Equality constraints in multiple correspondence analysis

Stef van Buuren
Department of Psychology, University of Leiden

Jan de Leeuw
Departments of Psychology and Mathematics, UCLA

Stef van Buuren, Dept. of Psychology, University of Leiden
Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

Jan de Leeuw, Depts. of Psychology and Mathematics, University of California
405 Hilgard Ave., Los Angeles, CA 90024-1563, USA.

A previous version of this report appeared as UCLA Statistics Series # 54, April 1990.

Equality constraints in multiple correspondence analysis

The equality constraint is a simple extension of multiple correspondence analysis. It can be used to incorporate prior knowledge concerning relationships between categories that belong to different variables. In its simplest form, the restriction specifies that all variables receive identical data transformations. This is for example useful if the same variable is measured on different points of time. The paper discusses two variations on this idea: equality for unequal category numbers and equality among subsets of variables. Algorithms are given and the procedures are illustrated with examples.

Equality constraints in multiple correspondence analysis

1 Introduction

Multiple correspondence analysis (MCA) is a popular technique for analyzing categorical data. The idea of MCA is to scale categorical variables in such a way that the agreement among those scaled variables becomes maximal. Graphics play a dominant role. Standard references are Nishisato (1980), Lebart, Morineaux and Warwick (1984), Greenacre (1984) and Gifi (1990).

Every categorical variable consists of at least two categories. In MCA each of these categories is replaced by one or more numbers, here called scale values or category quantifications. Since there are always more variables, there are also more groups of categories. In conventional MCA categories of different variables are treated as if they were unrelated to each other. The goal of applying MCA is often just to uncover any structure among variables and categories.

Sometimes however, we know beforehand that categories are related. For example, for ordinal variables we know the order of the categories. If we wish to preserve this order in the quantifications we can require that the scale values are monotonely related to the category numbers. Such a restriction is defined on the categories *within* the same variable. It is an example of a within-variable restriction. Within-variable restrictions are quite common in MCA.

Another interesting but less familiar possibility is to have restrictions *across* variables. For example the OVERALS technique for multiset canonical correlation analysis is based on addition of categories that come from different variables (cf. Gifi, 1990) . In this paper we study another instance of across-variables restrictions, the equality constraint. Equality is both conceptually and computationally quite simple. The equality constraint demands that categories that belong to different variables receive identical quantifications.

Why do we need equality ? In general, equality restrictions can be helpful if variables are supposed to use 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. Clearly, the variable is the same at each time point and we may wish keep the scaling identical across all occasions.

Another example is a variable that appears more than once in the analysis but with its rows permuted in some way. This occurs frequently in time series analysis where all observations on a variable are shifted one or more positions and the relations between these lagged variables are studied.

A further example is a copied variable with altering values. Suppose we duplicate a variable and analyze both the original and the copy together. Now if an observation belongs to neither category one nor to category two, for example if the observation is missing, we may code it the first time into one and the second time into two. Then by requiring equality among these two variables we assure that this observation evenly contributes to both categories. In this way it is possible to analyze 'half categories'. But if we can handle half categories we can also handle threefold or fourfold categories by using three or four copies. In the limit, we can describe any rational point between two categories by a suitable ratio of copies. Copying variables under equality thus provides a means to analyze continuous variables.

Ranking data make up another class of applications. Suppose that 50 psychometricians rank 10 journals on, say, readability. We then obtain 50 readability variables on 10 objects. It would be interesting to derive a consensus ranking as well as some of the most typical deviations. As the variables can be seen as replications of each other we can assume the existence of an underlying common scale. An obvious way to implement such a common scale assumption is to restrict the obtained scale values per rank to be the same over 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, if many missing data are present in our observations they may considerably influence the MCA solution. The result is usually not very appealing. The trouble can be alleviated however if we specify that the scale values of all missing categories must be equal to each other. Likewise if we are not interested in certain

differences, 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.

There are two main advantages in using equality constraints. First, it allows us to incorporate prior knowledge (of the kind described above) into the analysis. This directly results into the second advantage: we need to estimate less parameters and hence it is likely that solutions will be stabler under arbitrary deletion of variables, categories or observations.

Some of the ideas presented here has been proposed before. De Leeuw (1973, p. 50, 160) illustrated the use of equality constraints applied to sorting data. Deville and Saporta (1980), Saporta (1981), De Leeuw, Van der Heijden and Kreft (1985) and Van der Heijden (1987) use equality constraints in the analysis of several types of longitudinal data. In these forms of MCA each columns represents a time point. In Van Buuren (1990) time points are in the rows and equality restrictions are placed on lagged variables. Gifi (1990) discusses the relationship with the method of successive intervals, a classic scaling technique (cf. Guilford, 1954).

Section 2 presents a concise description of MCA with a slight twist towards equality. Section 3, 4 and 5 discuss three varieties of equality: equality among all variables, equality among specific categories only, and equality among subsets of variables. These three sections all have a similar structure. The technical problems are defined and resolved first. After this each procedure is illustrated with a typical example. Section 6 draws conclusions and closes the paper.

2 Homogeneity Analysis

MCA can be formulated in many ways. In this paper we adopt the notation and terminology of the Gifi system (cf. De Leeuw, 1984; Gifi, 1990). In the Gifi system MCA goes under the name of homogeneity analysis.

2.1 *A loss function for homogeneity analysis*

Suppose we have n observations on m categorical variables, each with k_j distinct categories. Let G_j denote the $n \times k_j$ binary indicator matrix for variable j (j = 1,...,m). If observation i falls into category l of variable j then the l'th column of row i of G_j is equal to one. All remaining entries of this row are zero. Let Y_j denote the $k_j \times p$ matrix of scale values on p dimensions. Formally, the homogeneity problem can be formulated as minimizing

$$\sigma(X; Y_1,\ldots,Y_m) = \frac{1}{m}\sum_{j=1}^{m} SSQ\,(X-G_jY_j) \qquad (1)$$

over X and $Y_1,\ldots,Y_m$. We write SSQ(.) for tr(.)'(.). Equation (1) is known as the HOMALS loss function. In order to prevent degenerate and trivial solutions is $\sigma(.)$ often normalized with $1'X = 0$ and $X'X = nI$.

It is convenient to write (1) in a partitioned form. Let $D_j = G_j'G_j$ be the diagonal matrix of marginal frequencies and let $\tilde{Y}_j = D_j^{-1}G_j'X$ be the usual unrestricted least squares estimator. By inserting the identity $Y_j = \tilde{Y}_j + (Y_j - \tilde{Y}_j)$ into (1) and noting that the cross-product vanishes we obtain

$$\sigma(X; Y_1,\ldots,Y_m) = \frac{1}{m}\sum_{j=1}^{m} SSQ\,(X-G_j\tilde{Y}_j) + \frac{1}{m}\sum_{j=1}^{m} SSQ^{D_j}\,(Y_j-\tilde{Y}_j), \qquad (2)$$

where $SSQ^{D_j}(.) = tr(.)'D_j(.)$. The loss can be minimized over X and $Y_1,\ldots,Y_m$ by alternating least squares, that is, we fix one set of parameters while updating the other by least squares. Next, the roles are interchanged and the other set of parameters is updated, and so on until the configuration hardly changes anymore.

In this paper we will deal with restrictions on Y_j. Since Y_j appears exclusively in the term $SSQ^{D_j}(Y_j-\tilde{Y}_j)$ it suffices to be concerned with the minimization of the second portion of $\sigma(.)$ only. The results derived below all apply without changing anything in the update of X. This X-step can be summarized as $X = GRAM(\Sigma_j G_jY_j)$, where GRAM(X) denotes the Gram-Schmidt orthonormalization of X.

2.2 Equality and perfect homogeneity

The goal of homogeneity analysis is to make the induced scores G_jY_j as similar as possible. If X denotes the average of G_jY_j over j = 1,...,m then the perfect, but also rather trivial, homogeneity solution will satisfy $X = G_1Y_1 = \ldots = G_jY_j = \ldots = G_mY_m$. Observe that this solution will occur only if each category can be perfectly matched into one or more categories of all other variables.

An example of a perfectly fitting dataset is a consistent set of hierarchically nested variables. Here, higher level variables can be perfectly mapped into lower ones, which usually have more categories. For example, suppose we divide all people in our data into children and parents, and that we divide these two groups into males and females. Effectively, we have then classified everybody as being either a boy, a girl, a man or a woman. This defines a second, lower leveled variable with four 'type of people' categories. Performing correspondence analysis on these two nested variables (i.e. 'is child' and 'type of people') results in a solution in which the scores for 'boy' and 'girl' are equal to the score for the category 'child' (and hence they of course equal to each other).

Table 1 contains a small numerical example of a perfectly nested set. Since categories C and D are nested within A, and E is nested within B their quantifications are equal.

Data		Scale values	
A	C	A	0.817
A	C	B	-1.220
A	D	C	0.817
A	D	D	0.817
A	D	E	-1.220
A	D		
B	E		
B	E		
B	E		
B	E		

Table 1: A perfectly nested dataset with its optimal scale values

If, in addition, the number and the order of categories conform then $Y_1 = \ldots = Y_j = \ldots = Y_m$ will be true. For example think of 3 wine-tasters who rank 10 wines according to their preferences. Let these preference ranking scores be coded into three column variables, each with 10 categories. Thus there are 30 independent preference scale values. In the perfect case that each taster comes up with the same ranking the 30 distinct scale values will be located at no more than 10 different points. There is automatic equality now. There is absolutely no variation between the judges.

It will be clear that we can also require equality beforehand. We have seen above that a consistent hierarchically nested tree structure yields equal scale values. Provided that we choose a proper set of equality constraints, the reverse is also true. Thus, we can require that categories fit a hierarchical tree by choosing a suitable set of constraints. A rather

trivial instance of such a category tree is a tree without branches. If we want equal quantifications for all m variables there is a one-to-one correspondence among their k_j categories. We then have k_j branchless trees, each with m levels.

Throughout this paper we will see many uses of equality. The important point to remember is that equality can be used to analyze data as if certain parts of it were perfect. We act as if there was no variation between (parts of) variables. In this way we can 'model out' presumably uninteresting information. The likelihood of obtaining a succinct description of the data increases accordingly.

2.3 Graphical illustration of equality

To see what happens if we minimize (1) with and without equality constraints it is instructive to make a visualization of its individual constituents. Consider the small artificial data set in Table 2 consisting of 10 objects and 3 variables. Without any loss of information we may represent these data as in Figure 1a. We arbitrarily place all objects and all categories on a two-dimensional surface. For each score we connect the object and the category by a line. The resulting plot is quite a mess. Homogeneity analysis can now be used to make this display as clear as possible by relocating objects and categories. 'Messiness' is measured by the average (squared) length of all lines. This measure corresponds to (1). The simplest, but also least interesting, solution to this problem is to place all objects and categories on top of each other. To prevent this trivial solution we often require that the total distance between all objects is fixed in advance.

A	B	C
1	1	1
1	2	1
1	1	2
1	2	1
3	3	2
3	3	1
2	2	1
2	3	1
1	1	2
2	2	1

Table 2

Figure 1b displays the homogeneity analysis solution for this data set. The display is much clearer now and the average squared line length is the shortest possible in two dimensions. There is more than one way to interpret this kind of plot but a discussion of these possibilities would lead us too far away from the main topic. Figure 1c shows what happens if we require that variable A and B have equal scale values. For the sake of illustration let the three categories of A and B correspond with the outcomes 'pass', 'undecided' and 'fail' of two parallel tests. The scale values of both 'pass' categories, labeled 'a1' and 'b1' respectively, coincide under equality. Similarly, both 'undecided' and both 'fail' categories become equal.

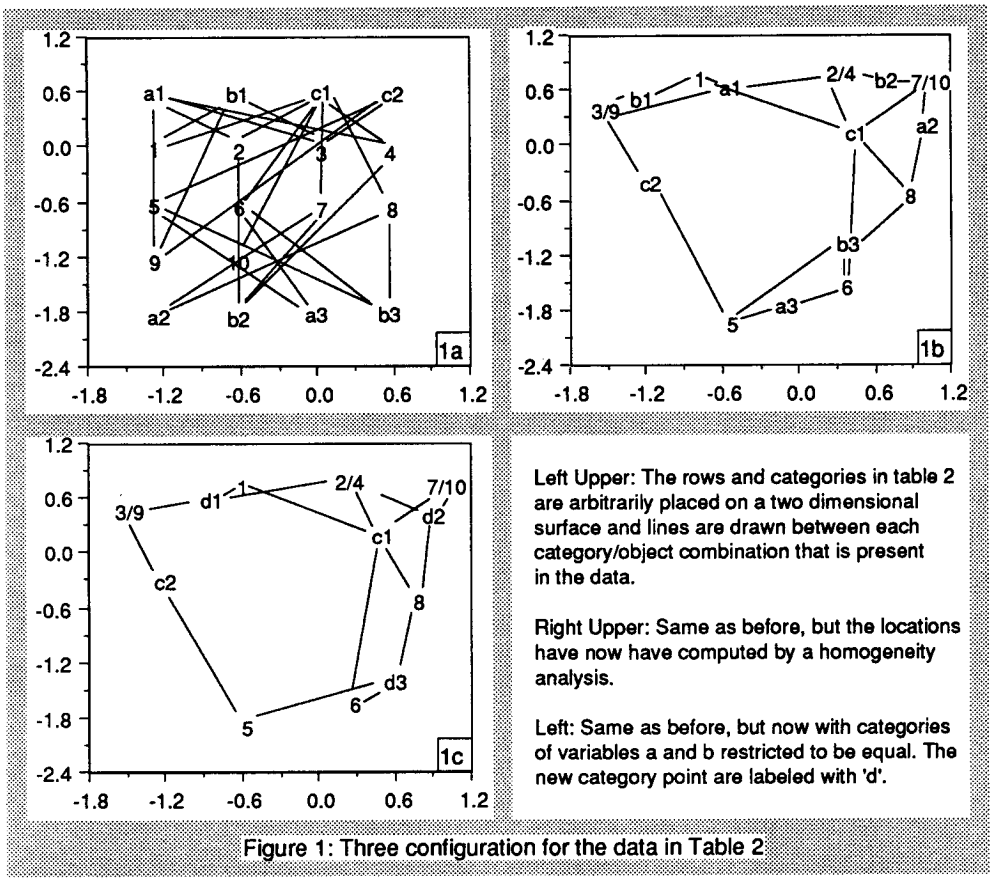

Figure 1: Three configuration for the data in Table 2

The average line length in 1c and thus our messiness measure is larger than in 1b. On the other hand many lines fall on top of each other so the overall picture actually appears cleaner. The number of different lines we need drops from 21 for the homogeneity solution to 16 for the equality solution. In practical data analysis the 'needed line ratio' is even more favorable since real data usually consist of more distinct profiles. The figure clearly illustrates the trade-off of the equality restriction. Although it leads to longer lines and a poorer fit, many lines fall on top of each other and the display becomes less crowded. Therefore equality is most appropriate if the plot simplifies considerably while the fit remains nearly the same.

3 Equal quantifications for all variables

In this section we study the problem in which we require that scale values of all variables coincide. In essence, we then treat the variables as if they were identical. In section 4 and 5 we describe two variations on this idea. First, we show how to handle situations in which the number of categories is not the same for each variable. Second, an extension of the method applied to subsets of variables is presented.

3.1 Theory

Suppose that we restrict all multiple quantification matrices Y_j to be equal, i.e.

$$Y = Y_1 = \ldots = Y_j = \ldots = Y_M. \tag{3}$$

This is of course only possible if all variables have an equal number of categories. Let $D = \Sigma_j D_j$ be the diagonal matrix that contains the total number of observations per category, and let $E = \Sigma_j D_j \tilde{Y}_j$ denote the weighted sum of intermediate scale values. Moreover define $\tilde{Y} = D^{-1}E$ as the centroids of these scale values. We write $\sigma(Y)$ for the second component of the loss function (2). Substituting Y for Y_j in $\sigma(Y)$ and inserting $Y = \tilde{Y} + (Y - \tilde{Y})$ then yields a second level between-within partitioning

$$\begin{aligned}\sigma(Y) &= \frac{1}{m}\sum_{j=1}^{m} SSQ^{D_j}(Y-\tilde{Y}_j) \\ &= \frac{1}{m}\sum_{j=1}^{m} SSQ^{D_j}(\tilde{Y}-\tilde{Y}_j) + \frac{1}{m} SSQ^{D}(\tilde{Y}-Y)\end{aligned} \tag{4}$$

If the scale vectors are not constrained any further then (4) attains its minimum for $Y = \tilde{Y}$. Thus equality requires us to add up over all indicator matrices, acting as if it were one new variable, and divide it by the total marginal frequencies D.

The number of different solutions will decrease under equality constraints. 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. The maximum dimensionality then diminishes by a factor of (s-1)(k-1). If s = m, i.e. all variables are restricted, the number of independent solutions is just k-1.

Note that minimizing (4) produces induced variables G_jY that are not necessarily in deviations from their means. It is in general not possible to require $1'D_jY = 0$ for all j simultaneously because the marginal frequencies D_j may be unequal. Since $1'X = 0$ it will be true that $1'DY = 0$, i.e. the grand average over all variables is still zero. This is entirely in the tradition of correspondence analysis. In many applications, for example in ranking data or in lagged data (i.e. the same variables though shifted in time) differences in marginal frequencies are negligible so the means of all separate induced variables will approximately be zero then.

3.2 An example on ranking data

A straightforward application of equality is in the analysis of ranking data. 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. In psychology, data are often collected in the form of preference rankings. These ranking order foods, beverages, odors, people, political parties id. from the least to the most preferred. In the case that rankings are replicated, for example over time, space 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 gathered in an $n \times m$ matrix. Notice that unlike normal MCA the individuals are in the columns. Every variable corresponds to an observed rank order and thus consists of n categories. By the nature of ranking data each category is observed only once. Suppose that the categories are coded as a permutation of the first n integers with the highest rank equal to 1 and the lowest rank equal to n. Homogeneity analysis on this data would produce a joint plot of objects and individuals in which individuals are located nearby their favorite objects. However, performing an unrestricted analysis is not very enlightening since the solution consists of solely trivial solutions (for any centered n vector x it is possible to obtain the perfectly homogeneous solution $y_j = G_j'x$). To avoid this degeneracy we need to impose some sort of restriction on the solution. It is possible to require that all y_j are monotonely related to the category numbers, a within-variables restriction. Degeneracy then usually disappears but we still have a separate scaling vector for each observation. Below we demonstrate a method that yields one or a few common ranking vectors.

Equality applied to ranking data implies that all categories corresponding to identical (preference) ranks will obtain identical scale values. Thus all m first choices are located at the same spot, all m second choices are located at the same spot and so on. As a result, favorite objects will be located near the top rank categories while disliked objects will be positioned towards later ones. If there is much agreement among rankers the objects can be scaled on a single line in one dimension. Under equality, the number of free scale values per dimension drops from nm to n. The price we pay for this stabler solution is that any differences between scale use of (groups of) individuals are ignored. All individuals are treated as replications of each other. Whether this is correct depends on the data of course. Small discrepancies between the overall and the individual ranking vectors indicate that the assumption is appropriate. Homogeneity analysis minimizes these differences.

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

Table 3: Rank-order frequency matrix for 10 odors ranked by 102 subjects

We analyzed a subset of odor preference data collected by Moncrieff (1966, pp. 124-129). The subjects are 36 males and 66 females between 20 - 40 years of age. Each subject was asked to rank 10 bottles of odors from most preferred to least preferred. Some odors were pleasant, some of them were not. The odors are:

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 essential odorant of the vanilla pod

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 antic fire-lighters
9. Rape oil, nutty and oily odor
10. Oil-soluble chlorophyll, strong and unpleasant

If we require equality over all subjects then the analysis amounts to performing correspondence analysis on the sum of all permutation matrices G_j. This sum matrix is known in psychophysics as the rank-order frequency matrix (cf. Guilford, 1954, Ch. 8). It is shown in Table 3. 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. Strawberry and lavender are popular, chlorophyll clearly is not.

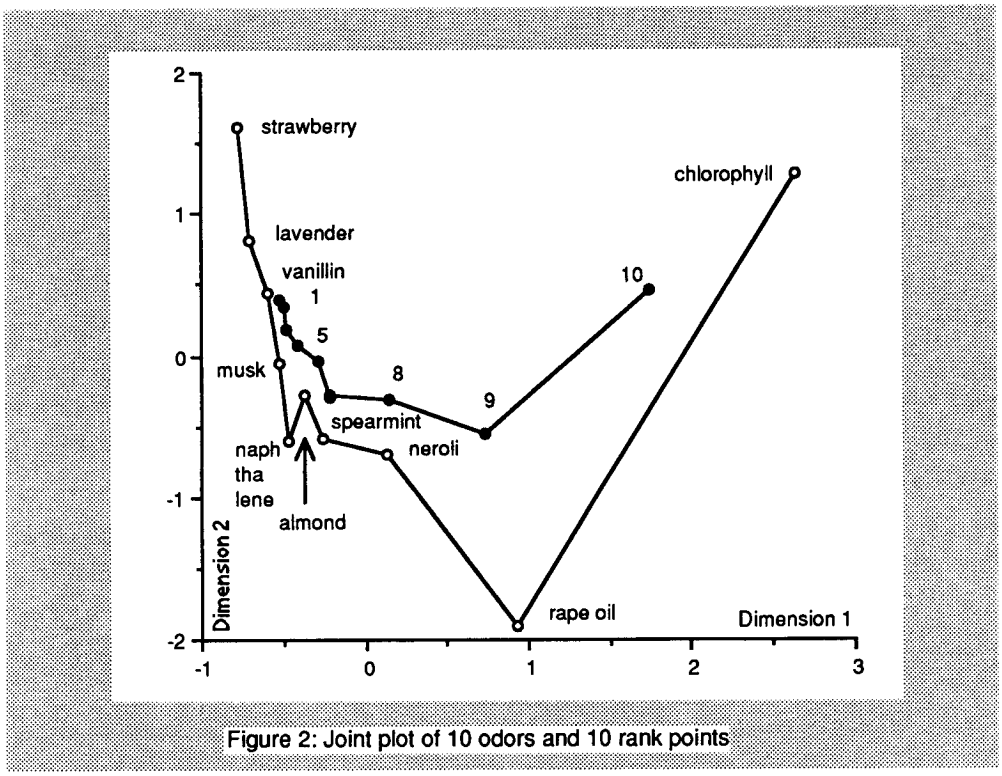

Figure 2: Joint plot of 10 odors and 10 rank points

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 the ten odors and the ten preference categories is given in Figure 2. The object and rank points are connected with each other. The object points correspond to the 10 odors. The distance between a pair of odors portrays the difference between of their rank profiles. If two profiles are very distinct, for example strawberry is very different for chlorophyll, then the distance will be large. Put in another way, distances between odor points reflect differences in their average ranking. The scale on which these differences are measured is given by the set of rank points. Rank points are restricted category quantifications. They are located in the centroid of all odors that score on that rank, weighted by their frequencies.

The distance between two rank points may be interpreted a measures of the average psychological distance between those two ranks. In the example, the lower rank numbers, corresponding to a large preference, are very close to each other. 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, 10 tend to be very dissimilar, not only from the lower rank scores, but also from one another. We may interpret this finding that, at least in the sample of odors studied here, the pleasant odors are difficult to distinguish from each other but can be very well distinguished from unpleasant ones. At the same time, the 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.

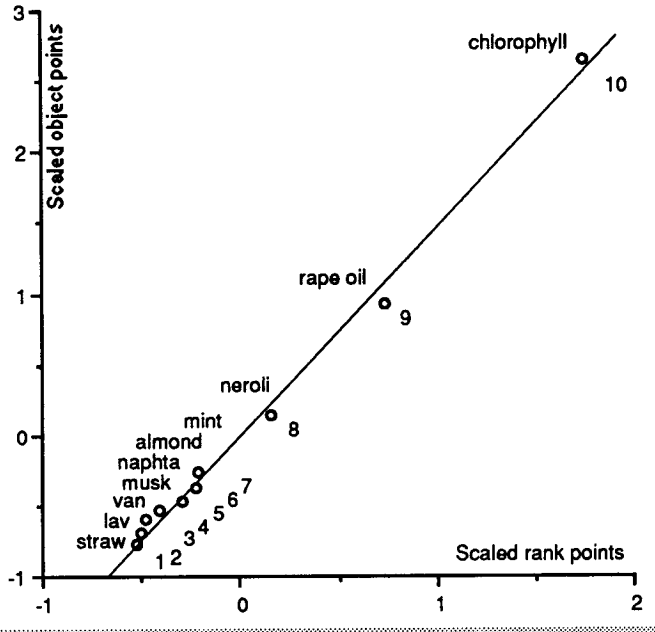

Figure 3: Matched rank-object plot for the odor preference data

We also plotted the odor values in the first dimension against their matched rank points. We ranked all odors from most to least preferred using the average preference curve drawn in Figure 2 and displayed their scores against the rank points (cf. Figure 3). In the ideal case, where all individuals come up with the same ranking, these points will be located on a line. Here this is almost the case. The correlation between the matched ranks and the odor point is no less than 0.994. The regression line plotted in Figure 3 can be used to infer the predicted object score given the rank point of the object. Since the agreement among both sets of parameters is very good it does not make a great difference whether we use rank points or object points for describing variations between odor preferences.

4 Equality of certain categories only

The equality condition (3) will be too restrictive in many practical cases since it requires an equal number of categories that, in addition, all have a one-to-one correspondence. If we deal with varying category numbers, or with variables in which categories are mixed up, then (3) will not work satisfactory. However, it is not difficult to link individual categories to the levels of a common scaling matrix Y. Through this common Y it is possible to equate any pair, triad, quartet id. of categories.

4.1 Theory

Let us study the case in which we require that

$$Y_j = S_j Y \tag{5}$$

for each j = 1,...,m. To keep things simple we suppose that S_j is a known indicator matrix. Let S_j have k_j rows, indexed by s = 1,...,k_j, and k columns indexed by t = 1,...,k. If category s of variable j is linked to level t of Y then element (s,t) in S_j is equal to one. Thus by setting a proper allocation pattern in each S_j category quantifications can be equated indirectly through a common Y. If all S_j are identity matrices then (5) reduces to (3).

Let $D = \Sigma_j S_j'D_jS_j$ denote the k × k diagonal matrix that holds the frequencies of each common level totalled over all variables. Let $E = \Sigma_j S_j'D_j\tilde{Y}_j$ be the k × p matrix that sums all intermediate category quantifications in the common category system defined by Y. Again let $\tilde{Y} = D^{-1}E$ be the centroid of these quantifications. Under (5) the loss component for Y then becomes

$$\begin{aligned}\sigma(Y) &= \frac{1}{m}\sum_{j=1}^{m} SSQ^{p_j}(S_jY-\tilde{Y}_j) \\ &= \frac{1}{m}\sum_{j=1}^{m} SSQ^{p_j}(S_j\tilde{Y}-\tilde{Y}_j) + \frac{1}{m} SSQ^{p}(\tilde{Y}-Y).\end{aligned} \tag{6}$$

If Y is not constrained any further the minimum of (6) can be found by setting $Y = \tilde{Y}$.

Observe that (5) allows for a much broader class of relationships than just equality. The number of levels in Y, that is k, can be chosen independently of k_j. If $k \geq k_j$ for all j then each category of G_j can be linked to a distinct level of Y and we can have a one-to-one correspondence among categories of different variables. For $k < k_j$ this is not possible, but one may choose to require that categories *within* a variable should be treated as identical. It is also simple to eliminate categories from the analysis by setting entire rows to zero.

For more general S_j it is possible to generate complex patterns of dependencies among categories and, at a higher level, variables. The interpretation of D and E becomes more complicated then. In general it is instructive to study the product matrix G_jS_j to see what happens to the data. If both G_j and S_j are indicator matrices then G_jS_j is also an indicator matrix. A thorough treatment of all possibilities requires a paper on its own so we will not explore this matter any further here. Nevertheless, all results derived in this paper apply to general S_j. The appealing centroid interpretation of MCA may be lost though.

4.2 An example on missing data

The majority of empirical research in the behavioral sciences is based on questionnaires. A typical problem in the analysis of these data is the occurrence of missing data. An observation can be missing for a variety of reasons. For example the subject was unwilling to fill in the question, an experimental animal died or the subject did not reach the item. Missing data can also be systematic. Many questionnaires contain items that only need to be answered if the subject fulfils a particular condition. For example, items on partners need only be answered by those who actually have a partner and can be skipped by others. Or, items on computer fatigue need only be answered by computer users.

An important consequence of this is that each variable contains an extra hidden category, namely 'missing'. Sometimes it is useful to imagine that there are even more of those extra categories. For example, one 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 observations. Then, 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 extra categories to have equal scale values.

			dimension					
			1	2	3	1	2	3
		data	unrestricted			object scores		restr
		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	design matrix S	unrestricted			quantifications		restr
1	1	10000000000	0.63	-0.09	-0.66	0.44	-0.50	-0.71
	2	01000000000	-0.15	-0.88	0.54	0.51	0.77	-0.07
	3	00000000001	-0.42	1.25	-0.06	0.89	-0.21	0.07
2	1	00100000000	-1.06	-0.23	-0.93	-1.10	-0.97	0.12
	2	00010000000	0.54	-0.80	0.57	0.18	0.87	-0.40
	3	00000000001	0.34	1.30	0.17	0.89	-0.21	0.07
3	1	00001000000	-0.73	0.61	0.84	-0.33	0.42	1.11
	2	00000100000	-0.78	-0.93	-0.46	-1.14	-0.14	-0.84
	3	00000000001	1.14	0.24	-0.29	0.89	-0.21	0.07
4	1	00000010000	-0.53	-0.23	0.67	-0.53	0.71	-0.30
	2	00000001000	-1.22	-0.01	-0.77	-1.15	-1.00	1.00
	3	00000000001	1.14	0.24	-0.29	0.89	-0.21	0.07
5	1	00000000100	-0.48	0.08	-0.59	-0.37	-0.53	-0.51
	2	00000000010	-0.27	0.28	0.65	-0.16	0.42	0.72
	3	00000000001	1.50	-0.72	-0.12	0.89	-0.21	0.07

Table 4: Data and results for the artificial missing data example. The restricted solution has equal quantifications on category 3.

As an example consider the small artificial dataset in Table 4. The first three eigenvalues of the MCA solution are 0.606, 0.426 and 0.320. Scale values and the object scores are also given in Table 4. 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 their quantifications. They exhibit the highest magnitudes in Dimensions 1 and 2. The result is that information on observed data is snowed under by the presence of missing data, at least in the first two dimensions.

If category 3 is restricted to receive identical scale values then missing scores attain only one position instead of five. The first three eigenvalues are now 0.533, 0.321 and 0.272. Dimension 1 stays more or less the same. The correlation between the unrestricted and restricted object scores is 0.89. The influence of missing data in the second dimension has diminished considerably. The old Dimension 2 has actually ceased to exist. Old Dimension 3 it almost equal to the new Dimension 2 (their correlation is 0.90). We won a dimension.

5 Equality in subsets of variables

Despite the ability to deal with varying category numbers our methodology is still somewhat limited from a practical point of view. Suppose it would be possible to construct several subsets of variables and that equality only applies within these subsets. Then we could for example make a set of, say, mood scores measured on 28 consecutive days and a set of scores that indicates the day of the week. Subsequently we could restrict the 28 variables within each set to retain a common scale. A method like this would considerably reduce the number of independent, but also largely redundant, scaling parameters and therefore it would probably provide stabler results. The reader will not be very surprised to learn that such a procedure in fact *is* possible.

5.1 Theory

Allowing for subsets is simple. The only thing we must do is change the index in the summation signs of (7). So instead of summing over all variables we just sum over those variables included into the set and do so for each set separately. To describe this more precisely, let the variables be partitioned into s sets C_r (r = 1,...,s). For variable j belonging to set r we have equality of the form

$$Y_j = S_j Y_r, \tag{7}$$

where Y_r is a the common scaling matrix of order $k_r \times p$. Thus instead of one common Y there are s set-wise common Y_r's. Define $D_r = \sum_{j \in C_r} S_j' D_j S_j$ as a $k_r \times k_r$ diagonal matrix of marginal frequencies summed over the variables of set r and let $E_r = \sum_{j \in C_r} S_j' D_j \tilde{Y}_j$ stand for

the totalled scale values of set r. Let $\tilde{Y} = D^{-1}E$ once more. The problem is now to minimize

$$\sigma(Y_1,\ldots,Y_s) = \frac{1}{m}\sum_{r=1}^{s}\sum_{j\in C_r} SSQ^{p_j}(S_jY_r-\tilde{Y}_j)$$
$$= \frac{1}{m}\sum_{r=1}^{s}\sum_{j\in C_r} SSQ^{p_j}(S_j\tilde{Y}_r-\tilde{Y}_j) + \frac{1}{m}\sum_{r=1}^{s} SSQ^{p_r}(\tilde{Y}_r-Y_r) \qquad (8)$$

over $Y_1,\ldots,Y_s$. Again, the second part vanishes if we set $Y_r = \tilde{Y}_r$.

Let us take a look at two extreme cases of (8). Suppose that there is only one set, i.e. s = 1. Then (8) immediately reduces to (6). If $S_j = I$ for all j then it further simplifies to (4). Thus category equality can be seen as a special case of subset equality where we have one set. Note that by choosing the right S_j matrices we can also obtain the reverse, that is, subset equality as a special case of category equality. In the other extreme case we have s = m with no equality restrictions at all. Then $Y_r = Y_j$ and we are back at the normal homogeneity loss function given in (2). Thus by using subsets we have the opportunity to proceed gradually from an entirely unrestricted to completely restricted analysis. Below we will see an example of how we may use this fact to our advantage.

It is also possible that a variable is allocated to more than one subset. Then (8) is still valid but (7) does not hold anymore. Instead an optimal Y_j will be located in the centroid of its corresponding set scale values. In formula

$$Y_j = 1/s_j \sum_{j\in C_r} S_{j_r}Y_r, \qquad (9)$$

where s_j is the number of subsets to which variable j belongs. Again, the centroid principle holds.

5.2 An example with event history data

Martin	15	?	21	50
Carol	15	19	37	50
Max	14	?	25	54
Virginia Ruth	?	21	41	54
Sibyl	?	22	37	58
David	19	27	34	60
James D.	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	?

Table 5: 21 babies

Table 5 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. In Table 6 the data are coded by using 71 successive weeks covering the whole observation period. In each week the babies are in one of six 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 analyzed the table with homogeneity analysis with 'missing data deleted' (cf. Meulman, 1982; Gifi, 1990). We deleted Doris from the analysis. The first three eigenvalues of this homogeneity analysis are 0.441, 0.361 and 0.246.

Martin	1	11111122222244444444444444444444444444444455555555555555555555555555
Carol	2	11111122223333333333333333344444444444445555555555555555555555555555
Max	3	11111222222222224444444444444444444444444444445555555555555555555555
Virginia	4	00000000000033333333333333333334444444444444555555555555555555555555
Sibyl	5	00000000000003333333333333334444444444444444444455555555555555555555
David	6	11111111112222222233333344444444444444444444444444455555555555555555
James D.	7	11111111112222222222233333333333333344444444444444445555555555555555
Harvey	8	11111222222222222233333333333333344444444444444444455555555555555555
Winifred	9	11111122222222222222333333333334444444444444444444455555555555555555
Quentin	10	11111122222222333333333333333444444444444444444444444455555555555555
Maurice	11	11111111122222333333333333333333333344444444444444444444455555555555
Judy	12	11111111122222222222333333333333333344444444444444444444455555555555
Irene May	13	11111111112222222222222223333333333344444444444444444444455555555555
Peter	14	11111122222222222222333333333333333333344444444444444444455555555555
Walley	15	11111111122222222222222333333333333333333344444444444444445555555555
Fred	16	11111122222222222222223333333333333444444444444444444444444455555555
Donovan	17	00000000000000333333333333333333333333344444444444444444444455555555
Patricia	18	11111122222222222222233333333333333344444444444444444444444455555555
Torey	19	00000000000033333333333333333333333333333333333333333333333333333445
Larry	20	11112222222222222222222222222233333333333344444444444444444444444444

Table 6: Data of table 5 in a recoded form, weeks 9 - 79

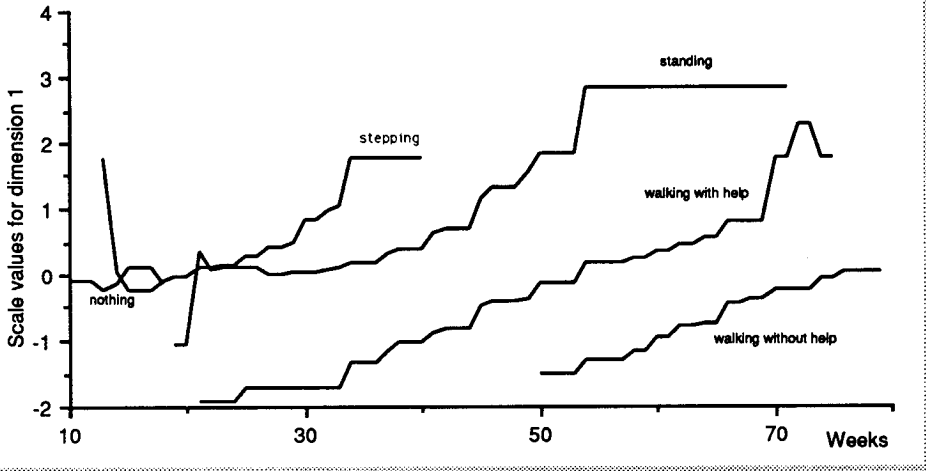

Figure 4: Quantifications plotted against time, unrestricted homogeneity analysis

Since each time point has five possible states Y_j contains five category quantifications. Most of these are zero however since at a given time point usually only two or three states are actually recorded. Figure 4 contains the first dimension of the Y_j for j = 1,...,71 plotted against time. To improve the display, zeroes were left out.

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. It 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 object score 2.84 causes the little jump in the beginning of the standing curve.

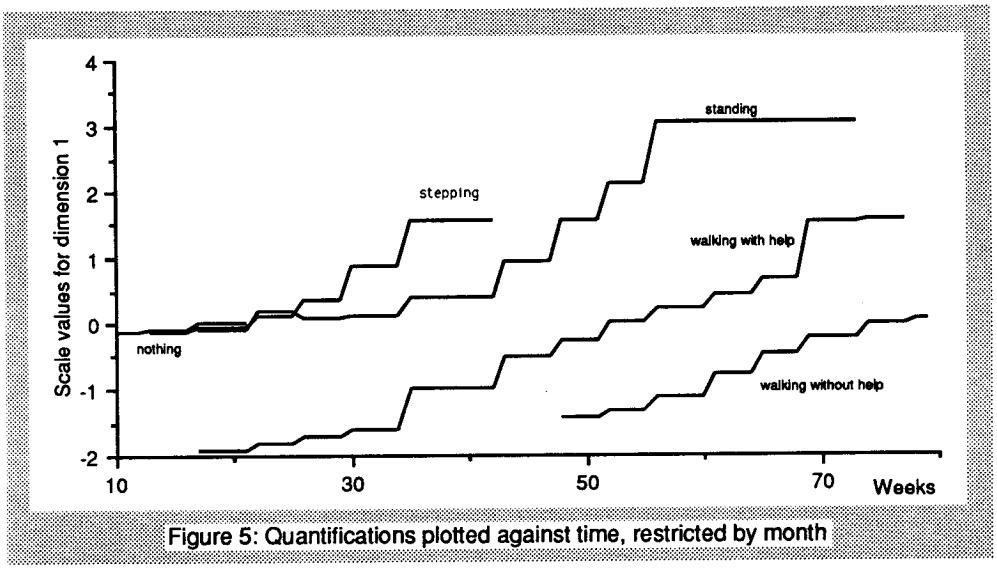

Figure 5: Quantifications plotted against time, restricted by month

Since baby development is typically measured in months we also performed an alternative 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 in those intervals.

Object scores differ little from the scores of the more general analysis given before. 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 5. The smoothing we have introduced removes most of the irregularities. 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. In this case we 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 6).

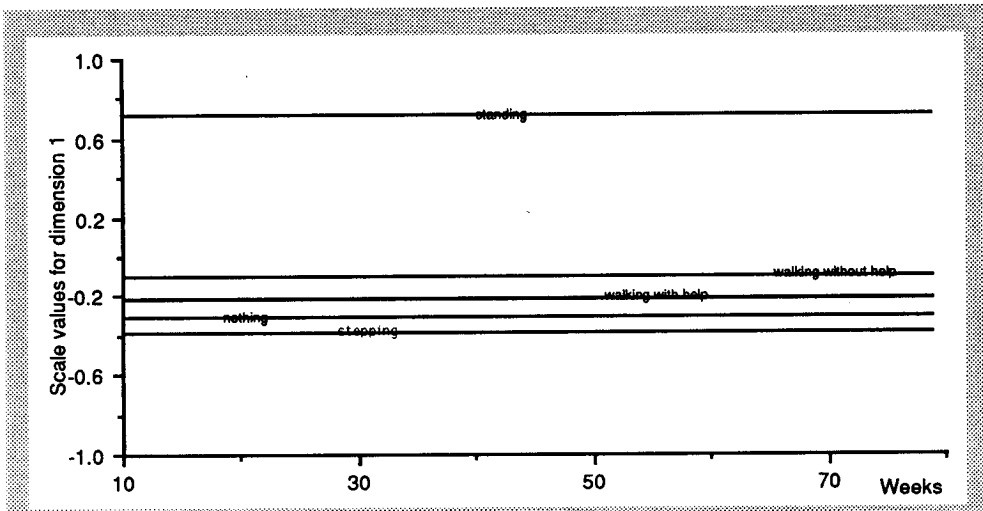

Figure 6: Quantifications plotted against time, restricted for all weeks

Since we neglect all growth variations here, the first three eigenvalues of the solution decline dramatically to respectively 0.159, 0.088 and 0.020. The configuration of object scores looks different now, though Torey still stands out with distinction.

Figure 7 displays the joint plot of babies and states for the first two dimensions. Dimension 1 is dominated by Torey and Donovan, who stood for a respectable 27 weeks and who later hit the charts for the same span of time. The '+' in the figure, indicating the location of the standing category, plainly 'loads' on this dimension. The top position is reserved for Larry. He 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 is a portrayal of the differences among motoric development of the babies.

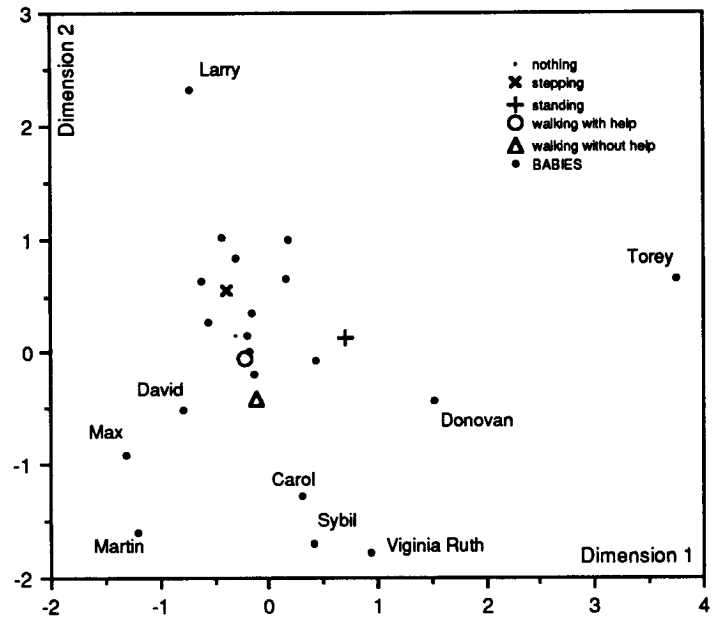

Figure 7: Joint plot for the completely restricted solution

6 Conclusion

Let us make some final observations on the equality constraint. Equality restrictions aim at equalizing scale values. As such they provide a simple and straightforward extension of MCA. This paper collects a number of applications and interpretations of equality. It also provides the accompanying computational procedures.

The work is easy to generalize. The equality constraint is actually a simple special case of the general linear constraint $Y_j = S_j Y$ given in (6) where S_j is an indicator matrix. If desired, our results can also be applied to analysis forms that go beyond equality by specifying other dependency patterns in S_j. Likewise, allowing for mixed measurement levels does not present many new problems. If variables are partitioned into subsets however, the loss function does not split nicely anymore and we need a majorization approach then.

An important property of the equality constraint is that it does not affect the standard ways of dealing with missing data. This is so because the second part of (2) remains unchanged if we introduce the so-called M_j matrix that codes missing responses (cf. Meulman, 1982; Gifi, 1990).

The formulas given above only involve the steps needed for finding quantifications. Therefore it should be relatively easy to modify existing MCA computer programs to integrate equality. We found that the resulting algorithms usually become somewhat faster, especially in the case of complete equality of Section 3. A computational alternative is to use SVD on suitably summed indicator matrices. This approach is taken by Van der Heijden (1987, Ch. 6, 7) in the analysis of panel data.

Most important of all however, there are useful applications. Whenever categories can be interpreted as (almost) identical it makes sense to consider equality. If there is no reason to treat categories separately 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 is likely to increase.

References

De Leeuw, J. (1984). The Gifi-system of nonlinear multivariate analysis. In E. Diday et al. (Eds.), *Data analysis and informatics III*, 415-424. North-Holland, Amsterdam.

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

Deville, J.-C. & Saporta, G. (1980). Analyse harmonique qualitative. In E. Diday et al. (Eds.), *Data analysis and informatics*. North-Holland, Amsterdam.

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

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

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

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

Meulman, J.J. (1982). *Homogeneity analysis of incomplete data*. DSWO Press, Leiden.

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

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

Saporta, G. (1981). *Méthodes exploratoires d'analyse de données temporelles*. Unpublished doctoral dissertation, l'université P. et M. Curie, Paris.

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

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

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