Multivariate Behavioral Research, 27 (4), 567-583

Equality Constraints in Multiple Correspondence Analysis

Stef van Buuren
TNO Institute of Preventive Health Care, Leiden

Jan de Leeuw
University of California, Los Angeles

The application of equality constraints on the categories of a variable is a simple but useful extension of multiple correspondence analysis. Equality can be used to incorporate prior knowledge about the relations between categories. Categories may belong to the same variable, to different variables, or both. The simplest form of equality specifies that all variables receive identical data transforms. This is useful, for example, if the same variable is measured on many points of time. This article outlines a procedure to deal with unequal category numbers and with subsets of variables. Though the technical results are not difficult to derive, they are not very well-known. Some applications illustrate the method.

Multiple correspondence analysis (MCA) is a popular technique for analyzing multivariate categorical data. Standard references are Benzécri (1973), Nishisato (1980), Lebart, Morineau and Warwick (1984), Greenacre (1984) and Gifi (1990). This article deals mainly with homogeneity analysis, a form of MCA popularized by Gifi, but the main results can be easily translated into other incarnations of the technique.

MCA is usually applied to reveal systematic patterns among the categories of the variables of interest. In conventional MCA the category quantifications are usually not restricted, possibly except for normalization. In practice however, we may know in advance that categories are related in some way. For example, for ordinal measurements we know the order of the categories. If we wish to preserve this order we may require that the category quantifications are a monotone function of the category numbers. This restriction is defined within the same variable, so we call it a within-variable constraint.

Another useful possibility is to impose constraints across variables. For example, the OVERALS technique for polyset canonical analysis can be

Correspondence should be addressed to Stef van Buuren, TNO Institute of Preventive Health Care, Wassenaarseweg 56, 2333 AL Leiden, The Netherlands (Email: buuren@nipg.tno.nl), or to Jan de Leeuw, Department of Psychology and Mathematics, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024-1536 USA.

viewed as a constrained form of homogeneity analysis in which the categories that belong to one set of variables should conform to a specific summation pattern (cf. Gifi, 1990, p. 205). In this article we study another across-variable restriction, the equality constraint. Equality is both conceptually and computationally quite simple. The type of equality we study stipulates that different categories should receive identical quantifications. Categories may belong to the same variable or to different variables.

Why do we need equality? In general, equality can be helpful if variables are supposed to be measured on the same scale, that is, if the interpretation of their categories is identical. This is often the case if variables are comparable in some sense. An obvious example is if we have scores on the same variable at different points of time, as in event history data. Here it makes sense to fix the scaling across all occasions. Another example emerges if a variable appears more than once in the analysis but with its rows permuted in some way. This frequently occurs in time series analysis where the values on a variable are shifted one or more positions and the dependencies between the resulting lagged variables are studied.

Ranking data give rise to another class of applications. Suppose that 50 psychometricians rank 10 journals on, say, readability. We thus obtain 50 readability variables on 10 objects. It would be interesting to derive a consensus ranking as well as to gain insight into the most typical deviations. Because the variables are replications of each other we may assume the existence of an underlying common scale. An obvious way to implement such a common scale assumption is to restrict the quantifications per rank to be the same for all psychometricians. And this is nothing more than requiring equality.

Last but not least, equality provides a means to avoid degenerate, redundant and uninteresting solutions. For instance, heterogeneity of missing data may have a profound impact on the solution because homogeneity analysis often places rare categories near the border. The resulting configuration is usually not very appealing. One way to alleviate the trouble is to require equality of all missing categories, assuming that these have something in common across variables. Likewise if we are not interested in certain source of heterogeneity, say differences between all "don't know" answers, we may treat them as equal. The stability of the solution thereby enhances and the graphs simplify. Equality constraints can be used to incorporate prior knowledge (of the kind described above) into the analysis. A desirable consequence of this feature is that less parameters are needed and hence solutions will be more stable under arbitrary deletion of variables, categories or observations.

Some of the ideas presented here have been proposed before. De Leeuw (1973, pp. 50, 160) applies constrained homogeneity analysis to sorting data.

Deville and Saporta (1980), Saporta (1981), de Leeuw, van der Heijden and Kreft (1985) and van der Heijden (1987) use equality restrictions in the analysis of several types of longitudinal data in which each variable represents a time point. Van Buuren (1990) codes time points in the rows and applies equality to different lags of the same variable. Gifi (1990, p. 332) relates equality to the method of successive intervals, a classic scaling technique proposed by Guilford (1954), and shows how the singular value decomposition (SVD) can be applied if we want equality of *complete variables*. This article proposes an extension to the Gifi system that allows for equality of *individual categories*, which is more general and which is also more useful. An additional advantage of the method is that, compared to SVD, it is much easier to impose further rank- or order restrictions on the solution. See de Leeuw and van Rijckevorsel (1988) for a detailed description of methods that deal with these types of constraints.

This article is organized as follows: the *Method* section describes how equality can be specified in terms of homogeneity analysis and presents two methods to compute the constrained solution. Next, applications are discussed. *Ranking Data* deals with the most elementary case of complete equality among all variables. The example of *Missing Data* focuses on equality for one specific category, and *Event History Data* shows how equality can be applied within subsets of variables. The *Conclusion* summarizes the main results and deals with some practical issues.

Method

Using the notation of Gifi (1990), suppose that n observations on m categorical variables, each with k_j categories, are coded into indicator matrices $\mathbf{G}_j$ ($j = 1, ..., m$) of order $n \times k_j$. Let $\mathbf{Y}_j$ denote a $k_j \times p$ matrix of p-dimensional category quantifications and let $\mathbf{X}$ be an $n \times p$ matrix of object scores. Homogeneity analysis can then be formulated as minimizing

$$(1) \qquad \sigma(\mathbf{X}; \mathbf{Y}_1, ..., \mathbf{Y}_m) = m^{-1} \sum_{j=1}^{m} \mathrm{SSQ}\,(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)$$

over $\mathbf{X}$ and $\mathbf{Y}_1, ..., \mathbf{Y}_m$. We write SSQ(.) for tr(.)'(.). Equation 1 is known as the HOMALS loss function. Minimization procedures and theoretical properties are thoroughly discussed in Gifi (1990, p. 105). The normalization constraints $\mathbf{1}'\mathbf{X} = \mathbf{0}$ and $\mathbf{X}'\mathbf{X} = n\mathbf{I}$ prevent degenerate and trivial solutions.

The simplest form of equality is to require $\mathbf{Y} = \mathbf{Y}_j$ for all $j = 1, ..., m$ which indicates a one-to-one correspondence between the categories of all variables.

Note that this will only work if all variables possess an equal number of categories. It will be clear that this condition is often too restrictive in practice. A more flexible method, which allows for equality on a category level, is to constrain the solution by $\mathbf{Y}_j = \mathbf{S}_j\mathbf{Y}$ for all j, where $\mathbf{S}_j$ is a known $k_j \times k$ indicator matrix that links the individual categories to the levels of a common $k \times p$ quantification matrix $\mathbf{Y}$. A typical use of this more liberal form of equality is the construction of common "not applicable" or "missing" categories. It is also possible to require that categories within a variable should receive identical scale values, or to require that some categories do not enter the analysis at all by coding the entire row to zero. Also, equality in subsets of variables can be specified quite easily.

To see how Equation 1 can be minimized under $\mathbf{Y}_j = \mathbf{S}_j\mathbf{Y}$ let us define some auxiliary matrices. Let $\mathbf{D}_j = \mathbf{G}_j'\mathbf{G}_j$ be the diagonal matrix of marginal frequencies, let $\mathbf{D} = \Sigma_j\mathbf{S}_j'\mathbf{D}_j\mathbf{S}_j$ denote a $k \times k$ diagonal matrix containing the number of observations per common category and let $\tilde{\mathbf{Y}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{X}$ be the usual unrestricted update of category points. Then $\mathbf{E} = \Sigma_j\mathbf{S}_j'\mathbf{D}_j\tilde{\mathbf{Y}}_j$ is the $k \times p$ matrix that sums these intermediate quantifications into the common category system so that $\tilde{\mathbf{Y}} = \mathbf{D}^{-1}\mathbf{E}$ contains the corresponding centroids. Now Equation 1 may be partitioned as

$$(2)\quad m\sigma(\mathbf{X}; \mathbf{Y}_1, ..., \mathbf{Y}_m) = \sum_{j=1}^{m} \mathrm{SSQ}(\mathbf{X} - \mathbf{G}_j\tilde{\mathbf{Y}}_j) + \sum_{j=1}^{m} \mathrm{SSQ}^{\mathbf{D}_j}(\tilde{\mathbf{Y}}_j - \mathbf{S}_j\tilde{\mathbf{Y}}) + \mathrm{SSQ}^{\mathbf{D}}(\tilde{\mathbf{Y}} - \mathbf{Y})$$

where $\mathrm{SSQ}^{\mathbf{D}}(.) = \mathrm{tr}(.)'\mathbf{D}(.)$. Both $\tilde{\mathbf{Y}}_j$ and $\tilde{\mathbf{Y}}$ are least squares estimators so the remaining problem is to minimize $\mathrm{SSQ}^{\mathbf{D}}(\tilde{\mathbf{Y}} - \mathbf{Y})$. If $\mathbf{Y}$ is not restricted any further then the solution is found by setting $\mathbf{Y} = \tilde{\mathbf{Y}}$. In other cases, the last term must be minimized over any additional constraints, like rank or order restrictions. The steps for $\mathbf{X}$ can be found in Gifi (1990).

It is well known that the minimum of Equation 1 can also be obtained by performing correspondence analysis on the super-indicator $\mathbf{G} = [\mathbf{G}_1, ..., \mathbf{G}_j, ..., \mathbf{G}_m]$. One of the referees pointed out that it is also possible to create equality by replacing the relevant columns by their sum, followed by a CA on this condensed matrix. This result may be derived from the so-called "principle of distributional equivalence" (Benzécri, 1973; Greenacre, 1984, pp. 65, 95). The principle postulates that proportional profiles may be supplanted by their sum without affecting the analysis in any way. Conversely, and this appears to be less well-known, if we want equal scores, we only have to add the columns. So if $\mathbf{S} = [\mathbf{S}_1', ..., \mathbf{S}_j', ..., \mathbf{S}_m']'$ and if $\tilde{\mathbf{G}} = \mathbf{G}\mathbf{S}$ is the condensed matrix, then the constrained solution is equal to the nontrivial components of $\mathbf{X} = \sqrt{n}\tilde{\mathbf{D}}^{-1/2}\mathbf{W}\mathbf{\Lambda}^{-1}$ and $\mathbf{Y} = \tilde{\mathbf{D}}^{-1}\tilde{\mathbf{G}}'\mathbf{X}$ where $\tilde{\mathbf{D}} = \mathrm{diag}\tilde{\mathbf{G}}'\tilde{\mathbf{G}}$ and where $\mathbf{W}$ and $\mathbf{\Lambda}$ derive from the eigenvector/value decomposition $\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{G}}'\tilde{\mathbf{G}}\tilde{\mathbf{D}}^{-1/2} = \mathbf{W}\Lambda^2\mathbf{W}'$. This procedure generalizes equation (10.3) of Gifi (1990) to individual categories. Summation of indicator columns has probably been practiced for years in France, but the fact that it is equivalent to MCA under equality constraints is not mentioned in any of the standard references.

The number of different solutions will decrease under equality. The total number of solutions in the unrestricted solution is equal to $\Sigma_j (k_j + 1)$. Suppose we impose equality constraints on s variables, each having k categories, then the maximum dimensionality diminishes by a factor of $(s + 1)(k + 1)$. If $s = m$ (i.e., if all variables are restricted), the number of independent solutions is just $k - 1$.

Note that minimizing Equation 2 induces variables $\mathbf{G}_j\mathbf{Y}_j$ that are not necessarily in deviations from their means. It is not possible to require $1'\mathbf{D}_j\mathbf{Y}_j = 0$ for all j simultaneously because the marginal frequencies $\mathbf{D}_j$ may be unequal for different j. Because $1'\mathbf{X} = 0$, it will still be true that $1'\mathbf{DY} = 0$ (i.e., the grand average over all variables is zero), which is entirely in the tradition of correspondence analysis. In some applications, for example if we analyze ranking data or lagged variables, differences in marginal frequencies are negligible so the means of the induced variables then will be approximately zero.

Ranking Data

The analysis of ranking data is a straightforward application of equality. Ranking data are very common: athletes can be ranked from fast to slow, events can be ranked from most to least likely, universities can be ranked from most to least prestigious, stocks and shares can be ranked from most to least profitable, cities can be ranked from most to least rainy, and so on. Psychological research is often concerned with preference rankings that order items like foods, beverages, odors, people, and political parties from the least to the most preferred. If rankings are replicated over time or individuals we can use homogeneity analysis with equality constraints to determine one or more common latent ranking vectors.

To see how MCA applies, suppose that m individuals rank n objects and that these data are collected into an $n \times m$ matrix. Each variable corresponds to a particular ranking and consists of n categories. Categories are coded by the first n integers, with the highest preference being coded as "1". Ideally, homogeneity analysis on these data should produce a joint plot of objects and individuals in which individuals are located nearby their favorite objects. However, performing an unrestricted analysis is not appropriate, because each category is observed only once so that it is possible to obtain a perfectly homogeneous solution $\mathbf{y}_j = \mathbf{G}_j'\mathbf{x}$ for any centered, but otherwise arbitrary n-vector $\mathbf{x}$. To avoid this degeneracy we may impose an equality restriction: the ranking scale is considered to be the same for every judge. Thus all m first choices are located at the same spot, all m second choices are located at the same spot and so on. If there is much agreement among the judges the objects can be scaled on a single line; if not, more dimensions will be needed to describe the data adequately. The common scale assumption is appropriate if the discrepancies between the consensus ranking and the individual preferences are small. Homogeneity analysis aims to find the consensus ranking that minimizes these differences.

We analyzed a set of odor preference data collected by Moncrieff (1966, pp. 124-129). The subjects are 36 males and 66 females between 20 and 40 years of age. Table 1 contains the data, summed over the 102 individuals.

The task of each judge was to order 10 bottles of odors from the most to the least preferred. Some odors were pleasant, some of them were not. The 10 odors were:

1. Strawberry, excellent flavoring essence, high proportion of natural material.
2. Spearmint, excellent oil with a fine characteristics note.
3. French lavender, oil with high ester contents, a little fruity.
4. Musk lactone, synthetic, powerful smell, closely related to muscone.
5. Vanillin, synthetic, chemically identical to the odorant of the vanilla pod.

Table 1

Rank-order Frequency Matrix of 10 Odors Ranked by 102 Subjects

Odor	Rank									
	1	2	3	4	5	6	7	8	9	10
Strawberry	25	22	19	13	5	4	6	6	0	2
Spearmint	8	11	6	12	10	19	12	12	8	4
Lavender	22	16	12	11	13	7	11	8	1	1
Musk	12	10	11	14	13	15	14	8	3	2
Vanillin	6	20	16	16	15	11	9	2	5	2
Neroli	8	10	6	8	11	6	12	16	17	8
Almond	12	6	12	11	17	13	12	9	7	3
Naphthalene	8	6	17	13	11	14	12	14	7	0
Rape oil	1	1	3	3	4	11	13	17	33	16
Chlorophyll	0	0	0	1	3	2	1	10	21	64

6. Neroli oil, natural, highly priced, bright note with a "dazzling" smell.

7. Almond, flavoring essence, a very fine flavor.

8. Naphthalene, chemical, reminiscent of moth-balls and antique fire-lighters.

9. Rape oil, nutty and oily odor.

10. Oil-soluble chlorophyll, strong and unpleasant.

The restricted analysis amounts to performing correspondence analysis on the sum of all permutation matrices G_j, which is exactly what is shown in Table 1. A row in this table lists how many times the odor was ranked as first, as second, and so on. Hence all rows sum to 102. The first three eigenvalues for the restricted solution are 0.472, 0.110 and 0.034, so the first two dimensions capture most common information. The joint plot of odors and preferences is given in Figure 1.

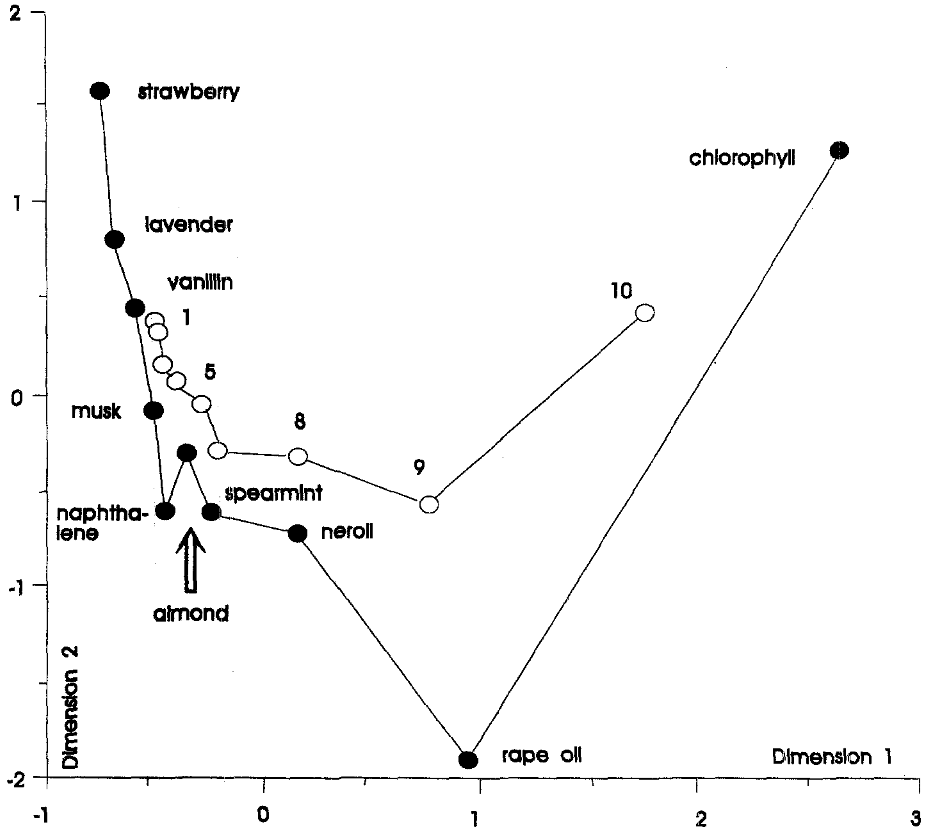

Figure 1
Joint plot of 10 odors and 10 rank points.

The distance between a pair of odors portrays the difference between their rank profiles that are given by the rows of Table 1. If two profiles are very dissimilar, for example, strawberry and chlorophyll, then their distance will be large. Distances between odor points thus reflect differences in their average ranking. The scale on which these differences are measured is defined by the set of rank points, which are equal to the restricted category quantifications. Rank points are located in the centroid of all odors that score on that rank, weighted by their frequency. Because the scores of all human raters consist of the same ten ranks the subject points coincide at the centre of the rank points, which is also the origin. It is therefore not possible to study differences between raters in this configuration.

The distance between two rank points may be interpreted as a measure of the average psychological difference between those ranks. In the example, the lower rank numbers, corresponding to a larger preference, are placed closely together. This indicates that although there certainly are perceptional differences between pleasant stimuli, these are not very large. On the other hand, the rank scores 8, 9 and 10 tend to be very distinct, not only from the lower rank scores, but also from one another. We interpret this finding, at least in the sample of odors studied here, that the pleasant odors are difficult to distinguish from each other but can be very well distinguished from unpleasant ones. At the same time, unpleasant scents themselves are also easy to separate. Apparently, nature has equipped us with an instrument that easily recognizes hazardous smells. Delicate odors have subtle distinctions. They are often confusing.

Missing Data

The occurrence of missing data is an important empirical problem. Homogeneity analysis allows for several strategies to deal with missing data, grossly subdivided into deletion and imputation (cf. Gifi, 1990, p. 73; van Buuren & van Rijckevorsel, 1992). It is sometimes imperative to distinguish between several types of missing data; we may have a separate category for "unanswered," one for "not reached, "one for "does not apply," one for "don't know," and so on. Unfortunately, it is not unusual to find that these categories dominate the MCA solution. One only needs three or four variables that have some common missing data, and chances are large that their category points will be located towards the periphery. Fortunately, this type of degeneracy can be avoided a great deal by restricting such missing categories to receive equal scale values. Of course, the restriction only makes sense if the constrained categories are comparable in some way.

As an example consider the small artificial data set in Table 2. Scale values and object scores are given in the same table. The first three eigenvalues of the

Table 2
Data and Results for an Artificial Missing Data Example

		Dimension					
		1	2	3	1	2	3
Data		Object Scores					
		Unrestricted			Restricted		
31121		-1.29	0.80	-0.94	-0.77	-1.42	1.32
33112		-0.53	1.51	1.43	0.28	0.70	1.24
33331		0.57	1.46	-0.66	0.20	-0.85	-0.15
13332		0.98	0.92	-0.26	0.11	-0.44	0.17
12333		1.63	-0.53	-0.49	0.23	-0.16	-0.65
11211		-0.73	-0.65	-1.23	-1.01	-0.90	-1.65
22333		1.37	-0.90	0.26	0.88	0.63	-0.18
22112		-0.38	-0.48	2.05	-0.51	0.98	0.78
22211		-0.46	-1.29	0.45	-0.89	0.04	-1.56
21222		-1.15	-0.83	-0.61	-1.52	-0.58	0.68
Var	Cat	Category Quantifications					
1	1	0.63	-0.09	-0.66	0.44	-0.50	-0.71
	2	-0.15	-0.88	0.54	0.51	0.77	-0.07
	3	-0.42	1.25	-0.06	0.89	-0.21	0.07
2	1	-1.06	-0.23	-0.93	-1.10	-0.97	0.12
	2	0.54	-0.80	0.57	0.18	0.87	-0.40
	3	0.34	1.30	0.17	0.89	-0.21	0.07
3	1	-0.73	0.61	0.84	-0.33	0.42	1.11
	2	-0.78	-0.93	-0.46	-1.14	-0.14	-0.84
	3	1.14	0.24	-0.29	0.89	-0.21	0.07
4	1	-0.53	-0.23	0.67	-0.53	0.71	-0.30
	2	-1.22	-0.01	-0.77	-1.15	-1.00	1.00
	3	1.14	0.24	-0.29	0.89	-0.21	0.07
5	1	-0.48	0.08	-0.59	-0.37	-0.53	-0.51
	2	-0.27	0.28	0.65	-0.16	0.42	0.72
	3	1.50	-0.72	-0.12	0.89	-0.21	0.07

Note. The restricted solution requires that category 3 (= missing) receives identical quantifications.

solution are 0.606, 0.426 and 0.320. Suppose that missing values are indicated by "3"s. These missing categories dominate both the first and the second dimension. This can most easily be seen from the quantifications which are large in dimensions 1 and 2. As a result, these most important axes hardly contain information on the actually observed, non-missing data.

If we require equality of category 3 across all variables then the missing categories obtain only one score instead of five. The effect on the solution is substantial. The first three eigenvalues are now 0.533, 0.321 and 0.272. Dimension 1 stays more or less the same (unrestricted and restricted object scores correlate 0.89), however the influence of missing data onto the second dimension reduces considerably. Dimension 2 of the unrestricted solution actually ceases to exist, whereas the new dimension 2 is almost equal to the former dimension 3 (their correlation is 0.90). It thus appears that equality suppresses the largely irrelevant previous dimension 2.

Event History Data

The last illustration applies equality to subsets of variables. We analyze the set of event history data given in Table 3 under three options: without equality, with equality on subsets and with equality on all variables simultaneously.

Table 3 contains data taken on 25 babies from Shirley (1931, Appendix 8). The data indicate the age in weeks of the babies when they started, respectively, stepping, standing, walking with help, and walking alone. Question marks indicate missing data. If there is a question mark in the first column this means that the babies were already stepping when the observation started. Max and Martin, who have a question mark in the second column, skipped standing and went directly from stepping to walking with help. Doris has a question mark in the last column, because she died before she could walk alone.

Table 4 codes the data by using 71 successive weeks covering the whole observation period. In each week the babies are in one of five states. State one is *not yet stepping*, state two is *stepping*, state three is *standing*, four is *walking with help*, five is *walking alone*, and zero is *missing*. We first analyzed the table with homogeneity analysis with "missing data deleted" (Gifi, 1990). We excluded Doris from the analysis. The first three eigenvalues of this analysis are 0.441, 0.361 and 0.246.

The five possible states correspond to five category quantifications. However most of these are zero at a given time point because usually only two or three states will be actually recorded. Figure 2 contains the first dimension of the quantifications for all variables plotted against time. To improve the display, zeroes were left out.

Table 3
Walking Data for 21 Babies

Infants	Weeks			
Martin	15	?	21	50
Carol	15	19	37	50
Max	14	?	25	54
Virginia	?	21	41	54
Sibyl	?	22	37	58
David	19	27	34	60
James	19	30	45	60
Harvey	14	27	42	62
Winifred	15	30	41	62
Quentin	15	23	38	64
Maurice	18	23	45	66
Judy	18	29	45	66
Irene May	19	34	45	66
Peter	15	29	49	66
Walley	18	33	54	68
Fred	15	32	46	70
Donovan	?	23	50	70
Patricia	15	30	45	70
Torey	?	21	72	74
Larry	13	41	54	76
Doris	?	23	44	?

The irregularity at the beginning of the stepping curve is easily explained. Larry was a slow walker, in fact the last baby to walk alone. But at the same time he was the first to go into the stepping phase. Thus the stepping curve at 13 weeks of age is determined only by Larry, and is this equal to his object score, which is 1.75. The large platform in the standing curve is the work of Torey. After 54 weeks Larry and Walley started to walk with help, leaving only Torey in the standing state. He remained there for another 18 weeks as the only baby, which is exactly the length of the platform. The height of the platform is 2.84, which is Torey's object score. There is also another small irregularity in the standing curve at the beginning. This is also due to Torey. In weeks 19 and 20 Carol was the only one standing. She is a very quick walker, and has the low object score -1.08. In week 21 Carol was joined by Virginia Ruth, another quick walker with score -0.71, and by Torey (of all babies) who was quite quick to stand, but continued in the standing state for about a year. His

Table 4
Data of Table 3 Recorded as Stage per Week, Weeks 9 -79

Martin	11111122222244444444444444444444444444444555555555555555555555555555555
Carol	11111122223333333333333333334444444444444555555555555555555555555555555
Max	11111222222222224444444444444444444444444444455555555555555555555555555
Virginia	00000000000033333333333333333333444444444444455555555555555555555555555
Sibyl	00000000000003333333333333334444444444444444444445555555555555555555555
David	11111111112222222233333334444444444444444444444444455555555555555555555
James	11111111112222222222333333333333333344444444444444455555555555555555555
Harvey	11112222222222222233333333333333344444444444444444444555555555555555555
Winifred	11111122222222222222333333333333444444444444444444444555555555555555555
Quentin	11111122222222333333333333333344444444444444444444444445555555555555555
Maurice	11111111122222333333333333333333333333444444444444444444444555555555555
Judy	11111111122222222222333333333333333344444444444444444444455555555555555
Irene May	11111111112222222222222223333333333344444444444444444444455555555555555
Peter	11111122222222222222222233333333333333334444444444444444455555555555555
Walley	11111111122222222222222222223333333333333333344444444444444555555555555
Fred	11111122222222222222222333333333333344444444444444444444444445555555555
Donovan	00000000000003333333333333333333333333333444444444444444444445555555555
Patricia	11111122222222222222333333333333333344444444444444444444444445555555555
Torey	00000000000033333333333333333333333333333333333333333333333334455555555
Larry	11112222222222222222222222222222333333333333344444444444444444444445555

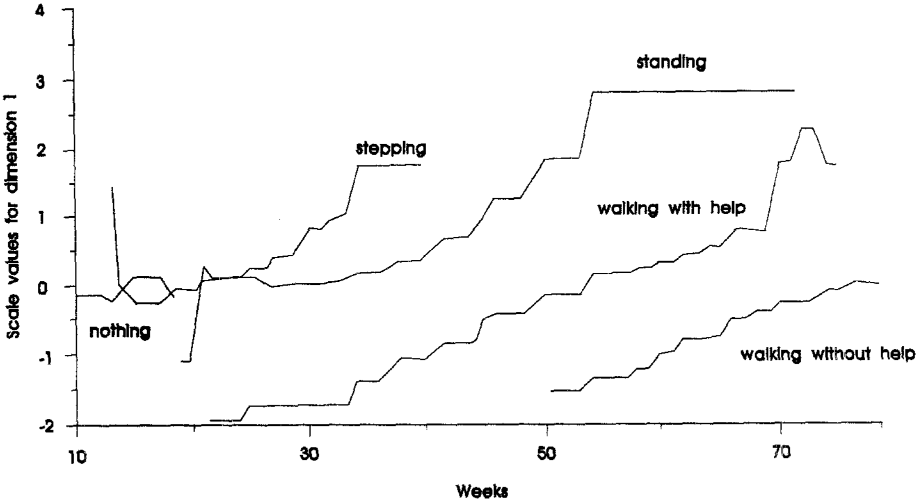

Figure 2
Quantifications plotted against time, unrestricted homogeneity analysis.

object score 2.84 causes the little jump in the beginning of the standing curve. Because baby development is typically measured in months rather than weeks we performed a second analysis in which time is grouped into 17 intervals, each of which approximately spans a month (each quarter was divided into 4, 4 and 5 weeks). The curves are restricted to be constant within months.

The object scores differ little from the scores found above. The first three eigenvalues are now 0.420, 0.333 and 0.216. These are smaller of course, but only slightly, and so, grouping weeks into months does not notably change the solution. The time curves are plotted in Figure 3 (next page). Equality smooths the curves such that most irregularities are removed. Of course "Torey's Plateau" is still visible.

The time curves reflect how the transition between states like stepping and standing varies over time. From a developmental point of view it is also interesting to study how babies differ for each other in choosing their routes from crawling to walking. Going one step further then, suppose we restrict all scale values to be equal across time. We then mimic the hypothetical situation in which all babies begin to step, stand and walk at the same time. Clearly, in such a situation all curves will be flat, and this is just what the equality restriction leads to (see Figure 4).

Because we neglect all growth variation now, the first three eigenvalues of the solution decline dramatically to 0.159, 0.088 and 0.020, respectively. The

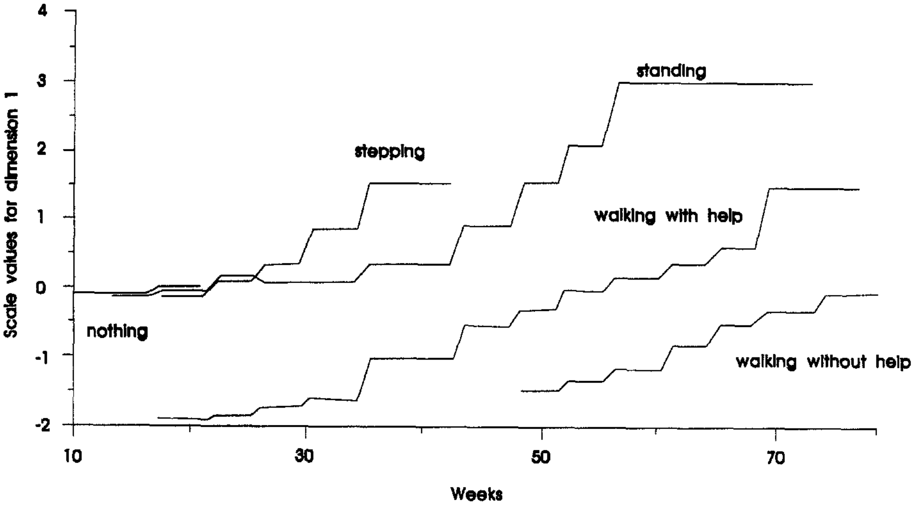

Figure 3
Quantifications plotted against time, restricted by month.

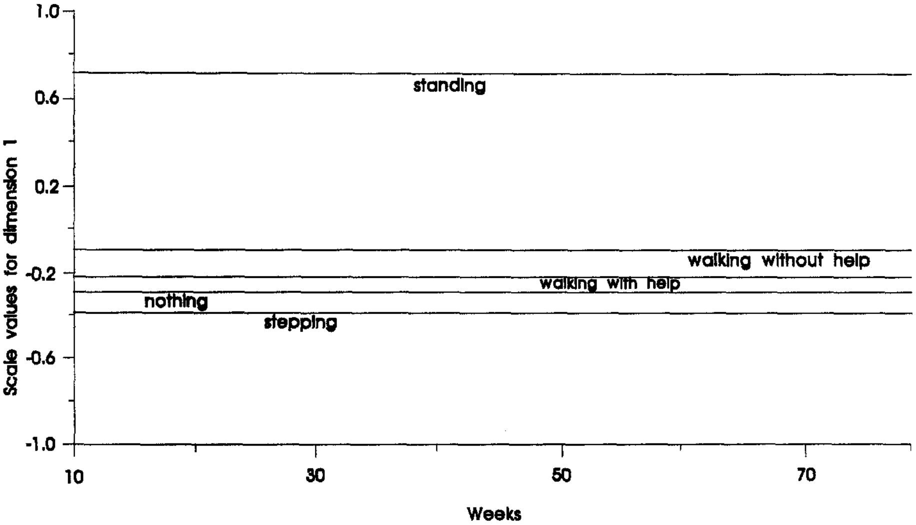

Figure 4
Quantifications plotted against time, restricted for all weeks.

configuration of object scores looks different now, though Torey still stands out.

Figure 5 displays the joint plot of babies and states for the first two dimensions. Dimension 1 is dominated by Torey and Donovan, who stood for 27 weeks. The "+" in the figure, indicating the location of the standing category, plainly loads on this dimension. The top position is reserved for Larry, who was the first to step but the last to walk. Martin, Max, Carol, Sybil and Virginia Ruth all could walk alone before they were 60 weeks of age. Differences between these early walkers are caused mainly by the time they needed to master the last phase. Despite the low eigenvalues associated with the axes the plot is easy to interpret. It portrays the salient differences in motoric development.

Conclusion

Let us make some final observations on equality. Equality provides a simple and straightforward extension of MCA. This article sketches some applications and interpretations of equality and outlines the accompanying computational procedures.

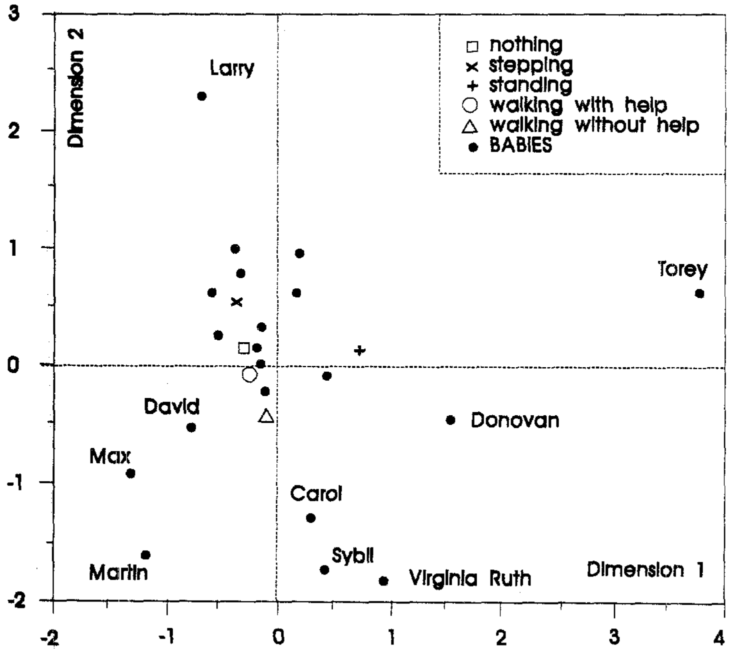

Figure 5
Joint plot for the completely restricted solution.

Because the method naturally fits into the Gifi system of nonlinear multivariate analysis, it is not difficult to extend the technique to more general situations. The equality restriction is actually a simple special case of the general linear constraint $\mathbf{Y}_j = \mathbf{S}_j\mathbf{Y}$ with arbitrary $\mathbf{S}_j$. If desired, our results can extended to other dependency patterns in $\mathbf{S}_j$. Likewise, allowing for mixed measurement levels does not present new problems. If variables are partitioned into subsets however, the loss function does not split nicely anymore and then we need a more complicated majorization approach. See van Buuren (1990) for details. Equality can be used in conjunction with the "missing data deleted" option without any problems. This is so because the $\mathbf{M}_j$ matrix that codes missing responses (Gifi, 1990) does not appear into the second and third components of Equation 2, where the weighting of non-missing entries is accounted for by $\mathbf{D}_j$.

It should be relatively easy to implement equality into existing software packages. We found that the resulting algorithm usually becomes somewhat faster, especially if equality is applied to all variables simultaneously. Most important of all however, there exist useful applications. Whenever categories can be interpreted as (nearly) identical it makes sense to consider equality. If there is no reason to treat categories one by one why should one do so? Equality provides an easy way to incorporate prior knowledge. Less parameters are needed and hence the stability and clarity of the solution are likely to increase.

References

Benzécri, J. P. (1973). *L'Analyse des données*. Vol 2. L'Analyse des correspondances. Paris: Dunod.

de Leeuw, J. (1973). *Canonical analysis of categorical data*. Unpublished doctoral dissertation, University of Leiden. Published in 1984 by DSWO Press, London.

de Leeuw, J. (1984). The Gifi-system of nonlinear multivariate analysis. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone (Eds.), *Data analysis and informatics III* (pp. 415-424). Amsterdam: North-Holland.

de Leeuw, J., van der Heijden, P. G. M., & Kreft, I. (1985). Homogeneity analysis of event history data. *Methods of Operations Research*, *50*, 299-316.

de Leeuw, J. and van Rijckevorsel, J. L. A. (1988). Beyond homogeneity analysis. In J. L. A. van Rijckevorsel and J. de Leeuw (Eds.), *Component and correspondence analysis*. New York: Wiley.

Deville, J. C., and Saporta, G. (1980). Analyse harmonique qualitative. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone (Eds.), *Data analysis and informatics*. Amsterdam: North-Holland.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

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

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw Hill.

Lebart, L., Morineau, A. & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis*. New York: Wiley.

Moncrieff, R. W. (1966). *Odour preferences*. London: Leonard Hill.

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

Saporta, G. (1981). *Méthodes exploratoires d'analyse de données temporelles*. Cahiers du BURO 37-38, ISUP. Paris: L'université P. et M. Curie.

Shirley, M. M. (1931). *The first two years. A study of twenty-five babies, Volume 1: Postural and locomotor development*. Minneapolis: University of Minnesota Press.

van Buuren, S. (1990). *Optimal scaling of time series*. Dissertation Utrecht. Leiden, Netherlands: DSWO Press.

van Buuren, S., and van Rijckevorsel, J. L. A. (1992). Imputation of missing categorical data by maximizing internal consistency, *Psychometrika, 57*, 567-580.

van der Heijden, P. G. M. (1987). *Correspondence analysis of longitudinal categorical data*. Dissertation Leiden. Leiden, Netherlands: DSWO Press.

Accepted December 10, 1991.