2159	*Not severe*	V_2
					1529	0	1237	292	119	145	1265	*Moderately severe*	
						688	451	237	45	56	587	*Severe*	
							4124	0	274	398	3452	*Not ejected*	V_3
								707	76	72	559	*Ejected*	
									350	0	0	*Small*	V_4
										470	0	*Compact*	
											4011	*Standard*	

V_1: *Accident type*
V_2: *Accident severity*
V_3: *Driver ejected*
V_4: *Car type*

Table 7.1. Bivariate and univariate marginal frequencies.

$\lambda_2 = 0.639$ $\lambda_3 = 0.533$	gof(2) = 0.346

Variable	Category	Category scores: first non-trivial component	Category scores: second non-trivial component	Goodness of fit: first two non-trivial components	Discrimination: first non-trivial component	Discrimination: second non-trivial component
V_1: *Accident type*	*Collision with vehicle*	0.700	-0.001	0.219	0.364	0.420
	Collision with object	0.569	-0.456	0.063		
	Rollover without collision	-2.042	3.646	0.564		
	Other rollover	-2.318	-1.691	0.472		
V_2: *Accident severity*	*Not severe*	0.998	0.078	0.481	0.310	0.302
	Moderately severe	-0.918	0.994	0.289		
	Severe	-1.751	-2.505	0.503		
V_3: *Driver ejected*	*Not ejected*	0.452	0.070	0.495	0.298	0.007
	Ejected	-2.637	-0.408	0.495		
V_4: *Car type*	*Small*	-1.161	3.386	0.297	0.027	0.263
	Compact	-0.076	1.011	0.025		
	Standard	0.110	-0.414	0.259		
		$y_{j(k),2}$	$y_{j(k),3}$	$gof(\tilde{n}_{j(k)})$	$discr(V_j,2)$	$discr(V_j,3)$

Table 7.2 Results of HA with $\tilde{q} = 2$.

In many examples involving one or more ordinal variables, the natural ordering of the categories is reflected by HA scores on the first non-trivial component. Theoretical discussions on the retrieval of order relations and dependence structure appear in SCHRIEVER (1982 a,b). We first inspect the ordering of categories of $V_1,\ldots,V_4$ implied by the scores $y_{j(k),2}$. For *accident type* (V_1), the two *collision* categories receive nearly equal positive scores whereas the two *rollover* categories receive similar negative scores. The scale for V_2 reproduces the expected ordering of the three categories, with *not severe* scoring near the *collision* categories of V_1, and with *severe* near the *rollover* categories, and with *moderately severe* intermediate. Similarly, the category *not ejected* of V_3 scores near *collisions* while *ejected* scores near *rollovers*. The scale for V_4, *car type*, also reproduces the expected ordering, although the spread of these scores is relatively small. The discrimination measures for the first non-trivial component show that the first dimension involves V_1,V_2 and V_3 but not V_4. Thus the first dimension suggests that V_2,V_3 and V_1 (at least contrasting *collisions* with *rollovers)* are more strongly associated with each other than with V_4. For further detail, we investigate higher dimensions.

Figure 7.1, a plot of the points $\tilde{\eta}_{j(k)}$ for the first two non-trivial components, is constructed with $t = 1$. Inspection of this two dimensional representation provides additional insight. For example, a line through the three points for V_4, *car type*, is roughly orthogonal to a line through two points for V_3, *ejected* or *not*. Thus V_3 and V_4 are nearly independent. Furthermore, if V_1 was reduced to two categories, *collisions* and *rollovers*, this variable would be nearly independent of V_4, but strongly associated with V_3. The points for *rollover without collision* (V_1) and *small* (V_4) are both rather extreme in the upper left quadrant, indicating strong association between these categories. Upon inspection of the contingency table or bivariate marginals, this association is also evident; however, its detection is simplified by figure 7.1. Analogously, there is evidence of positive association among *ejected* (V_3), *other rollover* (V_1) and *severe* (V_2), all in the lower left quadrant. Five points to the right of the origin in figure 7.1, *not ejected, not severe, standard* and the two *collision* categories also seem to show positive association. But goodness of fit measures must be inspected to help validate this interpretation, since these points are all close to the origin. Two of the twelve categories are poorly represented in the plane, *collision with object* (V_1) and *compact* (V_4). The proximity to the origin of four of the five category points mentioned above is primarily due to large marginal frequencies rather than to poor fit. Thus very strong positive association among *not ejected* (V_3), *not severe* (V_2) and *collision with vehicle* (V_1) is indicated. These categories are positively but less strongly related to *standard* (V_4). The three categories of V_1 with reasonable goodness of fit in the plane are each strongly posively associated with one of the categories of V_2.

With gof(2) = 0.346 it is reasonable to inspect the third and higher dimensions for additional features of interest. The next two proportionality constants, λ_4 = 0.504 and λ_5 = 0.500, are similar to λ_3. Table 7.3 gives discrimination measures for these dimensions.

	discr(V_j,2)	discr(V_j,3)	discr(V_j,4)	discr(V_j,5)
V_1	0.364	0.420	0.573	0.004
V_2	0.310	0.302	0.231	0.056
V_3	0.298	0.007	0.004	0.005
V_4	0.027	0.263	0.189	0.756

Table 7.3. Discrimination measures

The third dimension (i.e. α=4) involves V_1 , and to a lesser extent, V_2 and V_4. The main effect of this dimension is to fit the category of V_1 which was poorly fitted in the plane. The next dimension (α=5) involves almost exclusively V_4; this dimension only provides a better fit for the category *compact*. Thus these dimensions give us no further insight into the associations among the variables.

Figure 7.2 is a plot of the points $\tilde{\xi}_i$ for the first two non-trivial components. The 71 points represent the profiles of the 4831 observed accidents; one possible profile (*collision with object-severe-ejected-small*) was not observed. It is important for correct interpretation of figure 7.2 to keep in mind that each plotted point represents a number of identical observations; thus in application of HA-i.r. 7, each $\tilde{\xi}_i$ must be weighted by the frequency of the profile (a cell frequency in the original four dimensional contingency table). Figure 7.2 shows seven nearly vertical bands of points, with the number of points in each band a multiple of three (exept where the profile which was not observed should be). Each band is characterized by one or more combinations of the categories of V_1,V_2 and V_3 with all levels of V_4. For example, the left band corresponds to profiles *other rollover - severe - ejected.* The next band corresponds to profiles *other rollover - moderately severe - ejected* and *rollover without collision - severe - ejected.* At the other extreme, the band furthest toward the right corresponds to profiles with *collision with vehicle or object - not severe - not ejected.* Figure 7.3 is a copy of figure 7.2 with lines indicating the bands. The profiles of accidents may be ordered by a partial ordering into seven groups from *other rollover - severe - ejected* at one extreme to *collision - not severe - not ejected* at the other.

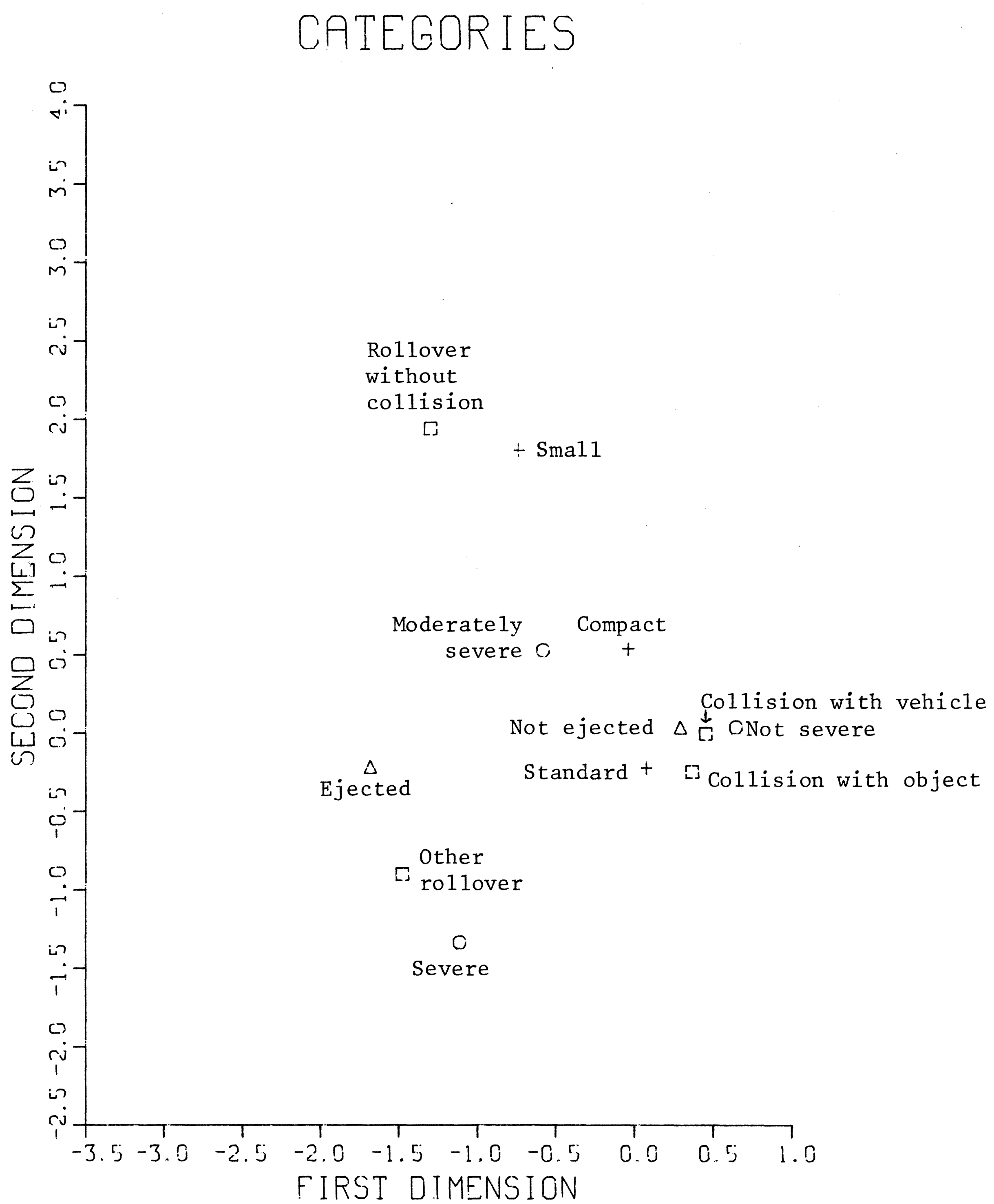
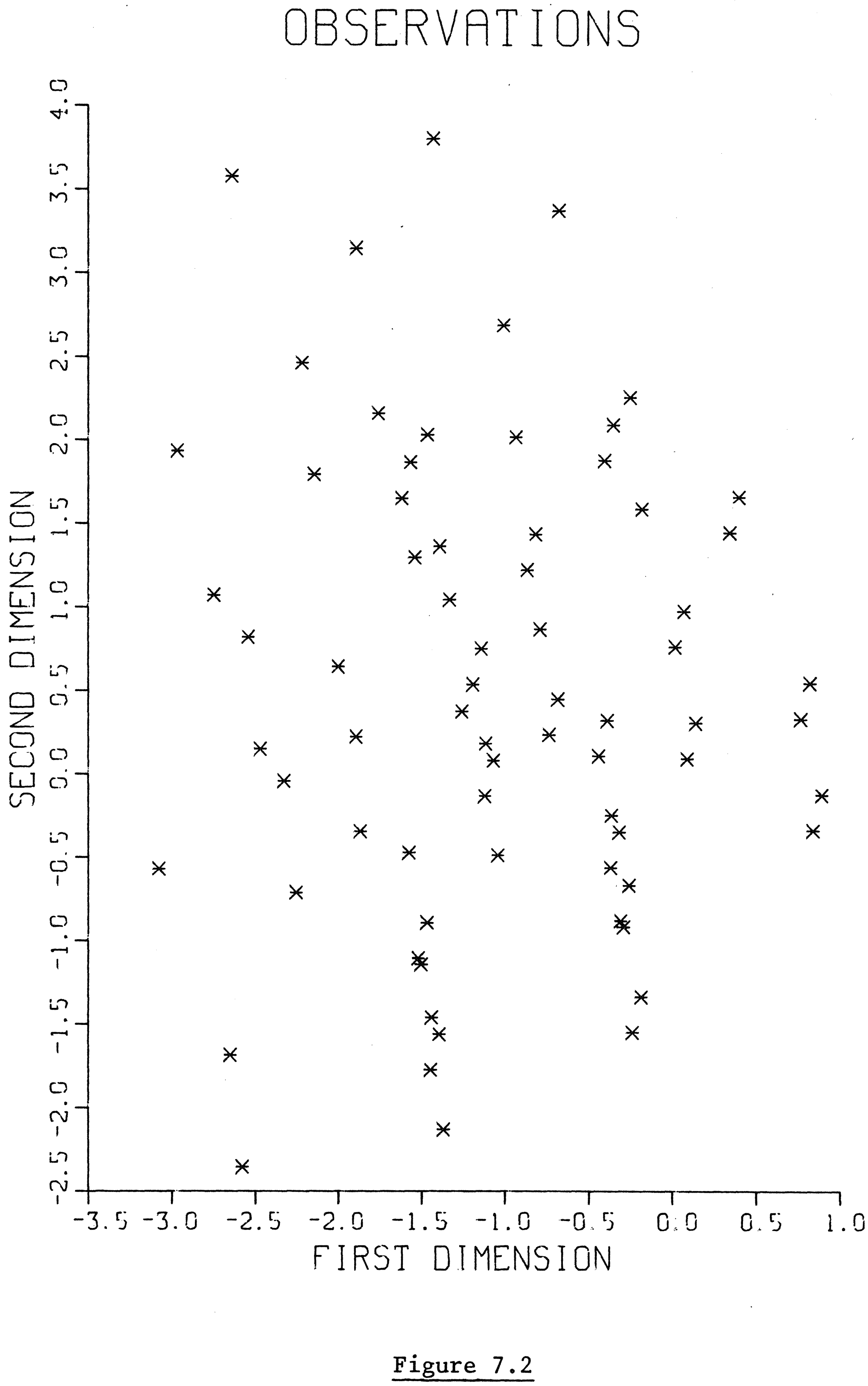
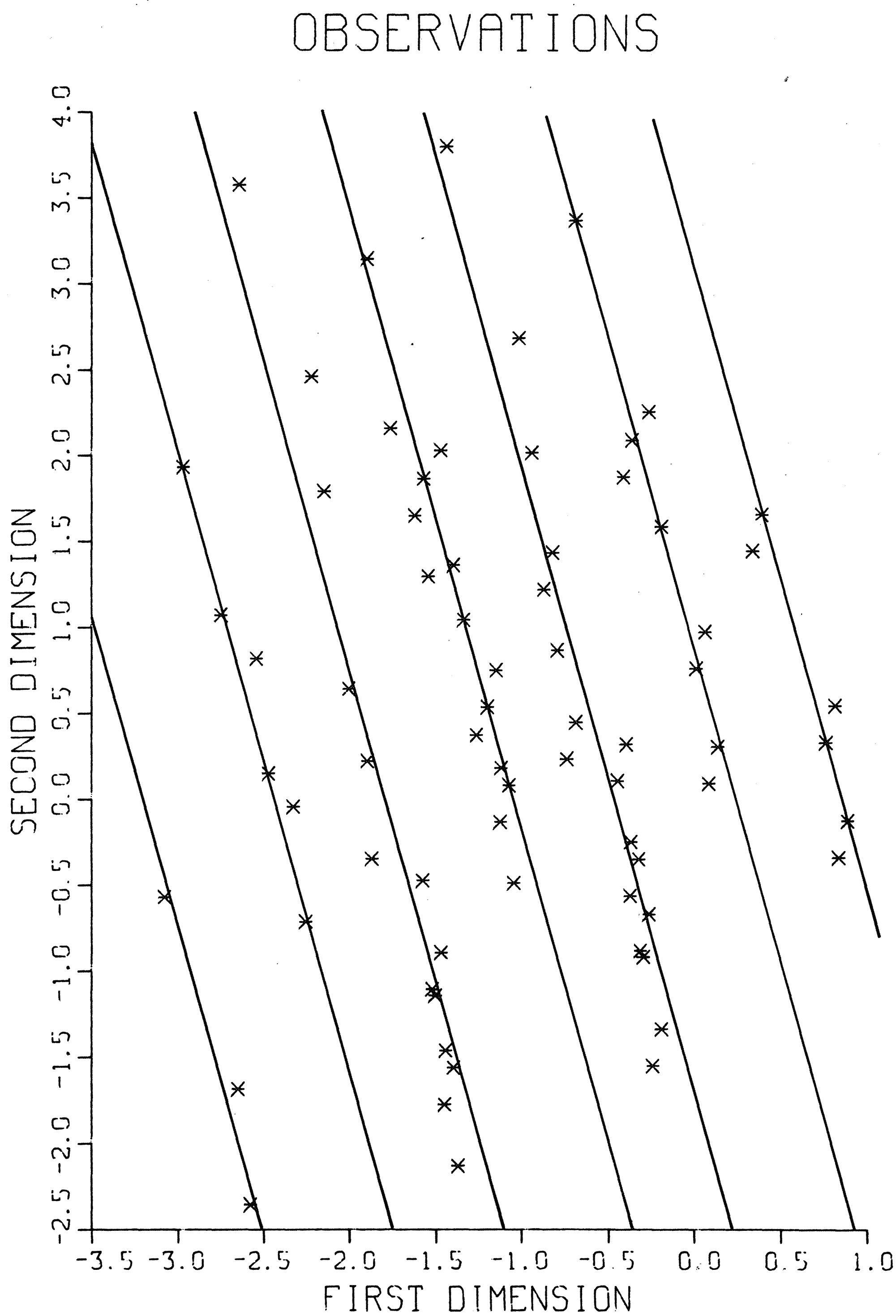

CATEGORIES
SECOND DIMENSION
FIRST DIMENSION
Rollover without collision
Small
Moderately severe
Compact
Collision with vehicle
Not ejected
Not severe
Standard
Collision with object
Ejected
Other rollover
Severe
-3.5 -3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0
-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

Figure 7.1


OBSERVATIONS
SECOND DIMENSION
FIRST DIMENSION
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

<u>**Figure 7.2**</u>



OBSERVATIONS
SECOND DIMENSION
FIRST DIMENSION
-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
-3.5 -3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0

<u>Figure 7.3</u>

ACKNOWLEDGEMENT. *The authors are grateful to* Dr. R.D. GILL *and* Prof.Dr. J. OOSTERHOFF *for their valuable remarks. We are also grateful to* BELL TELEPHONE LABORATORIES *for the first author's leave of absence.*

REFERENCES

[1] BENZÉCRI, J.P. (1973), *L' Analyse des données* II: *l' Analyse des correspondances*, Dunod, Paris.

[2] FIENBERG, S.E. (1977), *The analysis of cross- classified categorical data*, The MIT Press, Cambridge, Massachusetts.

[3] FISHER, R.A. (1940), *The precision discriminant functions*, Ann. Eugen. London, 10, p. 422-429.

[4] GABRIEL, K.R. (1971), *The biplot graphic display of matrices with application to principal component analysis*, Biometrika, 58, p. 453-467.

[5] GIFI, A. (1981), *Non-linear multivariate analysis*, Dept. of data theory University of Leiden.

[6] HILL, M.O. (1974), *Correspondence analysis: a neglected multivariate method*, Appl. Statist. 23, p. 340-354.

[7] HIRSCHFELD, H.O. (1935), *A connection between correlation and contingency*, Proc. Camb. Phil. Soc., 31, p. 520-524.

[8] HORST, P. (1935), *Measuring complex attitudes*, Journal of Soc. Psychology, 6, p. 369-374.

[9] HOUSEHOLDER, A.S. & YOUNG, G. (1938), *Matrix approximation and latent roots*, Am. Math. Monthly, 45, p. 165-171.

[10] KIHLBERG, J.K., NARRAGON, E.A. & B.J. CAMPBELL, (1964), *Automobile crash injury in relation to car size*, Cornell Aero. Lab. Report No. VJ -1823- R11.

[11] KENDALL, M.G. & A. STUART, (1979), *The advanced theory of statistics*, vol II, 4-th ed., Hafner, New York.

[12] LEBART, L., MORINEAU, A. & N. TABARD (1977), *Techniques de la déscription statistique*, Dunod, Paris.

[13] DE LEEUW, J. & J. VAN RYCKEVORSEL (1980), *Homals and princals: some generalisations of principal component analysis*, in Data analysis and informatics, ed. E. DIDAY et.al., North Holland, Amsterdam, p. 231-242.

[14] MAAS-DE WAAL, C., SCHRIEVER, B.F. & D. SIKKEL, *Toepassing van niet-lineaire technieken bij de analyse van maatschappelijke participatie*, C.B.S. report in preparation, Staatsuitgeverij 's-Gravenhage.

[15] NISHISATO, S. (1980), *Analysis of categorical data: dual scaling and its applications*, Univ. of Toronto Press, Toronto.

[16] RAO, C.R. (1973), *Linear statistical inference and its applications*, Wiley, New York.

[17] SCHRIEVER, B.F. (1982 a), *Ordering properties in correspondence analysis*, Mathematisch Centrum report SW 80-82, Amsterdam.

[18] SCHRIEVER, B.F. (1982 b), *Scaling of order dependent categorical variables with correspondence analysis*, Mathematisch Centrum report SW 83-82, Amsterdam.

[19] TENENHAUS, M. (1977), *Analyse en composantes principales d'un ensemble de variables nominales ou numériques*, Revue de Stat. Appliquée, 25, p. 39-56.

[20] WILKINSON, J.H. (1965), *The algebraic eigenvalue problem*, Clarendon, Oxford.

MC NR

35235