RR 87-13

ALGEBRA AND GEOMETRY OF OVERALS

John P. Van de Geer

Rijksuniversiteit Leiden
Vakgroep Datatheorie
Middelstegracht 4
2312 TW Leiden

INDEX

1 INTRODUCTION

This monograph contains a discussion of the form of data analysis called OVERALS. Such an analysis can be performed by a computer program also called OVERALS. This program analyzes relations between K sets of variables, where it is assumed that the variables are categorical. A variable is said to be categorical if it sorts objects into a limited number of distinct categories.

Objective of this monograph is to explain the various *expressions* which appear in application of OVERALS, to give their *algebraic definitions*, and to show how the properties of an OVERALS solution appear in *graphs*.

Sections 2 to 10 describe the *classical* or *numerical* OVERALS. Here it is assumed that the categories of each variable have pre-assigned numerical values, also called the *a priori quantification* of the variables. The OVERALS solution then produces *weights*, such that for each of the K sets a weighted sumvector can be formed, and such that these K weighted sumvectors have optimal interrelations.

After these first sections, the monograph continues with sections about what is often called a *non-linear* OVERALS solution. The basic feature then becomes that the OVERALS criterion function is optimized not only by a suitable choice of weights, but also by the optimal choice of the *category quantifications* themselves. Such an OVERALS solution has essentially the same properties as a numerical solution that would appear *if* the optimal quantification is used for the pre-assigned values instead of the a priori quantification. Understanding of the numerical solution therefore is a pre-requisite for the understanding of the non-linear applications.

2 NOTATION

2.1 Introduction

OVERALS requires a very extensive repertory of symbols. They are collected in this chapter which thereafter can serve as a reference text. Our listing of symbols is,

somewhat arbitrarily, divided in a Section 2.2 on "general symbols" and 2.3 on "specific symbols".

2.2 General symbols

The point of departure of a numerical OVERALS analysis is a data matrix $\mathbf{Q}$, with n rows for the n objects, and m columns for the m variables. It will be assumed that columns of $\mathbf{Q}$ are standardized. I.e.: there is an a priori quantification that satisfies the requirement that columns of $\mathbf{Q}$ have zero mean, and also that the diagonal elements of $\mathbf{Q}'\mathbf{Q}/n = \mathbf{R}$ are equal to unity. It follows that $\mathbf{R}$ is a matrix of correlations between the m variables.

The m variables are partitioned into K sets, so that we can write

$$\mathbf{Q} = (\mathbf{Q}_1,...,\mathbf{Q}_k,...\mathbf{Q}_K)$$

where k is a running index to identify the kth set ($k = 1,...K$). The number of variables in $\mathbf{Q}_k$ is indicated by m_k, so that $m = \Sigma m_k$.

An individual column of $\mathbf{Q}_k$ is indicated by $\mathbf{q}_{kj}$ ($j = 1,...,m_k$): the j^{th} variable in the k^{th} set. The number of categories in $\mathbf{q}_{kj}$ is symbolized by k_{kj}.

OVERALS solves for weights in such a way that for each dimension of the OVERALS solution weighted sums can be derived for each of the K sets. The number of dimensions of the solution will be symbolized by p, and s will be used as a running index to identity a particular dimension: $s = 1,...,p$.

For dimension s, the weights for set $\mathbf{Q}_k$ are given the notation $\mathbf{w}_{k.s}$: a vector with m_k elements. Weights for $\mathbf{Q}_k$ on the p dimensions can be collected in a matrix $\mathbf{W}_k$, with columns $\mathbf{w}_{k.s}$ ($s = 1,...,p$), and with m_k rows. Such matrices $\mathbf{W}_k$, in turn, can be concatenated in a matrix $\mathbf{W}$, in which the submatrices $\mathbf{W}_k$ appear in vertical order. So $\mathbf{W}$ has m rows, and p columns.

A weighted sum for $\mathbf{Q}_k$ on dimension s can be written as $\mathbf{Q}_k\mathbf{w}_{k.s}$ (we shall always place the dimension index after a dot). Such weighted sumvectors can be collected in a matrix $\mathbf{Q}_k\mathbf{W}_k$, with p columns $\mathbf{Q}_k\mathbf{w}_{k.s}$, and n rows for the objects. It follows that

$$\mathbf{QW} = \Sigma\mathbf{Q}_k\mathbf{W}_k \ (k = 1,...K).$$

The OVERALS solution also refers to "object scores". For dimension s they are given in a vector $\mathbf{x}_{.s}$ with n elements. Those column vectors can be collected in a matrix $\mathbf{X}$, with

p columns $\mathbf{x}_{.s}$. It will be assumed that columns of $\mathbf{X}$ are standardized and that they are uncorrelated. This implies

$$\mathbf{X}'\mathbf{X}/n = \mathbf{I},$$

where $\mathbf{I}$ the p x p identity matrix.
For an arbitrary matrix $\mathbf{A}$ with n rows we shall use the notation

$$SSQ(\mathbf{A}) = \mathbf{A}'\mathbf{A}/n.$$

E.g., $SSQ(\mathbf{X}) = \mathbf{I}$. Also, $SSQ(\mathbf{Q}) = \mathbf{R}$ (a correlation matrix). Moreover, $SSQ(\mathbf{Q}_k)$ obtains the symbol $SSQ(\mathbf{Q}_k) = \mathbf{Q}_k'\mathbf{Q}_k/n = \mathbf{R}_{kk}$, where $\mathbf{R}_{kk}$ is a correlation matrix that appears as a diagonal block of $\mathbf{R}$. $\mathbf{R}_{kk}$ has dimension m_k x m_k.

The symbol $\mathbf{D}_R$ will be used for the block-diagonal matrix of $\mathbf{R}$. This matrix $\mathbf{D}_R$ can be derived from $\mathbf{R}$ by leaving all diagonal blocks $\mathbf{R}_{kk}$ as they are, but by replacing the off-diagonal blocks $\mathbf{R}_{kh} = \mathbf{Q}_k'\mathbf{Q}_h/n$ $(r \neq h)$ by zero matrices.

2.3 Special symbols

We also will use some special notations which depend on the fact that the variables are categorical.

Take variable $\mathbf{q}_{kj}$, with k_{kj} categories. For this variable an indicator matrix $\mathbf{G}_{kj}$ is defined as a matrix with n rows and k_{kj} columns. A row of the indicator matrix $\mathbf{G}_{kj}$ contains just one element equal to 1: it appears where the row object belong to the column category.
A matrix $\mathbf{D}_{kj}$ is defined by

$$\mathbf{D}_{kj} = \mathbf{G}_{kj}'\mathbf{G}_{kj},$$

from which it follows that $\mathbf{D}_{kj}$ is a diagonal matrix, with on the diagonal the marginal frequencies of each category of $\mathbf{q}_{kj}$.
Indicator matrices $\mathbf{G}_{kj}$ can be concatenated to a matrix $\mathbf{G}_k$:

$$\mathbf{G}_k = (\mathbf{G}_{k1},...,\mathbf{G}_{kj},...\mathbf{G}_{km_k}).$$

Such matrices $\mathbf{G}_k$ can in turn be aligned in a "super indicator matrix" $\mathbf{G}$:

$$\mathbf{G} = (\mathbf{G}_1,...,\mathbf{G}_k,...\mathbf{G}_K).$$

Suppose that $\mathbf{z}_{kj}$ is a column vector which gives the quantification of the k_{kj} categories of $\mathbf{q}_{kj}$. It then follows that

$$\mathbf{q}_{kj} = \mathbf{G}_{kj}\mathbf{z}_{kj}.$$

In addition, it will be true that

$$\mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{q}_{kj} = \mathbf{z}_{kj}.$$

The latter equation just says that if we take the average of all objects in $\mathbf{q}_{kj}$ within the same category (and therefore with identical quantification), we must obtain the quantification of that category.

3 GENERAL CRITERION OF THE OVERALS SOLUTION

The general criterion of the OVERALS solution can be elaborated in many different ways, as will be shown in the sequel of this monograph. But on the whole, the criterion is quite simple. It is to identify weights w (we now drop the dimension index *s*) such that weighted sumvectors $\mathbf{Q}_k\mathbf{w}_k$ (one for each of the K sets) are as much as possible "similar" to each other.

If the vectors $\mathbf{Q}_k\mathbf{w}_k$ are very similar, they also must be very similar to their average vector $\mathbf{Q}\mathbf{w}/K$.

Then let $\mathbf{x}$ be a vector proportional to the average vector $\mathbf{Q}\mathbf{w}/K$, and let $\mathbf{x}$ be standardized (SSQ($\mathbf{x}$) = 1). The aim of an OVERALS solution then becomes: let all individual weighted sum vectors $\mathbf{Q}_k\mathbf{w}_k$ be as similar as possible to their standardized average vector $\mathbf{x}$. Or: select weight $\mathbf{w}$ such that this criterion is satisfied.

More precisely: select, for the first dimension, weights $\mathbf{w}_{.1}$ such that this criterion is satisfied unconditionally. Select for the second dimension weight $\mathbf{w}_{.2}$ such that the criterion is maximized under the condition that $\mathbf{x}_{.2}$ must be uncorrelated with $\mathbf{x}_{.1}$.

4 FIRST INTERPRETATION OF THE OVERALS CRITERION

4.1 Introduction

In this chapter we give a purely *algebraic* formulation of the OVERALS criterion. Later on we shall indicate other ways to interpret the criterion. In fact, there are many different ways to explain the OVERALS criterion, and one of the difficulties of OVERALS is to show that these various explanations arrive at the same result.

4.2 The algebraic criterion

Consider the difference vector

$$\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k, \tag{1}$$

where, for the time being, we drop the dimension index (one should read $\mathbf{x} = \mathbf{x}_{.s}$ and $\mathbf{w}_k = \mathbf{w}_{k.s}$). This difference vector has average squared element:

$$\begin{aligned} SSQ(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k) &= (\mathbf{x}'\mathbf{x} - 2\mathbf{x}'\mathbf{Q}_k\mathbf{w}_k + \mathbf{w}_k'\mathbf{Q}_k'\mathbf{Q}_k\mathbf{w}_k)/n \\ &= 1 - 2\mathbf{x}'\mathbf{Q}_k\mathbf{w}_k/n + \mathbf{w}_k'\mathbf{R}_{kk}\mathbf{w}_k. \end{aligned} \tag{2}$$

Added over all K sets, we obtain:

$$\begin{aligned} K - 2\mathbf{x}'(\Sigma\mathbf{Q}_k\mathbf{w}_k)/n \quad &+ \mathbf{w}'\mathbf{D}_R\mathbf{w} = \\ K - 2\mathbf{x}'\mathbf{Q}\mathbf{w}/n \quad &+ \mathbf{w}'\mathbf{D}_R\mathbf{w}. \end{aligned} \tag{3}$$

The solution for $\mathbf{w}$ then must satisfy the following equation:

$$\mathbf{Q}_k'\mathbf{Q}\mathbf{w} = \mathbf{Q}_k'\mathbf{Q}_k\mathbf{w}_k(K\gamma) \tag{4}$$

in which γ is called an *eigenvalue* corresponding to the dimension s of the solution (one should read: $\gamma = \gamma_{.s}$). Why the OVERALS solution should obey this equation (4) will not

be proved here: this would entail far too much algebraic details.*

Equation (4) can be re-written as

$$\mathbf{Q}_k'\mathbf{Q}\mathbf{w}/n = \mathbf{R}_{kk}\mathbf{w}_k(K\gamma) \tag{5}$$

and this latter expression can be generalized to

$$\mathbf{R}\mathbf{w} = \mathbf{D}_R\mathbf{w}(K\gamma). \tag{6}$$

Equation (6) implies that

$$\mathbf{w}'\mathbf{R}\mathbf{w} = \mathbf{w}'\mathbf{D}_R\mathbf{w}(K\gamma). \tag{7}$$

However, equation (6) does not specify how the vector **w** should be normalized. Equation (6) remains valid when **w** is multiplied by an arbitrary constant. So we can take a normalization of **w** such that

$$\mathbf{w}'\mathbf{R}\mathbf{w} = (K\gamma)^2, \tag{8}$$

which implies

$$\mathbf{w}'\mathbf{D}_R\mathbf{w} = K\gamma. \tag{9}$$

Now define object scores **x** by taking

$$\mathbf{x} = \mathbf{Q}\mathbf{w}/(K\gamma) \tag{10}$$

so that $SSQ(\mathbf{x}) = \mathbf{w}'\mathbf{Q}'\mathbf{Q}\mathbf{w}/n(K\gamma)^2 = \mathbf{w}'\mathbf{R}\mathbf{w}/(K\gamma)^2 = 1$, as was required in Section 2.3. Equation (2) now can be re-written as

$$1 - 2\mathbf{w}_k'\mathbf{Q}_k'\mathbf{Q}\mathbf{w}/_n(\mathrm{K}\gamma) \quad + \quad \mathbf{w}_k'\mathbf{R}_{kk}\mathbf{w}_k \quad =$$

$$1 - 2\mathbf{w}_k'\mathbf{R}_{kk}\mathbf{w}_k \quad + \quad \mathbf{w}_k'\mathbf{R}_{kk}\mathbf{w}_k \quad =$$

$$1 - \mathbf{w}_k'\mathbf{R}_{kk}\mathbf{w}_k \tag{11}$$

---

* For a justification of the equation, see Van de Geer (1984, 1986). In both publications OVERALS is identified as "analysis based on **P**, with **t't**=K, with MAXBET criterion function".

with sum (added over the K sets, and equivalent to equation (3)):

$$K - \mathbf{w}'\mathbf{D}_R\mathbf{w} = K - K\gamma = K(1-\gamma). \tag{12}$$

The latter implies that, averaged over all K sets, the average squared distance between $\mathbf{x}$ and all $\mathbf{Q}_k\mathbf{w}_k$ becomes equal to $(1-\gamma)$. And so the solution for the first dimension, with largest eigenvalue $\gamma_{.1}$ gives the corresponding solution for $\mathbf{x}_{.1}$ with smallest average squared distance to all $\mathbf{Q}_k\mathbf{w}_{k.1}$.
Equation (6) can be extended to include more dimensions. It then must be written as

$$\mathbf{RW} = \mathbf{D}_R\mathbf{W}K\Gamma, \tag{13}$$

where $\mathbf{W}$ has p columns $\mathrm{w}_{.s}$, and where Γ is a diagonal matrix with on its diagonal the p eigenvalues $\gamma_{.s}$. Equations (8) and (9) now become

$$\mathbf{W}'\mathbf{R}\mathbf{W} = K^2\Gamma^2, \tag{14}$$

$$\mathbf{W}'\mathbf{D}_R\mathbf{W} = K\Gamma, \tag{15}$$

whereas equation (10) can be re-written as

$$\mathbf{X} = \mathbf{Q}\mathbf{W}K^{-1}\Gamma^{-1}, \tag{16}$$

with the consequence that

$$\mathrm{SSQ}(\mathbf{X}) = \mathbf{X}'\mathbf{X}/n = \mathbf{I}. \tag{17}$$

5 GRAPHIC INTERPRETATION OF OVERALS CRITERION

5.1 Introduction

The algebraic interpretation of the OVERALS criterion, given in Chapter 4, asks for a counterpart in geometrical (graphic) terms. We shall not produce this geometrical interpretation in abstract form, but will give just an example. It is an illustration with

$K = 3$ sets of variables, $m = 6$ categorical variables, $m_k = 2$ variables in each set. There are $n = 15$ objects. Each variable has $k_{kj} = 3$ categories.

The basic data is given in Table 1. In this table the categories of each variable are indicated with letters as labels. Table 2 gives the same data, but now with numerical values assigned to each category, in such a way that Table 2 corresponds to a matrix $\mathbf{Q}$ in which each column is standardized (zero mean, sum of squares equal to $n = 15$).

5.2 Numerical OVERALS solution of the example

For the example we take a numerical OVERALS solution in $p = 2$ dimension. The solution for weights $\mathbf{W}$ is given in the 6 x 2 matrix of Table 3. This table also shows the two eigenvalues (the diagonal elements of $\text{SSQ}(\mathbf{QW}) = \mathbf{W'Q'QW}/n = \mathbf{W'RW} = K^2\Gamma^2$ give the squares of those eigenvalues after multiplication with K).

For the further geometrical illustration we shall concentrate on the second set $\mathbf{Q}_2$, with weighted sumvectors $\mathbf{Q}_2\mathbf{W}_2$. Table 4 gives the solution for vectors $\mathbf{q}_{2j}\mathbf{w}_{2j.s}$ ($j = 1,2$; $s = 1,2$), and for the sumvectors $\mathbf{Q}_2\mathbf{w}_{2.s}$ collected in a matrix $\mathbf{Q}_2\mathbf{W}_2$.

In a column $\mathbf{q}_{2j}\mathbf{w}_{2j.s}$ there are only three different values, because $\mathbf{q}_{2j}$ contains only three categories. Columns $\mathbf{q}_{2j}\mathbf{w}_{2j.1}$ and $\mathbf{q}_{2j}\mathbf{w}_{2j.2}$ are proportional to each other (because they are both proportional to $\mathbf{q}_{2j}$ itself). In a column $\mathbf{Q}_2\mathbf{w}_{2.s}$ we may find 3 x 3 = 9 different values (by taking the 3 x 3 combinations of the values in $\mathbf{q}_{21}\mathbf{w}_{21.1}$ and $\mathbf{q}_{22}\mathbf{w}_{22.1}$ or of $\mathbf{q}_{21}\mathbf{w}_{21.2}$ and $\mathbf{q}_{22}\mathbf{w}_{22.2}$). Some of those 9 combinations do not actually occur in the data.

Figure 1 illustrates the results of Table 4. This figure, firstly, shows points for the categories p q r of variable $\mathbf{q}_{21}$, by taking the values in three different rows of ($\mathbf{q}_{21}\mathbf{w}_{21.1}$ $\mathbf{q}_{21}\mathbf{w}_{21.2}$) as coordinates. Since these values are proportional to each other, the category points, labelled p1, q1, r1, are located on a straight line. These points are called *single category quantifications*. In this expression the word "single" just means that the category points are on a straight line.

Similarly we have single category points for the categories of $\mathbf{q}_{22}$, with label p2 q2 r2, and coordinates given in ($\mathbf{q}_{22}\mathbf{w}_{22.1}$ $\mathbf{q}_{22}\mathbf{w}_{22.2}$).

Figure 1 also shows the 3 x 3 = 9 points obtained by taking rows of $\mathbf{Q}_2\mathbf{W}_2$ as coordinates. These 9 points are located on a regular 3 x 3 *lattice*, or grid.

For all $K = 3$ sets, Table 5 shows values of $\mathbf{Q}_k\mathbf{W}_k$, $k = 1,2,3$. Table 5 therefore partly repeats the data of Table 4, for the second set ($k = 2$). Table 5 also shows sumvectors $\mathbf{QW}$, with two columns $\mathbf{Qw}_{.1}$ and $\mathbf{Qw}_{.2}$. These sumvectors have sum of squares equal to, respectively, $n(K\gamma_{.s})^2$. They become standardized after division by

($K\gamma_{.s}$) (s = 1,2), and then become equal to $\mathbf{x}_{.1}$ and $\mathbf{x}_{.2}$. These two columns then are the columns of $\mathbf{X} = \mathbf{Q}\mathbf{W}\Gamma^{-1}/K$.

The reader is invited to check all these numerical operations: it will help greatly to understand not only what OVERALS is about, but also how the notation is utilized.

Figure 1, in addition, show 15 *object points*, using the rows of $\mathbf{X}$ as coordinates. These 15 object points have been connected to their corresponding lattice points by dotted lines. The average squared length of these dotted lines is found by adding the diagonal elements of

$$SSQ(\mathbf{X} - \mathbf{Q}_2\mathbf{W}_2) = \mathbf{I} - \mathbf{W}_2'\mathbf{R}_{22}\mathbf{W}_2.$$

Averaged over all K = 3 sets (i.e., if one would draw such dotted lines not only for the second lattice, given in Figure 1, but also for similar figures for the first and the third set, in which object points have the same location) one obtains an average squared length of the dotted lines equal to the sum of the diagonal elements in

$$\Sigma(\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k)/K = \mathbf{I} - \Gamma.$$

Their sum is equal to the trace of ($\mathbf{I}$ - Γ), and therefore equal to $K - \Sigma\gamma_{.s}$. The larger the p = 2 eigenvalues, the smaller the average squared length of the dotted lines in all K = 3 possible figures.

The value of ($K - \Sigma\gamma_{.s}$) is called the *loss* of the numerical OVERALS solution. The smaller the loss, the better the solution.

5.3 Graph of all lattices

Figure 2 gives a graph of all K = 3 lattices. This figure also contains the object points for the first three objects. These three objects are the objects in category a of variable $\mathbf{q}_{11}$. The three object points are connected to their corresponding lattice points, which results into 3 (object points) x 3 (lattice points = 9 dotted lines. If one would do the same for all object points, 15 x 3 dotted lines would appear, with average squared length equal to

$$\Sigma(1-\gamma_{.s}) = (1-.815) + (1-.592) = .593.$$

So the OVERALS solution minimizes this average squared length between object points and their corresponding lattice points.

5.4 Properties of lattices

The OVERALS interpretation given in Section 5.3 has its counterpart in the interpretation of the values of $\mathbf{Q}_k\mathbf{W}_k$ (coordinates of the lattice points). For given set k (k = 1,2,3) the average squared distance of lattice points to the origin is found by adding the diagonal elements of $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$. Averaged over all K sets we obtain an average squared distance of lattice points to the origin equal to

$$\Sigma(\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k)/K = \mathbf{W}'\mathbf{D}_R\mathbf{W}/K = \Gamma.$$

This shows that the OVERALS solution is better to the extent that lattice points (averaged over all sets) have larger spread around the origin.

The value $\gamma_{.s}$ could be called the *fit* of the OVERALS solution for dimension s. Fit is the spread of lattice points around the origin, fit and loss add up to 1, where it should be noted that 1 is the spread (variance) of the object points in the direction of dimension s.

5.5 Summary of geometrical results

In sum, one can say that the OVERALS solution in p dimensions has the following characteristic properties.

(i) Object points have equal spread in all directions ($\mathbf{X}'\mathbf{X}/n = \mathbf{I}$).

(ii) Distances between object points and lattice points are made as small as possible.

(iii) Distances between lattice points and the origin are made as large as possible.

Properties (ii) and (iii) are complementary. The larger the average squared distance between object points and lattice points, the smaller the average squared distance between lattice points and the origin.

But we want to conclude this section with a warning. A "perfect" solution implies that object points coincides with their lattice points, and that distances between object points and lattice points are zero. Such a solution implies that eigenvalues are equal to 1 (their maximum value). But at the same time the solution will probably be very trivial, because it shows that all $\mathbf{Q}_k\mathbf{W}_k$ are exactly the same (and all equal to $\mathbf{X}$). This implies that variables in different sets nevertheless sort objects in exactly the same way, and that they just replicate each other.

6 INTERPRETATION OF THE OVERALS CRITERION IN TERMS OF VARIANCES AND COVARIANCES

6.1 Interpretation of the OVERALS criterion in terms of optimal sum of covariances

Take the weighted sums of $\mathbf{Q}_k\mathbf{W}_k$ and the object scores $\mathbf{X}$. Their covariance matrix is found in

$$\mathbf{W}_k'\mathbf{Q}_k'\mathbf{X}/n = \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k.$$

It follows that fit for dimension s and set k is just the covariance between $\mathbf{Q}_k\mathbf{w}_{k.s}$ and $\mathbf{x}_{.s}$, and that this fit is found on the diagonal of $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$. Averaged over the K sets, this covariance is found to be equal to diagonal elements of

$$\Sigma\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k/K = \mathbf{W}'\mathbf{D}_R\mathbf{W}/K = \Gamma.$$

For the numerical example, results of $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$ are given in Table 6. These results are graphed in Figure 3. This figure is based on $\mathbf{x}_{.1}$ and $\mathbf{x}_{.2}$ as two orthogonal axes with unit length. Projections of vectors $\mathbf{Q}_k\mathbf{w}_{k.s}$ then have coordinates which are shown as the covariances in the rows of $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$ (Table 6).

Figure 3 shows that the vectors $\mathbf{Q}_k\mathbf{w}_{k.1}$ form a rather narrow bundle, or fan, around $\mathbf{x}_{.1}$, whereas vectors $\mathbf{Q}_k\mathbf{w}_{k.2}$ form a (somewhat wider) fan around $\mathbf{x}_{.2}$. This is characteristic of a good OVERALS solution. In fact, a good OVERALS solution implies that the vectors $\mathbf{Q}_k\mathbf{w}_{k.s}$ ($k = 1,...,K$) are similar to each other, and therefore are close to their average vector, of which $\mathbf{x}_{.s}$ is the standardized version. Figure 3 pictures this property that vectors $\mathbf{Q}_k\mathbf{w}_{k.1}$ have projections on $\mathbf{x}_{.1}$ with *maximum sum*. The vectors $\mathbf{Q}_k\mathbf{w}_{k.2}$ have projections on $\mathbf{x}_{.2}$ with maximum sum under the condition that $\mathbf{x}_{.2}$ is orthogonal to $\mathbf{x}_{.1}$.
Moreover, the sum of the projections of $\mathbf{Q}_k\mathbf{w}_{k.1}$ on $\mathbf{x}_{.1}$ have sum

$$\Sigma\mathbf{Q}_k\mathbf{w}_{k.1} = \mathbf{Q}\mathbf{w}_{.1} = \mathbf{x}_{.1}(K\gamma_{.1})$$

so that their sum has length $(K\gamma_{.1})$ - because $\mathbf{x}_{.1}$ has length 1. On the other hand, the projections of $\mathbf{Q}_k\mathbf{w}_{k.1}$ on $\mathbf{x}_{.2}$ have zero sum. This sum is found in an off-diagonal cell of $\mathbf{W}'\mathbf{D}_R\mathbf{W}$, and such cells are all 0.

6.2 Interpretation in terms of correlations

Elements of $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$ stand for covariance between the p weighted sumvectors in $\mathbf{Q}_k\mathbf{W}_k$ and the p vectors of object scores in $\mathbf{X}$. The reason why they are covariances is that the columns of $\mathbf{Q}_k\mathbf{W}_k$ are not standardized - their variances are found on the diagonal of $\mathbf{W}_k\mathbf{R}_{kk}\mathbf{W}_k$. It follows that this matrix becomes a matrix of correlations between $\mathbf{Q}_k\mathbf{W}_k$ and $\mathbf{X}$ if each cell of the matrix is divided by the square root of the diagonal element in the same row. Such correlations are given in Table 7. They also might be called the component loadings of the vectors $\mathbf{Q}_k\mathbf{w}_{k.s}$.

It then follows that the component loading of $\mathbf{Q}_k\mathbf{w}_{k.s}$ on $\mathbf{x}_{.s}$ is just the square root of the covariance. In Section 6.1 it was shown how the OVERALS criterion is related to the *sum of the covariances*. We now see that this sum is identical with the sum of *squared* component loadings.

6.3 Interpretation in terms of principal components

The reasoning in Section 6.2 implies that $\mathbf{x}_{.s}$ is a principal component of the standardized variables $\mathbf{Q}_k\mathbf{w}_{k.s}$ (by definition of a principal component: a weighted sum of variables such that its correlations with the variables have stationary value for the sum of squares).

The result is also illustrated in Table 7. This table gives the two 3 x 3 correlation matrices $\mathbf{R}_{.1}$ and $\mathbf{R}_{.2}$ of correlations between the vectors $\mathbf{Q}_k\mathbf{w}_{k.1}$, and $\mathbf{Q}_k\mathbf{w}_{k.2}$, respectively ($k$ = 1,2,3). The matrix $\mathbf{R}_{.s}$ (s = 1,2) has a first eigenvector corresponding to the component loadings of $\mathbf{Q}_k\mathbf{w}_{k.s}$ on $\mathbf{x}_{.s}$, and first eigenvalue equal to ($K\gamma_{.s}$) (sum of the squared component loadings).

However, a warning must be added. In the example, the OVERALS solution corresponds to the *first* principal component of $\mathbf{R}_{.1}$ and of $\mathbf{R}_{.2}$. But if we take subsequent OVERALS dimensions, it must happen that there is a shift. In the example six OVERALS dimensions are possible; the sixth one would be the *worst* OVERALS solution, and would correspond to the third (last) principal component of $\mathbf{R}_{.6}$ with smallest eigenvalue. And for the third, fourth, fifth OVERALS dimension it must happen that one or more of them correspond to the second principal component of $\mathbf{R}_{.s}$ (s = 3,4,5).

7 COMPONENT LOADINGS OF ORIGINAL VARIABLES

Evidently, one also might calculate component loadings for the original variables **Q**. They are found in the correlation matrix

$$\mathbf{Q}'\mathbf{X}/n = \mathbf{Q}'\mathbf{Q}\mathbf{W}\Gamma^{-1}/(Kn) = \mathbf{R}\mathbf{W}\Gamma^{-1}/K = \mathbf{D}_R\mathbf{W}$$

so that for variables within the k^{th} set they are found in the matrix $\mathbf{R}_{kk}\mathbf{W}_k$.

Let this matrix be called $\mathbf{L}_k$. Table 7 gives the numerical results for the example. Figure 4 shows a graph, in which vectors **Q** are plotted on the basis of coordinates given in their component loadings.

For this artificial example it happens that the variables in **Q** fall apart in two groups: $\mathbf{x}_{.1}$ is mainly correlated with ($\mathbf{q}_{12}$ -$\mathbf{q}_{22}$ $\mathbf{q}_{31}$), whereas $\mathbf{x}_{.2}$ has large correlations with ($\mathbf{q}_{11}$ $\mathbf{q}_{21}$ -$\mathbf{q}_{32}$). But this result is just an accidental property of the example. In fact, the OVERALS criterion concentrates on the loadings of the weighted sums $\mathbf{Q}_k\mathbf{W}_k$, and is not at all interested in the loadings of **Q** itself. In theory, it may happen that the best OVERALS dimension shows large loadings for the weighted sums $\mathbf{Q}_k\mathbf{w}_{k.1}$, together with very small loadings for all **Q**. In the analysis of real data such a result is rather unlikely - but if a researcher would find such a result it does not mean that there is something wrong with the program.

Of course, component loadings of original variables **Q** have their use for the substantive interpretation of the object score vectors $\mathbf{x}_{.s}$. To the extent that such loadings are large (or large in the negative direction) it becomes easier to identify what $\mathbf{x}_{.s}$ seems to stand for.

8 AVERAGE RANK ONE QUANTIFICATIONS

8.1 Definition of average rank one quantifications

The j^{th} row of $\mathbf{L}_k$ gives the component loadings of original variable $\mathbf{q}_{kj}$. It is also the j^{th} row of $\mathbf{R}_{kk}\mathbf{W}_k$. We shall now use the notation $\mathbf{l}_{kj}'$ for such a row. It has p cells; they could be indicated by $l_{kj.s}$.

Earlier we have defined $\mathbf{z}_{kj}$ as the column of the k_{kj} category quantifications of the k_{kj} categories in variable $\mathbf{q}_{kj}$. In other words, $\mathbf{z}_{kj}$ gives the standard scores of the categories of $\mathbf{q}_{kj}$.

Average rank-one quantifications now are defined as the elements of the matrix

$$\mathbf{A}_{kj} = \mathbf{z}_{kj}\mathbf{l}_{kj}'.$$

In words: average rank-one category quantifications $\mathbf{a}_{kj.s}$ for variable $\mathbf{q}_{kj}$ are obtained by taking the component loadings of $\mathbf{q}_{kj}$, and by multiplying them by the standard scores of the categories of $\mathbf{q}_{kj}$ (these standard scores are given in $\mathbf{z}_{kj}$).

Results for the numerical example are given in Table 8. In Figure 5 the average rank-one category points are plotted for variable $\mathbf{q}_{22}$. This figure shows that these points are located on a straight line (which justifies why they are called "rank-one" points), where this straight line passes through the point with the component loadings ($\mathbf{l}_{kj}'$ = (-.748 .209), and where this line intersects the edges of the lattice.

(When there are more than two variables in $\mathbf{Q}_k$, the average rank-one category points are located where the line through $\mathbf{l}_{kj}$ intersects (hyper)planes of the more-dimensional lattice.)

8.2 Geometric meaning of average rank-one category points

The geometric meaning of the points $\mathbf{A}_{kj}$ is that, given that they must be located at the intersection of the lattice by a straight line, the average squared distance between object points $\mathbf{X}$ and average rank-one category points $\mathbf{A}_{kj}$ is minimized. This property is illustrated in Figure 5, where it is shown by dotted lines that the 15 object points are (approximately) divided into three "clusters" on the bases of their distance to their corresponding average rank-one category points for category (p q r) of variable $\mathbf{q}_{21}$.

8.3 Average rank-one quantification does not define the OVERALS solution

The optimalization mentioned in Section 8.2 has no direct relation with the OVERALS criterion. Rather, the optimalization in 8.2 depends on a *secondary* optimalization criterion: *given* the OVERALS solution, average rank-one quantification can be optimized, as a sort of "afterthought". It makes it possible to compare the OVERALS

solution with a straightforward PCA solution, where the criterion described in 8.2 is decisive (but where the partitioning of variables in K groups will be ignored).

9 MULTIPLE CATEGORY QUANTIFICATION IN SINGLE OVERALS SOLUTION

9.1 Introduction

The numerical OVERALS solution discussed thus far, is a solution for single quantification. It implies that category points have coordinates $\mathbf{q}_{kj}w_{kj.s}$, and those vectors are all proportional to $\mathbf{q}_{kj}$ itself, so that lattice points become points on a *regular lattice* (shaped as a parallelogram for any two variables within a set).

Given this solution, one may raise the question: would it be possible to replace the regular lattice by a *broken lattice,* based on category points which are *not* located on a straight line? To anticipate upon such a solution, one might have a look at Figure 8, where the regular lattice of Figure 1 has been replaced by a broken lattice for the second set.

A broken lattice can be found in such a way that object points have smaller average squared distance to the points of the broken lattice than to the points of the regular lattice. The reason is simple enough: the broken lattice is not hampered by the restriction that category points must be on straight lines. By taking this restriction away, one can *improve* the solution for loss within the set.

The new loss is called the *multiple loss* per set. *Loss* per set is the average squared distance between object points and lattice points; *single loss* refers to these points of the regular lattice, multiple loss to the points of the broken lattice.

The category quantifications from which the broken lattice can be constructed (category quantifications not necessarily located on a straight line) are called the *multiple category quantifications*.

But note that the numerical OVERALS criterion is interested only in the single loss and ignores multiple loss. In fact, the (single) numerical solution defines the object points on the basis of minimized average squared distance to regular lattice points. *Given* that OVERALS solution for object points we may find a multiple category quantification which results in reduced (multiple) loss per set. Again, the OVERALS criterion is concerned with single category quantification, not with multiple category quantification.

9.2 Centroids and pseudo-centroids

Still taking the numerical OVERALS solution, two new concepts can be introduced. They are called *centroids* and *pseudo-centroids*.

(i) *Centroids* $\mathbf{C}_{kj}$ are defined by taking the averages of rows of $\mathbf{X}$ within a same category of a variable $\mathbf{q}_{kj}$. So $\mathbf{C}_{kj}$ has k_{kj} rows (k_{kj} is the number of categories in $\mathbf{q}_{kj}$), and p columns (p is the number of dimensions of the OVERALS solution).
Algebraically, the expression becomes

$$\mathbf{C}_{kj} = \mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{X}$$

and this implies that

$$\mathbf{G}_{kj}\mathbf{C}_{kj} = \mathbf{G}_{kj}\mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{X}$$

replaces all elements of $\mathbf{X}$ by their category averages.

(ii) Pseudo-centroids $\underline{\mathbf{C}}_{kj}$. They are defined in the same way as $\mathbf{C}_{kj}$, with the difference that the process of averaging is applied to $\mathbf{Q}_k\mathbf{W}_k$:

$$\underline{\mathbf{C}}_{kj} = \mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{Q}_k\mathbf{W}_k.$$

Figure 6 shows the location of centroids and pseudo-centroids for variable $\mathbf{q}_{22}$. Obviously, pseudo-centroid points are located on the sides of the lattice. Centroid-points are not (unless by chance).

Figure 6 also contains the single category points of $\mathbf{q}_{22}$, here with labels (s_p s_q s_r). In general, single category points of $\mathbf{q}_{kj}$ have coordinates in the rows of

$$\mathbf{S}_{kj} = \mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{q}_{kj}\mathbf{w}_{kj}' = \mathbf{z}_{kj}\mathbf{w}_{kj}'.$$

In Figure 6, this implies that single category points (s_p s_q s_r) are located on a straight line, through the points with coordinates

$$\mathbf{w}_{22}' = (-.893\ -.070).$$

Numerical values for centroids and pseudo-centroids for the example, are given in Table 9.

9.3 Multiple category quantification

Coordinates of multiple category points are defined by

$$\mathbf{M}_{kj} = \mathbf{S}_{kj} + \mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj}.$$

Figure 6 shows the geometric meaning of this equation. Multiple category points appear as angular points of parallelograms, of which s_g c_g $\underline{c}_g$ are the other three angular points.

The figure also suggests that lines connecting points m_g and s_g, or c_g and $\underline{c}_g$, are orthogonal to the line on which the single category points are located. This orthogonality will not be generally true for a numerical solution (why it is true for this particular example will be explained in Section 11.5).

Multiple category quantification can be used to construct a *broken lattice*. In the illustration, points of the regular lattice are obtained by adding a row of $\mathbf{S}_{21}$ to a row of $\mathbf{S}_{22}$, with 3 x 3 = 9 possible combinations. Figure 7 shows a *partly* broken lattice, the points of which are obtained by adding rows of $\mathbf{S}_{21}$ to rows of $\mathbf{M}_{22}$ (numerical values of $\mathbf{M}_{kj}$ can be found in Table 10).

Object points **X** will have smaller average squared distance to their corresponding points of the partly broken lattice, than to their points of the regular lattice. This will apply even more so for the fully broken lattice, based on combinations of rows of $\mathbf{M}_{21}$ and $\mathbf{M}_{22}$, and shown in Figure 8.

10 LOSS AND FIT

10.1 Introduction

This chapter recapitulates earlier definitions of loss or fit, and introduces some new ones. All definitions are related to the distinction between:

(i) loss and fit, either *per variable* or *per set*;

(ii) loss and fit, for *single* or for *multiple* quantification.

The terminology in this chapter is somewhat different from that used by Verdegaal (1986). We come back to that in Sections 10.7 and 11.9

10.2 Single loss and fit per set

For set k, the single loss on dimension s is defined by the average squared difference between object $\mathbf{x}_{.s}$ and coordinates $\mathbf{Q}_k\mathbf{w}_{k.s}$ of the lattice points. So, single loss for set k on dimension s scores is defined by

$$SSQ(\mathbf{x}_{.s} - \mathbf{Q}_k\mathbf{w}_{k.s}) = 1 - \mathbf{w}_{k.s}'\mathbf{R}_{kk}\mathbf{w}_{k.s}.$$

Generalizing over all p dimensions ($s = 1,...,p$) one obtains the matrix of *single loss per set:*

$$SSQ(\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k) = \mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k.$$

Averaging over all K sets, one obtains the matrix of *total single loss:*

$$\Sigma(\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k)/K = \mathbf{I} - \mathbf{W}'\mathbf{D}_R\mathbf{W}/K = \mathbf{I} - \Gamma.$$

The OVERALS criterion is to minimize the trace of the latter matrix. This trace corresponds to the average squared distance between object points and their corresponding regular lattice points.
Single fit per set then is defined by the matrix

$$SSQ(\mathbf{Q}_k\mathbf{W}_k) = \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$$

so that matrices of single fit per set and single loss per set add up to **I**. Single fit per set on dimension s is found on the diagonal of this matrix. The trace of this matrix corresponds to the averaged squared distance between the lattice points of set k and the origin.
Total single fit is obtained by averaging over the K sets:

$$\Sigma(\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k)/K = \mathbf{W}'\mathbf{D}_R\mathbf{W}/K = \Gamma$$

so that, again, the objective of the numerical OVERALS solution is to maximize the trace of this diagonal matrix.

10.3 Single fit and loss per variable

Single fit for variable $\mathbf{q}_{kj}$ is defined as

$$SSQ(\mathbf{q}_{kj}\mathbf{w}_{kj}') = \mathbf{w}_{kj}\mathbf{q}_{kj}'\mathbf{q}_{kj}\mathbf{w}_{kj}'/n = \mathbf{w}_{kj}\mathbf{w}_{kj}'$$

which is a matrix of rank one. Verdegaal (1986) uses *discrimination* for single fit per variable. The matrix $\mathbf{w}_{kj}\mathbf{w}_{kj}'$ has trace $\mathbf{w}_{kj}'\mathbf{w}_{kj}$, and this value corresponds geometrically to the average squared distance of single category points to the origin.

Added over all variables within the k^{th} set, the single fit values add up to $\mathbf{W}_k\mathbf{W}_k'$, which is *not* equal to the single fit per set (equal to $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$) unless $\mathbf{R}_{kk} = \mathbf{I}$. It may be useful to point out that $\mathbf{w}_{kj}'$ is meant as a row of $\mathbf{W}_k$ (not a row of $\mathbf{W}'$).

Single loss per variable has thus far not been defined in OVERALS literature. Neither shall we introduce a definition here. One might take $\mathbf{I} - \mathbf{w}_{kj}\mathbf{w}_{kj}'$, but this seems rather superfluous.

10.4 Dispersion

In Verdegaal (1986) the concept of *dispersion* for $\mathbf{q}_{kj}$ is introduced as the sum of single fit (discrimination) of $\mathbf{q}_{kj}$ and single loss of set *k:*

$$\text{dispersion of } \mathbf{q}_{kj} = (\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k') + \mathbf{w}_{kj}\mathbf{w}_{kj}'.$$

The dispersion matrix gives the SSQ of the difference matrix

$$\mathbf{X} - (\mathbf{Q}_k\mathbf{W}_k - \mathbf{q}_{kj}\mathbf{w}_{kj}')$$

which shows that in the form between brackets the contribution of $\mathbf{q}_{kj}\mathbf{w}_{kj}'$ is omitted from $\mathbf{Q}_k\mathbf{W}_k$. Geometrically, this matrix shows differences between object points and the points of a "reduced" lattice (reduced: because it is a lattice based on the category quantifications of all variables in $\mathbf{Q}_k$, except variable $\mathbf{q}_{kj}$). Dispersion therefore becomes the same as loss per set with respect to the reduced lattice. This loss will be larger than the loss for the complete lattice: in fact, the single fit of $\mathbf{q}_{kj}$ has been changed into loss.

Figure 9A shows the relations in a graph where the *n*-dimensional columns appear as line-vectors. One basic property of the graph is that the difference vector $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k)$ is orthogonal to the plane spanned by $\mathbf{Q}_k\mathbf{w}_k$ and $\mathbf{q}_{kj}\mathbf{w}_{kj}$. The reason is that the basic equation of OVERALS (equation 4 in Section 4.2) implies:

$$\mathbf{Q}_k'(\mathbf{x} - \mathbf{Q}_k\mathbf{W}_k) = \mathbf{Q}_k'(\mathbf{QW}/K\gamma - \mathbf{Q}_k\mathbf{W}_k) = \mathbf{Q}_k'\mathbf{Q}_k\mathbf{W}_k - \mathbf{Q}_k'\mathbf{Q}_k\mathbf{W}_k = \mathbf{0}.$$

In Figure 9A the squared length of $\mathbf{q}_{kj}\mathbf{w}_{kj}$ is equal to the single fit of this variable. The squared length of $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k)$ is the single loss per set. The squared length of $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k + \mathbf{q}_{kj}\mathbf{w}_{kj})$ is the dispersion. The latter vector appears as a diagonal of the rectangle with sides $\mathbf{q}_{kj}\mathbf{w}_{kj}$ and $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k)$. It follows that dispersion is the sum of single fit of $\mathbf{q}_{kj}$ and single loss of the set.

10.5 Multiple fit and loss per variable

Multiple fit for variable $\mathbf{q}_{kj}$ is defined by

$$\mathrm{SSQ}(\mathbf{G}_{kj}\mathbf{M}_{kj}) = \mathbf{M}_{kj}'\mathbf{D}_{kj}\mathbf{M}_{kj}/n.$$

To define multiple loss per variable is slightly more complicated. Let $\mathbf{S}_{kj}$ be the single category quantification such that $\mathbf{q}_{kj}\mathbf{w}_{kj}' = \mathbf{G}_{kj}\mathbf{S}_{kj}$. Replacement of single category coordinates by multiple coordinates requires

$$\begin{aligned}\mathbf{Q}_k\mathbf{W}_k - \mathbf{G}_{kj}\mathbf{S}_{kj} + \mathbf{G}_{kj}\mathbf{M}_{kj} &= \mathbf{Q}_k\mathbf{W}_k - \mathbf{G}_{kj}(\mathbf{S}_{kj} - \mathbf{M}_{kj}) \\ &= \mathbf{Q}_k\mathbf{W}_k + \mathbf{G}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj}).\end{aligned}$$

Such a replacement is illustrated in a comparison of Figures 6 and 7.
Multiple loss then will be defined by the SSQ of the difference matrix:

$$\mathbf{X} - (\mathbf{Q}_k\mathbf{W}_k + \mathbf{G}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj}))$$

with SSQ:

$$(\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k)'(\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k)/n - (\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})/n.$$

The first term in this expression is the single loss of set k, equal to $(\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k)$. The second term depends entirely on the squared differences between centroids and pseudo-centroids. This second term therefore shows the difference between single loss per set and multiple loss per variable. So, multiple loss per variable never can be larger than single loss per set. The second term is obviously equal to

$$SSQ(\mathbf{G}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj}))$$

and will be called the *relative loss* of variable $\mathbf{q}_{kj}$.

Relative loss also is equal to the difference of multiple and single loss per variable. To show this, write $\mathbf{M}_{kj} = \mathbf{S}_{kj} + \mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})$, and the expression for multiple fit per variable becomes the SSQ of the latter vector, equal to

$$\mathbf{S}_{kj}'\mathbf{D}_{kj}\mathbf{S}_{kj}/n + (\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})/n.$$

The simplification used here depends on the fact that $\mathbf{S}_{kj}'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj}) = 0$ (this can be derived from equation (4) in Section 4.2). Moreover, we have:

$$\mathbf{S}_{kj}'\mathbf{D}_{kj}\mathbf{S}_{kj}/n = \mathbf{S}_{kj}'\mathbf{G}_{kj}'\mathbf{G}_{kj}\mathbf{S}_{kj}/n = \mathbf{w}_{kj}\mathbf{q}_{kj}'\mathbf{q}_{kj}\mathbf{w}_{kj}/n = \mathbf{w}_{kj}\mathbf{w}_{kj}'$$

which is the expression for single fit per variable. The conclusion is that for each variable: multiple fit = single fit + relative loss. This agrees with the fact that relative fit is

$$SSQ(\mathbf{G}_{kj}(\mathbf{C}_{kj} - \mathbf{C}_{kj})) = SSQ(\mathbf{G}_{kj}(\mathbf{M}_{kj} - \mathbf{S}_{kj})).$$

Verdegaal (1986) uses the term *single loss* in stead of relative loss. This is a bit confusing perhaps. But remember that we have left the term "single loss per variable" undefined and it therefore is perfectly legitimate to define this term in the way Verdegaal does. Relative loss indicates what we loose in fit if a multiple quantification is replaced by a single quantification. Or, the other way round: it shows what we gain if the single quantification is replaced by a multiple quantification. The latter phrasing is perhaps the better one, because relative loss is calculated by the OVERALS program for variables which are treated as single (not for variables which are treated as multiple). The amount of relative loss then gives a hint about how much gain can be expected if the variables were treated as multiple nominal instead of single.

Figure 9B is an attempt to visualize the results. The figure is the same as Figure 9A, but with new vectors added to it. Figure 9A is a picture of a three dimensional vector constellation. The difficulty becomes that in Figure 9B the vector $\mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj})$ is not contained in the three dimensions of Figure 9A: this vector requires a fourth dimension. What we do know, however, is that $\mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj})$ is orthogonal to $\mathbf{G}_{kj}\mathbf{s}_{kj} = \mathbf{q}_{kj}\mathbf{w}_{kj}$. What we also know is that the projection of $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k)$ on $\mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj})$ is identical to $\mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj})$ itself. It follows that $\mathbf{G}_{kj}\mathbf{s}_{kj}$ is orthogonal to the plane spanned by $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k)$ and $\mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj})$, and therefore also orthogonal to the vector $(\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k - \mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj}))$.

It is useful to make a list that indicates for each vector the meaning of its squared length (corresponding to the SSQ of the vector).

$\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k$	single loss per set
$\mathbf{G}_{kj}\mathbf{s}_{kj} = \mathbf{q}_k\mathbf{w}_k$	single fit per variable
$\mathbf{Q}_k\mathbf{w}_k$	single fit per set
$\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k + \mathbf{G}_{kj}\mathbf{s}_{kj}$	dispersion
$\mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj}) = \mathbf{G}_{kj}(\mathbf{c}_{kj} - \mathbf{c}_{kj})$	relative loss per variable
$\mathbf{x} - \mathbf{Q}_k\mathbf{w}_k - \mathbf{G}_{kj}(\mathbf{m}_{kj} - \mathbf{s}_{kj})$	multiple loss per variable
$\mathbf{G}_{kj}\mathbf{m}_{kj}$	multiple fit per variable.

Figure 9B also shows that

dispersion	=	multiple fit per variable + multiple loss per variable.
	=	single fit per variable + single loss per set.

10.6 Multiple fit and loss per set

In the OVERALS program the terms multiple fit or multiple loss per set are not used. Obviously, multiple fit per set could be defined as

$$SSQ(\mathbf{G}_k\mathbf{M}_k) = \mathbf{M}_k'\mathbf{G}_k'\mathbf{G}_k\mathbf{M}_k/n$$

where $\mathbf{G}_k\mathbf{M}_k = \Sigma\mathbf{G}_{kj}\mathbf{M}_{kj}$ ($j = 1,...,m_k$) whereas $\mathbf{M}_k$ is the (vertical concatenation of all m_k matrices $\mathbf{M}_{kj}$. Note that $\mathbf{G}_k'\mathbf{G}_k$ is not a diagonal matrix.
Although we can substitute

$$\mathbf{G}_k\mathbf{M}_k = \mathbf{G}_k(\mathbf{S}_k + \mathbf{C}_k - \underline{\mathbf{C}}_k) = \mathbf{Q}_k\mathbf{W}_k + \mathbf{G}_k(\mathbf{C}_k - \underline{\mathbf{C}}_k)$$

this does not lead to a simplification of the SSQ in terms of additive components.

Similarly multiple loss per set could be defined as

$$SSQ(\mathbf{X} - \mathbf{G}_k\mathbf{M}_k) = SSQ((\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k) - \mathbf{G}_k(\mathbf{C}_k - \underline{\mathbf{C}}_k))$$

where again a further simplification appears impossible.

10.7 Summary of results for numerical OVERALS

The results obtained thus far, are summarized in Table 11.

In this table the first column gives the name of a vector (where the vector has a specific name; sometimes the vector has no name and the first columns then is left open). The name in the first column is underlined if the result is printed by the OVERALS program. In those case where Verdegaal (1986) uses a different name than in this monograph, her naming is added between parentheses.

The second column gives the symbol, in such a way that in this notation there will be $\mathbf{k}_{kj}$ rows ($\mathbf{k}_{kj}$ is the number of categories in variable j of set k).

The third column gives the notation with n rows. The relation between third and second column is quite simple: to obtain the symbol in the third column, one should multiply the symbol in the second column by $\mathbf{G}_{kj}$.

E.g.: $\mathbf{q}_{kj} = \mathbf{G}_{kj}\mathbf{z}_{kj}$; $\mathbf{q}_{kj}\mathbf{w}_{kj}' = \mathbf{G}_{kj}\mathbf{S}_{kj}$; etc.

The fourth column gives the name of the SSQ matrix of the symbol in the third column, if such an SSQ matrix has a special name. Again this name is underlined if the SSQ matrix is printed by the OVERALS program, and if Verdegaal uses a different name than in this monograph, her naming is added between parentheses.

The fifth column gives the symbol for the SSQ matrix. Sometimes particular equalities are mentioned in this fifth column.

As an anticipation on Chapter 11 it can be said that Table 11 remains valid for an OVERALS solution in which all variables are treated as *single* (single numerical, single ordinal, or single nominal).

11 OVERALS AND OPTIMAL CATEGORY QUANTIFICATION

11.1 Introduction

Thus far OVERALS has been described in this monograph as a solution based on a priori quantification of the categories of each variables. This was called the "numerical" OVERALS solution. The criterion of this solution aims exclusively on finding optimal *weights* **W**, such that by this choice the OVERALS criterion is optimized. OVERALS then becomes just a classical solution for linear relations between K sets of a priori quantified variables (such as discussed in Van de Geer (1984) or Van de Geer (1986), under the name "analysis of $\mathbf{P\Phi}$, with $\mathbf{t't} = K$, and MAXBET criterion").

However, OVERALS is primarily meant to be a computer program in which the a priori quantification is not decisive. This means that the OVERALS program not only may give *weights,* but also an *optimal quantification of the categories* of each variable, in such a way that after this optimal quantification the OVERALS criterion becomes better (smaller loss, better fit) than on the basis of the a priori quantification.

11.2 Overview of OVERALS options

The basic feature of the OVERALS program is that the user may decide for each variable separately (as was the case in PRINCALS) how the variable must be treated. There are four possibilities.

(i) *Single numerical.* Category quantifications are given a priori. For such variables, OVERALS can only find optimal weights.

(ii) *Single ordinal.* The a priori quantification may be replaced by a different category quantification, but this new quantification should have *the same order* as the a priori quantification.

(iii) *Single nominal.* The a priori quantification may be replaced by any new category quantification; there are no restrictions.

(iv) *Multiple nominal.* Whereas in the previous three approaches it is required that category quantifications per dimension of the solution are proportional to each other, it is now allowed that category quantifications per dimension need not to be proportional to category quantification on any other dimension.

11.3 Numerical solution

The numerical solution has been discussed in the previous chapters. The only thing we want to add is that the numerical OVERALS solution is *nested.* I.e., a solution in p dimensions is identical to the result for the first p dimension in a solution with more than p dimensions.

11.4 Single nominal solutions

The single nominal solution is *not nested.* It gives optimal single category quantifications in such a way that the sum of the first p eigenvalues is maximized; the optimal quantification may change if one takes a different value for p.

When the solution for category quantification produced by single nominal treatment of all variables is taken as if it were the a priori quantification, then the result of a numerical OVERALS on the the variables thus quantified is identical to the single nominal solution.

In fact, the illustration of numerical OVERALS, used in the previous Chapter 5 to 10, takes the category quantification of single nominal OVERALS as "a priori quantification". So, if we had started with a different a priori quantification, and had asked for a single nominal OVERALS solution in $p = 2$ dimensions, the quantification used in the example would be found as the optimal one. The example, therefore, serves also an example of single nominal OVERALS.

The question arises whether the example has properties which are valid for the single nominal solution, but not for the numerical solution in general? There is at least one such property. It is found in Figure 6, where the dotted line connecting centroids and pseudo-centroids are orthogonal to the line on which the single category points are located. This property can algebraically be formulated as $(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})\mathbf{w}_{kj} = 0$ (see Section 9.3).

Suppose we have a numerical solution where this property is not valid. It then follows immediately that we could find another line, such that multiple category points $\mathbf{M}_{kj}$ are closer to their projection on this new line than to the original single category points. We then obtain a better single fit if we take those projections as the "improved" single category points. However, such an improvement may not be possible in a numerical solution, because it may violate the a priori category quantifications. But in a single solution there is no restriction, and the improvement is permissible. The property of Figure 6 (lines connecting points c_g and $\underline{c}_g$, or m_g and s_g, are orthogonal to the line on which the single category points s_g are located) is valid for a single solution, but not necessarily for a numerical solution.

The property above can be formulated as $(\mathbf{M}_{kj} - \mathbf{S}_{kj})\mathbf{w}_{kj} = 0$, or $(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})\mathbf{w} = 0$. It has the implication that $\mathbf{w}_{kj}$ must be an eigenvector of SSQ($\mathbf{M}_{kj}$).
In fact, we have seen in Section 11.5 that

$$\text{SSQ}(\mathbf{M}_{kj}) = \mathbf{w}_{kj}\mathbf{w}_{kj}' + (\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})/n$$

from which it follows that

$$\text{SSQ}(\mathbf{M}_{kj}) = \mathbf{M}_{kj}'\mathbf{D}_{kj}\mathbf{M}_{kj}\mathbf{w}_{kj}/n = \mathbf{w}_{kj}\mathbf{w}_{kj}'\mathbf{w}_{kj}$$

so that $\mathbf{w}_{kj}$ is an eigenvector and $\mathbf{w}_{kj}'\mathbf{w}_{kj}$ (equal to the trace of the matrix of single fit) the corresponding eigenvalue. For the numerical example, illustrated in Figure 6, SSQ($\mathbf{M}_{22}$) has eigenvector $\mathbf{w}_{22}'$ = (-.893 -.070) and eigenvalue $\mathbf{w}_{22}'\mathbf{w}_{22}$ = .802. This example also shows that the remaining eigenvalues of $(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})/n$ add up to the trace of the "relative loss" matrix (equal to .087, in the example). The example however has the special property that the latter matrix also has rank one (because there are only three categories in $\mathbf{q}_{22}$).

The reasoning above suggests a possible algorithm to "improve" upon a numerical solution, until it becomes a single nominal solution. The receipt would be: for any variable, starting with the numerical solution for single category points, replace them by projections of the multiple category on their best eigenvector. Do this in turn for all variables. Repeat the cycle until results become stabilized.

11.5 Single ordinal solution

The single ordinal solution is *not* nested. Such a solution is restricted by the requirement that optimal category quantifications should follow the same order as the a priori quantification.

It may happen that the single nominal solution has this required order. In that case there is no difference between the single ordinal and the single nominal solution. But it also may happen that the nominal solution does not have the same order. In that case things very often can be remedied by *grouping* of adjacent categories: where adjacent categories have a "wrong" order, group them together as if they are the same category. After grouping, the ordinal and nominal OVERALS single solution tends to become identical.

11.6 Multiple nominal solution

The first basic point is that this solution gives category quantifications which may be *different* for each dimension. In other words, the solution defines *broken* lattices in such a way that the average squared distances between objects points **X** and their corresponding points on the broken lattices is minimized.

In a solution where all variables are treated as single, points of the regular lattice are found in the rows of

$$\mathbf{G}_k\mathbf{S}_k = \Sigma\mathbf{G}_{kj}\mathbf{S}_{kj} \ (j = 1,...m_k)$$

(on the understanding that rows of $\mathbf{G}_k\mathbf{S}_k$ do not specify lattice points for category combinations that do not occur in the data). The solution requires that $\mathbf{S}_{kj}$ has rank one.

The solution with all variables treated as multiple, gives for each set a broken lattice, specified in the rows of $\mathbf{G}_k\mathbf{M}_k$. Submatrices $\mathbf{M}_{kj}$ now are no longer restricted to rank one. (But their rank has upper-bound $(k_{kj} - 1)$, where k_{kj} is the number of categories in the variable. In particular, if the variable is binary, so that $k_{kj} = 2$, then $\mathbf{M}_{kj}$ will have rank one, and there is no difference between $\mathbf{S}_{kj}$ and $\mathbf{M}_{kj}$.)

The second basic point relates a multiple solution to a possible numerical solution. The multiple category quantification on dimension s is found in the column $\mathbf{m}_{kj.s}$. Suppose we take this quantification *as if* it were an a priori quantification; i.e., define $\mathbf{z}_{kj} = \mathbf{m}_{kj.s}$. Perform a numerical OVERALS on those values $\mathbf{z}_{kj}$. Then this numerical solution will have some dimension with the same result as dimension s of the multiple solution.

To illustrate, let $p = 1$. The multiple solution then becomes identical to the single nominal solution with $p = 1$. If the category quantifications of this solution are treated as if they are a priori quantifications $\mathbf{z}_{kj}$, the first dimensions of this numerical solution will be identical to the one-dimensional ($p = 1$) single nominal solution, and also to the one-dimensional multiple nominal solution.

Or, take $p = 2$. Define $\mathbf{z}_{kj} = \mathbf{m}_{kj}$ and perform a numerical analysis on the basis of $\mathbf{z}_{kj}$. The first dimension will be identical to the first dimension of the multiple solution. However, define $\mathbf{z}_{kj} = \mathbf{m}_{kj.2}$. Now the numerical solution based on this choice of $\mathbf{z}_{kj}$ will also show up with a dimension identical to the second multiple solution, but this dimension may be the first one of the numerical solution, or the second.

In general, take $\mathbf{z}_{kj} = \mathbf{m}_{kj.s}$. The numerical analysis on the basis of $\mathbf{z}_{kj}$ will have *some* dimension identical to dimension s of the multiple solution. Compare Section 6.3, where a similar result was found for the principal components of standardized variables $\mathbf{Q}_k\mathbf{w}_{k.s}$.

The third point is that the multiple nominal solution is nested. The number of solutions has upper bound $\Sigma(k_{kj} - 1) = \Sigma k_{kj} - K$. A numerical analysis based on $\mathbf{z}_{kj} = \mathbf{m}_{kj.s}$ has at

most m solutions. In general, m will be much smaller than $\Sigma(k_{kj} - 1)$. (Unless all variables are binary: in that case the multiple nominal solution is the same as a single nominal solution).

The fourth point is more technical. It says that for a given dimension s centroids $\mathbf{c}_{kj.s}$ must be identical to pseudo-centroids $\underline{\mathbf{c}}_{kj.s}$. In what follows, we shall omit the index s for dimension.

Centroids are defined by $\mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{x} = \mathbf{c}_{kj}$. Pseudo-centroids are $\mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{G}_k\mathbf{m}_k$. Their equality implies:

$$\mathbf{G}_{kj}'\mathbf{x} = \mathbf{G}_{kj}'\mathbf{G}_k\mathbf{m}_k. \tag{18}$$

The basic equation (4) of Section 4.2 implies

$$\mathbf{q}_{kj}'\mathbf{Q}\mathbf{w} = \mathbf{q}_{kj}'\mathbf{Q}_k\mathbf{w}_k \ (K\gamma). \tag{19}$$

$\mathbf{Q}_{kj}\mathbf{w}_{kj}$ is the expression for the variable with single category quantification, and we may write $\mathbf{Q}_{kj}\mathbf{w}_{kj} = \mathbf{G}_{kj}\mathbf{s}_{kj}$. In addition, we have the definition $\mathbf{x} = \mathbf{Q}\mathbf{w}/K\gamma = \mathbf{G}\mathbf{s}/K\gamma$. This makes it possible to re-write (19) as

$$\mathbf{z}_{kj}'\mathbf{G}_{kj}'\mathbf{G}\mathbf{s} = \mathbf{z}_{kj}'\mathbf{G}_{kj}'\mathbf{G}_k\mathbf{s}_k \ (K\gamma) \tag{20}$$

or

$$\mathbf{z}_{kj}'\mathbf{G}_{kj}'\mathbf{x} = \mathbf{z}_{kj}'\mathbf{G}_{kj}'\mathbf{G}_k\mathbf{s}_k. \tag{21}$$

Equation (20) is valid for a single solution and depends upon the category quantification $\mathbf{z}_{kj}$. But in a multiple solution there is no such restriction, and $\mathbf{z}_{kj}$ can be omitted. The equation becomes:

$$\mathbf{G}_{kj}'\mathbf{G}\mathbf{m} = \mathbf{G}_{kj}'\mathbf{G}_k\mathbf{m}_k \ (K\gamma), \tag{22}$$

which shows that in the multiple solution a single variable $\mathbf{q}_{kj}$ is replaced by the k_{kj} binary variables of the indicator matrix $\mathbf{G}_{kj}$. So the multiple solution is actually a numerical solution, applied to the binary variables in $\mathbf{G}$.
In the same spirit, equation (21), or equation (22), can be re-written as

$$\mathbf{G}_{kj}'\mathbf{x} = \mathbf{G}_{kj}'\mathbf{G}_k\mathbf{m}_m. \tag{23}$$

Pre-multiplication of both sides in equation (23) by $\mathbf{D}_{kj}^{-1}$ then shows that $\mathbf{c}_{kj} = \underline{\mathbf{c}}_{kj}$.

The fifth point is about an improper single solution, given the proper multiple solution. We have seen that when a variable is treated as single, the OVERALS program also gives multiple category quantifications. They play no role in the OVERALS criterion, and therefore we could say that the single category quantifications are proper, whereas the multiple category quantifications are improper.

Now the other way round: if a variable is treated as multiple, the OVERALS criterion takes this quantification into account, and we may say that the multiple category quantification is proper. An improper single category quantification then would be obtained by using the same argument as in Section 11.4: define the improper single quantification by taking the first eigenvector of SSQ($\mathbf{M}_{kj}$). However, the OVERALS program does not give such an improper single solution for variables treated as multiple.

11.7 OVERALS output for variables treated as multiple

The OVERALS program has the following output for a variable treated as multiple.

Firstly, the *multiple category quantifications* are given in a table $\mathbf{M}_{kj}$ with k_{kj} rows and p columns. In addition, the table $\mathbf{C}_{kj}$ of *centroids* is given, where $\mathbf{C}_{kj} = \mathbf{D}_{kj}^{-1}\mathbf{G}_{kj}'\mathbf{x}$.

Secondly, the programs gives the matrix of multiple fit, called the *discrimination* matrix by Verdegaal (1986), and defined by SSQ($\mathbf{M}_{kj}$) = $\mathbf{M}_{kj}\mathbf{D}_{kj}\mathbf{M}_{kj}/n$.

In addition, the *dispersion* matrix is displayed. When all variables are treated as single, dispersion is defined by SSQ($\mathbf{x} - \mathbf{G}_k\mathbf{S}_k + \mathbf{G}_{kj}\mathbf{S}_{kj}$). But when all variables are treated as multiple, dispersion must be defined as SSQ($\mathbf{X} - \mathbf{G}_k\mathbf{M}_k + \mathbf{G}_{kj}\mathbf{M}_{kj}$), and this matrix becomes equal to the sum of SSQ($\mathbf{X} - \mathbf{G}_k\mathbf{M}_k$) and SSQ($\mathbf{M}_{kj}$). In words: dispersion is the sum of multiple loss per set and multiple fit per variable.

Multiple loss per set is defined as SSQ($\mathbf{X} - \mathbf{G}_k\mathbf{M}_k$), and is called total loss by Verdegaal. The matrix ($\mathbf{X} - \mathbf{G}_k\mathbf{M}_k$) contains the differences between object points $\mathbf{X}$ and the points of the broken lattice $\mathbf{G}_k\mathbf{M}_k$.

11.8 Mixed OVERALS solution

One of the attractive features of the OVERALS program is that the user can decide for each variable separately whether it should be treated as numerical, single ordinal, single nominal, or multiple nominal. If variables are treated in different ways, we obtain a *mixed* OVERALS solution. Such a solution is nested as long as all variables are treated as either numerical and/or multiple nominal. Once there are one or more variables treated as single

ordinal or single nominal, the solution no longer is nested, and will depend on the value of p (number of dimensions).

In principle, the OVERALS criterion remains simple enough: it depends on differences (distances) between object points and lattice points. When all variables are treated as single, regular lattice points are found in rows of $\mathbf{G}_k\mathbf{S}_k$, and when all variables are treated as multiple, broken lattice points are found in rows of $\mathbf{G}_k\mathbf{M}_k$.

The main problem now is to find a coherent *notation*. We suggest the notation $\mathbf{B}_{kj}$ for the category quantifications, where $\mathbf{B}_{kj} = \mathbf{S}_{kj}$ if a variable is treated as single, and where $\mathbf{B}_{kj} = \mathbf{M}_{kj}$ if a variable is treated as multiple. Lattice points then are given in the rows of $\mathbf{G}_k\mathbf{B}_k$, and they may be points of a regular lattice (if all variables within set k are treated as single), or of a partly broken lattice (if some variables within the sets are treated as single and the other variables as multiple), or of a completely broken lattice (if all variables within the set are treated as multiple).

We now also see the advantage of Verdegaal's (1986) usage of the terms *discrimination* and *dispersion*. Discrimination for variable $\mathbf{q}_{kj}$ is defined as $\mathrm{SSQ}(\mathbf{G}_{kj}\mathbf{B}_{kj})$, and is equal to single fit if the variable is treated as single, or to multiple fit if the variable is treated as multiple. Dispersion is defined as $\mathrm{SSQ}(\mathbf{X} - \mathbf{G}_k\mathbf{B}_k + \mathbf{G}_{kj}\mathbf{B}_{kj})$. Loss per set is defined as $\mathrm{SSQ}(\mathbf{X} - \mathbf{G}_k\mathbf{B}_k)$, by Verdegaal called the "total loss". It may be single loss per set (if all variables are treated as single, and $\mathbf{B}_{kj} = \mathbf{S}_{kj}$), or multiple loss per set (if all variables are treated as multiple, and $\mathbf{B}_{kj} = \mathbf{M}_{kj}$), or mixed loss per set (if some variables within the set are treated as single, and the other variables as multiple).

The OVERALS criterion now can be defined as minimizing the total loss defined by the trace of the diagonal matrix

$$\Sigma\mathrm{SSQ}(\mathbf{X} - \mathbf{G}\,\mathbf{B})/K = \mathbf{I} - \mathbf{G} \quad (k = 1,\ldots,K)$$

which becomes equivalent to maximizing the total fit, or discrimination, equal to the trace of

$$\Sigma\mathrm{SSQ}(\mathbf{G}_k\mathbf{B}_k)/K = \Gamma.$$

Object scores are defined by

$$\mathbf{X} = \mathbf{G}\mathbf{B}\Gamma^{-1}/K$$

and have the property that $\mathrm{SSQ}(\mathbf{X}) = \mathbf{I}$.

11.9 Summary

An attempt to summarize results is shown in Table 12. This table gives in its second column the name used by Verdegaal (1986). The third column gives the corresponding name used in this monograph, and the last column gives the algebraic expression. The upper part of the table is valid irrespective of whether a variable $\mathbf{q}_{kj}$ is treated as single or multiple. The lower half of the table refers to variables treated as single, and where $\underline{\mathbf{M}}_{kj}$ is the "improper" multiple category quantification.

REFERENCES

Van de Geer, J.P. (1984). Linear relations among k sets of variables. *Psychometrika,* 49, 79-94.

Van de Geer, J.P. (1986). *Introduction to linear multivariate data analysis*. Leiden: DSWO Press.

Van de Geer, J.P. (1987). *Analysis of linear relations among categorical variables*. Report RR-87-10, Dept. of Data Theory, University of Leiden.

Van der Burg, E., J. de Leeuw, and R. Verdegaal (1984). *Non-linear canonical correlation with m sets of variables*. Report RR-84-12, Dept. of Data Theory, University of Leiden.

Verdegaal, R. (1986). *OVERALS*. Report UG-86-01, Dept. of Data Theory, University of Leiden.

TABLE 1

Example with K=3 sets, m_k=2 variables in each set, k_{kj}=3 categories in each variable, n=15 objects.

		SET 1		SET 2		SET 3	
		q_{11}	q_{12}	q_{21}	q_{22}	q_{31}	q_{32}
	1	a	a	p	p	u	u
	2	a	b	p	q	v	v
	3	a	c	p	q	w	w
	4	b	a	q	p	u	u
	5	b	a	r	p	u	v
	6	b	b	p	r	v	u
	7	b	c	q	p	w	v
objects	8	c	a	p	p	u	w
	9	c	b	p	r	v	v
	10	c	b	q	p	v	w
	11	c	b	q	r	w	v
	12	c	c	q	q	w	u
	13	c	c	p	q	u	w
	14	c	c	r	p	v	u
	15	c	c	r	q	w	v

TABLE 2

Category quantifications z_{kj}

	q_{11}	q_{12}		q_{21}	q_{22}		q_{31}	q_{32}
a	-1.664	-1.569	p	-1.058	.970	u	-1.406	-.401
b	1.293	.984	q	.774	-.403	v	.832	-.759
c	-.022	.226	r	1.178	-1.592	w	.574	1.639

TABLE 3

Weights **W**

		dimension 1	dimension 2
	w_{11}	.148	.772
	w_{12}	.956	-.149
	w_{21}	.423	.818
	w_{22}	-.893	-.070
	w_{31}	.885	-.268
	w_{32}	-.089	-.715
eigenvalue		.815	.592

TABLE 4

Weighted variables $q_{2j}w_{2j.s}$ for second set and their sum $Q_2w_{2.s}$.
Last row gives averaged square.

$q_{21}w_{21.1}$	$q_{21}w_{21.2}$	$q_{22}w_{22.1}$	$q_{22}w_{22.2}$	$Q_2w_{2.1}$	$Q_2w_{2.2}$
-.448	-.866	-.866	-.068	-1.314	-.934
-.448	-.866	.360	.028	-.088	-.838
-.448	-.866	.360	.028	-.088	-.838
.328	.634	-.866	-.068	-.538	.565
.499	.964	-.866	-.068	-.367	.896
-.448	-.866	1.421	.111	.973	-.755
.328	.634	-.866	-.068	-.538	.565
-.448	-.866	-.866	-.068	-1.314	-.934
-.448	-.866	1.421	.111	.973	-.755
.328	.634	-.866	-.068	-.538	.565
.328	.634	1.421	.111	1.749	.745
.328	.634	.360	.028	.688	.662
-.448	-.866	.360	.028	-.088	-.838
.499	.964	-.866	-.068	-.367	.896
.499	.964	.360	.028	.859	.992
.179	.670	.797	.005	.718	.636

TABLE 5

Weighted sum per set $Q_k w_{k.s}$, their totals $Qw_{.k}$ and the standardized version x of these totals. Last row gives averaged square.

DIMENSION 1

$Q_1w_{1.1}$	$Q_2w_{2.1}$	$Q_3w_{3.1}$	$Qw_{.1}$	$x_{.1}$
-1.746	-1.314	-1.208	-4.268	-1.747
.694	-.088	.804	1.410	.576
-.030	-.088	.362	.244	.103
-1.309	-.538	-1.208	-3.055	-1.250
-1.309	-.367	-1.176	-2.852	-1.168
1.131	.973	.772	2.876	1.173
.407	-.538	.576	.445	.184
-1.503	-1.314	-1.390	-4.207	-1.722
.937	.973	.804	2.714	1.107
.937	-.538	.590	.989	.404
.937	1.749	.576	3.262	1.333
.213	.618	.544	1.445	.594
.213	-.088	-1.390	-1.166	-.516
.213	-.367	.772	.618	.253
.213	.859	.576	1.648	.676
.899	.718	.828	5.978	1.000

DIMENSION 2

$Q_1w_{1.2}$	$Q_2w_{2.2}$	$Q_3w_{3.2}$	$Qw_{.2}$	$x_{.2}$
-1.051	-.934	.663	-1.322	-.744
-1.431	-.838	.319	-1.950	-1.098
-1.319	-.838	-1.325	-3.482	-1.962
1.232	.565	.663	2.460	1.387
1.232	.896	.919	3.047	1.716
.852	-.755	.063	.160	.091
.964	.565	.388	1.917	1.082
.217	-.934	-.794	-1.511	-.852
-.163	-.755	.319	-.599	-.338
-.163	.565	-1.394	-.992	-.557
-.163	.745	.388	.970	.545
-.051	.662	.132	.743	.419
-.051	-.838	-.794	-1.683	-.949
-.051	.896	.063	.908	.512
-.051	.992	.388	1.329	.748
.648	.636	.491	3.154	1.000

TABLE 6

Matrices $\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$ (single fit per set) and their sum matrix (total single fit) $\Sigma\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k = \mathbf{W}'\mathbf{R}\mathbf{W} = K\Gamma$.

$\mathbf{W}_1'\mathbf{R}_{11}\mathbf{W}_1$	.899	-.121
	-.121	.648
$\mathbf{W}_2'\mathbf{R}_{22}\mathbf{W}_2$	.718	.149
	.149	.636
$\mathbf{W}_3'\mathbf{R}_{33}\mathbf{W}_3$	.828	-.029
	-.029	.491
	2.445	0
$\mathbf{W}'\mathbf{R}\mathbf{W} = K\Gamma$		
	0	1.775

TABLE 7

Correlations between $\mathbf{Q}_k\mathbf{W}_k$ and $\mathbf{X}$ (component loadings of $\mathbf{Q}_k\mathbf{W}_k$), between $\mathbf{Q}$ and $\mathbf{X}$ (component loadings of $\mathbf{Q}$), and correlations matrices $\mathbf{R}_{.s}$ between vectors $\mathbf{Q}_k\mathbf{w}_{k.s}$.

	$\mathbf{x}_{.1}$	$\mathbf{x}_{.2}$			$\mathbf{x}_{.1}$	$\mathbf{x}_{.2}$
$\mathbf{Q}_1\mathbf{W}_1$	.948	-.128		$\mathbf{Q}_1$	.024	.791
	-.150	.805			.937	-.249
$\mathbf{Q}_2\mathbf{W}_2$	.847	.176		$\mathbf{Q}_2$	.119	.795
	.187	.797			-.748	.209
$\mathbf{Q}_3\mathbf{W}_3$	.910	-.031		$\mathbf{Q}_3$	.906	-.098
	-.041	.701			-.300	-.651
	$\mathbf{R}_{.1}$				$\mathbf{R}_{.2}$	
1	.710	.844		1	.479	.345
.710	1	.606		.479	1	.333
.844	.606	1		.345	.333	1
.948	.847	.910	eigenvector	.805	.797	.701
	2.445		eigenvalue		1.776	

TABLE 8

Average rank-one quantifications

DIMENSION 1

	q_{11}	q_{12}		q_{21}	q_{22}		q_{31}	q_{32}
a	-.040	-1.470	p	-.126	-.726	u	-1.274	.120
b	.031	.922	q	.092	.301	v	.754	.228
c	-.001	.212	r	.140	1.191	w	.520	-.492

DIMENSION 2

a	-1.317	.391	p	-.841	.203	u	.137	.261
b	1.023	-.245	q	.616	-.084	v	-.081	.494
c	-.018	-.056	r	.937	-.333	w	-.056	-1.066

TABLE 9

Centroids and pseudo-centroids

CENTROIDS

dimension 1

	q_{11}	q_{12}		q_{21}	q_{22}		q_{31}	q_{32}
a	-.356	-1.472	p	-.146	-.721	u	-1.280	-.195
b	-.265	.919	q	.259	.287	v	.702	.451
c	.266	.216	r	-.080	1.204	w	.578	-.433

dimension 2

a	-1.268	.377	p	-.836	.364	u	.112	.333
b	1.069	-.271	q	.575	-.568	v	-.278	.442
c	-.059	-.025	r	.992	.099	w	.166	-1.080

PSEUDO-CENTROIDS

dimension 1

a	-.361	-1.467	p	-.135	-.711	u	-1.274	-.065
b	-.270	.927	q	.164	.257	v	.748	.360
c	.270	.205	r	.041	1.231	w	.526	-.457

dimension 2

a	-1.267	.408	p	-.842	.232	u	.132	.317
b	1.070	-.213	q	.621	-.172	v	-.126	.453
c	-.060	-.093	r	.929	-.255	w	-.006	-1.077

TABLE 11

Summary of notation for single OVERALS solution

cat. quant. (single cat. quant.)	$\mathbf{z}_{kj}$	$\mathbf{q}_{kj} = \mathbf{G}_{kj}\mathbf{z}_{kj}$	-----	1
single cat. quant. (rank-one cat. quant.)	$\mathbf{z}_{kj}\mathbf{w}_{kj}' = \mathbf{S}_{kj}$	$\mathbf{q}_{kj}\mathbf{w}_{kj}' = \mathbf{G}_{kj}\mathbf{S}_{kj}$	single fit per var. (discrimination)	$\mathbf{w}_{kj}\mathbf{w}_{kj}'$
centroids	$\mathbf{C}_{kj}$	$\mathbf{G}_{kj}\mathbf{C}_{kj}$	-----	$\mathbf{C}_{kj}\mathbf{D}_{kj}\mathbf{C}_{kj}/n$
aver. rank-one quant.	$\mathbf{z}_{kj}\mathbf{l}_{kj}'$	$\mathbf{q}_{kj}\mathbf{l}_{kj}'$	-----	$\mathbf{l}_{kj}\mathbf{l}_{kj}'$
mult. cat. quant.	$\mathbf{M}_{kj}$	$\mathbf{G}_{kj}\mathbf{M}_{kj}$	mult. fit per var.	$\mathbf{M}_{kj}\mathbf{D}_{kj}\mathbf{M}_{kj}/n$ = single fit per var. + rel. loss per var.
		$\mathbf{G}_{kj}(\mathbf{M}_{kj} - \mathbf{S}_{kj})$ $= \mathbf{G}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})$	rel. loss per var. (single loss)	$(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})/n$ = mult. fit per var. - single fit per var.
		$(\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k) - \mathbf{G}_{kj}(\mathbf{C}_{kj} - \mathbf{C}_{kj})$	mult. loss per var.	$(\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k) - (\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})'\mathbf{D}_{kj}(\mathbf{C}_{kj} - \underline{\mathbf{C}}_{kj})/n$ = single loss per set - rel. loss per var.
		$\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k + \mathbf{q}_{kj}\mathbf{w}_{kj}'$	dispersion (total dispersion)	$\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k + \mathbf{w}_{kj}\mathbf{w}_{kj}' =$ single loss per set + rel. loss per var. = mult. fit per var. + mult. loss per var.
regular lattice points		$\mathbf{Q}_k\mathbf{W}_k = \mathbf{G}_k\mathbf{S}_k$	single fit per set	$\mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$
		$\mathbf{X} - \mathbf{Q}_k\mathbf{W}_k$	single loss per set (total loss)	$\mathbf{I} - \mathbf{W}_k'\mathbf{R}_{kk}\mathbf{W}_k$
broken lattice points		$\mathbf{G}_k\mathbf{M}_k$	(improper) mult. fit per set	$\mathbf{M}_k'\mathbf{G}_k'\mathbf{G}_k\mathbf{M}_k/n$
		$\mathbf{X} - \mathbf{G}_k\mathbf{M}_k$	(improper) mult. loss per set	$\mathrm{SSQ}(\mathbf{X} - \mathbf{G}_k\mathbf{M}_k)$

TABLE 10

Multiple category quantifications

	q_{11}	q_{12}		q_{21}	q_{22}		q_{31}	q_{32}
				dimension 1				
a	-.242	-1.505	p	-.459	-.876	u	-1.250	-.094
b	.196	.932	q	.417	.390	v	.690	.159
c	-.007	.227	r	.378	1.394	w	.560	-.122
				dimension 2				
a	-1.286	.203	p	-.860	.064	u	.357	.302
b	.977	-.204	q	.588	-.368	v	-.375	.531
c	-.016	.034	r	1.027	.465	w	.018	-1.174

TABLE 12

Summary of OVERALS notation. Read $\mathbf{B}_{kj} = \mathbf{S}_{kj}$ if variable is treated as single. Read $\mathbf{B}_{kj} = \mathbf{M}_{kj}$ if variable is treated as multiple nominal, and $\mathbf{M}_{kj}$ contains the proper multiple category quantification. Read $\mathbf{B}_{kj} = \underline{\mathbf{M}}_{kj}$ for the improper multiple category quantification of a variable treated as single.

	Verdegaal	**This monograph**	**Formula**
no matter whether variable q_{kj} is treated as single or multiple	discrimination	fit per variable	$SSQ(\mathbf{G}_{kj}\mathbf{B}_{kj})$
	total dispersion	dispersion per var.	$SSQ(\mathbf{X} - \mathbf{G}_k\mathbf{B}_k + \mathbf{G}_{kj}\mathbf{B}_{kj})$
	total loss	loss per set	$SSQ(\mathbf{X} - \mathbf{G}_k\mathbf{B}_k)$
for variable q_{kj} treated as single	multiple fit	multiple fit per var.	$SSQ(\mathbf{G}_{kj}\underline{\mathbf{M}}_{kj})$
	multiple loss	mult.loss per var.	$SSQ((\mathbf{X} - \mathbf{G}_k\mathbf{B}_k) - \mathbf{G}_{kj}(\underline{\mathbf{M}}_{kj} - \mathbf{S}_{kj}))$
	single loss	relative loss per var.	$SSQ(\mathbf{G}_{kj}(\underline{\mathbf{M}}_{kj} - \mathbf{S}_{kj}))$

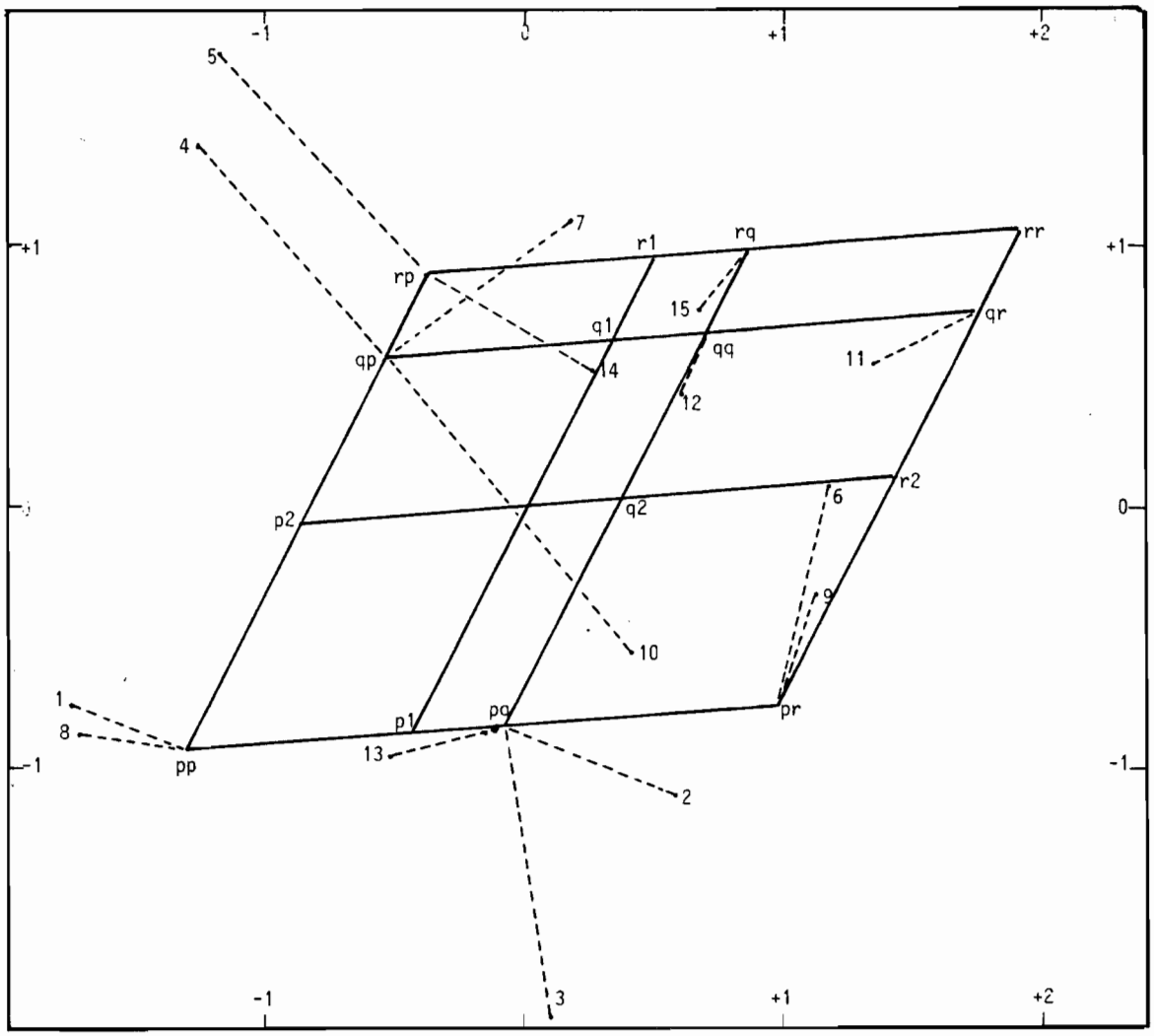

Figure 1.
Graph of lattice for second set. Single category points of $\mathbf{q}_{21}$ have labels p1 q1 r1; those of $\mathbf{q}_{22}$ have labels p2 q2 r2. Lattice points have label of two letters, the first one for the category of $\mathbf{q}_{21}$, the second for the category of $\mathbf{q}_{22}$. The n=15 object points are numbered. They are connected with their corresponding lattice points by dotted lines. The average squared length of these dotted lines is the loss of set 2.

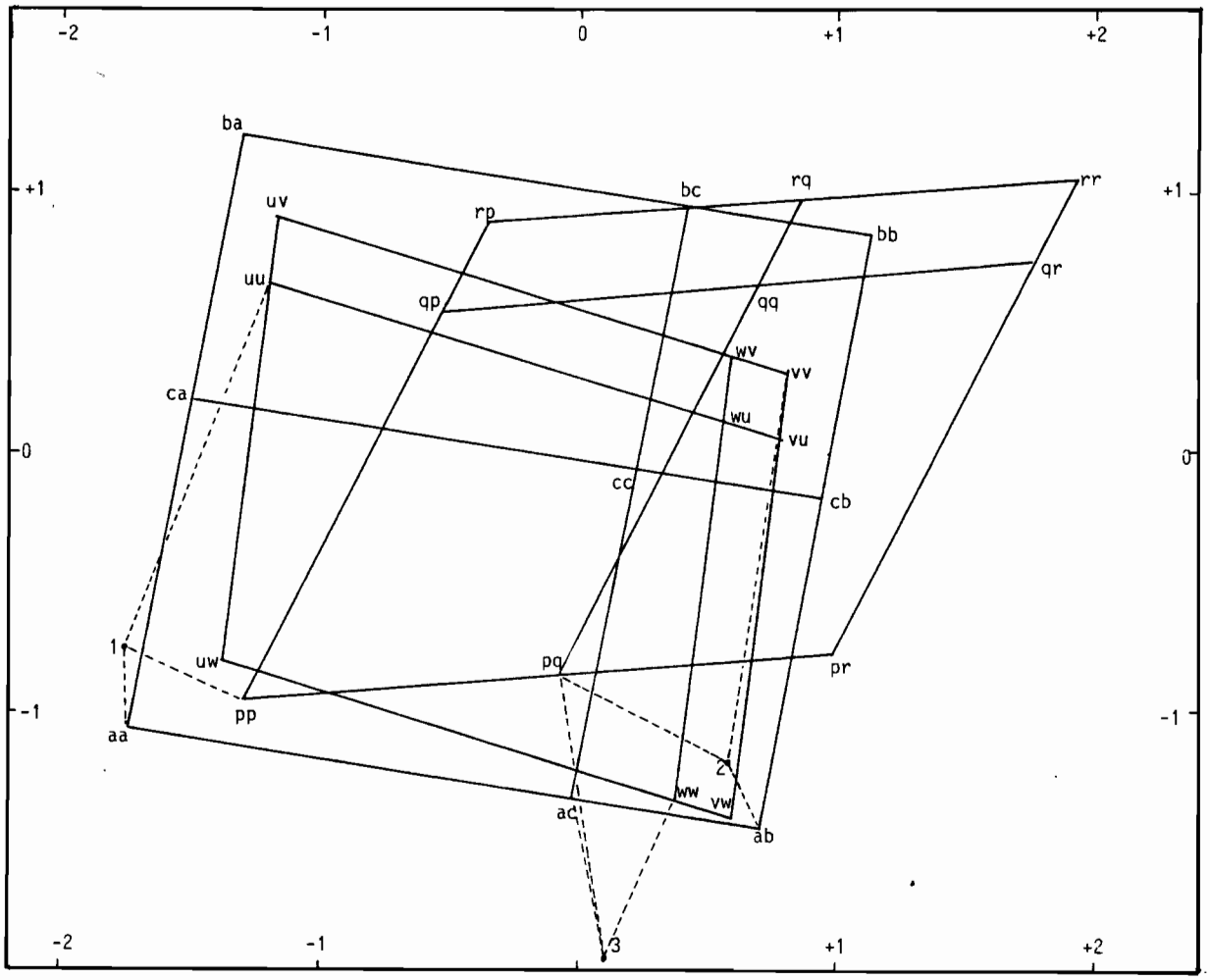

Figure 2.
Graph of all three lattices. Object points are shown only for objects 1 2 3 (all three in category a of $\mathbf{q}_{11}$). They are connected with their corresponding three lattice points by dotted lines. If such dotted lines were drawn for all 15 objects, their average squared length would represent the total loss.

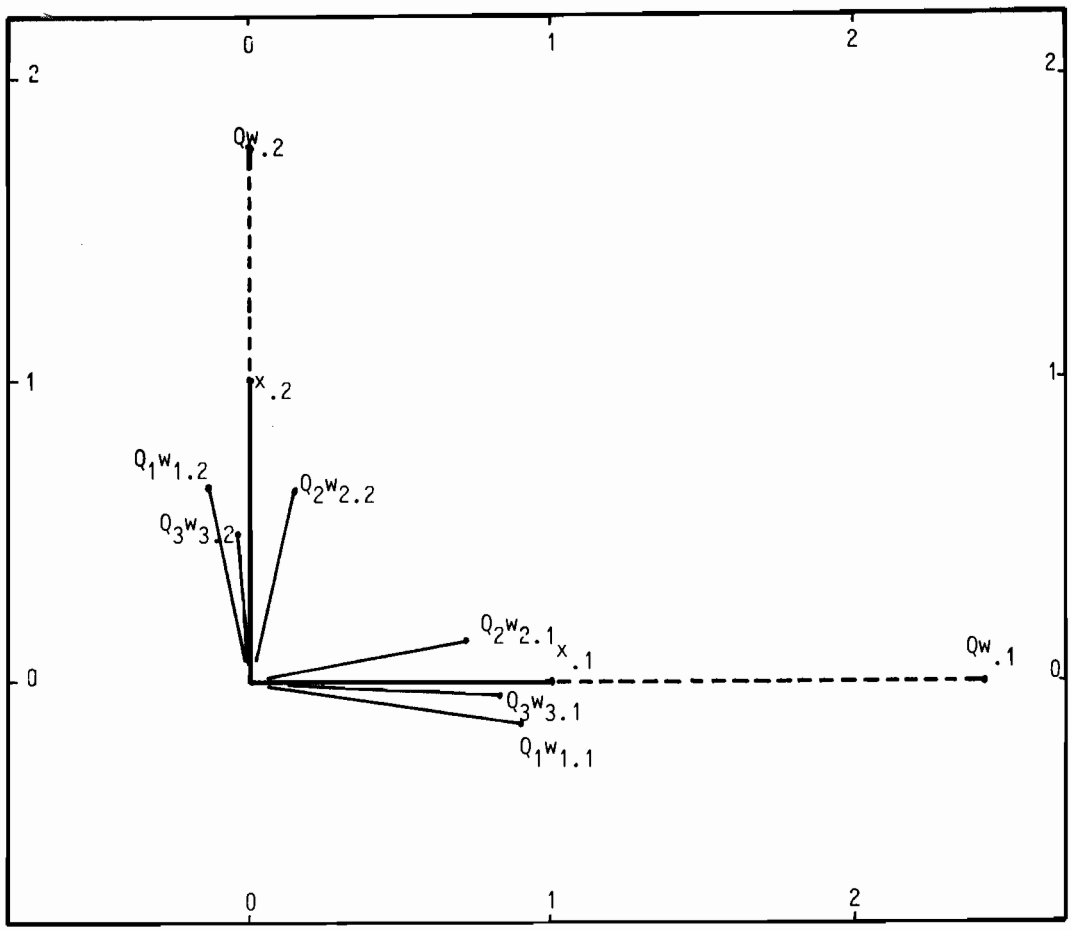

Figure 3.
Weighted sumvectors $\mathbf{Q}_k\mathbf{w}_{k.s}$ plotted with their component loadings as coordinates. The vector $\mathbf{Qw}_{.s}$ is the sum of the three vectors $\mathbf{Q}_k\mathbf{w}_{k.s}$; its length is equal to the sth eigenvalue multiplied by K=3. The vector $\mathbf{x}_{.s}$ is the same as $\mathbf{Q}_{w.s}$, but reduced to unit length. A good OVERALS solution will show that the vectors $\mathbf{Q}_k\mathbf{w}_{k.s}$ have small angles with $\mathbf{x}_{.s}$.

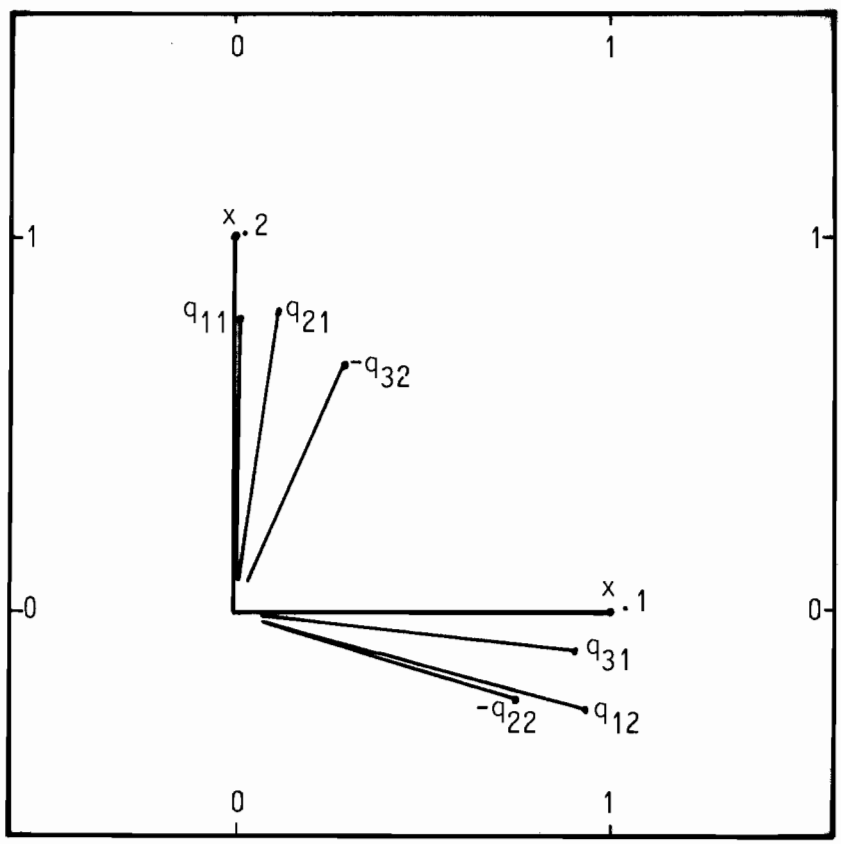

Figure 4.
Original vectors $\mathbf{q}_{kj}$ plotted on the basis of their component loadings, with sign reversal for $\mathbf{q}_{22}$ and $\mathbf{q}_{32}$. That the vectors $\mathbf{q}_{kj}$ fall apart into two fans around $\mathbf{x}_{.1}$ and $\mathbf{x}_{.2}$ is an accidental property of the example.

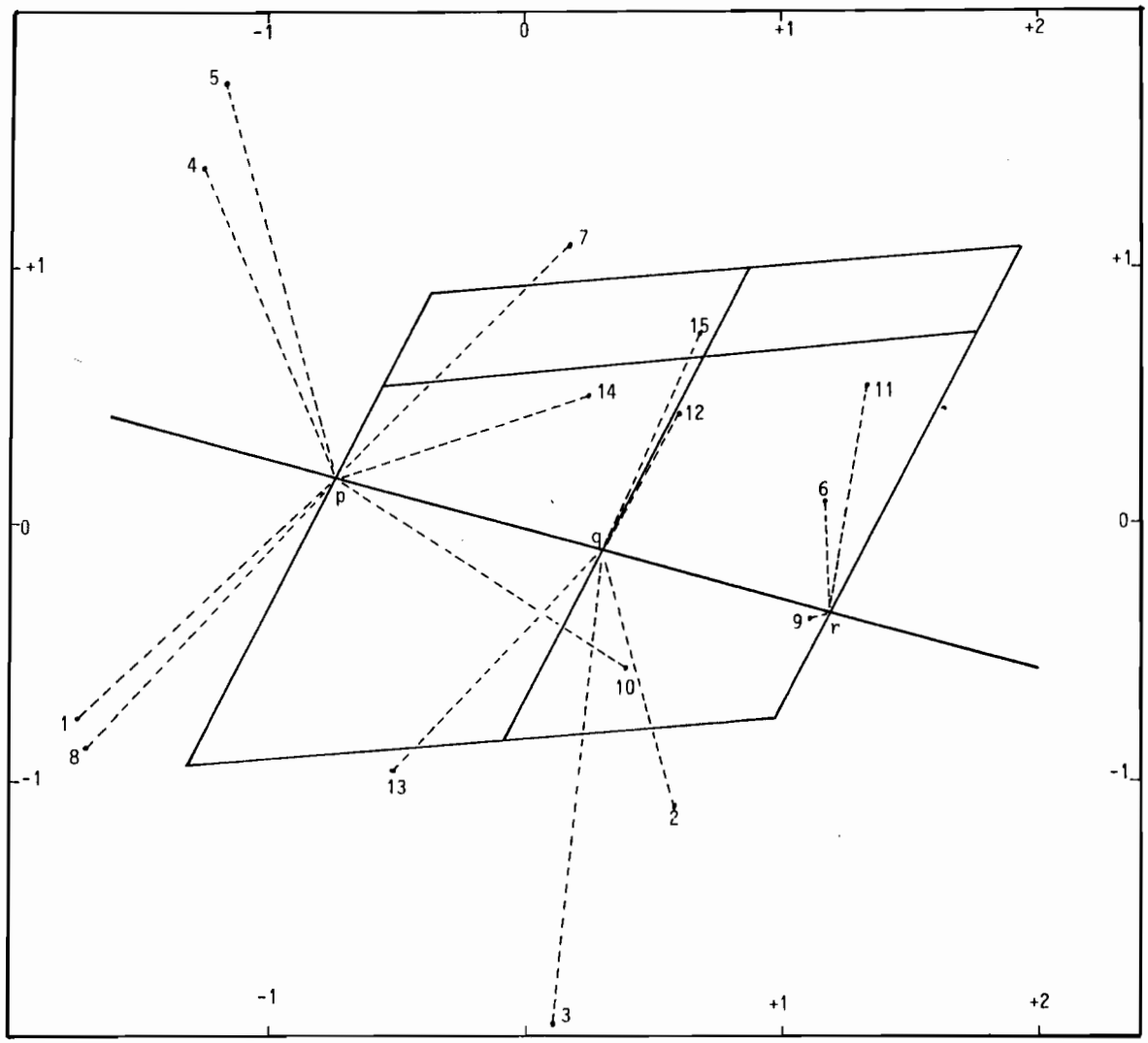

Figure 5.
Graph of second lattice (as in Figure 1), with average rank one points of $\mathbf{q}_{22}$ shown as the intersections of the lattice and straight line which passes through the point defined by the component loadings of $\mathbf{q}_{22}$, equal to (-.747 .209) - this point itself is not shown in the graph.
Average rank one points have label p q r. They are connected to corresponding object points by dotted lines. The average squared length of these dotted lines is minimized (it becomes larger if one takes intersections of the lattice by a different straight line).
The graph shows that object points tend to fall apart into three clusters (with object 10 as an explanation).

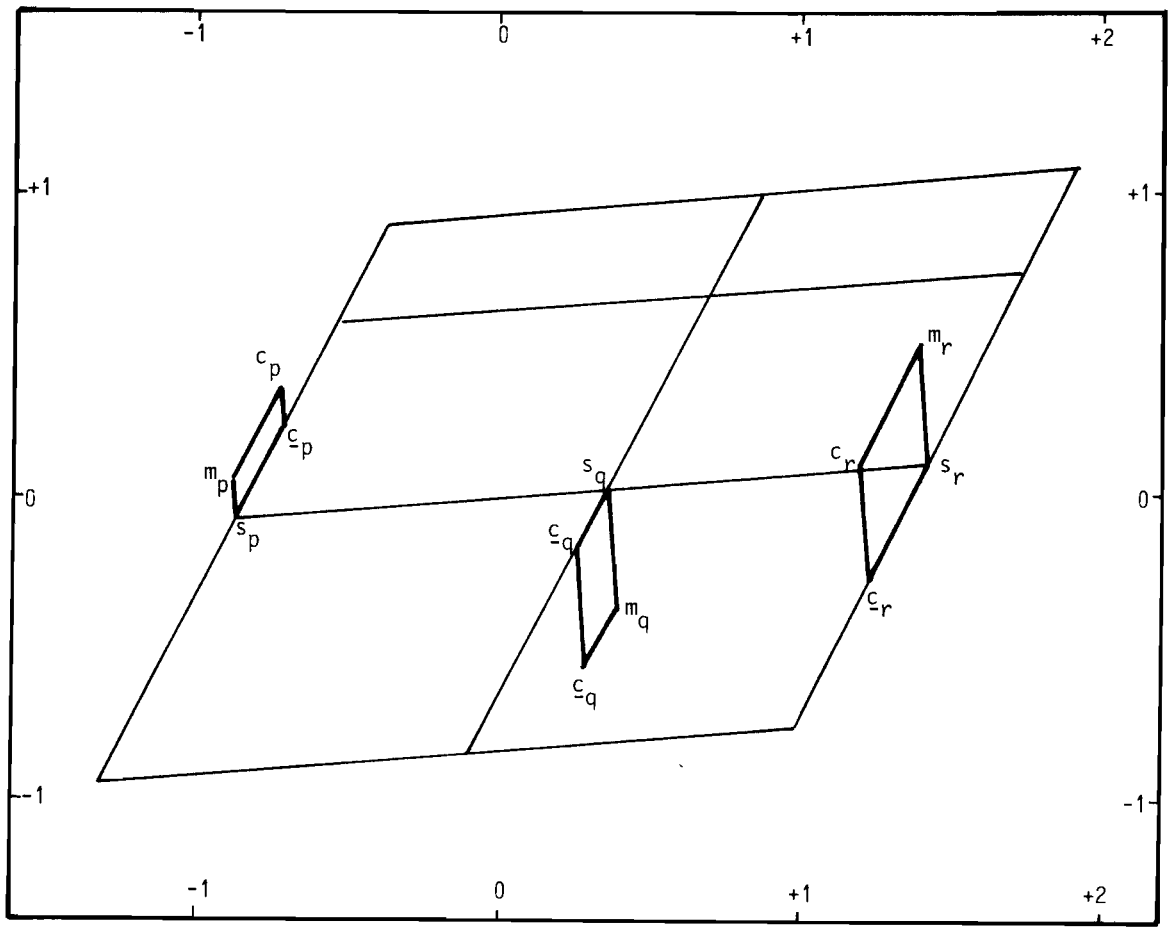

Figure 6.
The figure shows the lattice for the second set, with for each category of q_{22} the single category points s, the pseudo-centroids $\underline{c}$, the centroids c, and the multiple category points m. For each category those four points form a parallelogram.

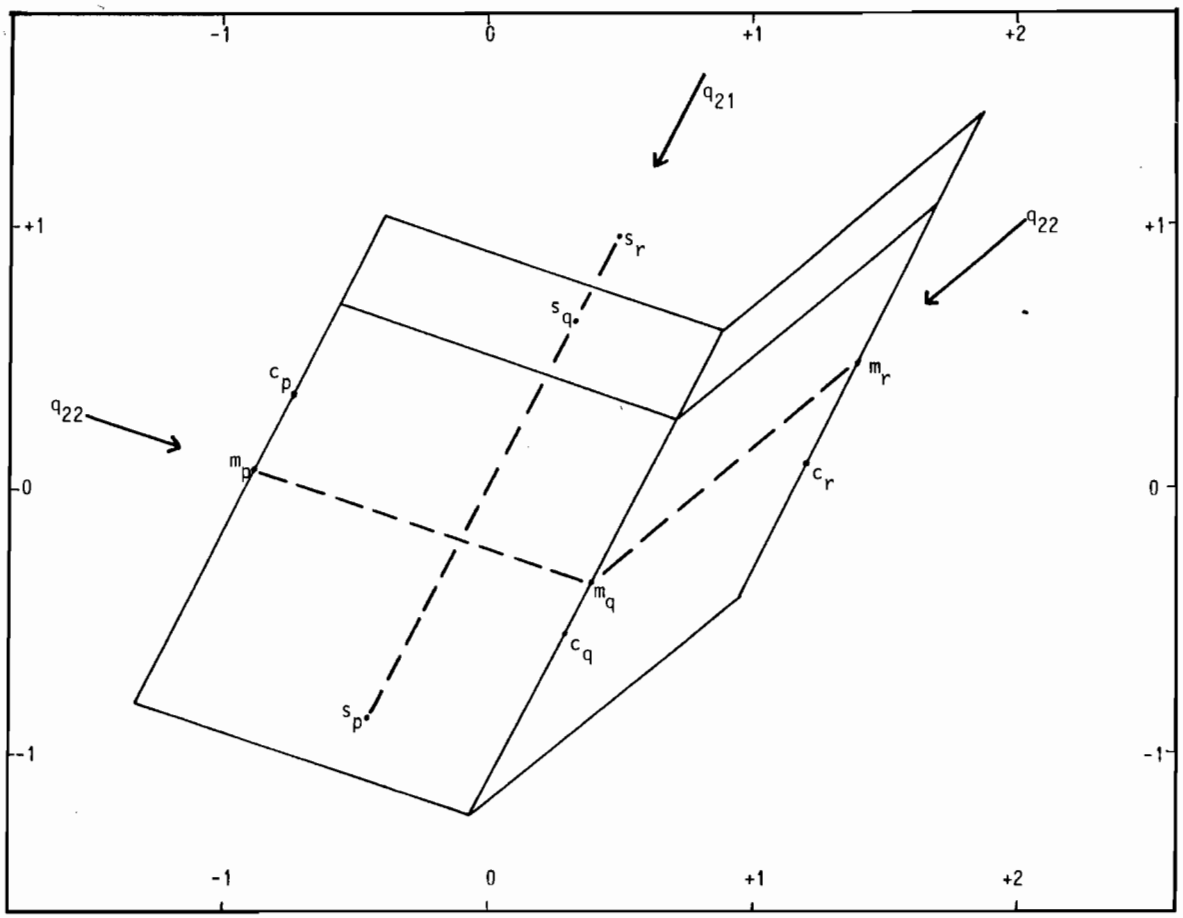

Figure 7.
Semi-broken lattice of second set, based on the single category points of $\mathbf{q}_{21}$ and the multiple category points of $\mathbf{q}_{22}$. A property of the semi-broken lattice is that centroids of the categories of $\mathbf{q}_{22}$ now are located on the sides of the lattice. But it is difficult to generalize this property when there are more than two variables in the set, some of which are taken as single, others as multiple.

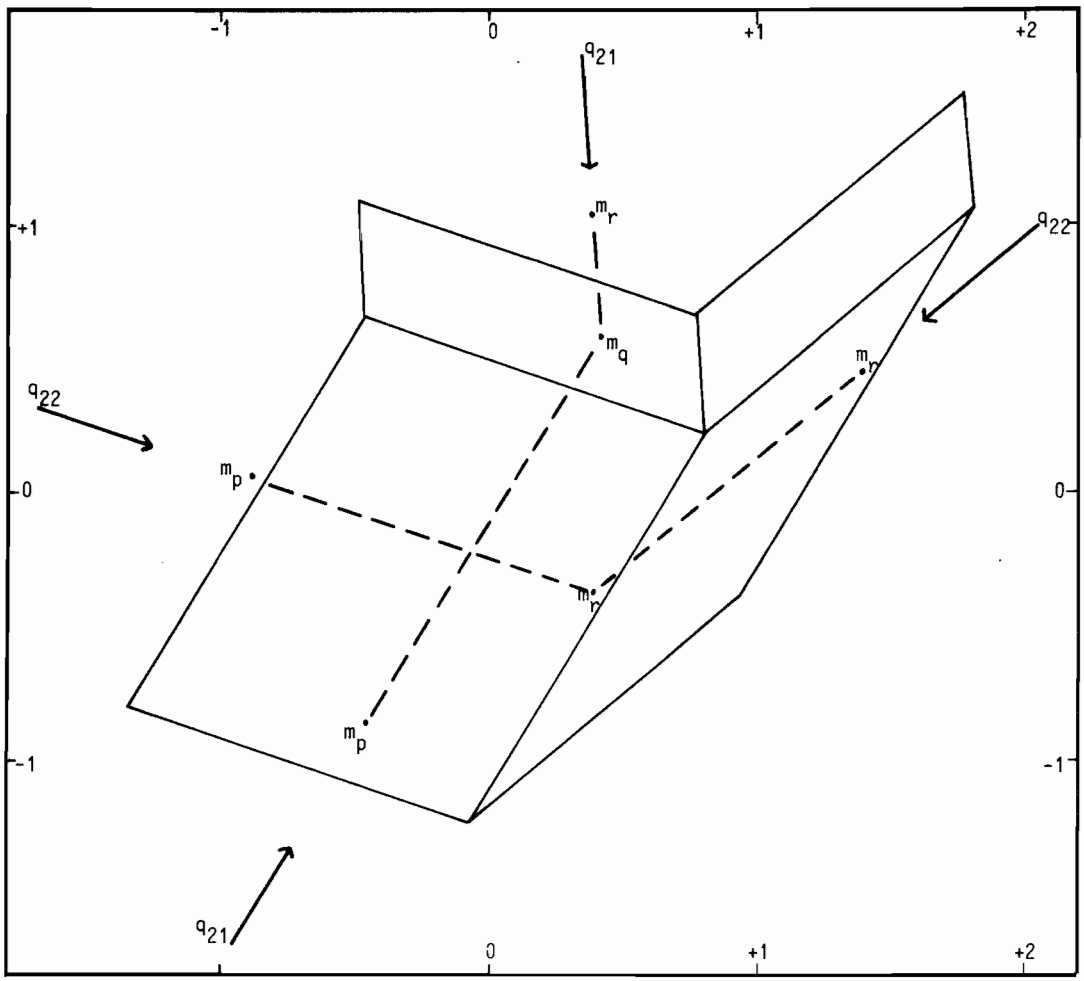

Figure 8.
Broken lattice of second set, based on the multiple category points of $\mathbf{q}_{21}$ and $\mathbf{q}_{22}$.

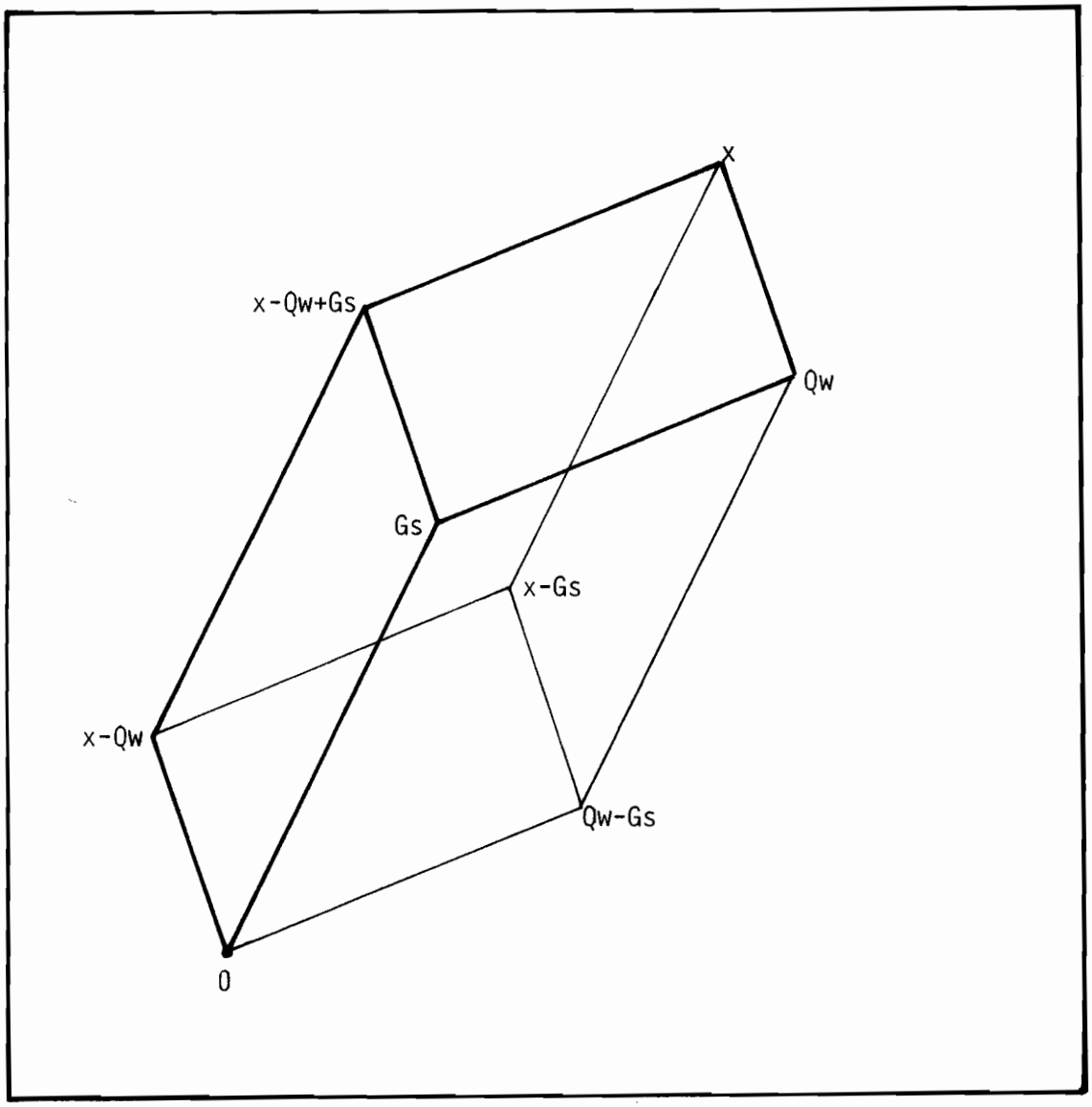

Figure 9A.
Scheme of various vectors for one dimension of the OVERALS solution. In the graph, the vectors have abbreviated labels. One should read:

x = : $\mathbf{x}_{.s}$,
Qw = : $\mathbf{Q}_k\mathbf{w}_{k.s}$,
G = : $\mathbf{G}_{kj}$,
s = : the s^{th} column of $\mathbf{S}_{kj}$, so that "Gs" becomes equal to $\mathbf{q}_{kj}\mathbf{w}_{kj.s}$.

The squared length of some of the shown vectors corresponds to definitions given in Table 11. In particular:

SSQ = 1,
SSQ(x-Qw+Gs) = dispersion of $\mathbf{q}_{kj}$ on dimension s,
SSQ(Gs) = single fit of $\mathbf{q}_{kj}$ on dimension s,
SSQ(x-Qw) = single loss of set on dimension s,
SSQ(Qw) = single fit of set on dimension s.

The graph represents a three-dimensional parallelopipedum which is in fact rectangular. According to Pythagoras' theorem it then follows that:

dispersion = single fit per variable + single loss per set ;
single fit per set + single loss per set = SSQ(x) = 1.