PSYCHOMETRIKA—VOL. 52, NO. 2, 207–215
JUNE 1987

THE CHOICE OF CONSTRAINTS IN CORRESPONDENCE ANALYSIS

HARVEY GOLDSTEIN

UNIVERSITY OF LONDON INSTITUTE OF EDUCATION

A discussion of alternative constraint systems has been lacking in the literature on correspondence analysis and related techniques. This paper reiterates earlier results that an explicit choice of constraints has to be made which can have important effects on the resulting scores. The paper also presents new results on dealing with missing data and probabilistic category assignment.

Key words: constraint systems, correspondence analysis, dual scaling, missing data, optimal scaling, probabilistic category assignment.

Introduction

The range of techniques variously known as optimal scaling, dual scaling, or multiple correspondence analysis, have been widely applied in a number of contexts including test item scoring and estimation of bone maturity. Nishisato (1980) and Greenacre (1984) present accounts and McDonald (1983) has contributed further results. De Leeuw (1984) describes various extensions known as the Gifi system. Healy and Goldstein (1976) considered the fundamental problem of which basic constraint systems should be used in order to obtain nontrivial solutions. They pointed out that the choice of an "end-point" or "linear" constraint gave different results to the usual choice of an "average" or "quadratic" constraint. They suggested that the nature of the problem should determine the constraint system, but this does not seem to have been followed up, and the choice of an average constraint typically is assumed to be a "natural" one. For example, in an attempt to synthesize existing procedures. Tenenhaus and Young (1985) fail to recognize the possibility of alternative constraint systems.

The main purpose of this paper is to reiterate the need for a careful choice of constraints and in particular to investigate multiple solutions with varying sets of constraints. In addition, the paper extends the results of Healy and Goldstein (1976) to the case of randomly missing data and probabilistic category assignment.

Notation

Consider a system with h attributes with the i-th attribute having p_i categories, with n categories in total. The score to be allocated to the j-th category of the i-th attribute is x_{ij}. We have a sample of size N with z_{im} the score of the m-th subject on the i-th attribute, so if the j-th category is observed for this subject $z_{im} = x_{ij}$. We define a mean score for the m-th subject as

$$\bar{z}_m = \Sigma_i w_i z_{im}, \qquad \Sigma_i w_i = 1$$

where the w_i are preassigned weights. Most applications of dual scaling have used $w_i = h^{-1}$. A measure of "disagreement" between the (weighted) components of a subject's score

I am most grateful to the following for their helpful comments. Arto Demirjian, Michael Greenacre, Michael Healy, Shizuhiko Nishisato, Roderick Mcdonald, and several anonymous referees.

Requests for reprints should be sent to Harvey Goldstein, Institute of Education, 20 Bedford Way, London WC1H OAL, ENGLAND.

0033-3123/87/0600-7105$00.75/0

can be defined as

$$d_m = \Sigma_i w_i(z_{im} - \bar{z}_m)^2,$$

and a class of optimum scoring systems can be defined as those which minimize

$$D = \Sigma_m d_m.$$

To avoid the trivial solution $x_{ij} =$ constant, we need to impose basic constraints. Healy and Goldstein considered two possibilities:

$$\Sigma \bar{z}_m^2 = 1,$$
$$\bar{z} = N^{-1}\Sigma \bar{z}_m = 0 \tag{1}$$

and

$$\mathbf{x}^T\mathbf{q} = 0, \qquad \mathbf{x}^T\mathbf{r} = 1 \tag{2}$$

where $\mathbf{x} =$ vec (x_{ij}), and N is the total number of subjects and q and r are $(n \times 1)$ with all elements equal to zero except for those corresponding to attribute categories 1 and p_i respectively, which are equal to w_i.

We define the following matrices: The nonnegative definite matrix A is $n \times n$ symmetric with diagonal elements $N_{ij}w_i(1 - w_i)$ and off-diagonal elements $-N_{ijkl}w_iw_k$, where N_{ij} is the number of subjects in category j of attribute i and N_{ijkl} is the number of subjects in category j of attribute i and l of attribute k.

The matrix Z is $n \times n$ symmetric with diagonal elements $N_{ij}w_i^2$ and off-diagonal elements $N_{ijkl}w_iw_k$. The matrix $S = A + Z$ is $n \times n$ diagonal with elements $N_{ij}w_i$, and the number of subjects $N =$ tr (S). J_{rt} is an $r \times t$ matrix of ones.

By equating coefficients of the x_{ij} we see that $D = \mathbf{x}^TA\mathbf{x}$. With the average constraints (1) the optimal scores are given by the solution of $2A\mathbf{x} - 2\lambda Z\mathbf{x} - \mu SJ_{n1} = 0$, that is, the latent vector corresponding to the smallest nonzero root of

$$|A - \lambda Z| = |A - \lambda(1 + \lambda)^{-1}S| = 0, \tag{3}$$

which in the case of equal weights is equivalent to the solution given by Guttman (1941). With the end point constraint (2) the optimal scores are given by the vector $\mathbf{x}$ satisfying both (2) and the set of linear equations,

$$2A\mathbf{x} - \mathbf{q}\lambda - \mathbf{r}\mu = 0 \tag{4}$$

Healy and Goldstein also considered the "canonical" scoring problem where each subject has p sets of attributes and we wish to find p sets of scores such that the weighted average disagreement between the p scores is minimized. This case is not considered in detail here, but the methods to be outlined below can be adapted readily to that case.

It is worth noting that there is sometimes a misunderstanding about the use of the term "constraint system." All derivations of the classical correspondence analysis results implicitly assume constraints. Thus, for example, the derivation which seeks to minimize the ratio $\mathbf{x}^TA\mathbf{x}/\mathbf{x}^TZ\mathbf{x}$ requires that $\mathbf{x}^TZ \lesssim 0$, and since, without loss of generality, we can set $\mathbf{x}^TZ\mathbf{x} = 1$, this becomes equivalent to the formulation given in (1). Likewise, the end point system of Healy and Goldstein can be derived by minimizing $\mathbf{x}^TA\mathbf{x}/(\mathbf{x}^T\mathbf{r} - \mathbf{x}^T\mathbf{q})^2$ with the requirement that $\mathbf{x}^T\mathbf{r} \neq \mathbf{x}^T\mathbf{q}$, and as before, without loss of generality setting $\mathbf{x}^T\mathbf{r} - \mathbf{x}^T\mathbf{q} = 1$. It should also be noted that a second constraint in each case, $(\mathbf{x}^TSJ_{n1} = \mathbf{x}^T\mathbf{q} = 0)$ is necessary to fix the location of the solution vector $\mathbf{x}$.

Probabilistic Category Assignment

In the case of probabilistic assignment to categories an individual subject has a score given by

$$Z_{im} = \Sigma_i p_{ijm} x_{ij}, \qquad \Sigma_j p_{ijm} = 1 \tag{5}$$

where p_{ijm} is the stated probability that for subject m an observation belongs to the j-th category of attribute i. The estimation equations are as above except that in the matrices A, S, and Z; N_{ij} is replaced by $\Sigma_m p_{ijm}^2$, and N_{ijkl} is replaced by $\Sigma_m p_{ijm} p_{klm}$.

Probabilistic assignment may be useful in a number of circumstances. For example, several raters may make category judgments which differ, and the p_{ijm} can then be taken to be the relative frequencies of categories chosen. In another case, an observation may possess some features typical of one category and other features typical of another category, and p_{ijm} can be defined in terms of relative frequencies or possibly a subjective assessment.

Further Constraints Among Scores

With the end point constraint given by (2) leading to (4) we can incorporate p, say, general linear constraints of the form:

$$\mathbf{x}^T C = 0 \tag{6}$$

where C is $n \times p$: which leads to the set of linear equations

$$2A\mathbf{x} - \mathbf{q}\lambda - \mathbf{r}\mu - C\mathbf{w} = 0 \tag{7}$$

together with (2) and (6), which can be solved using the same standard methods as before. $\mathbf{w}$ is $p \times 1$.

With the average constraint we obtain:

$$2A\mathbf{x} - 2Z\mathbf{x}\lambda - SJ_{n1}\mu - C\mathbf{w} = 0, \tag{8}$$

noting that $J_{1n}S\mathbf{x} = J_{1n}Z\mathbf{x} = J_{1n}A\mathbf{x} = 0$. Multiplying on the left by J_{1n}, we have

$$\mu + J_{1n}C\mathbf{w}\{J_{1n}SJ_{n1}\}^{-1} = 0,$$

so that 8 becomes:

$$A\mathbf{x} - Z\mathbf{x}\lambda - \{I - N^{-1}SJ_{nn}\}C\mathbf{w} = 0, \tag{9}$$

where N is the number of subjects and n is the number of categories. Equation (9) can be written as:

$$(S^* - \lambda^* B^*)\mathbf{x}^* = 0, \tag{10}$$

where the $(n + p) \times (n + p)$ matrix

$$S^* = \begin{Bmatrix} S & E \\ E^T & 0 \end{Bmatrix}$$

with $E = (I - N^{-1}SJ_{nn})C$, and the $(n + p) \times (n + p)$ matrix

$$B^* = \begin{Bmatrix} Z & 0 \\ 0 & 0 \end{Bmatrix},$$

and

$$\mathbf{x}^* = \begin{Bmatrix} \mathbf{x} \\ \mathbf{w} \end{Bmatrix}, \qquad \lambda^* = 1 + \lambda.$$

In the general case where $E^T S^{-1} E$ is nonsingular, (10) can be solved utilizing $(S^*)^{-1}$, with an appropriate procedure for dealing with an asymmetric matrix. We require in this case the smallest positive value of λ, which exists if and only if there is at least one latent root of $(S^*)^{-1} B^*$ within the interval (0, 1). Thus, not all constraint systems are necessarily admissible (McDonald & Goldstein, in preparation).

Missing Data

Several procedures have been proposed for dealing with the case where not all subjects have an observation on each attribute and responses are missing at random. Nishisato (1980) suggests creating an extra category for missing responses, but where there is random sampling of individuals from a well defined population, the expectations of these estimates will not equal those where the responses are not missing, and in this sense Nishisato's procedure can be said to be biased. He also suggests ignoring the missing responses when calculating the relevant matrices and then using these matrices as if the data were complete. This approach, however, while it may give a reasonable approximation when there are not many missing responses, is not exact since (3) and (4) are not satisfied strictly. The following procedure provides optimal estimates which minimize the disagreement for randomly missing data.

For the average constraint, minimization of D over the nonmissing data leads to

$$A\mathbf{x} - Z\mathbf{x}\lambda - SJ_{n1}\mu = 0.$$

We have $J_{1n} S\mathbf{x} = 0$, but $A^T J_{n1} = \{a_{ij}\}$, where $a_{ij} = w_i(N_{ij} - \Sigma_k M_{ijk} w_k)$, where $M_{ijk} = N_{ij} - L_{ijk}$ and L_{ijk} is the number of missing observations in attribute k for those subjects with the j-th category of attribute i. This gives

$$\mu = N^{-1} J_{1n} A(1 + \lambda)\mathbf{x}$$

leading to

$$\{[I - N^{-1} S J_{nn}]A - \lambda[Z + N^{-1} S J_{nn} A]\}\mathbf{x} = 0. \tag{11}$$

This can be solved by standard methods. For the end point constraint we solve the same set of equations as before with A calculated over nonmissing responses. Where there are constraints with missing data, the above results can be combined in straightforward fashion.

Multiple Components

With the average constraint system, the definition of second and subsequent components as mutually orthogonal leads to a straightforward extraction of eigenvectors of (3), (10), (11). For the end point constraint however the position is more complex.

First we note that Equation (2.6) in Healy and Goldstein (1976) is incorrect since $\mathbf{x}^T S J_{n1} \neq 0$. The correct additional constraint is

$$\mathbf{y}^T\{Z - N^{-1} S J_{nn} S\}\mathbf{x} = 0.$$

For the first component the use of the end point constraint implies that intermediate category scores lie in the range (0, 1) and this is indeed the case in the examples below. It

is not clear, however, that a second component should be orthogonal to the first, since this is likely to force intermediate category scores outside the range (0, 1), which somewhat invalidates the original choice of the end point constraints. Computations with the data in Healy and Goldstein (1976) verify that this is so.

Instead, it seems to us that we should reconsider the purpose of having multiple components. One function of a second component is to explain residual variation in observed responses, subject to the second component scores having a predefined relationship with the first component scores. For the average constraint, as in traditional multivariate analysis, the specified relationship is one of overall orthogonality of scores. This, however, is not necessary in order to obtain a solution. Thus, any two distinct constraints involving first component scores will lead to a solution, and likewise for the end point basic constraint. Thus, for either basic constraint system a general definition for such a second component can be written as minimizing D subject to a set of p constraints

$$\mathbf{x}^T C = \mathbf{k}^T, \tag{12}$$

where $\mathbf{k}$ is $p \times 1$ and includes first component scores.

There is no reason, however, why components should always be in order of importance of extraction. From a general standpoint each component is defined by a set of constraints which may or may not involve scores associated with other components.

In the case of an average constraint system, the definition of second and subsequent components as orthogonal to earlier ones seems reasonable if the use of an average constraint implies that no subset of scores is singled out for special attention. In the case of the end point (or other subset specific) constraints, however, we may often wish to define subsequent components in terms of further subset specific constraints. Thus a second component might involve constraints which forced early stages to be equal in order to produce better discrimination among later stages. In some cases a "hierarchical" second component might be chosen which incorporated constraints which used functions of first component scores. Also, the weights w_i may change from component to component.

Examples

The following example uses maturity data from X-rays of tooth development in French Canadian children. Each tooth can be observed in one of up to nine stages, labeled $0 - H$, where 0 is the least mature stage and H the most mature stage. The data are described by Demirjian and Goldstein (1976) who present end point constrained maturity scores for various sets of teeth. The data used here are those for boys on the following four teeth: first molar (M1), first premolar (PM1), second molar (M2), and second premolar (PM2). The scores for the first maturity component using equal attribute weights and the constraints (2) are given in Table 1. For convenience the estimated scores have been linearly rescaled so that all the first stage scores are zero and the final stage scores add to 100.

To illustrate the use of additional constraints, we now consider adding the constraints which set the first four stages of M2 and PM2 and the first three of PM1 to be equal in order to produce a scale which is more sensitive to the later period of development. In effect, this is equivalent to merging the related categories. The results are given in Table 2. The most marked effect is to reduce the scores for stages D-F of PM1 relative to the other teeth.

TABLE 1

Scores for tooth maturity stages for French Canadian boys using end point constraints

Tooth	Stage								
	O	A	B	C	D	E	F	G	H
M2	0.0	3.0	5.9	9.4	13.8	17.6	20.4	22.0	23.6
M1				0.0	7.5	11.9	16.1	20.9	27.1
PM2	0.0	2.9	5.4	9.1	13.3	16.8	19.9	21.3	22.8
PM1		0.0	5.4	10.1	14.9	19.8	23.0	24.8	26.6

NB: The scores given here differ slightly from those in Demirjian and Goldstein (1976) due to the use of a slightly different sample.

TABLE 2

Scores obtained by pooling first 4 stages of Table 1

Tooth	Stage					
	0-C	D	E	F	G	H
M2	0.0	8.5	14.8	18.9	21.2	23.6
M1	0.0	9.6	10.0	13.3	20.9	30.5
PM2	0.0	8.1	13.9	18.6	20.8	22.9
PM1	0.0	4.8	12.6	17.7	20.5	23.0

We next add the constraint which sets stages D-H for each tooth equal in order to produce a system which is sensitive in describing early development. In addition, this time, we weight M2, PM2 and PM1 twice as much as M1 to reflect the greater number of early stages. The results are given in Table 3. Again, the effect on PM1 is most marked with the scores being relatively much higher than for the other teeth compared to the first component solution.

Finally, we use the average constraint system, and the scaled scores are displayed in Table 4. The scores are similar to those in Table 1 with strictly increasing scores within attributes. The scales do not rank subjects in precisely the same order, however. For example, with the end point constraint, a subject in the initial stages of teeth M2, M1, PM2 and stage B of PM1 is given a higher maturity score than a subject in the initial

TABLE 3

Scores obtained by pooling last 5 stages of Table 1

Tooth	Stage				
	0	A	B	C	D-H
M2	0.0	9.7	14.4	17.3	20.2
M1				0.0	26.2
PM2	0.0	9.5	13.5	16.8	19.6
PM1		0.0	20.0	28.6	34.2

TABLE 4

Scores obtained for tooth maturity stages for French Canadian boys using average constraints

Tooth	Stage								
	0	A	B	C	D	E	F	G	H
M2	0.0	1.0	3.4	8.0	14.4	20.2	23.6	25.0	25.2
M1				0.0	1.0	4.7	10.2	17.3	24.7
PM2	0.0	0.9	2.8	7.6	13.7	19.1	23.0	24.5	24.9
PM1		0.0	0.6	4.2	10.6	17.9	22.4	24.6	25.3

stages of M2, M1, and PM1 and stage A of PM2. With the average constraint, however, this ordering is reversed. The correlation between the systems is given by

$$\mathbf{x}_1^T Z \mathbf{x}_2 \{\mathbf{x}_2^T (Z - N^{-1} S J_{nn} S) \mathbf{x}_2\}^{-1/2},$$

where x_1, x_2 are the score vectors for the average and end point constraint respectively. In the present case, using the standardizing sample gives a correlation of 0.99 which is very high and reflects the inherent strong ordering of the data. In the example given by Healy and Goldstein (1976), the correlation is 0.81, which is for a set of behavioral questionnaire data with no such strong inherent data ordering. Here, the use of an end point constraint seems desirable since it reflects the order assumptions involved in the design of the question categories.

Discussion

Clearly constraint systems and category weightings other than the ones considered in this paper are possible. Ideally, a particular choice should be justified by substantive arguments rather than any purely mathematical criterion. This applies to the choice of first component as well as to subsequent ones and is an issue which largely has been ignored in the literature. In practice, of course, it may be difficult to make a choice, and sometimes a range of different choices will anyway lead to similar conclusions. In other cases however, the choice of constraints will be crucial and it will often be a good idea to try several systems in an exploratory spirit. Such an examination would be assisted by graphical displays of individual subject "disagreement" scores, plotted against fitted values and each other.

It is also worth raising another issue alluded to by Healy and Goldstein (1976), namely the choice of reference population. In some applications this may be obvious but in others less so, and since the estimates will reflect the population structure, the issue is very relevant. In the above examples, we would wish to include "all ages" equally and to sample as many as necessary fully mature and immature subjects to give good estimates of the end point scores. There is, however, no obvious age cut-off and for example, sampling too many mature adults will have the effect of giving too much weight to the extreme category estimates. One solution is to fix the end points of each attribute, for example to be 0.0 and $(hw_i)^{-1}$, but this then introduces extra constraints which may not

be reasonable. In the present case a reasonable solution would be to sample ages in proportion to the number of nonmature individuals at each age.

Finally, let us reiterate the main point of this paper. Correspondence analysis, like other scaling techniques, depends on certain assumptions in order to produce nontrivial estimates. These assumptions involve the choice of loss function, D, as well as constraints, although we have focused on the latter. Thus, given the data, the estimates obtained are in effect defined by the choice of assumptions, a fact which these techniques share with so-called latent variable models. The choice of loss function is also a feature of the class of response/explanatory variable models such as linear regression, but the latter do not depend on the further choice of constraints to achieve a nontrivial solution, and given a loss function there are objective procedures for choosing between model equations in terms of "closeness" to the data. In the former case, however, such procedures are specific to the constraints chosen so that there can be no completely objective means of choosing between constraint systems. The implication is that careful attention should be paid to any choice and the dependence of any solution on different choices should be clearly understood. Furthermore, the examples suggest that the choice of constraint system may be especially important for data lacking a strong inherent structure. In cases where the estimates depend strongly on the choice of constraints, it might be reasonable to ask whether a scaling procedure is appropriate at all. From this viewpoint we could regard the comparison of constraint systems as a means of ascertaining the appropriateness of any of them.

References

de Leeuw, J. (1984). The GIFI system of nonlinear multivariate analysis. In E. Diday, et al. (Eds.), *Data analysis and informatics IV* (pp. 415–424). Amsterdam: North Holland.

Demirjian, A., & Goldstein, H. (1976). New systems for dental maturity based on seven and four teeth. *Annals of Human Biology, 3*, 411–421.

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

Healy, M. J. R., & Goldstein, H. (1976). An approach to the scaling of categorised attributes. *Biometrika, 63*, 219–229.

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika, 48*, 377–392.

McDonald, R. P., & Goldstein, H. (In preparation). Comparative properties of alternative constraint systems in optimal scaling.

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

Tenenhaus, M., & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika, 50*, 91–119.

Manuscript received 6/24/85
Final version received 6/30/86