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

5

Linear Constraints in Correspondence Analysis

Ulf Böckenholt and Yoshio Takane

5.1 INTRODUCTION

In any correspondence analysis (CA) study, external information is of paramount importance. External information may refer, for example, to prior knowledge about the row and column categories as well as expectations about their multidimensional representation. At a minimum, external information is necessary for an interpretation of the graphical representation, for instance, when labeling the dimensions. We may also include external information in the graphical display by fitting supplementary profiles (Greenacre 1984, p. 70) or by constraining the configuration. Both approaches are important tools in highlighting interesting features of the data that otherwise may be overlooked. However, in contrast to the use of supplementary points which do not affect the CA solution, the direct incorporation of external information in the form of constraints often leads to simplified multidimensional displays (Bentler and Weeks 1978, Carroll *et al.* 1980, Heiser 1981, p. 235, Ramsay 1982, Takane 1981). As a result, when analyzing the dependence between rows and columns, constraints are useful in the search for meaningful patterns, particularly in large data sets. The implementation and application of such constraints is the topic of this chapter.

Perhaps the simplest application of constrained CA is to explore whether row and column scores satisfy the order implied by the categories' labels (see Table 5.1), or, more parsimoniously, follow a linear order (Goodman 1991, Gilula and Haberman 1988). Although CA makes no assumptions about the spacing of the row and/or column scores, it is straightforward to estimate the scores under a linear order constraint and to compare the constrained solution

TABLE 5.1
Cross-classification of mental health status and parental socioeconomic status (Srole *et al.* 1962, p. 213).

	Parental socioeconomic status stratum						
Mental health category	A	B	C	D	E	F	Row totals
Well	64	57	57	72	36	21	307
Mild symptom formation	94	94	105	141	97	71	602
Moderate symptom formation	58	54	65	77	54	54	362
Impaired	46	40	60	94	78	71	389
Column totals	262	245	287	384	265	217	1,660

Note: A is the highest and F is the lowest socioeconomic status category

with its unconstrained counterpart. If the expectation about the spacing proves adequate, differences to the optimal solution are minimal and more than offset by the simplified and more parsimonious representation of the data. More generally, when the row or column categories form an (incomplete) factorial design, dummy or contrast variables used to code that design may be applied to constrain the graphical representation (Nishisato 1980). For example, Delbeke (1978) constructed different family compositions by factorially combining the number of sons and the number of daughters (which ranged from 0 to 3), and asked 82 students to rank order the 16 compositions according to their preference. In a constrained correspondence analysis of this data set, Takane *et al.* (1991a; see also Heiser 1981, p. 167) recoded the 16 family types as a factorial combination of the number of children and gender bias (defined as the difference between number of sons and daughters). Takane *et al*'s results showed that interactions between both factors can be ignored and that, in support of theoretical notions about family composition preferences (Coombs *et al.* 1973), subjects arrived at their preference judgments for the 16 family types by adding their separate utilities for the two factors gender bias and number of children. Many other applications with constrained configurations can be found in the multidimensional scaling literature. For example, Ekman's (1954) similarity ratings among pairs of 14 spectral hues are well described by a two-dimensional representation of the color circle. An analysis of Torgerson's (1958, p. 286) similarity data obtained for nine Munsell colors yields also a two-dimensional representation that corresponds closely to the colors' brightness and saturation (Takane 1978, Takane *et al.* 1991a). Clearly, the inclusion of known physical properties of the row and column categories in CA may not only reduce considerably the number of parameters to be estimated but may also lead to a much simplified interpretation of the data.

In this chapter we distinguish three general applications for imposing constraints in a CA. First, concomitant variables may be used to explain the association structure in the table. The equidistant spacing constraint is a simple example for this approach because it yields a readily interpretable representation of the results. Second, it may prove beneficial to partial out the effect of concomitant variables from a CA solution (Böckenholt and Böckenholt 1990, Gilula and Haberman 1986; van der Heijden *et al.* 1989). Third, in some applications it may be important to first partial out the effects of a subset of the concomitant variables and then to relate the residual information to the association structure in the table (ter Braak 1988). By incorporating external information through linear constraints on the row and/or column scores in these various ways, a representation of a contingency table is obtained that is not only more parsimonious but is also easier to understand. As a result, applications of constrained CA may prove especially useful in exploratory analyses of a contingency table (Escoufier and Junca 1986).

This chapter reviews and illustrates these three approaches for imposing linear constraints in a CA. Most of the theoretical results presented here can be found in Böckenholt and Böckenholt (1990), Golub and Underwood (1970), Rao (1964), Takane and Shibayama (1991), and Takane *et al.* (1991b). In particular, we refer to the last reference for rigorous derivations and comparisons between seemingly different approaches for incorporating linear constraints in the analysis of a contingency table.

5.2 CORRESPONDENCE ANALYSIS WITH LINEAR CONSTRAINTS

Correspondence analysis is a useful tool for obtaining a graphical display of the dependence between the rows and columns of a contingency table (e.g. Benzécri *et al.* 1980, Gifi 1990, Greenacre 1984, Lebart *et al.* 1984, Nishisato 1980). We first describe CA without constraints to introduce the notation used in this chapter. Consider an I by J contingency table $\mathbf{P}$ with proportions p_{ij} describing the joint distribution of two random categorical variables, X and Y, with I and J categories, respectively. Let $\mathbf{D_r}$ and $\mathbf{D_c}$ be diagonal matrices containing the row and column sums of $\mathbf{P}$, respectively. CA is the generalized singular value decomposition (GSVD) of

$$\mathbf{A} = \mathbf{D_r}^{-1}(\mathbf{P} - \mathbf{E})\mathbf{D_c}^{-1} = \mathbf{R}\mathbf{D}_\lambda\mathbf{C}' \tag{5.1}$$

with $\mathbf{E} = \mathbf{D_r}\mathbf{1}\,\mathbf{1}'\mathbf{D_c}$ (where $\mathbf{1}$ is a unit vector), and $\mathbf{D}_\lambda$ is a diagonal matrix with $\min(I-1, J-1)$ singular values λ in descending order. The sum of the squared

singular values is called the total inertia, and is equal to the χ^2-statistic for 'independence' divided by the sample size, n:

$$\sum \lambda_i^2 = \frac{\chi^2}{n}$$

The standard row and column coordinates (Greenacre 1984, p. 94), $\mathbf{R}$ and $\mathbf{C}$, satisfy the restrictions $\mathbf{R}'\mathbf{D}_r\mathbf{R} = \mathbf{I} = \mathbf{C}'\mathbf{D}_c\mathbf{C}$ and $\mathbf{1}'\mathbf{D}_r\mathbf{R} = \mathbf{0} = \mathbf{1}'\mathbf{D}_c\mathbf{C}$. In practice, the coordinates are computed by an ordinary SVD of the matrix $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{E})\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}' \tag{5.2}$$

with $\mathbf{U}'\mathbf{U} = \mathbf{I} = \mathbf{V}'\mathbf{V}$, and

$$\mathbf{R} = \mathbf{D}_r^{-1/2}\mathbf{U} \qquad \text{and} \qquad \mathbf{C} = \mathbf{D}_c^{-1/2}\mathbf{V}$$

The principal coordinates (Greenacre 1984, p. 90) are obtained by post-multiplying the standard scores by $\mathbf{D}_\lambda$. Usually, the interpretation of the data is based on a low-dimensional, graphical representation of the standard or the principal coordinates and is guided by the available background information about the row and column categories. In many applications, however, it may prove useful to explicitly take into account this background information when estimating the scores. As a result, the interpretation of a constrained representation is straightforward and differences between constrained and unconstrained solutions may point to unexplained features of the data. Linear constraints may be imposed by either the null-space or the reparametrization method (Böckenholt and Böckenholt 1990, Takane *et al.* 1991b). Both approaches can give identical results but because in some applications one method may be easier to use than the other, we review both procedures in the next two subsections.

5.2.1 The null-space method

According to the null-space method linear row and column constraints are defined by

$$\mathbf{G}'\mathbf{R}^* = \mathbf{0} \qquad \text{and } \mathbf{H}'\mathbf{C}^* = \mathbf{0}$$

where $\mathbf{G} = (\mathbf{D}_r\mathbf{1} \mid \mathbf{G}_*)$ is a known $I \times K$ matrix of rank K. Similarly, $\mathbf{H} = (\mathbf{D}_c\mathbf{1} \mid \mathbf{H}_*)$ is a known $J \times L$ matrix of rank L. The effects defined by the matrices $\mathbf{G}$ and $\mathbf{H}$ are partialed out from the standard row and column scores denoted by $\mathbf{R}^*$ and $\mathbf{C}^*$, respectively, by computing the complementary projection operators $\mathbf{Q}_r$ and $\mathbf{Q}_c$

$$\mathbf{Q}_r = \mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{D}_r^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{D}_r^{-1}$$

and

$$\mathbf{Q}_c = \mathbf{I} - \mathbf{H}(\mathbf{H}'\mathbf{D}_c^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}_c^{-1}$$

The constrained standard scores are then obtained by the SVD of

$$\mathbf{Z}^* = \mathbf{D}_r^{-1/2}\mathbf{Q}_r(\mathbf{P} - \mathbf{E})\mathbf{Q}_c'\mathbf{D}_c^{-1/2} = \mathbf{U}^*\mathbf{D}_\lambda^*\mathbf{V}^{*\prime} \tag{5.3}$$

with $\mathbf{U}^{*\prime}\mathbf{U}^* = \mathbf{I} = \mathbf{V}^{*\prime}\mathbf{V}^*$, yielding

$$\mathbf{R}^* = \mathbf{D}_r^{-1/2}\mathbf{U}^* \quad \text{and} \quad \mathbf{C}^* = \mathbf{D}_c^{-1/2}\mathbf{V}^*$$

and, consequently, $\mathbf{R}^{*\prime}\mathbf{D}_r\mathbf{R}^* = \mathbf{I} = \mathbf{C}^{*\prime}\mathbf{D}_c\mathbf{C}^*$ and $\mathbf{1}'\mathbf{D}_r\mathbf{R}^* = \mathbf{0} = \mathbf{1}'\mathbf{D}_c\mathbf{C}^*$. If constraints are imposed only on the row scores $\mathbf{H} = \mathbf{D}_c\mathbf{1}$ and, similarly, if constraints are imposed only on the column scores $\mathbf{G} = \mathbf{D}_r\mathbf{1}$. Thus, the matrices, $\mathbf{Z}^*$ in (5.3) and $\mathbf{Z}$ in (5.2), are identical when $\mathbf{G} = \mathbf{D}_r\mathbf{1}$ and $\mathbf{H} = \mathbf{D}_c\mathbf{1}$.

5.2.3 *The reparametrization method*

A second approach for imposing linear constraints on the standard row and column scores is given by

$$\mathbf{MR}_s = \mathbf{R}^* \quad \text{and} \quad \mathbf{NC}_s = \mathbf{C}^*$$

where $\mathbf{M} = (\mathbf{1} \mid \mathbf{M}_*)$ is a known $I \times K$ matrix of rank K and $\mathbf{N} = (\mathbf{1} \mid \mathbf{N}_*)$ is a known $J \times L$ matrix of rank L. The matrices $\mathbf{R}_s$ and $\mathbf{C}_s$ contain the reduced set of the scores. Thus, in contrast to the null-space method the constrained standard row and column scores are obtained by directly reparametrizing the unconstrained scores. The constrained standard scores are determined by computing the projection operators $\mathbf{O}_r$ and $\mathbf{O}_c$ as

$$\mathbf{O}_r = \mathbf{D}_r\mathbf{M}(\mathbf{M}'\mathbf{D}_r\mathbf{M})^{-1}\mathbf{M}'$$

and

$$\mathbf{O}_c = \mathbf{D}_c\mathbf{N}(\mathbf{N}'\mathbf{D}_c\mathbf{N})^{-1}\mathbf{N}'$$

and performing the SVD of

$$\mathbf{Z}^* = \mathbf{D}_r^{-1/2}\mathbf{O}_r(\mathbf{P} - \mathbf{E})\mathbf{O}_c'\mathbf{D}_c^{-1/2} = \mathbf{U}^*\mathbf{D}_\lambda^*\mathbf{V}^{*\prime} \tag{5.4}$$

with $\mathbf{U}^{*\prime}\mathbf{U}^* = \mathbf{I} = \mathbf{V}^{*\prime}\mathbf{V}^*$, and

$$\mathbf{R}^* = \mathbf{D}_r^{-1/2}\mathbf{U}^* \quad \text{and} \quad \mathbf{C}^* = \mathbf{D}_c^{-1/2}\mathbf{V}^*$$

By setting $\mathbf{M}$ and $\mathbf{N}$ equal to an identity matrix we obtain the unconstrained CA solution.

5.2.3 *Relationships between both methods*

Because one can determine $\mathbf{N}(\mathbf{M})$ for a given $\mathbf{H}(\mathbf{G})$ and vice versa both the null-space and the reparametrization method can yield the same or orthocomplement results by appropriately defining the constraint matrices (Takane *et al.* 1991b). For example, the reparametrization and the null-space method

give identical results in the case of row constraints when $\mathbf{O}_r = \mathbf{Q}_r$, and in the case of column constraints when $\mathbf{O}_c = \mathbf{Q}_c$. In contrast, if we impose constraints on the row scores, a residual analysis of the reparametrization method with $\mathbf{MR}_s = \mathbf{R}^*$ is equivalent to the null-space method with $\mathbf{G} = \mathbf{D}_r\mathbf{M}$, and a residual analysis of the null-space method with $\mathbf{G}'\mathbf{R}^* = \mathbf{0}$ is equivalent to the reparametrization method with $\mathbf{M} = \mathbf{D}_r^{-1}\mathbf{G}$. In either case, $\mathbf{O}_r = \mathbf{I} - \mathbf{Q}_r$, or

$$\mathbf{D}_r\mathbf{M}(\mathbf{M}'\mathbf{D}_r\mathbf{M})^{-1}\mathbf{M}' = \mathbf{G}(\mathbf{G}'\mathbf{D}_r^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{D}_r^{-1}$$

Similarly, in the case of column constraints, the null-space and the reparametrization methods yield complementary results when $\mathbf{NC}_s = \mathbf{C}^*$ and $\mathbf{H} = \mathbf{D}_c\mathbf{N}$, or when $\mathbf{H}'\mathbf{C}^* = \mathbf{0}$ and $\mathbf{N} = \mathbf{D}_c^{-1}\mathbf{H}$.

Clearly, the actual choice of the reparametrization or the null-space method depends only on the empirical application. The reparametrization method seems more natural when we want to directly constrain the coordinates, while the null-space method seems more natural when we want to exclude the effects of certain variables in interpreting a CA solution.

To summarize, we can decompose the $\mathbf{A}$ matrix in (5.1) into four components,

$$\begin{aligned}\mathbf{A} = \mathbf{D}_r^{-1}(&\mathbf{O}_r(\mathbf{P} - \mathbf{E})\mathbf{O}_c' + \mathbf{O}_r(\mathbf{P} - \mathbf{E})(\mathbf{I} - \mathbf{O}_c') \\ &+ (\mathbf{I} - \mathbf{O}_r)(\mathbf{P} - \mathbf{E})\mathbf{O}_c' + (\mathbf{I} - \mathbf{O}_r)(\mathbf{P} - \mathbf{E})(\mathbf{I} - \mathbf{O}_c)')\mathbf{D}_c^{-1}\end{aligned}$$

Each component refers to a particular effect of the constraints. To quantify the effects of these constraints the total inertia, $\Sigma\lambda_i^2$, may be decomposed into the corresponding four parts:

$$\begin{aligned}\Sigma\lambda_i^2 = &\operatorname{tr}(\mathbf{A}'\mathbf{O}_r\mathbf{D}_r\mathbf{A}\mathbf{O}_c\mathbf{D}_c) + \operatorname{tr}(\mathbf{A}'\mathbf{O}_r\mathbf{D}_r\mathbf{A}(\mathbf{I} - \mathbf{O}_c)\mathbf{D}_c) \\ &+ \operatorname{tr}(\mathbf{A}'(\mathbf{I} - \mathbf{O}_r)\mathbf{D}_r\mathbf{A}\mathbf{O}_c\mathbf{D}_c) + \operatorname{tr}(\mathbf{A}'(\mathbf{I} - \mathbf{O}_r)\mathbf{D}_r\mathbf{A}(\mathbf{I} - \mathbf{O}_c)\mathbf{D}_c)\end{aligned}$$

The first component gives the part of the inertia obtained when both row and column scores are constrained, the sum of the first and second component equal the part of the inertia when only the row scores are constrained, and the sum of the first and third component equal the part of the inertia when only the column scores are constrained. Thus, the ratio of the first component and the total inertia gives the proportion of the χ^2-statistic that is accounted for when both row and column constraints are imposed.

In some applications it may prove useful to combine the ideas underlying the reparametrization and the null-space method. For example, a set of concomitant variables for the row scores may be divided into two subsets denoted by $\mathbf{X}_1$ and $\mathbf{X}_2$ and one may be interested in examining the effects of $\mathbf{X}_2$ while statistically controlling for the effects of $\mathbf{X}_1$. This is accomplished by first partialing out the effects of $\mathbf{X}_1$ from $\mathbf{X}_2$:

$$\mathbf{X}_2^* = (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{D}_r\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{D}_r)\mathbf{X}_2$$

In the next step the residual information is related to the association structure

in the table by setting $\mathbf{M}_*$ equal $\mathbf{X}_2^*$ in (5.4) (ter Braak 1988). Obviously, a similar procedure can be applied for the analysis of the column scores.

Only one set of constraints can be imposed by the reparametrization or the null-space method. Occasionally, it may be more appropriate to impose different sets of constraints on the scores corresponding to each singular value. For instance, it may be useful to impose uniform spacing on the scores of the first singular vector but equality constraints on the scores of the second singular vector (for an application see Gilula and Haberman 1986). Different constraints can be introduced by extracting the row and column scores corresponding to the first singular value λ_1^* and computing the rank-one reduced matrix $\mathbf{Z}_1^*$:

$$\mathbf{Z}_1^* = (\mathbf{I} - \mathbf{u}_1^* \mathbf{u}_1^{*\prime})(\mathbf{P} - \mathbf{E})(\mathbf{I} - \mathbf{v}_1^* \mathbf{v}_1^{*\prime})$$

where $\mathbf{u}_1^*$ and $\mathbf{v}_1^*$ are the vectors corresponding to λ_1^*. In the next step, we substitute $\mathbf{Z}_1^*$ for $(\mathbf{P} - \mathbf{E})$ in (5.3) or (5.4) and apply the different constraint matrices for the row and column scores corresponding to the second singular value. Although this approach is computationally straightforward it does not satisfy a global fitting criterion and different solutions may be obtained depending on the order by which the constraints are imposed. Consequently, it may be more appropriate to use an algorithm that allows for the simultaneous fitting of different constraint sets (Takane *et al.* 1991a).

5.3 APPLICATIONS

To illustrate the null-space and the reparametrization method, we report two examples in this section. For the sake of simplicity and reproducibility of the results, the selected data sets are rather small and do not represent typical applications of CA. Procedures for imposing the linear constraints are easily implemented, particularly when matrix commands (such as in SAS) can be used. For example, Blasius and Rohlinger (1989) provide a comprehensive documentation of a CA program written in SAS PROC MATRIX. The necessary modifications for constrained CA are straightforward and involve only the computation of the projection matrices.

5.3.1 Mental health status and parental socioeconomic status

The first data set (in Table 5.1) is taken from a study about the relationship between mental health status and parental socioeconomic status (Srole *et al.* 1962, p. 213). Subjects were assigned to one of four health categories and to one of six socioeconomic status strata (SES) according to composite scores derived from their fathers' schooling and occupational level. The SES were designated A through F to describe a sequence from highest to lowest position.

Previous analyses of this table can be found, for example, in Haberman (1979), Gilula (1986), Gilula and Haberman (1986), and Goodman (1985). This example illustrates in a simple way the equivalence between the null-space and the reparametrization method.

The independence model for the 4×6 table yields a Pearson χ^2-statistic of 46 with 15 degrees of freedom. To examine the relationship between the rows and columns of this table an unrestricted CA was computed which yielded three singular values $\lambda_1 = 0.161$, $\lambda_2 = 0.037$, and $\lambda_3 = 0.017$. The corresponding proportions of the χ^2-statistic are 0.94, 0.04, and 0.02. Clearly, a one-dimensional solution is sufficient for representing this data set. The first column of Table 5.2 contains the standard row and column scores obtained from (5.2). These scores reveal that the row and column category orders are natural. However, the scores corresponding to the second and third row are close together, indicating that the prevalence of 'mild' and 'moderate' symptoms is similar across socioeconomic statuses. The first and second as well as third and fourth column scores are also poorly distinguished. A simplified representation of this data set may be thus obtained by constraining

TABLE 5.2
Standard row and column scores for Table 5.1 by unconstrained and constrained CA.

	(1)	(2)	(3)	(4)	(5)
λ_1	0.161	0.156	0.157	0.150	0.158
%	94	88	89	81	91
No. of parameters	15	5	3	1	1
r_{11}	−1.609	−1.439	−1.617	−1.439	−1.625
r_{21}	−0.183	−0.481	−0.149	−0.481	−0.077
r_{31}	0.088	0.477	0.037	0.477	−0.077
r_{41}	1.472	1.436	1.472	1.436	1.472
c_{11}	−1.122	−1.067	−1.539	−1.539	−1.130
c_{21}	−1.147	−1.153	−0.918	−0.918	−1.130
c_{31}	−0.366	−0.343	−0.298	−0.298	−0.117
c_{41}	0.055	0.005	0.323	0.323	−0.117
c_{51}	1.025	0.952	0.944	0.944	0.896
c_{61}	1.783	1.874	1.565	1.565	1.909

Note: (1) Standard scores obtained from the first dimension of the unconstrained CA solution. (2) Row scores are restricted to follow a linear order. (3) Column scores are restricted to follow a linear order. (4) Both row and column scores are equidistant. (5) Row and column scores are equidistant and satisfy some equality constraints.
The number of parameters for unrestricted CA are equal to the degrees of freedom for the independence model. The remaining solutions are one-dimensional and the number of parameters is determined by {(I + J − 3) − number of linear restrictions}. For example, in columns (4) and (5) λ_1 is the only parameter to be estimated.

the row scores, the column scores, or both to follow a linear order which takes into account that some of the categories are so similar that they can be combined.

To impose these constraints we make use of orthogonal polynomials which are convenient for subdividing the total variation of the scores into linear, quadratic, cubic, etc., components. Although higher-degree polynomials may be difficult to interpret, polynomials are quite useful in describing or approximating general forms of relationships within a limited value range. As discussed in the previous section, the reparametrization method for restricted CA is identical to simple CA when the constraint matrices **M** and **N** are set equal to an identity matrix. The basis vectors spanning the vector space of **I** may be changed without affecting the results of CA. For example, $\mathbf{M}_*$ may be equal to a matrix of orthogonal polynomials,

$$\mathbf{M}_* = \begin{bmatrix} -3 & 1 & -1 \\ -1 & -1 & 3 \\ 1 & -1 & -3 \\ 3 & 1 & 1 \end{bmatrix}$$

with the first, second, and third columns corresponding to the linear, quadratic, and cubic effects, respectively (see Bock 1975, p. 585). However, by considering the one-dimensional subspace spanned by the first column vector with

$$\mathbf{M}_*' = [-3 \quad -1 \quad 1 \quad 3] \tag{5.5}$$

we restrict the standard row scores obtained from Table 5.2 to be equally spaced. Thus, the constrained standard scores conform to the linear ordering

$$r_{11}^* - r_{21}^* = r_{21}^* - r_{31}^* = r_{31}^* - r_{41}^*$$

To satisfy the additional equality constraint between the scores of the second and the third mental health categories:

$$r_{21}^* = r_{31}^* \qquad (\text{and } r_{11}^* - r_{21}^* = r_{31}^* - r_{41}^*)$$

we set

$$\mathbf{M}_*' = [-1 \quad 0 \quad 0 \quad 1] \tag{5.6}$$

Note that the equality constraint, $r_{21}^* = r_{31}^*$, is tantamount to combining the second and third categories. Thus, an equivalent approach for estimating the scores under the equality and linear spacing constraints is to group the second and third mental health category and to apply the linear constraint,

$$\mathbf{M}_*' = [-1 \quad 0 \quad 1]$$

to the collapsed table.

In a similar fashion we may constrain the column scores to follow a linear order. To facilitate the comparison between the unconstrained and constrained representation, we specify first $\mathbf{N}_*$ to be a matrix of orthogonal polynomials for the unconstrained estimation of the column scores

$$\mathbf{N}_* = \begin{bmatrix} -5 & 5 & -5 & 1 & 1 \\ -3 & -1 & 7 & -3 & 5 \\ -1 & -4 & 4 & 2 & -10 \\ 1 & -4 & -4 & 2 & 10 \\ 3 & -1 & -7 & -3 & -5 \\ 5 & 5 & 5 & 1 & 1 \end{bmatrix}$$

(see Bock 1975, p. 585). The linear spacing of the column scores is obtained by using only the first column of $\mathbf{N}_*$:

$$\mathbf{N}_*' = [-5 \quad -3 \quad -1 \quad 1 \quad 3 \quad 5] \tag{5.7}$$

Because the scores corresponding to the A and B and the C and D categories are poorly distinguished, we restrict the corresponding scores to be equal,

$$c_{11}^* - c_{21}^* = c_{31}^* - c_{41}^* = 0$$

This equality constraint in combination with the linear ordering constraint defined by (8),

$$c_{11}^* - c_{31}^* = c_{31}^* - c_{51}^* = c_{51}^* - c_{61}^*$$

is obtained by setting

$$\mathbf{N}_*' = [-7 \quad -7 \quad -1 \quad -1 \quad 5 \quad 11] \tag{5.8}$$

Because equality constraints are equivalent to collapsing the A and B as well as the C and D categories, we obtain the same results by applying the linear constraint

$$\mathbf{N}_*' = [-3 \quad -1 \quad 1 \quad 3] \tag{5.9}$$

to the collapsed data table.

The same constraints can be imposed by the null-space method. For example, we obtain a linear order for the standard row scores by partialing out the effects of the quadratic and the cubic trends. In this case $\mathbf{G}_*$ may be specified as

$$\mathbf{G}_*' = \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{bmatrix} \tag{5.10}$$

and it is easy to see that $\mathbf{G}_*$ specifies the ortho-complement space of $\mathbf{M}_*$ in (5.5). Thus, by eliminating the effects of the quadratic and cubic trends, only the linear effect remains and the standard row scores conform to an

equidistant spacing. The matrix $\mathbf{G}_*$ can be specified in different ways. For instance, an equivalent formulation of the linear spacing constraint given by (5.10) is

$$\mathbf{G}_*^{\prime} = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix}$$

The two rows of $\mathbf{G}_*^{\prime}$ stipulate that there is no quadratic trend for any triple of consecutive categories. To satisfy the additional constraint for the data set that the scores corresponding to the second and third mental health categories are equal, we obtain

$$\mathbf{G}_*^{\prime} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \tag{5.11}$$

or, equivalently,

$$\mathbf{G}_*^{\prime} = \begin{bmatrix} 1 & -2 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix}$$

In both cases, the second column of $\mathbf{G}_*$ stipulates equality between the second and third row score.

A linear ordering of the column scores is obtained by defining the matrix $\mathbf{H}_*$ to contain the quadratic, cubic, quartic, and quintic contrasts,

$$\mathbf{H}_*^{\prime} = \begin{bmatrix} 5 & -1 & -4 & -4 & -1 & 5 \\ -5 & 7 & 4 & -4 & -7 & 5 \\ 1 & -3 & 2 & 2 & -3 & 1 \\ -1 & 5 & -10 & 10 & -5 & 1 \end{bmatrix} \tag{5.12}$$

or, equivalently:

$$\mathbf{H}_*^{\prime} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

To examine the linear spacing constraint with equated first and second and third and fourth column scores, one may set $\mathbf{H}_*$ to

$$\mathbf{H}_*^{\prime} = \begin{bmatrix} 1 & 0 & -2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -2 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix} \tag{5.13}$$

While the first two columns specify that there is no quadratic trend, the remaining columns imply the relevant equality constraints between the column scores.

Table 5.2 lists some of the results obtained when applying these constraints to the data in Table 5.1. The second column of Table 5.2 contains the standard row scores that satisfy a linear ordering defined by (5.5) or (5.10), the third column contains the constrained column scores defined by (5.7) or (5.12), and the fourth column contains the scores obtained by simultaneously constraining the row and column scores. The last column of Table 5.2 contains the constrained scores obtained by simultaneously imposing (5.6) and (5.8) (or (5.11) and (5.13)). In the second and third rows of Table 5.2 we also give the number of estimated parameters and the percentage of the χ^2-statistic accounted for by the first singular value, respectively.

Overall, the solution given in the fifth column is the most preferable. It requires the estimation of only one parameter and accounts for about 91% of the χ^2-statistic. From this representation, we conclude that the prevalence of well-being decreases with the socioeconomic status groups and that the opposite pattern is observed for the impaired mental health category. There is no appreciable difference between the mild and moderate symptom formation categories. Both categories' prevalences are similar across the status groups and can be combined. Moreover, the first and the second as well as the third and the fourth socioeconomic status categories relate to the mental health categories in a similar way and can also be combined with little loss in information. When combining these categories equally spaced row and column scores are obtained. As a result, the constrained CA yields a parsimonious and compact representation of the relationship between the mental health and socioeconomic status categories.

5.3.2 *Magazine reading habits*

In this study, reported by Böckenholt and Böckenholt (1991), 347 students were asked about their reading habits for the eight magazines: *People, Rolling Stone, Time, Sports Illustrated, Scientific American, National Geographic, Readers' Digest*, and *TV Guide*. These students were assigned to four groups and Table 5.3 lists how many members in each group read the magazines on a regular basis. For example, 31 out of a group of 91 students read regularly the magazine *People*. A subset of the respondents (53 students) was also asked to evaluate the magazines on several five-point rating scales. The total inertia of the data in Table 5.3 is 0.355 and the percentages of the inertia are 52.4, 41.1, and 6.4. A two-dimensional representation seems adequate and Figure 5.1 contains the principal coordinates obtained from the CA of these data. To obtain a simultaneous representation of the groups' selection and non-selection frequencies of the magazines we also analyzed the doubled data matrix but obtained virtually the same results.

We interpret the graphical display by inspecting groupings and contrasts in the configuration (Greenacre and Hastie 1987). Thus, differences between

TABLE 5.3
Tabulation of selection frequencies for eight magazines.

Group	PE	RS	TI	SI	SA	NG	RD	TV	Total	Size
1	31	55	1	55	24	16	6	47	235	91
2	32	20	0	3	4	1	15	14	89	57
3	71	59	66	28	11	23	79	39	376	150
4	8	6	30	10	23	32	12	5	126	49
Total	142	140	97	96	62	72	112	105	826	347

Note: PE = *People*, RS = *Rolling Stone*, TI = *Time*, SI = *Sports Illustrated*, SA = *Scientific American*, NG = *National Geographic*, RD = *Readers' Digest*, TV = *TV Guide*

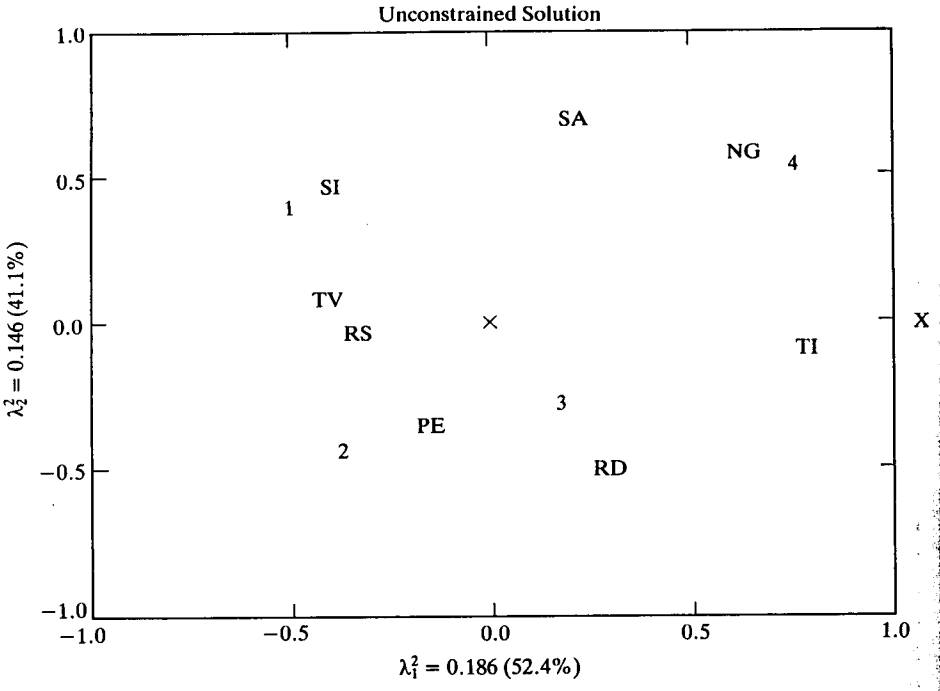

FIGURE 5.1 Two-dimensional display of the principal row and column coordinates obtained by unconstrained CA.

Note: PE = *People*, RS = *Rolling Stone*, TI = *Time*, SI = *Sports Illustrated*, SA = *Scientific American*, NG = *National Geographic*, RD = *Readers' Digest*, TV = *TV Guide*. The four groups are distinguished by the numbers 1 to 4.

magazines along an axis are small to the extent they were read on a similar recurrent basis. We note that the first axis separates *Scientific American, National Geographic, Time*, and *Readers' Digest* from the remaining four magazines, and that the second axis separates *Scientific American, National Geographic*, and *Sports Illustrated* from *People* and *Readers' Digest*. To guide the interpretation of these principal axes, the ratings of the magazines were averaged over the 53 respondents and included in the matrix $\mathbf{N}_*$ to constrain the column scores. This analysis showed that the two rating scales 'educational' and 'specialized' were particularly useful in distinguishing the magazines in the two-dimensional representation. Setting $\mathbf{N}_*$ equal to the (8×2) matrix of averaged ratings (for the eight magazines and the two attributes 'educational' and 'specialized') accounts for 87.6% of the total inertia. Figure 5.2 depicts the corresponding two-dimensional representation obtained from the column-constrained CA. This analysis indicates that *National Geographic* and *Scientific American* are perceived as more and *TV Guide* as less

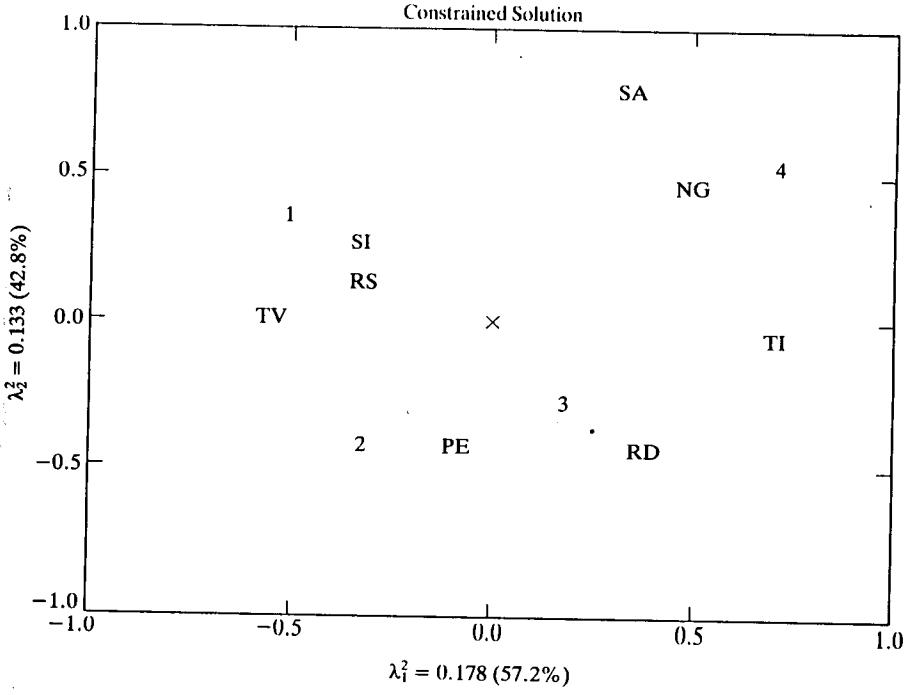

FIGURE 5.2 Two-dimensional display of the principal row and column coordinates obtained by reparametrization-method constrained CA.

educational magazines. *Readers' Digest* and *Scientific American* are at the two extreme positions on a continuum of increasing specialization.

A powerful comparison of the differences between the unconstrained and the constrained solution is provided by the null-space method with $\mathbf{H} = \mathbf{D}_c\mathbf{N}$. This residual analysis examines the effects not accounted for by the two rating scales. Figure 5.3 depicts the first two dimensions which account for 10.3% of the total inertia. The first axis indicates that the constraints represent the relationships between *Scientific American* and *Sports Illustrated* less well than they represent the relationships among the other magazines. As a result, group 2, whose position is most affected by this result, is placed at the lower end of this axis. Overall, however, the residual analysis provides further support for the usefulness of the constraints. With the one minor exception, the mean ratings capture well the multidimensional structure of the magazines, and, consequently, facilitate a straightforward interpretation of the data.

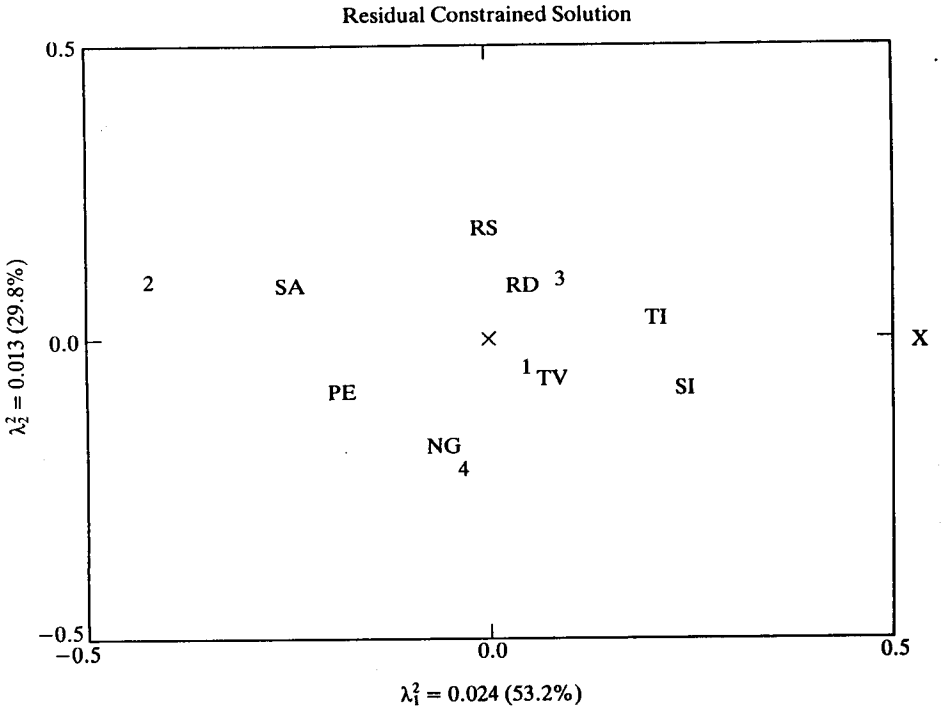

FIGURE 5.3. Two-dimensional display of the principal row and column coordinates obtained by null-space-method constrained CA.

5.4 CONCLUDING REMARKS

This chapter presented two approaches for imposing linear constraints on the row and/or column scores by restricting the CA solution to lie either in a particular subspace (reparametrization method) or to be orthogonal to a subspace (null-space method). In both cases, we obtain a solution not in the full-space as in unconstrained CA but in a subspace that is directly or indirectly specified by external information. We see three major advantages in applying constraints in a CA. First, because of its simplicity and low computational cost for large problems, CA provides frequently an excellent starting point for further refinements and model building. Constraints that are formulated on the basis of some theoretical considerations can be examined for their empirical validity before more complicated models are applied. In particular, by incorporating hypothesized data structures in the singular value decomposition we gain much additional flexibility in developing a meaningful multidimensional representation of a data table. Similarly, the option to partial out the effect of certain variables in constrained CA should prove especially beneficial for the decomposition and graphical display of residuals obtained from models other than the log-linear independence model (de Leeuw and van der Heijden 1988, van der Heijden *et al.* 1989, Novak and Hoffman 1990). Second, constrained representations of a data matrix are more parsimonious and stable than unconstrained solutions and less sensitive to undesirable effects produced by, for example, outliers or coding errors. Third, by imposing constraints the search for patterns in the data may be considerably simplified. Although in some applications data structures may reveal themselves in an unconstrained analysis, in general, it is not trivial to separate 'noise' from 'signals' in a large data set. In these cases, constraints can prove helpful by either eliminating certain effects from the data or by directly imposing a certain structure. The resulting gains in interpretability may far exceed the loss in information as a result of the constraints.