NONLINEAR METHODS FOR THE ANALYSIS OF HOMOGENEITY AND HETEROGENEITY

Willem J. Heiser
Jacqueline J. Meulman

Department of Data Theory
University of Leiden

RR-93-03

Nonlinear Methods for the Analysis of Homogeneity and Heterogeneity

Introduction

Although nonlinear phenomena are ubiquitous in science and technology, the majority of methods for analyzing statistical data is linear. Part of an answer to the puzzling question of how such a discrepancy could have come into existence would seem to be that first approximations are hard to beat. Statistics often enters the stage when the subject of study is badly understood, and then it is sensible to use linear approximation as a first try. If linear approximation does reasonably well, not unfrequently after being aided by quite some effort on data cleaning and readjustment of residuals, the returns from bringing in nonlinearity are expected to be low. Another reason is the fact that linearity is a versatile and universal concept in statistics: a model may predict nonlinear regression (as in curve fitting), yet it will be called linear as long as it is linear in the parameters; an estimate may determine the shape of a nonlinear function (as in density estimation), yet the estimator will be called linear as long as it is linear in the data. In multivariate analysis, we often work in a linear space, with linear operators, under linear constraints, and with computation methods that have linear convergence rate.

The nonlinear methods to be discussed in this chapter share the characteristic that they try to catch certain nonlinear phenomena, even though they may have linear aspects too. This goal is achieved by invoking some family of nonlinear functions or transformations, and selecting one or a few members of that family in such a way that a particular statistical (or data analytical) aim is served by it. Many statistical aims involve a partitioning of variability into different homogenous components, and therefore it is useful to look at methods that maximize homogeneity or try to articulate heterogeneity. Our discussion starts with a closer look at these two related, fundamental concepts.

Homogeneity and heterogeneity of distributions

The distinction between homogeneity and heterogeneity is most frequently used in statistics in connection with samples from different natural populations, which may – or may not – exhibit identical behavior, or display similar characteristics. More generally, however, it refers to a qualitative comparison of *distributions*, which can be called, for example, homogeneous in their

variance and heterogeneous in their means. A distribution represents the variability (of either some sample or some population) in a certain aspect or characteristic. If a number of distributions are regarded as being the same, most often as the result of a hypothesis or a decision on the part of the data analyst (not necessarily a *formal* decision – it may well be by judgment), they are collectively called *homogeneous*, meaning *of one kind.* The most familiar example is the "independent identically distributed" assumption for random samples from a single source. If not regarded as being the same, distributions – or their elements – are called *heterogeneous*, meaning *of different kinds.* As is evident from the earlier example of the idealized circumstances in which the F-test applies, a set of distributions can be homogeneous in one aspect and heterogeneous in another, underlining the qualitative character of the distinction.

The relation between homogeneity and heterogeneity is complex. First, it is asymmetric, in the sense that there are fewer possibilities to be "of one kind" than there are to be "of different kinds". In addition, it is a distinction that cannot be drawn on purely empirical grounds. The simple fact that two statistical units (or *objects*) have different measured values does not tell us much in this respect. Some observed difference of whatever magnitude does not contradict homogeneity – stochastic variation of a single kind – when initially assumed to be from one sample or population, while the same difference could confirm heterogeneity – systematic plus stochastic variation – when considered under the assumption of two or more samples or populations.

In multivariate analysis, the abstract concept of a distribution is to be preferred over sample or population, because it allows us to avoid the ambiguous use of these concrete terms when other aggregates than populations of individuals – for instance, sets of residuals or sets of variables – are being regarded as stochastic entities. Generally, the elements of any data analysis problem can be either fixed or stochastic; it is proper to use the latter term when and only when considering a distribution. The distinction between homogeneity and heterogeneity can be seen as a refinement of the more basic distinction between fixed and stochastic. Multivariate data are standardly arranged in a rectangular data matrix of units by variables, but such an arrangement leaves wide open the question of what should be regarded as fixed and what as stochastic in some specific form of data analysis. Indeed, it turns out that by applying this distinction not only to the set of units but also to the set of

variables, we obtain a novel classification of multivariate methods. The classification enables us to outline the material in the three parts of this chapter, and to put them into a broader perspective.

Fixed and stochastic (homogeneous or heterogeneous) sets of units and variables

All multivariate problems are characterized by the presence of several variables of interest, the analysis variables h_j, with $j = 1,...,m$. Each variable h_j refers to a group of measurements or observations on the same N units, individuals, or objects. Suppose that the set of units may be further partitioned into subsets, in such a way that the units are exchangeable within subsets, while between subsets they are not. Exchangeability is used here as an intuitively clear notion, but can be given a more formal treatment in terms of similarity (Draper *et al.*, 1993). Ideally, the partitioning forms – borrowing a phrase from Fisher (1956) – a *recognizable stratification* of the units. This condition implies that it is always possible to tell beforehand to which well-defined subset some designated unit belongs. Less ideally, there is merely a hypothetical stratification that still has to be verified or recognized. Whenever some stratification is at issue, the set of units is said to be *heterogeneous.*

Two interesting extreme cases may be noted. If there is only one subset, the units are without recognizable stratification and are completely mutually exchangeable, thus forming a *homogeneous* distribution. If there are N subsets, all units are recognizable and not exchangeable, and therefore they no longer constitute a distribution. In this case the set of units is said to be *fixed.*

To regard the set of variables as fixed means no more than that they serve to span the space in which the units vary. In the process of analysis, the variables may be transformed, represented in some other space, deemed to be non-discriminating, and so on, but they never disappear as recognizable elements. The combination of a fixed set of units and a fixed set of variables brings us to the upper-left box in Table 1. Methods operating under these premises aim at *data display* and

Insert Table 1 about here

data approximation. Usually, the units and variables are represented in a way that enhances accessibility by visual inspection, and enables recognition of special features or structural patterns. A

good example is Tucker's vector model (Tucker, 1960), which is a precursor of the *biplot* technique (Gabriel, this book). Another, perhaps less obvious example is *Procrustes analysis* of two sets of variables (Green, 1952), which can be generalized to K sets of variables (Gower, 1975). In Procrustes analysis, the variables are linearly transformed in each set, but not reduced. The K sets define K locations for each unit, which are superimposed, and the set-wise linear transformations are chosen in such a way that the K locations after transformation approximately coincide. Thus Procrustes analysis is a *matching* technique, in which both units and variables remain intact, and the only stochastic elements – like in the biplot – are the *residuals*, or approximation errors. Part II of this chapter will discuss some recent developments in this area.

Remaining at the left side of Table 1, now suppose that the set of units is assumed to be homogeneous. In the case of a sample, this situation is more commonly characterized by saying that it is assumed that there are N independent identically distributed vector-valued stochastic variables $\{X_1,...,X_j,...,X_N\}$. Much of multivariate distribution theory starts here. Loglinear modelling (Bishop *et al.*, 1975) also belongs to this group, because it analyses an m-dimensional contingency table, which is scrutinized for patterns of proportionality in each of the edges and faces of the table, without ever reducing its dimensions. We have little to add to the extensive literature in this area, except to remark that it is often possible and useful to define what is to be called "data" at a higher aggregation level. This switch of perspective involves first reducing the *raw data* (coded observations) to a table of *sufficient statistics* under some very general type of distribution, such as the multinormal or the multinomial, and then treating these sufficient statistics (e.g. in the form of a correlation matrix or a set of bivariate contingency tables) by the methods in the previous box.

In making the step from univariate to bivariate or multivariate, it is not always realized that one steps into a space with more empty spots than data points, and that the data points really are not free to move anywhere outside local regions. The methods in the lower-left box of Table 1 are geared to empty spots and uneven spread of data points. Certainly in the social and biological sciences, many research paradigms entail an assumption of heterogeneity of the set of units or populations (Gittins *et al.*, 1987). Although heterogeneity of variance has its place, and differential skewness is sometimes observed, the focus on subpopulations most often tries to spot *heterogeneity in location*. When the

subpopulations can be identified by some *a priori* rule, the question of whether they are also *recognizable* in the multivariate data is answered by classical discriminant analysis (Hand, 1981). When heterogeneity (and limited exchangeability) is merely assumed in a general fashion, we aim for identification by some *a posteriori* rule, and enter the area of cluster analysis and classification (for a recent overview, see Bock, 1988). Interesting recent work centers around the *mixture model* formulation of classification tasks, and this approach will be among the new developments to be discussed in Part III.

Turning to the right column of Table 1, the term *stochastic set of variables* needs some further explication. It is simply the name proposed here for a concept that is very familiar in psychometrics and econometrics (and perhaps outside these fields as well), but which unfortunately has been linked rather too closely to sampling considerations in homogeneous sets of units. The concept is to regard any particular variable in a given set of variables as a single random element from a distribution. Being a random element of a distribution implies that there is a definite operational sense in which the variables are exchangeable among themselves.

A clear example from psychometrics is the situation in which a group of students constitutes the set of units, and a collection of test items constitutes the set of variables. The leading idea is to select the test items as a random sample from all conceivable diagnostic test problems in some domain of cognitive achievement, and the prototypical psychometric mission is to reduce the item-wise responses to a single indication of relative achievement: the test score. Another example is to use the cost of several typical wage-earner purchases of goods and services in a given economy to obtain a single consumer price index, which can be traced in time or across economies. Generally, the aim is to reduce a stochastic set of variables to one or a few *factors*, also called *components*. Given the component(s), there remains only unidentifiable, random variation – however, not necessarily *small* variation. A stochastic set of variables can be homogeneous (one component) or heterogeneous (more components), but this level of detail is avoided in Table 1.

Note that the variables themselves can be fixed or stochastic, which in the present scheme is expressed by calling the set of units fixed or stochastic. In the above example of achievement testing, the group of students should be regarded fixed when individual decisions about them are required, as

in entrance examinations. As indicated in the upper-right box of Table 1, this situation is typical for psychometric test theory (Lord and Novick, 1968), practical use of classical factor analysis (Horst, 1965), and for exploratory methods such as projection pursuit (Friedman and Tukey, 1974; Huber, 1985). In the latter method, the aim is to reduce the variables to components that strike a balance between being (1) *representative* for the variance in all variables, (2) *insensitive* to outlying units, and (3) *responsive* to some structural feature in the set of units, like *clumpiness*. These requirements all involve distributional aspects of the set of units, but there is no explicit objective to reduce this set to its distributional characterization. However, in contrast with the data display methods neighbouring at the left, individual variables are of no concern, neither in terms of approximating the initial data, nor in terms of predicting the profile of new units.

If the units are regarded as exchangeable elements, we arrive in the next box in Table 1. The oldest example is the *Spearman hierarchy* (Spearman, 1904), in which the whole data matrix is summarized in terms of one quantity, the mean squared correlation of the variables with a single common factor. In the Spearman hierarchy, the variables are ordered according to their correlation with the common factor, conditionally upon which they are uncorrelated. Other examples are generally subsumed under the name *linear structural relations* (LISREL) modelling (Jöreskog and Wold, 1982), in which several components are in turn linked by a so-called *path model*, but it should be noted that the LISREL family, even though it most typically belongs here, also includes methods that are to be classified in the neighbouring cells of Table 1.

The lower-right box in Table 1, finally, contains methods – to be discussed in Part I – that combine the idea of variable reduction with the concept of potential heterogeneity in the set of units, actively working with information on recognizable subpopulations. The central technique in this area is, since the seminal paper by Hotelling (1936), *canonical correlation analysis*, which looks for one or more components with optimal predictability from some other set of variables. These components are called *canonical*, because they remain the same regardless of any preliminary linear combination of the analysis variables that one might contemplate. Like in the other Parts of this chapter, the focus in Part I will be on nonlinear methods, in particular on methods that make nonlinear regressions linear; they are organized in a hierarchy called the *Gifi system.*

Part I. Homogeneity analysis as a general framework for multivariate analysis of categorical data: the Gifi system

The *Gifi system* (Gifi, 1990) is a framework of nonlinear multivariate analysis methods that is built around the central theme of homogeneity of variables. Although variable reduction and transformation are central concepts, in view of the ambitious scope of the system it takes quite a lot of other ideas to get the whole framework together. Moreover, there are at least three major ways of introduction to the Gifi system:

(1) by generalizing principal components analysis as a *differential weighting technique* – an approach based upon Guttman (1941) and De Leeuw (1973), where the differential weighting idea goes back to Galton, Pearson and Spearman (*cf.* Gifi, 1990, chapter 3), and classical psychometrics (*cf.* Heiser and Meulman, 1993));

(2) as *principal coordinates analysis* (classical MDS) of χ-squared distances (which relates it to the French school, e.g. Lebart *et al.*, 1984; *cf.* Heiser and Meulman, 1983, and Meulman, 1986);

(3) in terms of *linearizing the regression* through quantification and transformation of categorical variables (De Leeuw, 1989; the approach goes back to Hirshfeld, 1935, and Fisher, 1938, 1940).

There are still other ways of introduction, such as through the unfolding model (Heiser, 1981), but to obtain some reduction of conceptual complexity, we will concentrate upon (3) here. The Gifi system has been built to analyze dependencies in categorical data, and therefore the first problem is how to characterize the relationship between two categorical variables when they are not independent.

Linearizing the regression of two categorical variables

Consider the case of two categorical variables, h_1 and h_2, with joint distribution $F(h_1,h_2)$, where h_1 has k_1 levels or *categories*, and h_2 has k_2 categories. An example from political voting in the Netherlands is given in Figure 1, in which the joint frequencies are plotted at a regularly spaced grid,

Insert Figure 1 about here

defined by some *a priori* ordering of the variables POLITICAL PARTY and URBANIZATION (source: CBS (1987) records of the 1986 elections; the figures are given in thousands). Since we are not interested in the univariate marginals at this point, these are omitted from Figure 1; instead, the labels of the categories are given. The labels for the variable URBANIZATION are self-evident. For POLITICAL PARTY, several parties have been taken together at the extreme right and the extreme left of the Dutch political spectrum, respectively; PvdA is the Dutch Labour party; CDA is the name of the Christian-Democrats; there are two kinds of liberal party in the Netherlands: both emphasize civilian rights and individual freedom, but one is economically conservative (VVD), while the other is economically undogmatic (D'66). Dutch politics will return in other examples of this chapter.

The regression of POLITICAL PARTY upon URBANIZATION is also displayed in Figure 1, as a series of connected weighted means. Without further restricting specifications, the regression of a discrete variable h_1 with integer levels $1,...,k_1$ upon a similar variable h_2 is nothing more than the set of conditional expectations $E(h_1 \mid h_2 = a)$, where $a = 1,...,k_2$. Figure 1 illustrates the general phenomenon that the unconstrained regression is *nonlinear* for the *a priori* chosen value of the categories of h_1 and h_2. The same thing is true, of course, for the unconstrained regression of h_2 on h_1 (not shown in Figure 1).

By choosing another quantification for the categories of the variable POLITICAL PARTY we can get more heterogeneity (variation in the means) in the vertical direction. Suppose a different quantification is selected by permuting the inital one. The mean and the variance of the variable remain the same when the integer category values are merely permuted, but the conditional means may be affected by such an operation, because they depend on the strength of the relationship in the joint distribution $F(h_1,h_2)$. Figure 2 shows the permutation of the political parties that gives maximal heterogeneity of

Insert Figure 2 about here

the conditional expectations in the vertical direction. In addition, it shows that the relationship can be made monotonically increasing too (the horizontal variable need not be changed to achieve monotonicity in this example, but generally both axes may have to be adjusted). By some well-chosen (monotonic) transformation of the variable URBANIZATION, while keeping its mean and

variance constant, the line connecting the conditional means – the *regression line* – could be made *exactly straight*. By further readjustment of the variable POLITICAL PARTY the other regression line can be made straight too, and this joint iterative process is called (reciprocally) *linearizing the regression*.

When the cross-classification is not independent (in the voting example, the normalized sum of squared deviations from independence (Cramér's statistic) is 0.097 – an indication for a modest interdependence in the voting population), and when it is really the only thing known about the units, linearization has a number of clear advantages. Reciprocal linearized regression leads to a characterization in terms of the Pearson correlation coefficient (Hirshfeld, 1935); the scores will give maximal mutual discrimination (Fisher, 1940); *optimal scaling* (Bock, 1960) of the categories removes the arbitrariness of the *a priori* quantification; and finally, as will be shown shortly, there is a dual relation between the scale position of the categories of h_1 and the categories of h_2 – hence the alternative name *dual scaling* (Nishisato,1980). Instead of nonlinear regression between the variables, we obtain a simple relationship and *nonlinear transformations* of the initial scores.

Loss function for reciprocal linearization of the regression

The goal of reciprocal linearization, i.e. simultaneously linearizing the regression of h_1 on h_2, and of h_2 on h_1 can be formulated by defining the *quantified* variables $\mathbf{q}_1$ and $\mathbf{q}_2$, two unknown N-element vectors, in terms of the *category quantifications* $\mathbf{y}_1$ and $\mathbf{y}_2$, two unknown vectors of length k_1 and k_2, by the equations

$$\mathbf{q}_1 = \mathbf{G}_1\mathbf{y}_1 ,$$
$$\mathbf{q}_2 = \mathbf{G}_2\mathbf{y}_2 ,$$

where $\mathbf{G}_1$ and $\mathbf{G}_2$ are binary *indicator matrices*, of size $N\times k_1$ and $N\times k_2$, respectively, which code for each unit (or *object*) its presence (1) or absence (0) in the categories of h_1 and h_2. It is assumed that the classifications are exhaustive and mutually exclusive, so that the indicator matrices have row sums equal to $\mathbf{1}$, an N-element vector of ones, and diagonal cross product matrices $\mathbf{D}_1 = \mathbf{G}_1'\mathbf{G}_1$ and $\mathbf{D}_2 = \mathbf{G}_2'\mathbf{G}_2$. Since linear regression preserves the mean, we are free to choose the origin of the quantified variables so that $\mathbf{1}'\mathbf{q}_1 = 0$ and $\mathbf{1}'\mathbf{q}_2 = 0$, i.e. they are centered at zero. The scale is fixed by requiring

$\mathbf{y}_1'\mathbf{D}_1\mathbf{y}_1 = N$ and $\mathbf{y}_2'\mathbf{D}_2\mathbf{y}_2 = N$, so that $\mathbf{q}_1$ and $\mathbf{q}_2$ are in *standard scores*. Under these centering and standardization conditions, minimization of the loss function

$$\delta^2(\mathbf{y}_1,\mathbf{y}_2) = \| \mathbf{G}_1\mathbf{y}_1 - \mathbf{G}_2\mathbf{y}_2 \|^2 , \tag{1}$$

where $\| \cdot \|$ denotes the Euclidean norm, yields a reciprocal linear regression. To understand more fully why this property holds, let us first note that $\delta^2(\cdot)$ can be alternatively expressed as

$$\delta^2(\mathbf{y}_1,\mathbf{y}_2) = 2N [1 - \rho(\mathbf{G}_1\mathbf{y}_1,\mathbf{G}_2\mathbf{y}_2)] . \tag{2}$$

Here $\rho(\cdot)$ is the *canonical correlation* function, which generally measures the strength of the relationship between linear combinations of two sets of variables – in this case the orthogonal binary ("dummy") variables in $\mathbf{G}_1$ and $\mathbf{G}_2$. While $\delta^2(\cdot)$ in (1) is a *squared distance* loss function, the problem thus turns out to be equivalent to *maximizing a correlation* in (2), i.e. to minimizing an angle between the two standardized vectors $\mathbf{q}_1$ and $\mathbf{q}_2$.

The indicator matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ are formally equivalent to design matrices of a one-way analysis of variance, and the loss function $\delta^2(\cdot)$ can be decomposed in the usual way as

$$\delta^2(\mathbf{y}_1,\mathbf{y}_2) = \| \mathbf{G}_1\mathbf{y}_1 - \mathbf{G}_2\mathbf{D}_2^{-1}\mathbf{G}_2'\mathbf{G}_1\mathbf{y}_1 \|^2 + \| \tilde{\mathbf{y}}_2 - \mathbf{y}_2 \|^2_{\mathbf{D}_2} , \tag{3}$$

where $\tilde{\mathbf{y}}_2$ is the unconstrained minimizer defined as $\tilde{\mathbf{y}}_2 = \mathbf{D}_2^{-1}\mathbf{G}_2'\mathbf{G}_1\mathbf{y}_1$, and where $\| \cdot \|^2_{\mathbf{D}_2}$ denotes the squared Euclidean norm in the metric $\mathbf{D}_2$. Similarly, we have

$$\delta^2(\mathbf{y}_1,\mathbf{y}_2) = \| \mathbf{G}_2\mathbf{y}_2 - \mathbf{G}_1\mathbf{D}_1^{-1}\mathbf{G}_1'\mathbf{G}_2\mathbf{y}_2 \|^2 + \| \tilde{\mathbf{y}}_1 - \mathbf{y}_1 \|^2_{\mathbf{D}_1} , \tag{4}$$

with $\tilde{\mathbf{y}}_1 = \mathbf{D}_1^{-1}\mathbf{G}_1'\mathbf{G}_2\mathbf{y}_2$. The second terms on the right-hand side of (3) and (4) are simple least distance functions in the diagonal metrics $\mathbf{D}_2$ and $\mathbf{D}_1$, and these are the only terms to be considered for the constrained optimization of $\mathbf{y}_2$ and $\mathbf{y}_1$, respectively. By elementary arguments it follows that the *standardized category quantifications* $\hat{\mathbf{y}}_1$ must be proportional to the unconstrained quantifications $\tilde{\mathbf{y}}_1$, with proportionality factor $\rho(\mathbf{G}_1\hat{\mathbf{y}}_1,\mathbf{G}_2\hat{\mathbf{y}}_2)$; thus we obtain $\hat{\mathbf{y}}_1 = \tilde{\mathbf{y}}_1 / \rho(\mathbf{G}_1\hat{\mathbf{y}}_1,\mathbf{G}_2\hat{\mathbf{y}}_2)$, and analogously $\hat{\mathbf{y}}_2 = \tilde{\mathbf{y}}_2 / \rho(\mathbf{G}_1\hat{\mathbf{y}}_1,\mathbf{G}_2\hat{\mathbf{y}}_2)$ for the second variable. Suppose the quantifications of the second variable h_2 are fixed at some values $\mathbf{y}_2$; then the conditional expectation of h_2 as a function of

the optimally scaled values of the first variable can be expressed as

$$E(h_2 \mid \hat{\mathbf{y}}_1) = \mathbf{D}_1^{-1}\mathbf{G}_1'\mathbf{G}_2\mathbf{y}_2 = \tilde{\mathbf{y}}_1 = \rho(\mathbf{G}_1\hat{\mathbf{y}}_1,\mathbf{q}_2)\,\hat{\mathbf{y}}_1 \ , \tag{5}$$

showing that $E(h_2 \mid \hat{\mathbf{y}}_1)$ is proportional to $\hat{\mathbf{y}}_1$. So the regression is linear, even when h_2 is not (yet) optimally quantified. In the latter case, of course, $\rho(\mathbf{G}_1\hat{\mathbf{y}}_1,\mathbf{q}_2) < \rho(\mathbf{G}_1\hat{\mathbf{y}}_1,\mathbf{G}_2\hat{\mathbf{y}}_2)$; therefore, it should be noted that criterion (2) looks not just for any regression line, but for the *steepest* one.

The decompositions in (3) and (4) also show how to construct an ALS (Alternating Least Squares) algorithm by alternating between finding improved estimates for $\mathbf{y}_1$ given previous estimates of $\mathbf{y}_2$ and finding improved estimates for $\mathbf{y}_2$ given previous estimates of $\mathbf{y}_1$. Due to the special structure of $\mathbf{G}_1$ and $\mathbf{G}_2$, the unconstrained updates $\tilde{\mathbf{y}}_1$ and $\tilde{\mathbf{y}}_2$ can be obtained by averaging the appropriate scores in $\mathbf{q}_2$ and $\mathbf{q}_1$, respectively, and for this reason the process has been called *reciprocal averaging* (Horst, 1936). Both averaging operations involve the contingency table $\mathbf{G}_1'\mathbf{G}_2$, and it can be shown that there is a close connection with the French omnibus technique *correspondence analysis* (Gifi, 1990, chapter 8). Inserting the definition of $\tilde{\mathbf{y}}_1$ into the formula for $\tilde{\mathbf{y}}_2$ (or the other way around), and keeping track of the necessary normalization, yields an expression for eigenvector calculation. In the ALS process, it is easy to handle further constraints, such as the *isotonicity* requirement that the category quantifications should be non-decreasing with respect to the initial quantifications, because by working with one set at a time it is enough to solve the weighted least distance problems defined in the second terms of (3) and (4).

Before proceeding to the case of m categorical variables, which brings us into the heart of the Gifi system, a remark on the number of solutions (or *dimensionality*) is in order. Although there generally are $min(k_1 - 1, k_2 - 1)$ stationary points of loss function (1), there is only one pair $(\mathbf{y}_1,\mathbf{y}_2)$ that minimizes it. The rationale of linearizing the regression with maximal correlation does not permit us to select more than one solution. However, if we would aim at an *approximation of the bivariate distribution* (Lancaster, 1957, 1958), which constitutes an alternative rationale for the same algebraic operations, then it certainly would be sensible to consider a higher dimensionality, because that would yield a higher-order approximation of a nonlinear function.

Simultaneous linearization of m categorical variables

Generally, it is not possible to extend the idea of reciprocal linearization to the case of m categorical variables, because it leads to inconsistency in the requirements that each of the variables should satisfy. However, we can try as hard as possible by generalizing (1) into the loss function

$$\delta^2(\mathbf{y}_1,...,\mathbf{y}_m) = (1 / 2m^2) \sum_j \sum_l \| \mathbf{G}_j\mathbf{y}_j - \mathbf{G}_l\mathbf{y}_l \|^2 , \qquad (6)$$

where the reason for the appearance of the factor $(1 / 2m^2)$ will become clear in a short while. First note that minimizing $\delta^2(\mathbf{y}_1,...,\mathbf{y}_m)$ over $\mathbf{y}_j$, keeping the other variables fixed, only involves terms in (6) with $l \neq j$, and that for the sum of these terms, Huygens' Theorem (*cf.* Lebart *et al.*, 1984) gives the decomposition

$$\sum_{l\neq j} \| \mathbf{G}_j\mathbf{y}_j - \mathbf{q}_l \|^2 = \sum_{l\neq j} \| \mathbf{q}_l - \bar{\mathbf{q}}_{(-j)} \|^2 + (m-1) \| \bar{\mathbf{q}}_{(-j)} - \mathbf{G}_j\mathbf{y}_j \|^2 ,$$

in which $\bar{\mathbf{q}}_{(-j)} = (m-1)^{-1} \sum_{k\neq j} \mathbf{q}_k$, the multidimensional mean (center of gravity) across all variables except $\mathbf{q}_j$. From the second term of this decomposition, it follows that the unconstrained minimum over $\mathbf{y}_j$ is obtained for $\tilde{\mathbf{y}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\bar{\mathbf{q}}_{(-j)}$. If we now look at the conditional expectation of any variable h_l as a function of the optimally scaled values of h_j, the analogon to (5) is

$$E(h_l \mid \tilde{\mathbf{y}}_j) = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{G}_l\mathbf{y}_l \neq \mathbf{D}_j^{-1}\mathbf{G}_j'\bar{\mathbf{q}}_{(-j)} = \tilde{\mathbf{y}}_j ,$$

showing that $E(h_l \mid \tilde{\mathbf{y}}_j)$ is *not* proportional to $\tilde{\mathbf{y}}_j$, except when $\mathbf{q}_l$ happens to coincide with the leave-one-out center of gravity $\bar{\mathbf{q}}_{(-j)}$. So the regression remains nonlinear, even when all variables would be optimally quantified by finding the global minimum of $\delta^2(\mathbf{y}_1,...,\mathbf{y}_m)$.

Fortunately, there still is a way to keep linearization possible: by introducing a new variable x, quantified as an unknown N-vector $\mathbf{x}$ in the space spanned by the $\{\mathbf{q}_j\}$, which does have linear regressions with all variables. Using the well-known equality of the sum of squared Euclidean distances among pairs of vectors, used in (6), and $2m$ times their sum of squared Euclidean distance towards the center of gravity (*cf.* Gower, 1975), in fact another application of Huygens' Theorem, we obtain

$$\delta^2(\mathbf{y}_1,\ldots,\mathbf{y}_m) = \min_{\mathbf{x}} \sigma^2(\mathbf{y}_1,\ldots,\mathbf{y}_m; \mathbf{x}) = m^{-1} \sum_j \| \mathbf{G}_j\mathbf{y}_j - \mathbf{x} \|^2 . \qquad (7)$$

While the function $\delta^2(\cdot)$ aims at reciprocal linearization (but cannot achieve it), the function $\sigma^2(\cdot)$, which is the basic loss function of the Gifi system, called *loss of homogeneity*, aims at simultaneous linearization with respect to $\mathbf{x}$. Since $\sigma^2(\cdot)$ measures the *mean squared deviation* from some central vector, it is a natural measure of multidimensional dispersion, and the fact that the sum of squared inter-variable distances is $2m$ times the sum of squared deviations from their center of gravity explains the appearance of the factor $(1 / 2m^2)$ in (6).

The transition from $\delta^2(\cdot)$ to $\sigma^2(\cdot)$ is illustrated in Figure 3, where at the left the inter-variable distances are indicated with white lines, while at the right the variables are all connected with their

Insert Figure 3 about here

center $\mathbf{x}$. Using the same arguments as before, the *unconstrained* optimal value of $\sigma^2(\cdot)$ over $\mathbf{x}$ is the mean across all variables: $\tilde{\mathbf{x}} = m^{-1} \sum_j \mathbf{G}_j\mathbf{y}_j$. But because the only interest is in its direction, not in its length, $\mathbf{x}$ is required to have some standard length, usually $\| \mathbf{x} \|^2 = N$, leading to $\hat{\mathbf{x}} = N^{1/2} \tilde{\mathbf{x}} / \| \tilde{\mathbf{x}} \|$ as the standardized solution. Normalizing $\mathbf{x}$ allows us to leave the quantifications $\mathbf{y}_j$ unnormalized. Similarly, partial minimization of $\sigma^2(\cdot)$ over $\mathbf{y}_j$ yields the relationship $\rho(\mathbf{G}_j\hat{\mathbf{y}}_j,\mathbf{x})\hat{\mathbf{y}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{x} = \tilde{\mathbf{y}}_j$ between the standardized and the unconstrained optimal quantifications, respectively, where $\rho(\mathbf{G}_j\hat{\mathbf{y}}_j,\mathbf{x})$ now is the *multiple correlation* function, called the *component loading* of the optimally quantified $\mathbf{q}_j$ on the *component* $\mathbf{x}$. Let us once more have a look at the conditional expectation, this time of x with respect to the optimal $\hat{\mathbf{y}}_j$, obtaining

$$E(x \mid \hat{\mathbf{y}}_j) = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{x} = \rho(\mathbf{G}_j\hat{\mathbf{y}}_j,\mathbf{x})\, \hat{\mathbf{y}}_j ,$$

a linear regression line of the component on any of the quantified variables, with slope equal to the component loading, demonstrating the correctness of the claim that simultaneous linearization is possible, even though reciprocal linearization is not. Insertion of the expression for $\tilde{\mathbf{y}}_j$ into loss of homogeneity leads to the conclusion that the mean squared component loading is maximized, and therefore the slopes of the linearized regressions will be as steep as possible, on average.

As before, simple linear relationships with the component are obtained by a process that is independent of the initially given quantifications of the variables, which makes the analysis invariant under preliminary nonlinear transformations that preserve class membership. The introduction of x is not just a matter of computational convenience, because it is crucial for extending the linearization concept to the case of m variables. When a new object would become available, of which the class membership on h_j is given, we could use all cross-tabulations with h_j to predict its class membership on the other variables, but that would in fact involve a highly parametrized model of prediction. Analysis methods based on linearization allow us to predict the mean of a subdistribution in x, which is the appropriate element of $\tilde{\mathbf{y}}_j$, and to estimate the pooled within-class variance by $1 - N^{-1}\tilde{\mathbf{y}}_j'\mathbf{D}_j\tilde{\mathbf{y}}_j$, i.e. one minus a diagnostic called the *discrimination measure*. Given that information, simple rules for predicting the class membership in the other variables can be constructed.

Homogeneity analysis in a strict sense: HOMALS

Minimization of the homogeneity loss function $\sigma^2(\cdot)$ implies seeking maximum homogeneity of variables and maximum heterogeneity of units. When there are only two variables, identification of a single component with its associated optimal scaling of the categories finishes the analysis, because – as explained earlier – there is no way to improve the quantification. But with more variables, there may be reasons to repeat the analysis with multiple components. Either the mean squared loading may be low, indicating wide dispersion in several directions, or the loadings may have large variance, indicating that only a subset of the variables determines the component (note that, with two variables, the loadings will always be equal, as the component will be exactly in between the two quantified variables). These properties of the loadings are indications that the variables are not homogeneous, but heterogeneous or partially independent.

Without too much complication, it is possible to extent the analysis into a multicomponential one, by generalizing loss of homogeneity into

$$\sigma^2(\mathbf{Y}_1,\ldots,\mathbf{Y}_m;\,\mathbf{X}) = m^{-1}\,\textstyle\sum_j \|\,\mathbf{G}_j\mathbf{Y}_j - \mathbf{X}\,\|^2\,, \tag{8}$$

with p uncorrelated components $\mathbf{x}_s$, collected in an $N\times p$ matrix $\mathbf{X}$, satisfying $\mathbf{X}'\mathbf{X} = N\mathbf{I}$; the $\mathbf{Y}_j$ are now matrices of order $k_j\times p$. Minimizing (8) without further qualifications constitutes homogeneity

analysis in a strict sense, and can be executed with the computer program HOMALS (SPSS, 1990). The multiple components each have different linear regressions on each of the variables, collected in the columns of the $\mathbf{Y}_j$, which are called *multiple quantifications*. It can be shown that homogeneity analysis in a strict sense is equivalent to an extension of correspondence analysis, called *multiple correspondence analysis*, but the definition, representation and interpretation of the distances among units and variables that are typical for this framework become problematical in its extension (Meulman, 1986; Greenacre, 1991).

By contrast, the primary focus in the Gifi framework as presented here is on quantification and discrimination. In the multiple case these concepts are associated, like in linear discriminant analysis, with variable-wise distances among *category points* (rows of $\mathbf{Y}_j$), and with unit-category distances between category points and *object points* (rows of $\mathbf{X}$). The weighted average sum of squares of the former is maximized, whereas the average of the latter is minimized (Heiser, 1981). The difference with discriminant analysis is that HOMALS uses multiple classifications rather than a single one, and the part of the covariates is played by $\mathbf{X}$, the unknown components, which in turn are linear combinations of the quantified classifications.

The scope for application of HOMALS is quite large, and since several computer implementations have been made widely available, the technique is used increasingly, sometimes even routinely, in many areas (for examples, see Nishisato, 1980, Lebart *et al.* 1984, Greenacre, 1984, Gifi, 1990).

Homogeneity analysis of variables with a priori ordered or integer-valued categories: PRINCALS

It often happens that one has a mixed collection of variables, some of which are unconstrained categorical (called *nominal variables*), some defined by *a priori* ordered categories (*ordinal*) and others having an equally spaced succession of levels (*numerical*). To incorporate this additional information, it is possible to switch from unconstrained, multiple quantification to constrained, *single quantification* (De Leeuw and Van Rijckevorsel, 1980). Loss of homogeneity can still be an average sum of squares across variables, as in (7) and (8), but now with variable-wise components defined as

$$\sigma^2(\mathbf{y}_j,\mathbf{a}_j \mid \mathbf{X}) = \| \mathbf{G}_j\mathbf{y}_j\mathbf{a}_j' - \mathbf{X} \|^2 , \tag{9}$$

where the notation $\sigma^2(\cdot \mid \mathbf{X})$ is used to indicate that $\mathbf{X}$ is temporarily regarded as fixed, and where

the p-vector $\mathbf{a}_j = \{a_{j1},...,a_{js},...,a_{jS}\}$ contains the regression coefficients for the regression of the components $\mathbf{x}_s$ on the standardized quantified variable $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$. Optimizing $\sigma^2(\mathbf{y}_j, \mathbf{a}_j \mid \mathbf{X})$ over $\mathbf{a}_j$ yields, instead of a single loading per variable, the multiple components loadings $a_{js} = \rho(\mathbf{G}_j\mathbf{y}_j,\mathbf{x}_s)$, so the major new problem is how to obtain conditionally optimal estimates of $\mathbf{y}_j$. Application of Huygens' Theorem on the rescaled components $\mathbf{x}_s / a_{js}$ leads to the basic decomposition

$$\sigma^2(\mathbf{y}_j \mid \mathbf{X},\mathbf{a}_j) = \sum_s a_{js}^2 \,\|\, \mathbf{G}_j\mathbf{y}_j - \mathbf{x}_s / a_{js} \,\|^2 = \sum_s \|\, a_{js}\tilde{\mathbf{q}}_j - \mathbf{x}_s \,\|^2 + [\, \sum_s a_{js}^2 \,] \,\|\, \tilde{\mathbf{q}}_j - \mathbf{G}_j\mathbf{y}_j \,\|^2. \quad (10)$$

Here the notation $\tilde{\mathbf{q}}_j$ is used to denote the *component mixture* variable

$$\tilde{\mathbf{q}}_j = \sum_t a_{jt}\,\mathbf{x}_t \,/\, [\, \sum_t a_{jt}^2 \,],$$

which is the optimal representation of variable h_j in the space of the components. Since only the second part of decomposition (10) depends on $\mathbf{y}_j$, it may be concluded that the mixture variable $\tilde{\mathbf{q}}_j$, instead of the components itself, will have a linear regression on the optimal quantifications. For nominal variables, the optimal $\mathbf{y}_j$ are the category means of the mixture variable. For ordinal variables, we have to perform an *isotonic regression* (Barlow *et al.*, 1972) of $\tilde{\mathbf{q}}_j$ on the given ordering of the categories during the ALS iterations. This process may create ties in $\mathbf{y}_j$, leading to a flat part in the transformation of the original category values, but will leave the conditional expectation of $\tilde{\mathbf{q}}_j$ linear. For numerical variables – which in the Gifi system always are variables with integer-valued categories $\{1, 2, 3,..., k_j\}$, some of which may be missing – the vector $\mathbf{y}_j$ has to be chosen as the appropriately centered and standardized version of the sequence $\{1, 2, 3,..., k_j\}$, and there are no degrees of freedom left to optimize.

Note that, since a_{js} is a function of $\mathbf{y}_j$, and $\tilde{\mathbf{q}}_j$ is a function of a_{js} and $\mathbf{x}_s$, while $\mathbf{X}$ is a function of the mean across $\mathbf{G}_j\mathbf{y}_j$, the real complexity of the loss function is much greater than the succession of quadratic problems that has been obtained by the convenient way the problem can be split into separate parameter sets. The computer program PRINCALS (Gifi, 1985; SPSS, 1990) can be used for these iterative computations. The reader is referred to Vlek and Stallen (1981) and Kerkhof *et al.* (1988) for examples of the method in action.

Linearization of cross classification variables and additivity

The approach as described so far would seem to be limited to a joint analysis of bivariate relationships, but this limitation is easily removed by the introduction of other analysis variables than the ones that correspond directly to empirical measurements. One categorical variable corresponds to a partitioning of a set of units, but the cross product of two categorical variables also corresponds to a partitioning, called a *cross classification*, which is a refinement of both initial partitionings. Thus, more generally, if there is interest in the higher-order effects, the cells of a two-dimensional or higher-dimensional contingency table could be coded as the categories of an indicator matrix $\mathbf{G}_k$, with total number of categories $k_1 \times ... \times k_{m_k}$, where m_k is the number of initial variables involved.

Homogeneity analysis of such cross classification variables – in conjunction with other, single classifications or not – can be carried through without further technical complications, and is often worth considering, especially when the original partitionings are extremely coarse (e.g., the binary variable "gender"), or when some variables are strongly associated for reasons not of interest to the current study (e.g., the variables "number of adults in household" and "number of children in household" in a study of consumer expenditure).

Because a cross classification variable can easily get too many categories, and because the linearization of the regression is not in terms of the original variables, it may be desirable to add restrictions to the quantification process. Some more notation needs to be introduced for this extension. Suppose $\mathbf{S}_k$ is a partitioned binary matrix with $k_1 \times ... \times k_{m_k}$ rows corresponding to the cells of the multidimensional contingency table, i.e., to the columns of the cross classification indicator matrix $\mathbf{G}_k$, and with $k_1 + ... + k_{m_k}$ columns corresponding to the categories of the original variables. The matrix $\mathbf{S}_k$ (formally equivalent to the design matrix of a $k_1 \times ... \times k_{m_k}$ factorial design with one replication per cell) codes which category of each variable is involved in each cell of the contingency table. Furthermore, suppose $\mathbf{G}^k$ denotes the partitioned indicator matrix, consisting of m_k column-wise concatenated simple indicator matrices, which code for each unit its membership of any category of each of the variables, and let $\mathbf{y}^k$ be the conformably partitioned $(k_1 + ... + k_{m_k})$-vector of quantifications. From these definitions it easily follows that $\mathbf{G}_k \mathbf{S}_k = \mathbf{G}^k$.

Let us now consider one component of loss of homogeneity (7) associated with a cross

classification variable $\mathbf{q}_k = \mathbf{G}_k\mathbf{y}_k$. Thinking of the earlier example of POLITICAL PARTY crossed with URBANIZATION, the data set could include a number of opinion variables to be summarized – in conjunction with the cross classification – in the component x. Then the *regression line* of x as a function of $\mathbf{y}_k$ would be straight, as always, whereas the *regression surface* of x as a function of the grid points formed by all pairs of categories of POLITICAL PARTY and URBANIZATION would generally still be *nonlinear*. The natural constraint on $\mathbf{y}_k$ therefore is to require that it linearizes the regression surface too.

In the example, let the two submatrices of $\mathbf{S}_k$ be $\mathbf{S}_P$ and $\mathbf{S}_U$, the submatrices of $\mathbf{G}_k$ be $\mathbf{G}_P$ and $\mathbf{G}_U$, and the subvectors of $\mathbf{y}_k$ be $\mathbf{y}_P$ and $\mathbf{y}_U$. Then the grid points formed by all pairs of categories can be collected in the matrix $[\mathbf{S}_P\mathbf{y}_P,\mathbf{S}_U\mathbf{y}_U]$, and the conditional expectation of x would form a linear regression surface if there would exist an $\mathbf{y}_P$ and $\mathbf{y}_U$ so that

$$E(\, x \mid [\mathbf{S}_P\mathbf{y}_P,\mathbf{S}_U\mathbf{y}_U]\,) = \mathbf{D}_k^{-1}\mathbf{G}_k'\mathbf{x} = \mathbf{S}_P\mathbf{y}_P + \mathbf{S}_U\mathbf{y}_U\,. \tag{11}$$

The equations in (11) hold if the unconstrained minimizer $\tilde{\mathbf{y}}_k = \mathbf{D}_k^{-1}\mathbf{G}_k'\mathbf{x}$ satisfies $\tilde{\mathbf{y}}_k = \mathbf{S}_k\mathbf{y}^k = \mathbf{S}_P\mathbf{y}_P + \mathbf{S}_U\mathbf{y}_U$, a condition that cannot be expected to be true in general. But a method that imposes the restrictions $\mathbf{y}_k = \mathbf{S}_k\mathbf{y}^k$ will pull the component $\mathbf{x}$ as much as possible to the subspace of vectors satisfying the constraints, so that the conditional expectations will be *close to* a linear surface. If $\mathbf{y}_k = \mathbf{S}_k\mathbf{y}^k$, then $\mathbf{G}_k\mathbf{y}_k = \mathbf{G}_k\mathbf{S}_k\mathbf{y}^k$, and because $\mathbf{G}_k\mathbf{S}_k = \mathbf{G}^k$, we obtain (Gifi, 1990, section 5.3)

$$\mathbf{G}_k\mathbf{y}_k = \mathbf{G}^k\mathbf{y}^k = \sum_{j\in J_k}\mathbf{G}_j\mathbf{y}_j\,, \tag{12}$$

where the submatrices of $\mathbf{G}^k$ and the subvectors of $\mathbf{y}^k$ are now indexed with $j\in J_k$, an index set of variable identifications. Thus the constraints in (12) require the quantifications of the cross classification variable to be *additive* with respect to the quantifications of the original variables.

Loss functions for getting the regression as additive as possible

The additivity constraints described in the previous section can be incorporated in a loss funtion in various ways, depending upon the further characteristics of the analysis problem. Probably the first one to consider optimal scaling under additivity constraints was Fisher (1938, section 49.2)), who – under the heading "The Discrimination of Groups by Means of Multiple Measurements; Appropriate

Scores" – treated the case of non-numerical serological readings, with categories {–, ?, w, (+), +}, characterizing the reaction in twelve samples of human blood tested with twelve different sera. Fisher's additive scoring system minimizes the loss function

$$\delta^2(\mathbf{y}_R; \mathbf{y}_j, j\in J_k) = \| \mathbf{G}_R\mathbf{y}_R - \sum_{j\in J_k} \mathbf{G}_j\mathbf{y}_j \|^2 , \tag{13}$$

which is equal to (1) with the first variable as the response variable $\mathbf{q}_R = \mathbf{G}_R\mathbf{y}_R$, and with additivity constraints on the second variable. The relationship of (13) with multiple correspondence analysis of the concatenated table $[\mathbf{G}_R, \mathbf{G}_1, \mathbf{G}_2]$ is discussed in detail by Gower (1990). Fisher (1938) solved an eigenvalue equation to minimize (13), but Kruskal (1965) showed how to construct an algorithm with iterative estimate improvement and optimal scaling steps, including the possibility for monotonicity constraints on $\mathbf{y}_R$, calling the method MONANOVA. Then De Leeuw *et al.* (1976) and Young *et al.* (1976) gave a full ALS method for minimizing (13) with possibilities for constraining the quantifications $\mathbf{y}_j$ with $j\in J_k$, in methods called ADDALS and MORALS. Probably the most active application area of these methods is marketing research, where they were introduced by Green (1973, 1974) and Green and Srinivasan (1978) under the name *(additive) conjoint analysis*; for state-of-the-art reviews, see Wittink and Cattin (1989), Green and Srinivasan (1990), and Van der Lans (1992).

The next step could be to introduce additivity at the side of the response variable too. In the one-componential case this extension is without complications, but some intricate difficulties arise as soon as one tries to fit more components, optimizing, by combination of (9) and (13),

$$\delta^2(\mathbf{y}_j, \mathbf{a}_j; j\in J_1\cup J_2) = \| \sum_{j\in J_1} \mathbf{G}_j\mathbf{y}_j\mathbf{a}_j' - \sum_{j\in J_2} \mathbf{G}_j\mathbf{y}_j\mathbf{a}_j' \|^2 , \tag{14}$$

which is the generalization of (1) with additivity constraints on both sides and multiple components, but single quantifications. The difficulties have to do with the confounding of various types of constraints, and were solved by Van der Burg and De Leeuw (1983), in a method called CANALS.

Homogeneity analysis comes in when we switch to more than two sets of variables, be it of the cross classification type with additivity restrictions, or ordinary ones. Suppose there are M sets, indexed by k, each with m_k edge variables, indexed by $j\in J_k$. The variables within sets are called *edge* variables here, because they form the edges of a m_k-dimensional hypercube, with respect to which

the regression of the others is to be linearized. If $m_k = 2$, the hypercube is a grid, and if $m_k = 1$ there is only one edge: a set of points on a line. With $M > 2$, we can make the same transition from the reciprocal $\delta^2(\cdot)$ function to the simultaneous $\sigma^2(\cdot)$ function with several central components $\mathbf{X}$ as was done for simple homogeneity analysis. Loss function (8) with additivity constraints becomes

$$\sigma^2(\mathbf{Y}_1,...,\mathbf{Y}_m; \mathbf{X}) = M^{-1} \sum_k \| \sum_{j \in J_k} \mathbf{G}_j \mathbf{Y}_j - \mathbf{X} \|^2 , \tag{15}$$

with p uncorrelated components $\mathbf{x}_s$, collected in an $N \times p$ matrix $\mathbf{X}$, satisfying $\mathbf{X}'\mathbf{X} = N\mathbf{I}$; there now are $m = \sum_k m_k$ $\mathbf{Y}_j$-matrices of order $k_j \times p$. Van der Burg *et al.* (1988) proposed an ALS algorithm for minimizing (15) with edge variables of the nominal, ordinal, or integer-valued type, and this method has been implemented – including a provision for missing data – in the computer program OVERALS (SPSS, 1990).

If single quantification is used for all variables, so that the additional rank-one condition $\mathbf{Y}_j = \mathbf{y}_j\mathbf{a}_j'$ holds, as in PRINCALS, the OVERALS loss function can be written as

$$\sigma^2(\mathbf{Q}; \mathbf{A}; \mathbf{X}) = M^{-1} \sum_k \| \sum_{j \in J_k} \mathbf{q}_j\mathbf{a}_j' - \mathbf{X} \|^2 = M^{-1} \sum_k \| \mathbf{Q}_k\mathbf{A}_k - \mathbf{X} \|^2 ,$$

with $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$, collected in the matrix of quantified edge variables $\mathbf{Q}_k$, which in turn are collected in the partitioned matrix $\mathbf{Q}$, and with $\mathbf{A}_k$ the conformable matrices of regression weights, collected in $\mathbf{A}$. This reparametrization shows that we deal with a nonlinear generalization (in the sense of optimal scaling) of *generalized canonical correlation analysis*, as proposed by Horst (1961a,b) and Carroll (1968), which is one of the possible generalizations of Hotelling's (1936) two-sets canonical correlation analysis (Gifi, 1990, section 5.1), and which is the reason that the variables defined in (12) are called *canonical variables*.

It is important to notice that the optimal scaling in OVERALS does not linearize the regression of the components in $\mathbf{X}$, but of the auxiliary matrix

$$\mathbf{U}_j = \mathbf{X} - \sum_{l \neq j} \mathbf{G}_l\mathbf{Y}_l ,$$

which corrects $\mathbf{X}$ for the quantifications of the other variables. Thus additivity enables us to unravel the separate contributions of the edge variables to the effect of the finest possible distinction in

subpopulations, as provided by a full cross classification, but the interpretation of the category quantifications as centers of gravity of subsets of object scores **X** no longer holds.

Summarizing, OVERALS is the most general technique of the Gifi system. As a method of dimension reduction, it contains no notion of reducing the edge variables; it is the number of sets that is reduced into a number of components, which in turn can be linearly predicted by the edge variables. When there is only one variable in each set, we obtain PRINCALS. When each variable in PRINCALS is treated as multiple nominal, we obtain HOMALS. When there are only two variables in HOMALS, we obtain ANACOR, the Gifi version of correspondence analysis (SPSS, 1990).

Application of OVERALS to Demographic, Quality of Life, and Educational variables for the US, illustrating the possibility of multidimensional solutions to multiple regression problems

When we have one set with a number of predictor variables, and another one with only a single criterion variable, we are in the multiple regression situation. In that case, only optimization over one component is sensible, even though technically more components could be calculated (Meulman and Heiser, 1988). But now suppose that we can distinguish the predictor variables into two groups, within each of which additive effects would be reasonable. Then we have a three-sets analysis in OVERALS, and subsequent components are no longer arbitrary. Thus in that situation we can look at a meaningful two- or more-dimensional space, inspect distances between object points, and investigate the relationship between the predictors and the citerion via the component loadings.

This idea is illustrated with an example of social indicator data characterizing the 50 states of the United States, in terms of 10 demographic, quality of life, and educational variables, partitioned into 3 sets (sources indicated in Table 2). Several other nonlinear analyses of these (or similar) data have been reported elsewhere (Meulman, 1986). The definition of the variables, and their partitioning into

Insert Table 2 about here

three sets can also be found in Table 2. The OVERALS analysis was done on the rank numbers, with single ordinal quantification for all variables, to allow for nonlinear transformations that preserve the order of the categories. The homogeneity per component can be expressed as an eigenvalue, and is equal to the mean squared correlation of the canonical variables with the component, while the fit per

variable can be expressed as the square root of the sum of squared correlations of the transformed variable across components (which is equal to the correlation between the transformed variable and its projection into the component space (see Table 3)). The eigenvalues associated with the components

Insert Table 3 and Figure 4 about here

for a two-dimensional analysis turn out to be .98 and .65, respectively. A convenient way to inspect the solution is by plotting the object points (rows of **X**), and then adding the edge variables by projection into the component space (Figure 4; the length of the projected edge variables is proportional to their fit). The 50 states are located in such a way that the variables SCHOOL, LIFE, ILLIT, and FAIL are very well represented by the components (variables with the highest fit values in Table 3). By contrast, the variables UNEMP, and TEACH fit rather badly into the canonical space. Mississipi, South Carolina, and some other southern states don't do well, while it must be better to live in North Dakota, Nebraska, or Wyoming. Since the variable FAIL has been given a special role in the analysis, it is interesting to inspect the correlations between this variable and the other transformed variables (also see Table 3). The angles between the vectors in Figure 4 are an approximation of these correlations. It is clear with respect to test failure that ILLIT is the most positively associated variable, and that SCHOOL is the most negatively associated variable.

Figure 5 shows the monotone transformations of the variables. Here the original values are plotted

Insert Figure 5 about here

on the horizontal axes, and the optimal quantifications are plotted on the vertical axis. Although the OVERALS program expects integer values as input, so that often recoding of the originally recorded values is necessary, it is advisable to return to the initial coding when plotting transformations. Also, it must be emphasized that there is no need to have a small number of categories with ordinal variables, in contrast to nominal variables where this is imperative. From the transformations in Figure 5 we learn that variables that fit badly, like TEACH and UNEMP, have obtained transformations with a few steps only. This phenomenon is often observed and implies that little is retained from the original information provided by the edge variable, because it takes aggregation of

subclasses to keep the categories monotonically increasing. SCHOOL, FAIL, INC, and LIFE, on the other hand, fit well into the component space, and their monotone functions are likewise informative.

Part II. Spatial reprentation and approximation of two-mode data

Data are called *two-mode*, in a terminology popularized by Carroll and Arabie (1980), if they describe the relations between a fixed set of units and a fixed set of variables (or another set of units). The term two-mode is used in contradistinction to *two-way*, which merely indicates that the data are indexed by rows and by columns. Thus a correlation matrix is two-way, but one-mode, because it describes the relationships between one set of entities, the variables.

A first-order approximation of two-mode data, which also enables us to make useful spatial representations, is provided by the *bilinear model.* Would the column parameters of this model be known variables, it is (multivariate) linear in the row parameters, whereas it is (multivariate) linear in the column parameters if the row parameters would be known variables; hence the terms *bilinear* for the model, and *biplot* for the spatial representation of the parameters. The bilinear model is in fact a family of models, because it involves products of parameters that themselves can be identified in a number of different ways, and there are several issues of standardization and weighting (Rao, 1980). Of course, the most prominent member of the family is *principal component analysis*, since Eckart and Young (1936) showed that the principal components can be used to make a nested series of bilinear approximations in the least squares sense.

Because bilinear approximation does not involve a reduction of the row or column elements, but a decomposition of the form *DATA* = *BILINEARITY* + *RESIDUAL*, the stochastic part is in the residuals. In practice, one should always check if the residuals are reasonably homogeneous. If the data exhibit nonlinear scatter, it may be possible to *bilinearize* them by row-wise or column-wise nonlinear transformations, so that the *RESIDUAL* component becomes more regular. Algorithms and computer programs for this purpose, like PRINCIPALS (Young *et al.*, 1978) and options within MULTIPALS (Verboon *et al.*, 1991), are implementations of Kruskal and Shepard's (1974) pioneering work.

Regardless of data transformation, the residuals may still remain large or heterogeneous for bilinear representations of reasonably small dimensionality. In the following two paragraphs, we will

first have a look at robust techniques for two-mode approximation, and next discuss second-order methods that are based on a particular non-bilinear model, the unfolding model.

The Procrustes problem with heterogeneous residuals

In models like the ones considered here, with many degrees of freedom and a two-way structure of residuals, the commonplace remark that least squares lacks robustness is easier said than verified. Generating error-perturbed data by adding bimodal random deviates to a low-rank matrix is often not sufficient to disturb the first few components, and it could even be conjectured that the principal components are robust against disturbances that are unrelated to the two-way structure. It is important to note here that – unlike the situation in the linear model, where we can always inspect various conditional distributions – outliers in the residuals need not manifest themselves in the data, and neither do outliers in the data contradict a low-dimensional bilinear model. As remarked by Green (1984), the traditional correspondence of "residuals" with "observations" has been lost. Because of these complexities, the discussion will be limited to a special case, the Procrustes problem.

In the *orthogonal Procrustes* problem (Green, 1952), the set of points $\{\mathbf{q}_i\}$, collected in the rows of data matrix $\mathbf{Q}$, is approximated by rotation of another set of points $\{\mathbf{p}_i\}$, collected in the rows of a matrix $\mathbf{P}$, which may be another data matrix or some *a priori* given set of characteristics. A rotation of the coordinate axes associated with $\mathbf{P}$ leaves all distances among the row points unchanged, and is effected by an *orthonormal* linear transformation, i.e. multiplication by some mxm matrix $\mathbf{Y}$ satisfying $\mathbf{Y'Y} = \mathbf{YY'} = \mathbf{I}$. Thus the least squares Procrustes problem is to minimize

$$\delta^2(\mathbf{Y}) = N^{-1} \sum_i \| \mathbf{q}_i - \mathbf{Y}\mathbf{p}_i \|^2 \qquad (16)$$

over the rotation matrix $\mathbf{Y}$. Verboon and Heiser (1992) showed how to fit a variety of robust or resistant alternatives to the squared Euclidean norm in (16), such as Tukey's biweight function (Beaton and Tukey, 1974), with an iteratively reweighted least squares algorithm, and proved why such a procedure converges monotonically (also see Heiser *et al.*, this volume).

Approaching the Procrustes problem under the assumption of *heterogeneous residuals* implies that some points $\mathbf{p}_i$ have to be rotated with rather different angles to match their corresponding point $\mathbf{q}_i$ than others. The sensitivity of least squares to this type of heterogeneity is illustrated with an example

from Verboon and Heiser (1992). In Figure 6 two sets of eight points on a square are given,

Insert Figure 6 about here

Q with sides parallel to the coordinate axes, and **P** with sides under a 45° rotation. Thus rotation towards perfect match is possible. When only one outlier is created by changing the second coordinate of $\mathbf{q}_1$ so that it would require an angle of minus 84° for $\mathbf{p}_1$ to have the same direction, the least squares result deteriorates as indicated by the dashed square in Figure 6. With Tukey's biweight, an angle of approximately 44° is found, creating much smaller residuals for the non-perturbed points and excessively large values for the first pair of points. In a resistant loss function excessive residuals do no harm, as they are radically downweighted.

Approximation with a two-sets distance model: the unfolding technique

The bilinear model works with two sets of points or vectors: one for the rows of the data matrix, and one for the columns. Suppose that we denote the row points by $\mathbf{x}_i$ and the column points by $\mathbf{y}_j$, both being vectors in p-dimensional space, where p is some pre-chosen parameter. According to the bilinear model, a data value q_{ij} is approximated by the inner product $\mathbf{x}_i'\mathbf{y}_j$. Thus a row of **Q**, viewed as a function of the configuration of points $\{\mathbf{y}_1,...,\mathbf{y}_j,...,\mathbf{y}_m\}$, forms a linear response surface. In the context of individual choice behavior, Coombs (1964) argued that preference data would better be modelled by *single-peaked* response surfaces, and suggested the so-called *unfolding model.* In a relatively simple form of the unfolding model (Heiser, 1981, 1987a) it is assumed that q_{ij} is approximated as a function of the Euclidean distance $d(\mathbf{x}_i,\mathbf{y}_j) = \| \mathbf{x}_i - \mathbf{y}_j \|$. Taking squared distances, the relation with the bilinear model is given as

$$d^2(\mathbf{x}_i,\mathbf{y}_j) = \mathbf{x}_i'\mathbf{x}_i + \mathbf{y}_j'\mathbf{y}_j - 2\,\mathbf{x}_i'\mathbf{y}_j \quad .$$

This quadratic function has a minimal value at $\mathbf{x}_i = \mathbf{y}_j$, which is characteristic for the unfolding model, and its curvature provides an example of "undesired" nonlinearities that cannot be linearized by the transformation methods of the Gifi system (Heiser, 1985).

A least squares fit of the two-sets distance model can be calculated iteratively by the program

SMACOF-3 (Heiser, 1981), and the type of representation obtained will be illustrated with an example of political preference data collected among Dutch Members of Parliament (MP's) in 1990 (Hillebrand and Meulman, 1992). Here the rows of the data matrix refer to 136 MP's, and will be labelled in the unfolding representation with the initial letter of their party allegiance ('g' = left wing, green party; 'p' = PvdA, social democrats; 'd' = D'66, liberal democrats; 'c' = CDA, christian democrats; 'v' = VVD, liberal-conservatives; 'k' = right-wing, calvinist parties). The columns of the data matrix refer to the four major parties PvdA, D'66, CDA, and VVD, and the MP's had indicated their "sympathy" on a response scale ranging from 0 to 100. The SMACOF-3 unfolding

Insert Figure 7 about here

representation in two dimensions is plotted in Figure 7. When one would pick up the plane at a particular party point, and *fold* it, the MP's would appear (approximately) in the order of their sympathy for that party – hence the name *unfolding* for a technique that has to reverse this process. The fitted distances account for 84% of the variance of the sympathy ratings, which is a very reasonable fit. CDA and D'66 are frequently called "center" parties, but this analysis shows that centrality is not at all reflected in the mutual sympathies in Parliament. Interpretation is aided by inspection of the *issue directions* that have been fitted in afterwards, by multiple regression of additional rating scale data concerning a number of political issues. One can see, for example, that the large separation between MP's from D'66 and the christian-democrats has to do with the EUTHANASIA and ABORTION issues, while the opposition between MP's from PvdA and VVD is related to more traditional concerns like the distribution of INCOME.

Summarizing, the unfolding model can be used for approximation and representation of nonlinear response surfaces of moderate complexity. Its major strength lies in the idea of having one family of functions with common shape that is shifted in location, much like the situation usually assumed in discriminant analysis (Takane, 1987). For further theoretical developments in modeling of single-peaked (or *unimodal*) response surfaces and applications in ecology, see Ter Braak (1987, 1988); in psychology, many unfolding models for binary data have been developed, e.g. Andrich (1988), Formann (1988), and Hoijtink (1990).

Part III. Classification of units and constrained latent class analysis

The third and final part of this chapter is concerned with methods for the analysis of heterogeneity of units. Many methods for grouping statistical units in subpopulations (types, or clusters) on the basis of multivariate data have been proposed. Rather than attempting to give an exhaustive overview of this steadily growing field, we will mention a few recent highlights and then discuss one particular approach in somewhat more detail.

When some *a priori* classification of units is available, we have in fact an additional categorical variable, and we could choose between a two-sets (of variables) analysis with the methods discussed in Part I, such as MORALS, CANALS and OVERALS (*cf.* Meulman *et al.*, 1992), or a one-set (of variables) *mixed* approach, with PRINCALS. In the latter case, the influence of the additional classification variable can be increased to a definite asymptotic level by including it several times, or by explicit weighting – a procedure called *forced classification* (Nishisato, 1984, 1988).

Alternatively, we could apply Meulman's (1986, 1992) distance approach, which aims at a representation of the units in a low-dimensional Euclidean space by multidimensional scaling, under optimal transformation of variables. Working with least squares distance approximations, this approach, which has the same scope of application as the Gifi system, avoids the projection of points from the high-dimensional observation space into low-dimensional classification space, an operation that is characteristic for most multivariate techniques, and that is bound to diminish the interunit distances unevenly. For classification purposes, it may be necessary to reduce the full dimensionality of observation space, but to keep small distances small and large distances large in the process of allocating objects to classes. A succesful case of classification analysis in this framework has been reported by Meulman (1992).

When the classification has to be done on *a posteriori* grounds, it may still be a good idea to use differential weighting of the variables (De Soete *et al*, 1985), or differential weighting and optimal quantification of variables (Van Buuren and Heiser, 1989). Many traditional clustering methods start in a fixed observation space, in which some measure of inter-unit dissimilarity is derived, and proceed from the dissimilarity matrix without ever referring back to the original data. Differential

weighting does not work with a fixed definition of dissimilarity, but iteratively downweights the influence in the dissimilarity function of variables that are detrimental to the clustering objective. Similarly, the GROUPALS method proposed by Van Buuren and Heiser (1989) is a version of PRINCALS in which the object scores $\mathbf{X}$ are constrained to K locations in space, under free choice of dimensionality p, where the number of groups K is some prechosen parameter larger than p (and of course smaller than N).

In recent years there has been a remarkable revival of Latent Class Analysis (LCA), a group of methods which originated in the fifties, culminating in the classic monograph of Lazarsfeld and Henry (1968). A good technical article on an application of LCA is Aitkin *et al.* (1981); many recent developments are covered in Langeheine and Rost (1988), and Lindsay *et al.* (1991). Perhaps one of the main reasons for its increased use has been that more flexible algorithms have become available than the ones used in the early days, as will also become evident in the work to be presented below.

The latent class approach to unfolding

As we have seen in Part II, the unfolding technique tries to describe the elements of a two-mode data matrix by allocating *row points* $\mathbf{x}_i$ to the units, or *row objects*, and *column points* $\mathbf{y}_j$ to the variables, or *column objects*, and then fitting a distance model. If the set of units is regarded as fixed, we need as many row points as there are units. But under the assumption of heterogeneity, it is supposed that the N units can be clustered into K homogeneous groups, so that units within each group exhibit a similar data profile that varies only randomly, while the systematic variation between groups would be described by a distance model. Then it would suffice to represent the units within the same homogeneous group by a single point, called the *ideal point* $\mathbf{x}_k$ with $k = 1,...,K$.

De Soete and Heiser (1993) have developed a method that *simultaneously* arrives at a clustering of the units into a small number of homogeneous groups and constructs a geometrical representation based on the unfolding model where each group is represented by a single ideal point. Their approach is based on a mixture model formulation. Since independence is assumed *within* each component distribution (i.e., local independence holds), the model can be considered as a *latent class model.* Let λ_k denote the unconditional probability that a unit belongs to latent class k, with $\Sigma_k \lambda_k = 1$, and let $\boldsymbol{\lambda}$ denote the vector $(\lambda_1,...,\lambda_k,...,\lambda_K)'$. It is assumed that the data in row $\mathbf{q}_i$ of $\mathbf{Q}$ for any unit i that is

allocated to latent class k are independently and multivariate normally distributed with means $\boldsymbol{\mu}_k = (\mu_{k1},...,\mu_{kj},...,\mu_{km})'$ and common variance σ^2. The class-specific mean vector $\boldsymbol{\mu}_k$ is *reparametrized* in terms of the vector of p-dimensional Euclidean distances $d(\mathbf{x}_k,\mathbf{Y})$ between the ideal point $\mathbf{x}_k$ and the column points $\mathbf{y}_j$, collected in the rows of the $m \times p$ matrix $\mathbf{Y}$. Thus for unit i in class k we have

$$\mathbf{q}_i \sim N(\boldsymbol{\mu}_k, \sigma^2\mathbf{I}) \quad \text{with} \quad \boldsymbol{\mu}_k = \alpha_k\mathbf{1} - d(\mathbf{x}_k,\mathbf{Y}) , \tag{17}$$

where $\mathbf{I}$ denotes the identity matrix, $\mathbf{1}$ denotes an m-vector of ones, and α_k is a class-specific additive constant, which picks up the general response level of class k that cannot be accounted for by the distance model. The independence assumption entailed in (17) is analogous to the assumption of local independence in the classical latent class model (Lazarsfeld and Henry, 1968). The reparametrization in (17) implies that the data are assumed to be *similarities* on an *interval scale*, i.e. inversely and linearly related to the distance, which in turn is related to a smaller number of spatial parameters.

Model assumption (17) is conditional upon the class membership of unit i. Because in fact we do not know to which latent class a particular unit belongs, the probability density function $\phi(\cdot)$ of the data of an arbitrary unit i is a *finite mixture* of multivariate normal densities $f(\cdot)$:

$$\phi(\mathbf{q}_i \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \sigma^2) = \sum_k \lambda_k f(\mathbf{q}_i \mid \mathbf{x}_k, \mathbf{Y}, \alpha_k, \sigma^2) ,$$

where the ideal points are collected in the rows of the $K \times p$ matrix $\mathbf{X}$, and the additive constants in the vector $\boldsymbol{\alpha}$. It should be noted that we are dealing here with a special case of the general mixture model with multivariate normal densities as discussed by Wolfe (1970); a similar specialization for fitting the bilinear model with latent classes has been proposed by De Soete and Winsberg (1993).

Generalized EM algorithm for unfolding with ideal points for latent classes

As is the case with most mixture distribution problems (*cf.* McLachlan and Basford, 1988), the parameters of the latent class unfolding model are most conveniently estimated by means of an EM algorithm (Dempster *et al.*, 1977). For this purpose, the unknown indicator matrix $\mathbf{G}$ with elements $\{g_{ik}\}$ is introduced, containing "non-observed data": g_{ik} indicates the class membership of unit i with respect to class k. Row i of $\mathbf{G}$, written as a column vector, is denoted by $\mathbf{g}_i$. Assuming that the $\mathbf{g}_i$ are independently and identically multinomially distributed with probabilities $\boldsymbol{\lambda}$, the likelihood of the

"complete data" can be written as

$$L_C(\mathbf{X}, \mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \sigma^2 \mid \mathbf{Q}, \mathbf{G}) = \prod_i \prod_k \lambda_k^{g_{ik}} f(\mathbf{q}_i \mid \mathbf{x}_k, \mathbf{Y}, \alpha_k, \sigma^2)^{g_{ik}} .$$

Using (17), the complete-data log-likelihood can therefore be given explicitly as

$$\log L_C = \sum_i \sum_k g_{ik} \log \lambda_k - mN \log \sigma(2\pi)^{1/2} - (1/2\sigma^2) \sum_i \sum_k g_{ik} \| \mathbf{q}_i - \alpha_k \mathbf{1} + \mathrm{d}(\mathbf{x}_k, \mathbf{Y}) \|^2 . \quad (18)$$

The generalized EM (GEM) algorithm for maximizing (18) alternates between an E-step (expectation step) and a *generalized* M-step (*partial* maximization step). Since $\log L_C$ is linear in g_{it}, the E-step of the algorithm amounts to determining the expected value of g_{it} given the observed $\mathbf{q}_i$ and the current fitted parameters. These estimated expected values $E(g_{ik} \mid \cdot)$ are equal to the current conditional probabilities that $\mathbf{q}_i$ belongs to each of the K classes (Dempster *et al.*, 1977).

Given the $E(g_{ik} \mid \cdot)$, which may be inserted in (18) to obtain the expected log-likelihood, the provisional parameter estimates are improved in the generalized M-step, and new expected values of g_{ik} are determined. As is well-known, the GEM procedure converges monotonically; for further details concerning the partial maximization of the expected log-likelihood, see De Soete and Heiser (1993). Here, just one more interesting feature is pointed out. Given estimates of $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, and σ^2, the maximization of the expected $\log L_C$ only involves the minimization of the third term in (18), which can be decomposed, with $E(g_{ik} \mid \cdot)$ inserted, into a *within-class* and a *between-class* component:

$$\sum_i \sum_t E(g_{ik} \mid \cdot) \| \mathbf{q}_i - \alpha_k \mathbf{1} + \mathrm{d}(\mathbf{x}_k, \mathbf{Y}) \|^2 = \sum_i \sum_k E(g_{ik} \mid \cdot) \| \mathbf{q}_i - \bar{\mathbf{q}}_k \|^2 + \sum_k \gamma_k \| \tilde{\mathbf{q}}_k - \mathrm{d}(\mathbf{x}_k, \mathbf{Y}) \|^2 .$$

In this decomposition, which can be verified by applying Huygens' Theorem K times on the vectors $\mathbf{q}_i$ with masses $E(g_{ik} \mid \cdot)$, we have used the notation $\bar{\mathbf{q}}_k = \sum_i E(g_{ik} \mid \cdot) \mathbf{q}_i \,/\, \sum_i E(g_{ik} \mid \cdot)$ for the weighted mean observation vector for class k, and $\tilde{\mathbf{q}}_k$ for the *pseudodistances* $\tilde{\mathbf{q}}_k = \alpha_k \mathbf{1} - \bar{\mathbf{q}}_k$ (*i.e.*, transformed data that are to be approximated by distances). The quantities $\gamma_k = \sum_i E(g_{ik} \mid \cdot)$ are the expected marginals of $\mathbf{G}$, and are estimates of $N\lambda_k$. The first component is a pooled mean squared deviation within classes, and estimates $Nm\sigma^2$, while the second component is a between-class residual sum of squares, which is equal to a weighted least squares unfolding loss function. So good fit of the unfolding model to the class-specific pseudo-distances will tend to annihilate the between

component, and the within component measures the homogeneity of the classes.

Latent class unfolding of 1979 sympathy ratings in Dutch Parliament

Rating party sympathies among Members of Parliament is a continuing collaborative research project in the Netherlands that started in 1968. As the number of political parties tends to decrease over time, it is interesting to analyse some of the older data to retrospectively predict party merging with the latent class approach. In Table 4 13 parties are given the that were present in Parliament in 1979; they are listed in order from political left to political right according to expert judgment, and

Insert Table 4 about here

grouped on *a priori* grounds in the same groups as the ones used in Figures 1 and 2. The 135 MP's that participated in the 1979 study belonged to eight of the thirteen parties listed in Table 4. MP's of the remaining (small) parties did not participate for a variety of reasons. A detailed description of the model selection process for constrained latent class methods in general, and of the results to be presented below can be found in De Soete and Heiser (1993); these data have also been analysed by Meulman and Verboon (1993) with their generalization of *points of view analysis.*

Two important parameters for model selection are K, the number of classes, and p, the dimensionality of the unfolding representation. They have been chosen, first K and then p, by a parametric bootstrap significance test based upon Monte Carlo simulation of the distribution of the likelihood ratio statistic. An elementary approach like that is necessary, because the regularity conditions for the usual asymptotic distribution of the likelihood ratio statistic are known to be violated when comparing two mixture models with a different number of component distributions (see McLachlan and Basford, 1988). The significance tests are given in Table 5; they indicate that 3 classes are sufficient for these

Insert Table 5 about here

data, leading to an unconstrained latent class model with 42 (3 + 3x13) degrees of freedom. For two two-dimensional unfolding models, one with three class-specific additive constants and the other with one common additive constant, the Monte Carlo significance tests are also reported in Table 5. The

model with the separate additive constants seems to fit the data equally well as the unconstrained 3-class model (but has only 35 degrees of freedom), while the common additive constant model must be discarded at a rejection level of .05.

The unfolding configuration is displayed in Figure 8. In the figure, the latent class ideal points are

Insert Figure 8 about here

labeled A, B, and C. Each MP can be assigned to one of these classes on the basis of the posterior probabilities that follow from the estimated parameters and Bayes' rule. Such a classification was carried out for the solution presented in Figure 8 and is summarized in the right part of Table 5. From the table, it is clear that latent class A groups most of the MP's of the leftists parties PSP, PPR, PvdA and D66. The members of the christian-democratic parties ARP, KVP and CHU are mainly classified in latent class B, while class C groups the members of the VVD, which is an economically conservative party. In Figure 8, an ellipse is drawn around the parties that correspond to each latent class according to this classification. Note that the major features of Figure 7 are very similar.

During the eighties, two mergers occured in the Dutch political scene: the christian-democratic parties KVP, ARP and CHU merged into the CDA (Christian Democratic Appeal), and the small left-wing parties CPN, PPR and PSP merged into the GL (Green Left). At the right, nothing much changed, except that D70 disappeared. The first actual merger is our latent class B, but the second was not along the lines of latent class A, as we could have predicted. Although the formation of a large progressive party around PvdA and D66 has been rather seriously contemplated, the process aborted. Meanwhile, according to recurrent opinion polls, the balance of seats in Parliament among these two parties would have to be dramatically altered in favour of D66.

Discussion

In this chapter, we have found it useful to distinguish nonlinear methods in several ways. A first major distinction is between methods that use nonlinear transformations of the data to obtain more simple relationships, and methods that are nonlinear in their description of systematic variability. A second major, three-fold distinction is between methods that focus on heterogeneity of samples or populations, methods that regard the set of units as homogeneous, and methods that take the units as a group of fixed, recognizable objects. Thirdly, we have seen that the distinction between fixed and stochastic (which in turn might be split into homogeneous and heterogeneous) is applicable to sets of variables too. The unavoidable occurrence, in many sciences, of stochastic sets of variables – or "indicators" with measurement errors – forms an important reason why the homogeneity analysis methods described in Part I exist, and why various forms of variable reduction, quantification, and transformation are of interest.

An obvious link upon which we have not been able to touch, is the relationship between analysis of variance and multiple regression with ordinal transformations, as in MONANOVA and CANALS, with generalized linear modelling (McCullagh and Nelder, 1983). There is also a close link of OVERALS with generalized Procrustes analysis (Gower, 1975) and matching of configurations of points (Commandeur, 1991), but here the similarity is more in terms of the algebra and the mechanics of algorithm construction, because the rationale of restricting cross classification variables in terms of linear combinations of edge variables has nothing to do with Procrustes analysis or matching. Our example in Part I presented a new possibility – which is a unique feature of OVERALS, not shared by its two-sets predecessors – to give a multidimensional solution with optimal dimension reduction in a situation of predicting one criterion variable.

Considerations of homogeneity and heterogeneity also extend to sets of residuals. In actual data analysis, heterogeneous residuals are the rule, rather than the exception – especially with the highly parametrized models considered here. So far, there have only been a few attempts to robustify these nonlinear methods along the lines presented in Part II (Heiser, 1987b,1988; Verboon and Heiser, 1992, 1994; Verboon, 1993). Frequently, a sensible thing to do when one has to work with least

squares methods is a *post-hoc* stability analysis (Gifi, 1990), to get an impression of the sensitivity of the results to changes in the data. A noteworthy theoretical study of apparent "anomalies" in the results of these nonlinear methods under specific circumstances is Buja (1990).

In Part III we discussed constrained latent class analysis in some detail, because it seems to have prototypical elements that can be adapted to a wide variety of other situations and models. A similar rationale for the case of unfolding binary and ordered categorical data, for example, was developed in Böckenholt and Böckenholt (1991). Generalized points-of-view analysis (Meulman & Verboon, 1993) is another classification approach of great flexibility, in which data vectors or matrices are clustered so that some parametric model is optimized *within* clusters, rather than *between* clusters, and there, too, a wide range of new developments is to be expected.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modeling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society, A, 144*, 419-461.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*, 33-51.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. New York: Wiley.

Beaton, A.E. and Tukey, J.W. (1974). The fitting of power series. *Technometrics, 16*, 147-185.

Birnbaum, S. (Ed.). *Get 'em and Go Travel Guide for the United States 1979*. Boston: Houghton Mifflin, 1979.

Bock, R.D. (1960). *Methods and Applications of Optimal Scaling*. Report 25. Chapel Hill, North Carolina: L.L. Thurstone Laboratories, University of North Carolina.

Bock, H.H. (Ed.), *Classification and Related Methods of Data Analysis*. Amsterdam: North-Holland, 1988.

Böckenholt, U. and Böckenholt, I. (1991). Constrained latent class analysis: simultaneous classification and scaling of discrete choice data. *Psychometrika, 56*, 699-716.

Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *The Annals of Statistics, 18*, 1032-1069.

Carroll, J.D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. *Proc. of the 76th Annual Convention of the Am. Psych. Ass.*, pp. 227-228.

Carroll, J.D. and Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31*,

607-649.

CBS (1987). *Statistiek der Verkiezingen 1986, Tweede Kamer der Staten-Generaal 21 mei.* Den Haag: Staatsuitgeverij.

Commandeur, J.J.F. (1991). *Matching Configurations.* Doctoral Dissertation. Leiden: DSWO Press.

Coombs, C.H. (1964). *A Theory of Data.* New York: Wiley.

De Leeuw, J. (1973). *Canonical Analysis of Categorical Data.* Doctoral Dissertation, University of Leiden.

De Leeuw, J. (1989). Multivariate analysis with linearization of the regressions. *Psychometrika, 53,* 437-454.

De Leeuw, J. and Van Rijckevorsel, J. (1980). HOMALS and PRINCALS: some generalizations of principal components analysis. In E. Diday *et al.* (Eds.), *Data Analysis and Informatics, Vol. I,* pp. 231-242. Amsterdam: North-Holland.

De Leeuw, J., Young, F.W., and Takane, Y. (1976). Additive structure in qualitative data: an alternating least squares method with optimal scaling features. *Psychometrika, 41,* 471-503.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from uncomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, B, 39,* 1-38.

De Soete, G., De Sarbo, W.S., and Carroll, J.D. (1985). Optimal variable weighting for hierarchical clustering: an alternating least squares algorithm. *Journal of Classification, 2,* 173-192.

De Soete, G. and Heiser, W.J. (1993). A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika, 58, in press.*

De Soete, G. and Winsberg, S. (1993). A latent class vector model for preference ratings. *Journal of Classification, 10,* in press.

Draper, D., Hodges, J.S., Mallows, C.L., and Pregibon, D. (1993). Exchangeability and data analysis. *Journal of the Royal Statistical Society, A, 156,* 9-37.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1,* 211-218.

Fisher, R.A. (1938). *Statistical Methods for Research Workers, 7th edition.* Edinburgh: Oliver and Boyd.

Fisher, R.A. (1940). The precision of discriminant functions. *Annals of Eugenics, 10,* 422-429.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* London: Hafner.

Formann, A.K. (1988). Latent class models for nonmonotone dichotomous items. *Psychometrika, 53,* 45-62.

Friedman, J.H. and Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, C-23,* 881-890.

Gabriel, R. (this book).

Gifi, A. (1985). *PRINCALS.* Research Report UG-85-03. Leiden: Department of Data Theory.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* New York: Wiley.

Gittins, R., Amir, S., Dupouey, J.-L, Heiser, W.J., Meyer, M., Sokal, R.R. and Werger, M.J.A.

(1987). Numerical methods in terrestrial ecology. In P. Legendre et al. (Eds.), *Developments in Numerical Ecology*, pp. 521-558. New York: Springer.

Gower, J.C (1975). Generalized Procrustes analysis. *Psychometrika, 40*, 33-51.

Gower, J.C (1990). Fisher's optimal scores and multiple correspondence analysis. *Biometrics, 46*, 947-961.

Green, B.F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika, 17*, 429-440.

Green, P.E. (1973). On the analysis of interactions in marketing research data. *Journal of Marketing Research, 10*, 410-420.

Green, P.E. (1974). On the design of choice experiments involving multifactor alternatives. *Journal of Consumer Research, 1*, 61-68.

Green, P.E. and Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research, 5*, 103-123.

Green, P.E. and Srinivasan, V. (1990). Conjoint analysis in marketing: new developments with implications for research and practice. *Journal of Marketing, 54*, 3-19.

Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, B, 46*, 149-192.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M.J. (1991). Interpreting multiple correspondence analysis. *Applied Stochastic Models and Data Analysis, 7*, 195-210.

Guttman, L. (1941). The quantification of a class of attributes: a theory and method of scale construction. In P. Horst (Ed.), *The Prediction of Personal Adjustment.* New York: SSRC.

Hand, D.J. (1981). *Discrimination and Classification.* London: John Wiley.

Heiser, W.J. (1981). *Unfolding Analysis of Proximity Data.* Doctoral dissertation, University of Leiden, The Netherlands.

Heiser, W.J. (1985). Undesired nonlinearities in nonlinear multivariate analysis. In E. Diday et al. (Eds.), *Data Analysis and Informatics, Vol IV*, pp. 455-469. Amsterdam: North-Holland.

Heiser, W.J. (1987a). Joint ordination of species and sites: the unfolding technique. In P. Legendre *et al.* (Eds.), *Developments in Numerical Ecology*, pp. 189-221. New York: Springer.

Heiser, W.J. (1987b). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis, 5*, 337-356.

Heiser, W.J. (1988). Multidimensional scaling with least absolute residuals. In H.H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, pp. 455-462. Amsterdam: North-Holland.

Heiser, W.J. and Meulman, J.J. (1983). Analyzing rectangular tables by joint and constrained multidimensional scaling. *Journal of Econometrics, 22*, 139-167.

Heiser, W.J. and Meulman, J.J. (1994). Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In M. Greenacre and J. Blasius (Eds.), *Correspon-*

dence Analysis in the Social Sciences: Recent Developments and Applications. London: Academic Press, in press.

Hillebrand, R. and Meulman, J.J. (1992). Afstand en nabijheid: verhoudingen in de Tweede Kamer. In J.J.A. Thomassen *et al.* (Eds.), *De Geachte Afgevaardigde*, pp. 98-128. Naarden: Coutinho.

Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society, 31*, 520-524.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55*, 641-656.

Horst, P. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika, 1*, 53-60.

Horst, P. (1961a). Relations among *m* sets of variables. *Psychometrika, 26*, 129-149.

Horst, P. (1961b). Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology, 17*, 331-347.

Horst, P. (1965). *Factor Analysis of Data Matrices*. New York: Holt, Rinehart and Winston.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika, 28*, 321-377.

Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics, 13*, 435-475.

Jöreskog, K.G. and Wold, H. (1982). *Systems under Indirect Observation. (2 Vols.)*. Amsterdam: North-Holland.

Kerkhof, A.J.F.M., Van der Wal, J., and Hengeveld, M.W. (1988). Typology of persons who attempted suicide with predictive value for repetition: a prospective cohort study. In H.-J. Möller, A. Schmidtke, and R. Welz, *Current Issues of Suicidology*, pp. 194-203. Berlin: Springer.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society, Series B, 27*, 251-263.

Kruskal, J.B. and Shepard, R.N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika, 39*, 123-157.

Lancaster, H.O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika, 44*, 289-292.

Lancaster, H.O. (1958). The structure of bivariate distributions. *Annals of Mathematical Statistics, 29*, 719-736.

Langeheine, R. and Rost, J. (Eds., 1988). *Latent Trait and Latent Class Analysis*. New York: Plenum Press.

Lebart, L., Morineau, A., and Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*. New York: Wiley.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lindsay, B., Clogg, C.C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86*, 96-107.

Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley Publ..

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Meulman, J.J. (1986). *A Distance Approach to Multivariate Analysis*. Doctoral Dissertation. Leiden: DSWO Press.

Meulman, J.J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika, 57*, 539-565.

Meulman, J.J. and Heiser, W.J. (1988). Second order regression and distance analysis. In: W. Gaul & M. Schader (Eds.), *Data, Expert Knowledge and Decisions*, pp. 368-380. Berlin: Springer.

Meulman, J.J. and Verboon, P. (1993). Points of view analysis revisited: fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika, 58*, 7-35.

Meulman, J.J., Zeppa, P., Boon, M.E., and Rietveld. W.J. (1992). Prediction of various grades of cervical preneoplasma and neoplasma on plastic embedded cytobrush samples: discriminant analysis with qualitative and quantitative predictors. *Analytic and Quantitative Cytology and Histology, 14*, 60-72.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.

Nishisato, S. (1984). Forced classification: a simple application of a quantification method. *Psychometrika, 49*, 25-36.

Nishisato, S. (1988). Forced classification procedure of dual scaling: its mathematical properties. In H.H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, pp. 523-532. Amsterdam: North-Holland.

Rao, C.R. (1980). Matrix approximation and reduction of dimensionality in multivariate statistical analysis. In P.R. Krishnaiah (Ed.), *Multivariate Analysis, Vol V*. Amsterdam: North-Holland.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

SPSS (1990). *Categories*. Chicago, Ill.: SPSS Inc.

Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika, 52*, 493-513.

Ter Braak, C.J.F. (1987). *Unimodal Models to Relate Species to Environment*. Doctoral Dissertation, Wageningen, The Netherlands.

Ter Braak, C.J.F. (1988). Partial canonical correspondence analysis. In H.H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, pp. 551-558. Amsterdam: North-Holland.

Tucker, L.R. (1960). Intra-individual and inter-individual multidimensionality. In H. Gulliksen and S. Messick (Eds.), *Psychological Scaling: Theory and Applications*, pp. 155-167. New York: Wiley.

Van Buuren, S. and Heiser, W.J. (1989). Clustering N objects into K groups under optimal scaling

of variables. *Psychometrika, 54*, 699-706.

Van der Burg, E. and De Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology, 36*, 54-80.

Van der Burg, E., De Leeuw, J., and Verdegaal, R. (1988). Homogeneity analysis with *K* sets of variables: an alternating least squares method with optimal scaling features. *Psychometrika, 53*, 177-197.

Van der Lans, I.A. (1992). *Nonlinear Multivariate Analysis for Multiattribute Preference Data.* Doctoral dissertation, Leiden: DSWO Press.

Verboon, P. (1993). Robust regression with optimal scaling. *British Journal of Mathematical and Statistical Psychology, 46, in press.*

Verboon, P. and Heiser, W.J. (1992). Resistant orthogonal Procrustes analysis. *Journal of Classification, 9*, 237-256.

Verboon, P., and Heiser, W.J. (1994). Resistant lower rank approximation of matrices by iterative majorization. *Computational Statistics & Data Analysis, 17, in press.*

Verboon, P., Van der Lans, I.A., and Heiser, W.J. (1991). *The MULTIPALS algorithm.* Research Report RR-91-04. Leiden: Department of Data Theory, University of Leiden.

Vlek, Ch. and Stallen, P.J. (1981). Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance, 28*, 235-271.

Wainer, H. and Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32*, 191-241.

Walberg, H.J. and Rasher, S.P. (1977). The ways schooling makes a difference. *Phi Delta Kappa, 58*, 703-707.

Wittink, D.R. and Cattin, P. (1989). Commercial use of conjoint analysis: an update. *Journal of Marketing, 53*, 91-96.

Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5*, 329-350.

Young, F.W., De Leeuw, J., and Takane, Y. (1976). Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. *Psychometrika, 41*, 505-529.

Young, F.W., Takane, Y., and De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. *Psychometrika, 43*, 279-281.

Table 1. Classification of multivariate approaches by distinguishing type of variability in both units and variables

SET(S) OF UNITS	SET(S) OF VARIABLES	
	fixed (no reduction of variables)	*stochastic* (variable reduction)
fixed (recognizable units)	*II* Data display and approximation Procrustes analysis	Projection pursuit Psychometric test theory Classical factor analysis
homogeneous (exchangeable units)	Multivariate distribution theory Loglinear modelling	Spearman hierarchy and LISREL modelling
heterogeneous (recognizable subpopulations)	*III* Mixture models Clustering/classification Classical discriminant analysis	*I* Canonical correlation Homogeneity analysis Reduced rank regression

Table 2. Variables of the social indicator data of the US, partitioned into three sets

	First set
URB	Ratio of urban to rural population (3)
INC	Per capita income in dollars (2)
LIFE	Life expectancy in years (2)
HOMIC	1976 homicide/non-negligent manslaughter rate (2)
UNEMP	1975 unemployment rate (3)
	Second set
SCHOOL	Pct of population who are high school graduates (1)
PUBLIC	Percent of public school enrollment (1)
TEACH	Ratio of public school pupils to teachers (1)
ILLIT	Illiteracy rate in percent of population (2)
	Third set
FAIL	Rate of failure on the Selective Service mental ability test (1)

Sources:
(1) Walberg & Rasher (1977); (2) Wainer & Thissen (1981); (3) Birnbaum (1979)

Table 3. Fit of the transformed variables in the component space (correlations with FAIL in brackets)

first set		*second set*		*third set*	
URB	.69 (.12)	SCHOOL	.87 (-.79)	FAIL	.99
INC	.82 (-.59)	PUBLIC	.78 (-.07		
LIFE	.86 (-.83)	TEACH	.31 (.19)		
HOMIC	.79 (.74)	ILLIT	.90 (.88)		
UNEMP	.16 (.16)				

Table 4. Parties included in the 1979 party sympathy study, and classification of MP's based on expert grouping and on posterior probabilities

Party	Description	Expert grouping	Latent class A	B	C	number of MP's
CPN	Communists		–	–	–	–
PSP	Pacifistic Socialists	left-wing	1	–	–	1
PPR	Radical Christians		3	–	–	3
PVA	Labour	social-	52	1	–	53
D70	Social Democrats (conservative)	democrats	–	–	–	–
D66	Liberals (econ. undogmatic)	democratic-liberal	8	–	–	8
ARP	Protestants (lower class)	christian-	–	12	–	12
KVP	Catholics	democrats	2	21	1	24
CHU	Protestants (upper-middle class)		–	9	–	9
VVD	Liberals (econ. conservative)	conservative-liberal	–	1	24	25
GPV	Very conservative Calvinists		–	–	–	–
SGP	Very conservative Calvinists	right-wing	–	–	–	–
BP	Farmers' Party		–	–	–	–
		Class Totals	66	44	25	135

Table 5. Results of the 1979 party sympathy study

Analyses with no constraints on the class means				
No. of Classes (K)	Model df	Log Likelihood	*Monte Carlo significance test of K versus K + 1 classes*	
			Likelihood Ratio	Probability
1	14	– 8025.5	758.1	< 0.01
2	28	– 7646.5	210.9	< 0.01
3	42	– 7541.1	109.0	0.17

Analyses with 3-class unfolding models					
No. of Dimensions (p)	Common α	Model df	Log Likelihood	*Monte Carlo significance test against 3-class unconstrained model*	
				Likelihood Ratio	Probability
2	yes	33	– 7582.2	82.3	0.02
2	no	35	– 7572.3	62.5	0.10

Note
The Monte Carlo significance procedure is based on $n = 500$

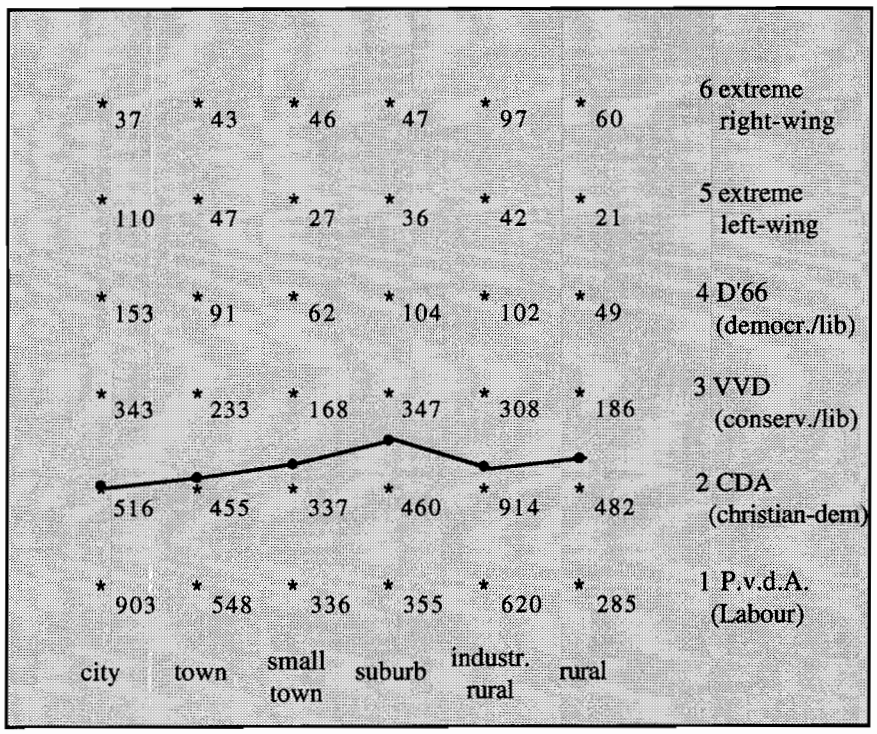

Figure 1.

Nonlinear regression of POLITICAL PARTY on URBANIZATION in Dutch 1986 elections (figures in thousands).

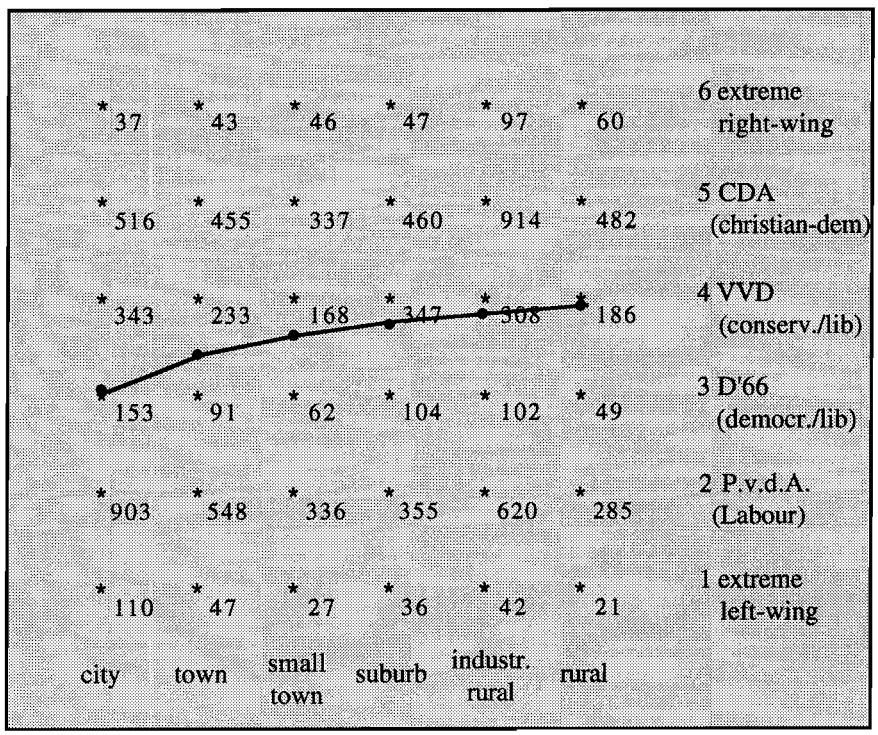

Figure 2.

Monotonized regression of POLITICAL PARTY on URBANIZATION in Dutch 1986 elections (figures in thousands).

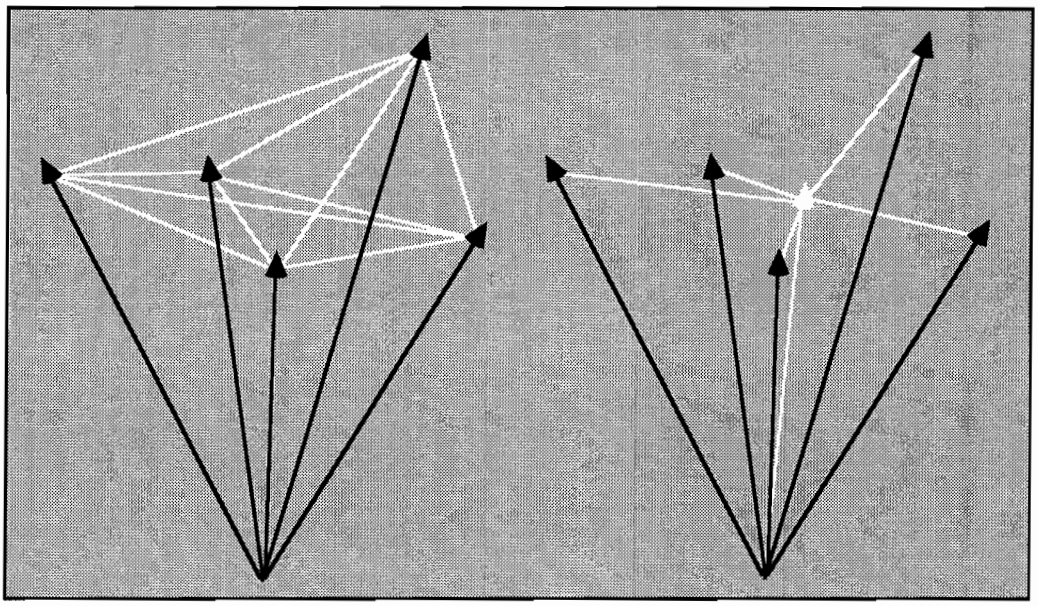

Figure 3.

Sum of squared distances among pairs of vectors (left) is equal to $2m$ times sum of squared distance towards their center (right).

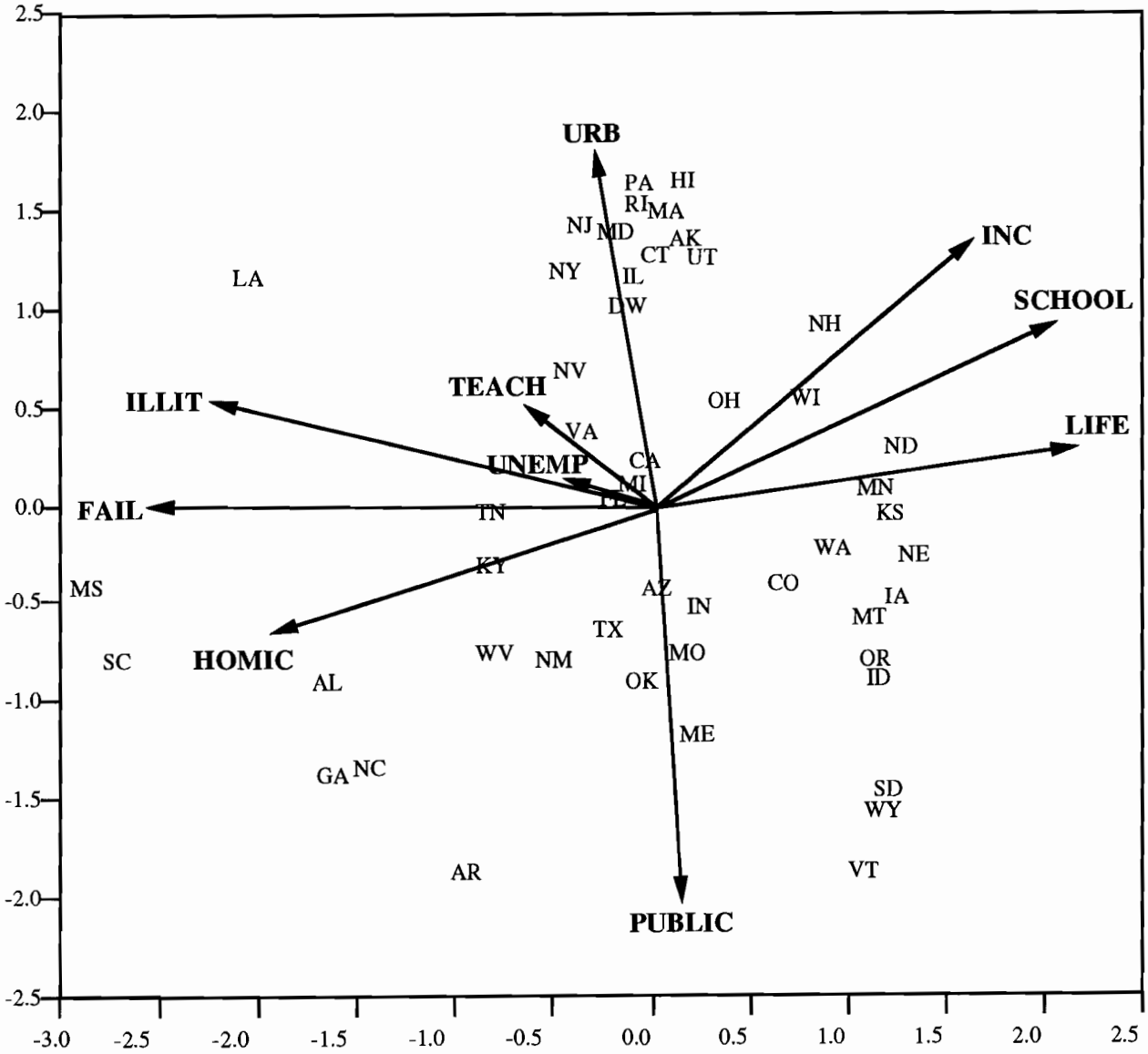

Figure 4.

Object points and projected edge variables in two-dimensional OVERALS solution for US data.

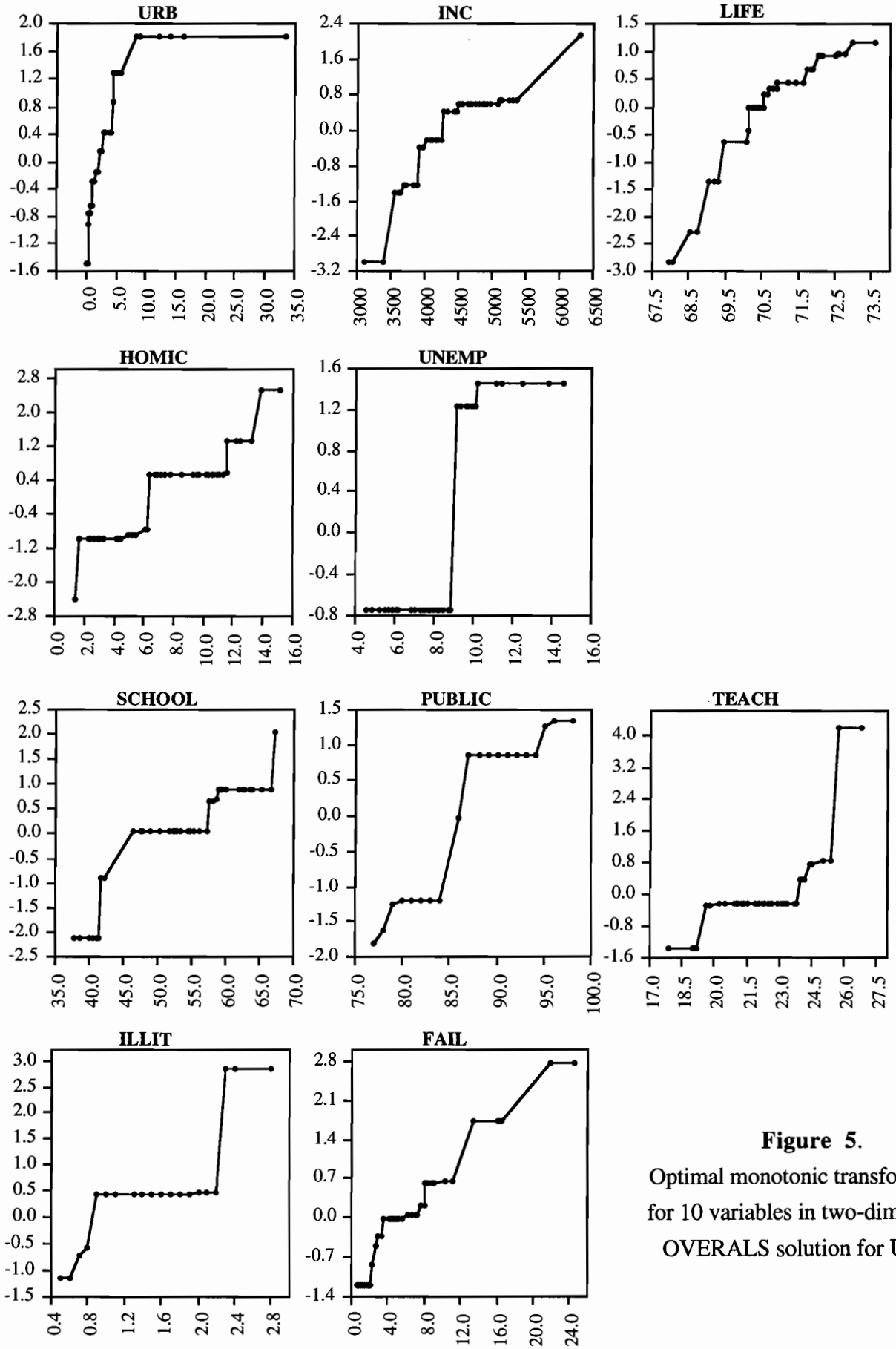

Figure 5.
Optimal monotonic transformations for 10 variables in two-dimensional OVERALS solution for US data.

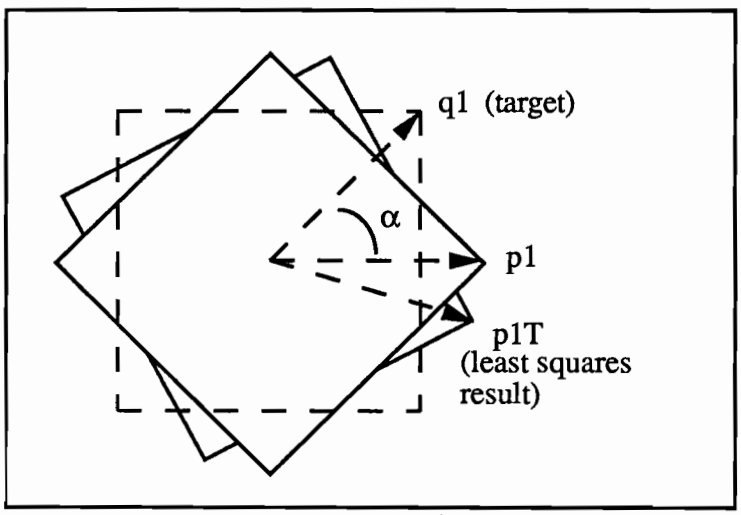

Figure 6.

The disturbing effect of one outlier in Procrustes analysis.

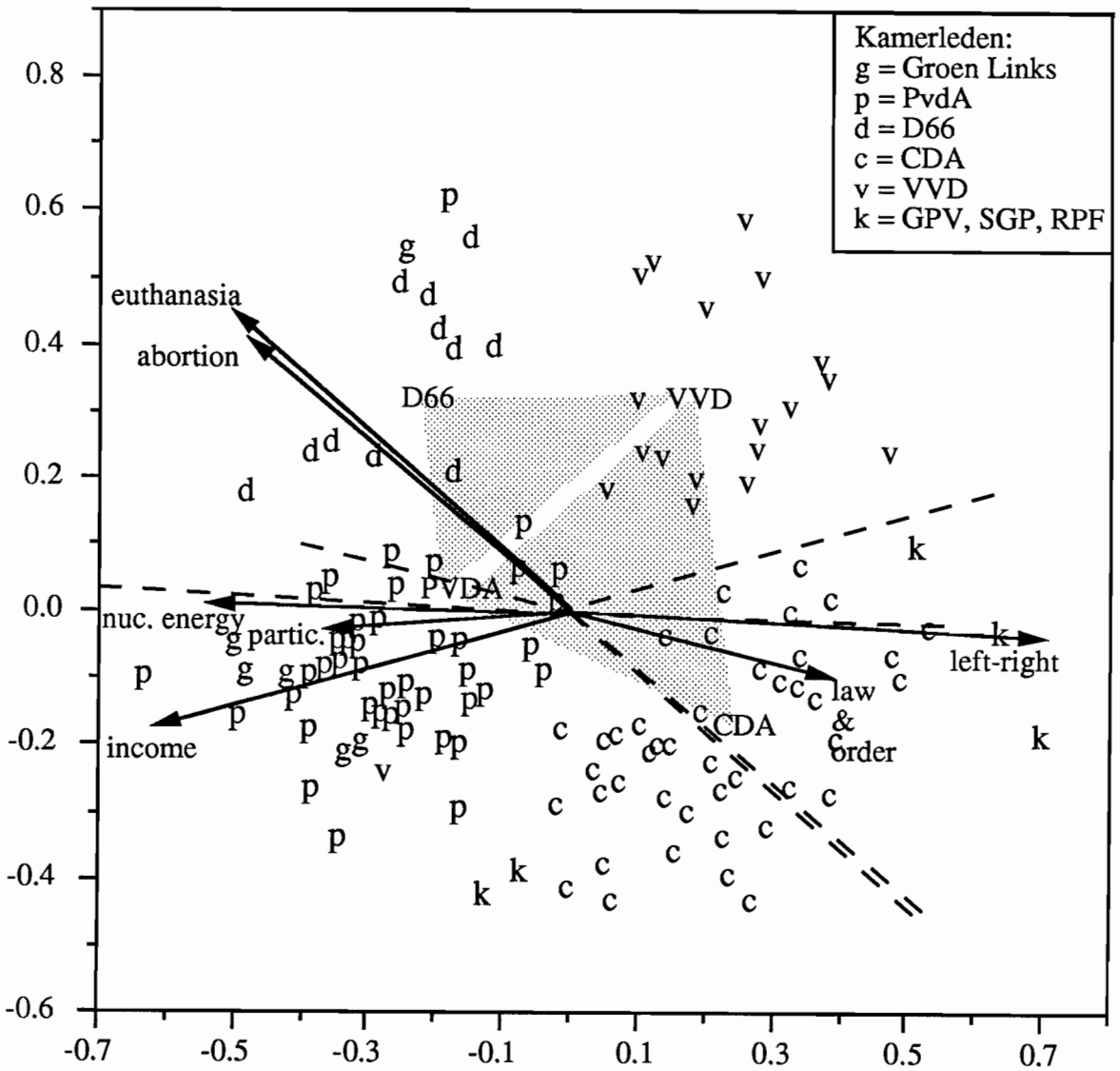

Figure 7.

Unfolding of 1990 party sympathies of Dutch Members of Parliament in two dimensions, with political issue directions fitted in afterwards.

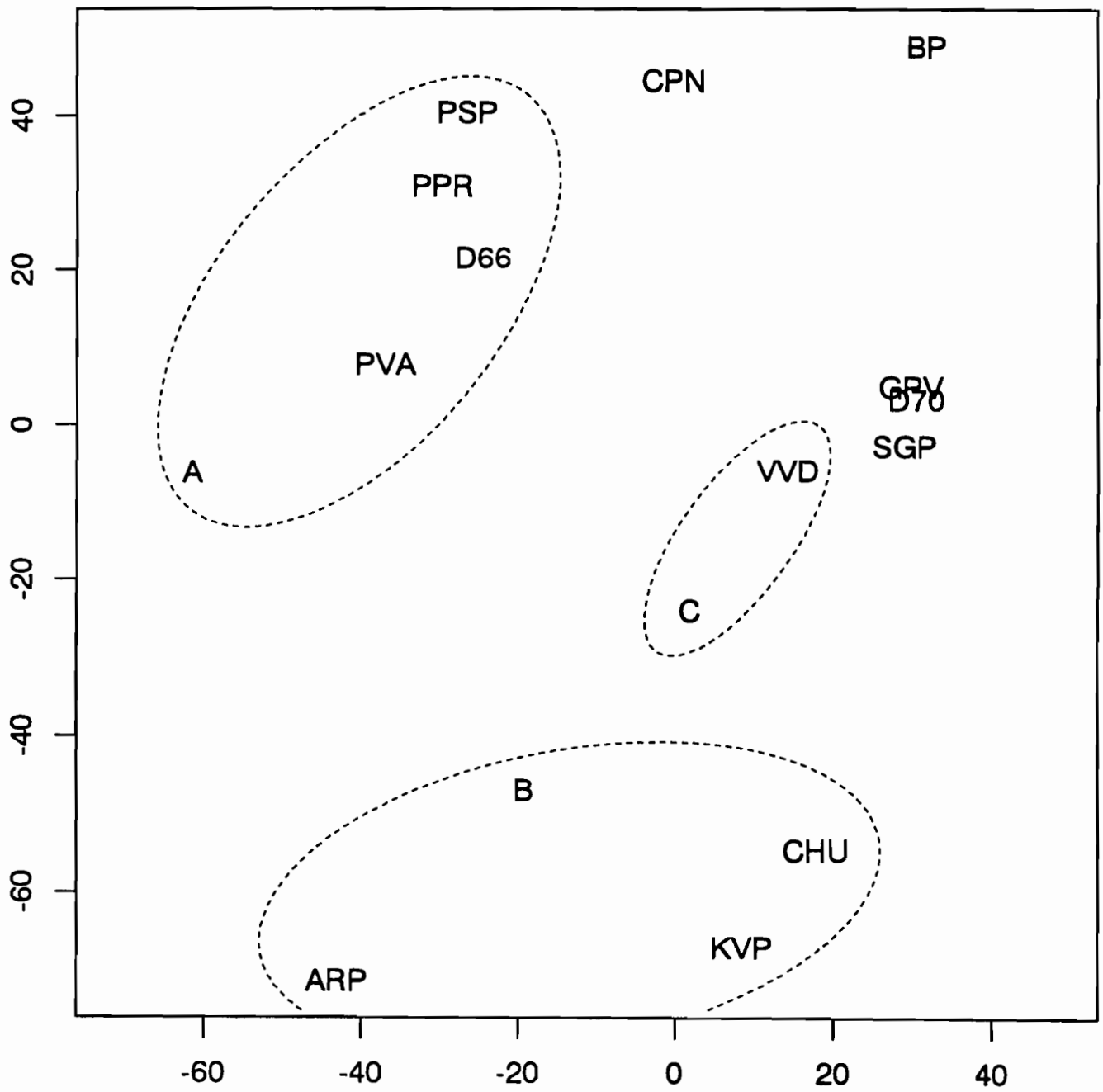

Figure 8.

Latent class unfolding of 1979 party sympathies of Dutch Members of Parliament in two dimensions (dashed ellipses indicate parties corresponding to each latent class).