I.5. MULTIVARIATE STATISTICS

Multivariate analysis is the analysis of observations on several (possibly) correlated random variables. There are two important aspects in multivariate statistics. The first is the description of a variable as a function of several other variables. In this context the term 'regression' is used when fitting observational data to a model. An obvious example is the number of library requests as a function of library size, price, response time, copying quality, etc. In general, we wish to describe a function Y, depending on k variables $X_1,\ldots,X_k$:

$$Y = Y(X_1,\ldots,X_k) \ .$$

The second aspect of multivariate statistics may be called 'dimensionality reducing techniques'. In this case we consider principal component analysis, multi-dimensional scaling and cluster analysis (to be explained further on). These methods are referred to as dimensionality-reduction methods because their aim is to simplify what is first a complex pattern of associations in many variables.

Geometrically, this process of simplification is done by projecting or representing an object in a higher-dimensional space in a space of a smaller number of dimensions (usually two). This is somewhat similar to projecting a three-dimensional globe onto a two-dimensional map (Kinnucan et al. (1987)). Orthogonal projection is illustrated by Fig.I.5.1.

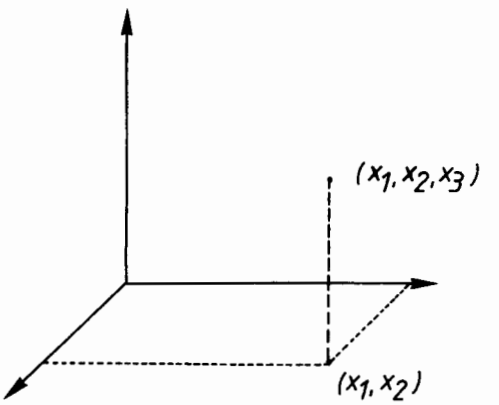

Fig.I.5.1 Orthogonal projection of a point in 3-space onto a two-dimensional plane

These methods operate on matrices of relations. A classical case is offered by citation analysis (to be discussed in more detail in Part III). For example, if one selects a group of journals $\{J_1,\ldots,J_n\}$ in a fixed subject area and studies the number of citations, c_{ij}, given in journal J_i to journal J_j, this yields a (square) matrix of raw data :

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & & c_{2n} \\ \vdots & \vdots & & \\ c_{n1} & c_{n2} & & c_{nn} \end{pmatrix} .$$

This is an example of a network study : an investigation of certain relations within one group. In general, one also studies n objects and k variables, giving rise to rectangular matrices.

I.5.1. Multiple regression and correlation

In *multiple regression*, we consider the relationship between a string of values of the dependent variable Y, and several strings of corresponding values of the so-called predictor variables $X_1,\ldots,X_k$. The simplest relation between these variables is the linear equation :

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k . \qquad [I.5.1]$$

In mathematics, vectors $(X_1,X_2,\ldots,X_k,Y)$ that satisfy this equation are said to form a *hyperplane* in the (k+1)-dimensional space. A hyperplane in $\mathbb{R}^3$ is what we usually call a plane. As in Subsection I.3.8.4, we require that hyperplane to fit the raw data (the vectors of values for the variables) best in the sense of least squares. Although this makes the equations for $k \geq 2$ intricate, this is not a serious drawback. There are numerous computer programs that quickly find the best fitting values for $a,b_1,b_2,\ldots,b_k$.

For k = 2, the best fitting plane

$$Y = a + b_1X_1 + b_2X_2$$

is obtained when applying the following equations given below, in which Σ represents the sum over all observed values (denoted by minuscules); the summation index is not written :

$$b_1 = \frac{(\Sigma\,(x_1-\bar{x}_1)(y-\bar{y}))(\Sigma\,(x_2-\bar{x}_2)^2) - (\Sigma\,(x_2-\bar{x}_2)(y-\bar{y}))(\Sigma\,(x_1-\bar{x}_1)(x_2-\bar{x}_2))}{(\Sigma\,(x_1-\bar{x}_1)^2(\Sigma\,(x_2-\bar{x}_2)^2) - (\Sigma\,(x_1-\bar{x}_1)(x_2-\bar{x}_2))^2} \quad [I.5.2]$$

$$b_2 = \frac{(\Sigma\,(x_2-\bar{x}_2)(y-\bar{y}))(\Sigma\,(x_1-\bar{x}_1)^2) - (\Sigma\,(x_1-\bar{x}_1)(y-\bar{y}))(\Sigma\,(x_1-\bar{x}_1)(x_2-\bar{x}_2))}{(\Sigma\,(x_1-\bar{x}_1)^2(\Sigma\,(x_2-\bar{x}_2)^2) - (\Sigma\,(x_1-\bar{x}_1)(x_2-\bar{x}_2))^2} \quad [I.5.3]$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \quad . \quad [I.5.4]$$

The best fitting plane is sketched schematically in Fig.I.5.2.

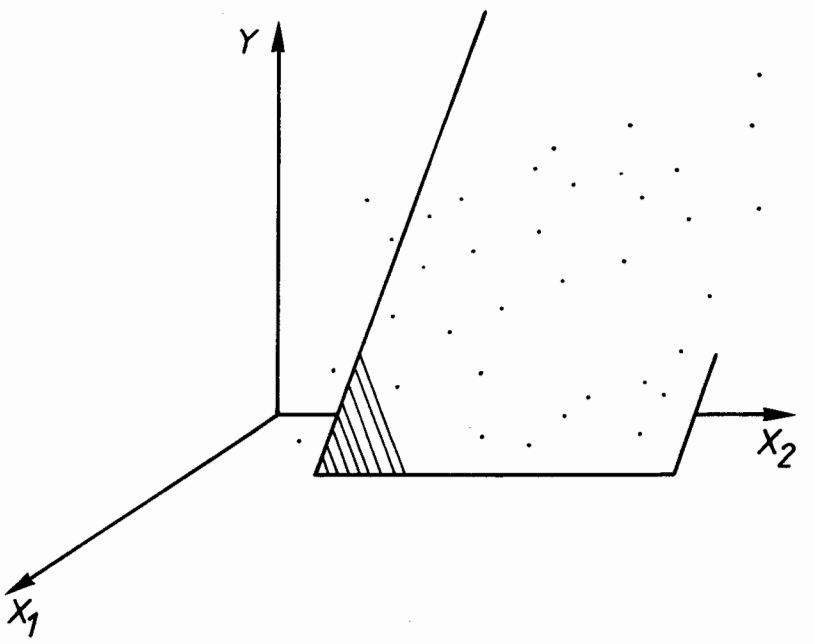

Fig.I.5.2 Best fitting plane of a three-dimensional scatterplot

The strength of this linear relationship is measured by the *two-dimensional correlation coefficient* :

$$r = \left(\frac{b_1\,\Sigma\,(x_1-\bar{x}_1)(y-\bar{y}) + b_2\,\Sigma\,(x_2-\bar{x}_2)(y-\bar{y})}{\Sigma\,(y-\bar{y})^2}\right)^{1/2} \quad . \quad [I.5.5]$$

An example :

Y : the number of interlibrary lending requests for books in a library (in hundreds)

X_1 : number of books in the library (in thousands)

X_2 : price (in \$).

Data are given in Table I.5.1.

Table I.5.1. Interlibrary lending data

Library	Y	X_1	X_2
A	23	10	7
B	7	2	3
C	15	4	2
D	17	6	4
E	23	8	6
F	22	7	5
G	10	4	3
H	14	6	3
I	20	7	4
J	19	6	3
SUM	170	60	40
MEAN	$\bar{y} = 17$	$\bar{x}_1 = 6$	$\bar{x}_2 = 4$

Further calculations can be done as illustrated in Table I.5.2.

Table I.5.2. Calculations for a two-dimensional regression analysis on Table I.5.1

Libr.	$Y-\bar{Y}$	$X_1-\bar{X}_1$	$X_2-\bar{X}_2$	$(X_1-\bar{X}_1)(Y-\bar{Y})$	$(X_2-\bar{X}_2)(Y-\bar{Y})$	$(X_1-\bar{X}_1)(X_2-\bar{X}_2)$	$(Y-\bar{Y})^2$	$(X_1-\bar{X}_1)^2$	$(X_2-\bar{X}_2)^2$
A	6	4	3	24	18	12	36	16	9
B	-10	-4	-1	40	10	4	100	16	1
C	-2	-2	-2	4	4	4	4	4	4
D	0	0	0	0	0	0	0	0	0
E	6	2	2	12	12	4	36	4	4
F	5	1	1	5	5	1	25	1	1
G	-7	-2	-1	14	7	2	49	4	1
H	-3	0	-1	0	3	0	9	0	1
I	3	1	0	3	0	0	9	1	0
J	2	0	-1	0	-2	0	4	0	1
				102	57	27	272	46	22

$$b_1 = \frac{102 \times 22 - 57 \times 27}{46 \times 22 - (27)^2} = 2.49$$

$$b_2 = \frac{57 \times 46 - 102 \times 27}{46 \times 22 - (27)^2} = -0.47$$

$$a = 17 - 2.49 \times 6 - (-0.47) \times 4 = 3.92$$

This yields the following 'best' linear relation :

$$Y = 3.93 + 2.49\ X_1 - 0.47\ X_2$$

We observe a positive correlation between Y and X_1 and a negative correlation between Y and X_2. This agrees with intuitive expectation.

Most uses of multiple regression can be classified into three categories : 1) for prediction, 2) for model specification and 3) for parameter estimation (Gunst and Mason (1980)). In prediction the emphasis is on estimating accurate values of the dependent variable (Y) for any combination of the independent variables (the X_i's). In model specification the emphasis shifts to finding the best combination of predictor variables and assessing their relative importance for prediction. Finally, in parameter estimation, regression analysis is used to provide accurate estimates of the parameters associated with the predictor variables (Kinnucan et al. (1987)).

Nevertheless such preditions are only reliable in the immediate neighbourhood of the actual situation; false predictions can result when predictor variables change too much. Let us take $x_1 = x_2 = 0$ in the preceding example. Then y = 3.93, meaning that an empty library would receive 393 requests. This is a clear example of undue extrapolation.

Numerous examples of applications of this technique are found in the literature on informetrics. See, for example, Virgo (1977), McDonough (1982), Bennion and Karschamroon (1984), Cooper (1984).

I.5.2. Principal components analysis (PCA)

I.5.2.1. Introduction

We consider an (n,k)-matrix of raw data :

$$C = \begin{pmatrix} c_{11} & c_{12} & & c_{1k} \\ c_{21} & & & \\ \vdots & & & \\ c_{n1} & \cdots & & c_{nk} \end{pmatrix} \qquad [I.5.6]$$

where n is not necessarily equal to k. This matrix will often be abbreviated to the notation (c_{ij}). Such matrices occur, for example, when studying n '*citing*' journals $(A_1,\ldots,A_n)$ and k '*cited*' journals $(B_1,\ldots,B_k)$. The matrix entry c_{ij} denotes the number of times journal A_i cites journal B_j in a fixed period. The functional relation studied in this example is 'to cite'. One can also study the relation 'to be cited', yielding a different matrix and maybe even different results, but from our technical point of view this is the same problem.

In every case we will view the matrix C as a representation of n points $C_i = (c_{i1},c_{i2},\ldots,c_{ik})$, $i = 1,\ldots,n$ in k-dimensional space $\mathbb{R}^k$. We wish to obtain more information about the configuration of the scatter diagram of the C_i's. The study of these kinds of interrelations helps library managers to make acquisition decisions : they can decide, for instance, to subscribe to journals often cited by the most popular journals of the library. In connection with scientometric studies such analyses may determine important groups of researchers (invisible colleges or 'schools', see further Part IV) or the scatter of different scientific fields in a country.

I.5.2.2. An intuitive approach to principal components analysis

The problem we face here is finding an adequate visualisation of n points in $\mathbb{R}^k$. Since humans are only capable of perceiving objects in at most three dimensions, this means we will have to find a method to reduce the dimension of the set of points under study. For practical reasons usually only two dimensional images are allowed. Projecting on a plane will, however, seriously deform the original scatter diagram. For example, consider two points A and B in 3-space and project them onto a plane perpendicular to the line joining A and B (cf. Fig.I.5.3).

This projection maps A and B onto the same point C, showing that this is the worst possible 2-dimensional representation of this set of points. In fact, any plane parallel to the line AB would represent this situation perfectly. See Fig.I.5.3 : the projection of A and B onto A' and B'. General scatter diagrams are much more complicated than the simple example above so that, no matter which plane we take, some information will be lost in the operation. Indeed, a projection reduces the distance between points (except in the case where the plane of projection is parallel to all points under consideration). We will look for whichever plane that avoids this reduction as much as possible.

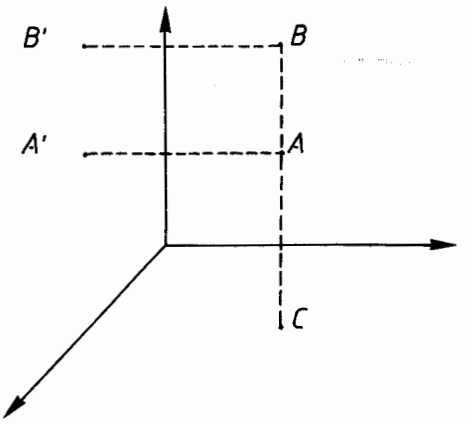

Fig.I.5.3 Projection

The internal variance of a scatter diagram is defined as the sum of the squares of all distances between any two (different) points. Since there are n points, there are $\binom{n}{2} = n(n-1)/2$ distances to consider. We will try to determine that plane which maximises the internal variance of the projected scatter diagram.

For easily recognisable objects it is immediately clear which plane is a maximising plane. For a perfect ball any plane is maximising, but for a pair of scissors we have less choice, as shown in Fig.I.5.4.

Note that any plane parallel to a maximising plane is also a maximising plane. Therefore, we can choose a plane passing through the origin $0 = (0,0,\ldots,0)$ in $\mathbb{R}^k$. This plane will be entirely determined by two perpendicular axes.

In practice, we will need a computer program to solve the problem satisfactorily. This program will work as follows. First, it finds an axis, call it x_1, such that the internal variance of the projected scatter diagram is maximal with respect to all other axes.

It then finds a second axis, x_2, orthogonal to x_1, maximising the internal variance among all axes orthogonal to x_1. The plane determined by x_1 and x_2 is then a maximising plane. The program continues in the same way, finding a third, a fourth, a fifth, ..., a k-th axis, all mutually orthogonal.

These k axes are said to be the principal components of the scatter diagram. Mathematically, the whole procedure is essentially a problem of finding eigenvalues and eigenvectors of the so-called variance-covariance matrix of C (see e.g. Kshirsagan (1972), Chapter 11). The first axis is then an eigenspace associated with the largest eigenvalue; the second axis is associated with the

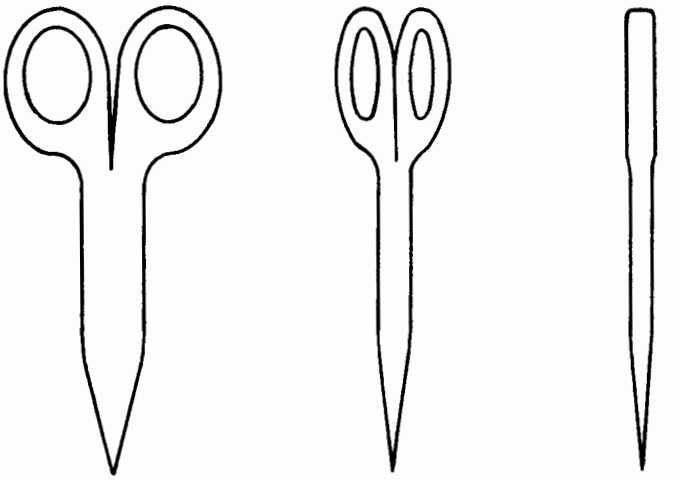

Fig.I.5.4 Projections of the same pair of scissors onto three different planes

second largest eigenvalue and so on. Intuitively speaking, the eigenvalue corresponding to an eigenvector contains the amount of variation that is retained after the k-dimensional scatter plot of observed data is projected onto the eigenspace associated with this eigenvector. The method of *principal components analysis* originated with Hotelling (1933), who developed the technique for his work in educational psychology.

In many cases the first plane (x_1,x_2) is the most important one. However, it is always a good practice to use several planes (e.g. (x_1,x_3) or (x_2,x_3) as well) to get a better idea of the scatter diagram. This may also help to detect anomalies in the data.

In practice we will retain as many axes as necessary - beginning, of course, with the most important ones - in order to recover a fixed percentage (say 75 %) of the total variance of the k-dimensional scatter diagram. Principal components are artificial variables and do not necessarily have any physical meaning or significance. While they are linear combinations of variables that can be measured, they themselves cannot generally be measured directly. Sometimes, however, they can be interpreted, as will be shown in the next example.

Table I.5.3. Citation matrix C of botanical journals (1983)

Cited	Citing 1	2	3	4	5	6
1. PLANT PHYSIOL	2906	382	97	682	204	1007
2. PHYTOCHEMISTRY	270	2115	9	119	38	134
3. PHYTOPATHOLOGY	42	36	1771	-	158	17
4. PLANTA	672	143	-	685	97	442
5. CAN J BOT	139	40	130	33	630	89
6. PHYSIOL PLANT	290	79	11	99	93	665
7. AM J BOT	64	26	27	40	254	60
8. NEW PHYTOL	118	24	18	54	312	75
9. ANNU REV PLANT PHYS	312	45	19	116	53	148
10. J EXP BOT	248	44	-	129	58	173
11. ANN BOT-LONDON	83	16	14	28	118	98
12. PLANT CELL PHYSIOL	174	62	-	65	28	141
13. Z PFLANZENPHYSIOL	129	63	-	60	33	183
14. PLANT SOIL	39	-	19	-	65	16
15. PLANT SCI LETT	168	44	-	80	26	102
16. J PHYCOL	21	13	-	9	40	-
17. WEED SCI	21	-	-	-	10	-
18. BOT GAZ	50	9	6	18	86	35
19. CAN J PLANT SCI	20	-	27	-	14	8
20. AUST J PLANT PHYSIOL	127	12	-	50	20	41
21. PHYSIOL PLANT PATHOL	58	24	83	8	23	8

I.5.2.3. An example : a network study of botanical journals (Keteleer (1986))

We report here on work involving a citation network of botanical journals, done by Ann Keteleer as an M.Sc. student of ours. One of the aims of this study was to find out whether this citation configuration was tight, indicating that botany exists as a strong field in itself, or whether it was loose, showing that other disciplines interfere in the field. Furthermore, principal components analysis was used to reveal possible subfields of botany.

Based on a citation criterion (for more details we refer to the study itself), 21 journals were selected. Although both the relations 'cite' and 'cited' were investigated, we report here only on the relation 'cited'. Hence citing journals are the variables and cited journals are the objects (points in $\mathbb{R}^{21}$). Data were collected from the 1983 Journal Citation Reports (JCR). (For more details concerning the JCR see Part III). The citation matrix is given in Table I.5.3.

Table I.5.3. - cont.

7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
76	166	288	481	165	524	418	148	383	35	125	39	35	97	79
20	23	50	48	15	82	117	20	60	8	12	13	9	18	39
-	47	-	6	-	-	-	40	8	-	11	16	31	-	236
84	91	115	212	117	178	334	32	200	22	19	39	7	36	9
130	105	19	82	57	34	63	54	34	27	19	49	42	8	81
36	77	40	113	104	133	215	94	114	-	19	10	9	15	21
450	73	35	47	116	37	44	21	21	25	16	105	14	-	11
63	467	13	84	79	26	54	201	20	25	-	16	12	10	14
28	49	63	77	56	81	86	42	58	6	19	11	7	22	11
33	79	48	339	111	52	98	60	50	7	14	9	7	36	-
86	89	21	94	326	14	47	34	19	12	6	32	13	24	11
9	6	28	28	34	466	68	-	56	6	6	14	-	13	-
21	16	30	64	51	60	302	16	74	7	6	9	-	-	-
-	53	-	16	21	-	24	319	10	-	7	-	12	-	-
7	18	38	32	25	44	128	10	177	-	7	-	-	10	-
23	36	-	-	6	-	13	-	-	132	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	662	-	80	-	-
114	23	10	16	48	11	20	9	12	-	13	61	7	-	-
-	-	-	-	13	-	-	13	-	-	38	-	227	-	-
-	11	21	49	33	21	22	20	22	-	-	-	-	81	-
-	13	-	-	7	-	6	-	-	-	-	-	-	-	22

Before the system begins the actual principal components analysis, data are first standardised, giving every variable a mean equal to zero and a variance equal to one. We will not go into detail here, but merely note that programs usually either do this automatically or have this option.

The following eigenvalues were found for the botanical journals (Table I.5.4). This table also indicates the percentage of variation they represent. Furthermore, the system draws the following projection on the plane formed by the first principal components (Fig.I.5.5).

Table I.5.4. Eigenvalues of the citation matrix of botanical journals (Table I.5.3)

principal components	eigenvalue	% variation	cum % variation
1	7.71	36.69	36.69
2	2.70	12.86	49.55
3	1.87	8.91	58.46
4	1.58	7.50	65.96
5	1.33	6.36	72.32
6	1.03	4.91	77.23
7	0.95	4.52	81.75
8	0.69	3.27	85.01
9	0.67	3.17	88.18
10	0.50	2.38	90.56
.	.	.	.
21	-9.11 E-08	0.00	100.00

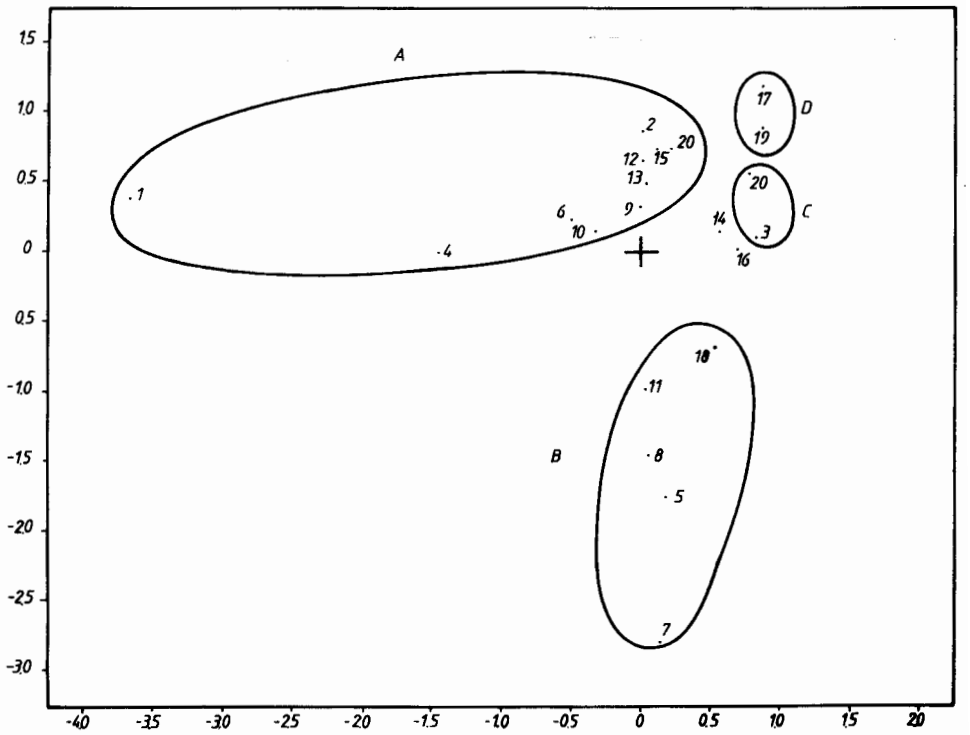

Fig.I.5.5 Projection of the scatterplot of citations to botanical journals onto the plane of the first two eigenvectors. The numbers have the following meaning :

1. PLANT PHYSIOL
2. PHYTOCHEMISTRY
3. PHYTOPATHOLOGY
4. PLANTA
5. CAN J BOT
6. PHYSIOL PLANT
7. AM J BOT
8. NEW PHYTOL
9. ANNU REV PLANT PHYS
10. J EXP BOT
11. ANN BOT - LONDON
12. PLANT CELL PHYSIOL
13. Z PFLANZENPHYSIOL
14. PLANT SOIL
15. PLANT SCI LETT
16. J PHYCOL
17. WEED SCI
18. BOT GAZ
19. CAN J PLANT SCI
20. AUST J PLANT PHYSIOL
21. PHYSIOL PLANT PATHOL

As can immediately be seen, this projection explains only 49.55 % of the variation. Still, we can draw some useful conclusions. The encircled areas in Fig.I.5.5 stand for recognisable groups of botanical journals : A = plant physiology, B = general journals, C = phytopathology, D = applied botany. As these areas are only partially separated, we are led to the tentative conclusion that no strong subdisciplines exist in botany. From other results of this work (Keteleer (1986)) we may conclude that botany uses quite a lot of results from other scientific fields (much more than vice versa). The same group of journals will be considered again when describing cluster analysis in Section I.5.4.

Further applications of PCA and related techniques such as factor analysis, correspondence analysis and quasi-correspondence analysis can be found, for example, in Bookstein and Podet (1986), Simeon et al. (1986), Cheney and Nelson (1988), Tijssen et al. (1987, 1988).

I.5.3. Multidimensional scaling

In the above section on PCA we considered n points in k-space $\mathbb{R}^k$, in which the problem was to find a 'best' low-dimensional representation. In informetric studies we also encounter more complex situations such as the following :

(1) We do not know the coordinates of the n points, but only their distance matrix, i.e. all n(n-1)/2 distances between any two different points. The objective is now the same as for PCA : to try to find a best two-dimensional representation.

(2) This distance matrix sometimes consists of distances measured in a different way (so-called non-Euclidian distances, see below). It also often happens that one considers similarity measures and, correspondingly, similarity matrices. The problem is still the same, but the greater the similarity between objects is, the closer to each other they have to be represented.

Techniques to deal with these situations are called '*multidimensional scaling techniques (MDS)*'.

I.5.3.1. Distances

Let $C = (c_{ij})$ be an (n,k)-matrix of raw data, where the n rows denote n points $C_i = (c_{i1},c_{i2},\ldots,c_{ik})$ in k-space, $i = 1,\ldots,n$. Let X be the set $\{C_1,C_2,\ldots,C_n\}$ of these n points.

A *metric* (or *distance function*) is a mapping $d : X \times X \to \mathbb{R}^+$ satisfying the following three requirements (axioms) :

(1) For every $x,y \in X$: $d(x,y) = 0$ if and only if $x = y$.
This axiom states that the distance between two points is zero only in the case in which these two points coincide.

(2) For every $x,y \in X$: $d(x,y) = d(y,x)$.
This equality expresses the requirement that the distance between point x and point y must be the same as the distance between point y and point x. This means that a distance function must be symmetric.

(3) For every $x,y,z \in X$: $d(x,y) \leq d(x,z) + d(z,y)$.
This inequality is referred to as the 'triangular inequality' and is illustrated in Fig.I.5.6.

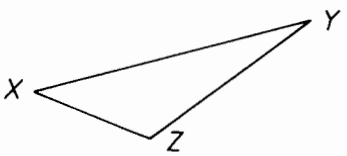

Fig.I.5.6 The distance between x and y is smaller than the sum of the distances between x and z, and y and z

A set X equiped with a metric d is denoted by (X,d) and is termed a '*metric space*'.

Examples.

a. The trivial distance function D_0 defined as $D_0(x,y) = 1$ if $x \neq y$ and $D_0(x,y) = 0$ if $x = y$. Although extremely simple, this is the underlying distance when perfectly matching pairs of vectors are sought (e.g. when searching for documents indexed by a fixed set of terms).

b. The Minkowski metric d_p, $p > 0$. This distance function is defined as :

$$d_p(C_i,C_j) = \left(\sum_{r=1}^{k} |c_{ir} - c_{jr}|^p \right)^{1/p} \quad . \qquad [I.5.7]$$

For $p = 2$ this produces the usual Euclidean metric. For $p = 1$, we obtain the so-called city-block metric :

$$d_1(C_i,C_j) = \sum_{r=1}^{k} |c_{ir} - c_{jr}| \quad , \qquad [I.5.8]$$

illustrated by Fig.I.5.7.

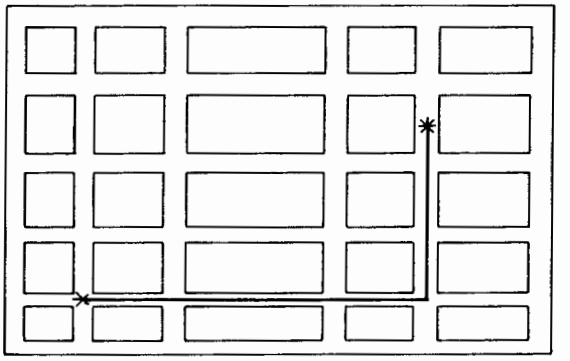

Fig.I.5.7 City-block metric

c. Chebycheff's distance, denoted by d_∞ :

$$d_\infty(C_i,C_j) = \max\{|c_{i1} - c_{j1}|,|c_{i2} - c_{j2}|,\ldots,|c_{ik} - c_{jk}|\} \ . \quad [I.5.9]$$

One can actually show that $\lim_{p\to\infty} d_p(C_i,C_j) = d_\infty(C_i,C_j)$. This explains the notation d_∞.

These generalised distances d_p and d_∞ are used by Salton, Fox and Wu (1983) in connection with a generalisation of the classical Boolean operators and by Egghe and Rousseau (1989, 1990) in studies on concentration and dispersion measures in informetrics and econometrics.

I.5.3.2. Similarities and dissimilarities

Distance measures can be considered as dissimilarity measures, in the sense that, intuitively speaking, the greater the distance between objects is the greater their dissimilarity will be. Stated formally, a function $d : X \times X \to \mathbb{R}^+$ is said to be a *dissimilarity function* if :

(1) for every x in X : d(x,x) = 0;

(2) for every x and y in X : d(x,y) = d(y,x) .

This clearly generalises the notion of distance since we have dropped the triangular inequality from the set of axioms. Moreover, two items can have a dissimilarity equal to zero without actually being the same.

A *similarity function* $s : X \times X \to [0,1]$ satisfies (2) above and $s(x,x) = 1$ (for every x in X). If s is a similarity function, $1 - s$ is a dissimilarity function. If d is a dissimilarity function, $\frac{2}{\pi}$ Arctg $(\frac{1}{d})$ is a similarity function.

Examples. Consider the following answers to a questionnaire of persons A and B (Table I.5.5; Y : yes, N : no).

Table I.5.5. Answers to a questionnaire

	Questions							
Person	1	2	3	4	5	6	7	8
A	N	Y	Y	N	N	Y	Y	N
B	Y	N	Y	N	Y	Y	N	Y

Table I.5.5 is then converted into the following contingency table (Table I.5.6).

Table I.5.6. Contingency table of data from Table I.5.5

		A	
		Y	N
B	Y	a = 2	b = 3
	N	c = 2	d = 1

Examples of measures describing the similarity between A and B are :

$$s_1(A,B) = \frac{a+d}{a+b+c+d} \qquad [I.5.10]$$

or

$$s_2(A,B) = \frac{a}{a+b+c} \qquad [I.5.11]$$

Other similarity measures such as Salton's cosine measure and the Jaccard index will be studied in connection with citation and cocitation analysis (Part III).

Standardisation

Different scalings of data may yield different results. Consider, for example, the following table (Table I.5.7).

Table I.5.7. Library data

Library	Number of loans (× 100)	Number of books (× 1000)
A	80	169
B	82	183
C	84	175

Using the Euclidean distance [I.5.7], p = 2) yields : $d_{AB} = 14.14$, $d_{AC} = 7.21$ and $d_{BC} = 8.25$ and hence $d_{AB} > d_{AC}$. This shows that libraries A and B are less similar than libraries A and C. However, using a different scaling results in Table I.5.8. Denoting the Euclidean distance for this situation by d' gives : $d'_{AB} = 2.005$, $d'_{AC} = 4.000$, $d'_{BC} = 2.002$. This leads to the contradictory result that $d'_{AB} < d'_{AC}$.

Table I.5.8. Library data - different scaling

Library	Number of loans (× 100)	Number of books (× 100000)
A	80	1.69
B	82	1.83
C	84	1.75

This contradiction (giving rise to useless results) is solved by standardising the data : every value in a column is divided by the standard deviation of this column. In the case of our example on library data we have : $s_{\{80,82,84\}} = 1.633$ and $s_{\{169,183,175\}} = 5.735$, resulting in Table I.5.9.

The same table would have been obtained when starting with the data in Table I.5.8. Now $d_{AB} = 2.73$ and $d_{AC} = 2.67$, showing that A and C are more similar than A and B. We wish to emphasise the fact that similarity is a relative notion. Similarity is determined with respect to the set of points (libraries, documents, persons) under study.

Table I.5.9. Library data : standardised values

Library	Number of loans (× 1.633)	Number of books (× 5735)
A	48.99	29.47
B	50.21	31.91
C	51.44	30.52

I.5.3.3. Principal coordinate analysis

This technique is also called *classical MDS* or *metric MDS*. It deals with problem (1) stated in the introduction to Section I.5.3. Basically, we can state the problem as follows : distances between cities (libraries, journals, scientists) are known and the problem is to reconstruct the map. More generally, this problem is posed in a space where the dimension k is also unknown. Part of the solution consists of finding a minimal k, such that the problem has a solution in $\mathbb{R}^k$.

We will not go into the mathematical details of the solution (requiring rather complex matrix techniques). In principle, one expects the solution to be a configuration of points in k-space, on which one can apply PCA. Most computer programs solve this in one step, immediately producing a two-dimensional representation. For more details on this method see Gower (1966) and Seber (1984).

I.5.3.4. Non-metric multidimensional scaling

This method tackles the second problem mentioned in the introduction to this section. In this case we have a dissimilarity matrix $D = (\delta_{ij})_{i,j=1,\ldots,n}$. Some trial and error quickly shows that finding a dimension k such that n points in $\mathbb{R}^k$ are situated exactly at distances δ_{ij} is asking too much. So we settle for the following : try to find n points in some $\mathbb{R}^k$ such that for every i,j,k,ℓ :

$$d_{ij} \leq d_{k\ell} \Rightarrow \delta_{ij} \leq \delta_{k\ell} \quad , \qquad [I.5.12]$$

where d_{ij} is the distance between the i^{th} and the j^{th} point in $\mathbb{R}^k$. The requirement in [I.5.12] is called a '*monotonicity constraint*'. Even this requirement may turn out to be too strong. In that case, try to satisfy [I.5.12] as well as possible. If [I.5.12] can be satisfied, we obtain an increasing graph in the (δ_{ij},d_{ij})-plane (see Fig.I.5.8).

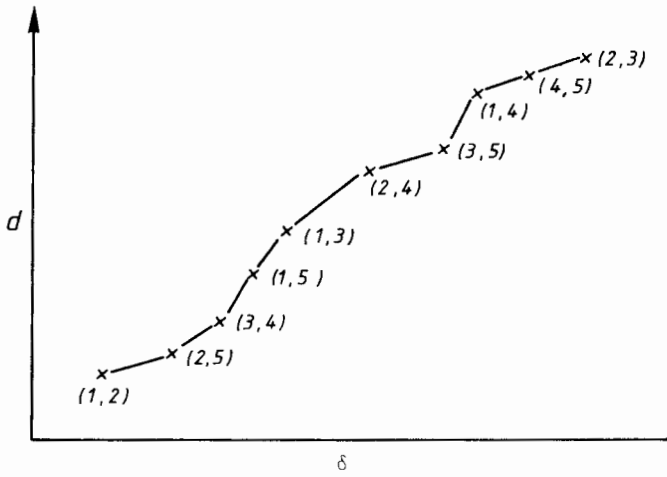

Fig.I.5.8 Scatter plot of distance versus dissimilarity where the monotonicity constraint is satisfied

How can this situation be obtained? We begin with n points $(Y_i)_{i=1,\ldots,n}$ in $\mathbb{R}^k$ (k also varies in reality, but for the sake of simplicity we will keep k fixed). All distances $d_{ij} = \|Y_i - Y_j\|$ are calculated and plotted against δ_{ij}. In general, this will not yield an increasing function. The intermediate values are taken, with same abscissae δ_{ij}, until an increasing function has been obtained. (In fact, this step is an application of some operations research techniques, cf. Part II). Several iterative steps are usually necessary. For more information we refer the reader to Shepard (1962a,b) or Kruskal (1964).

I.5.3.5. Examples

1. McGrath (1986) applied MDS to a problem in library design and library departmentalisation.

2. Small (1986) and Small and Garfield (1985) constructed maps of documents, scientific fields and researchers using MDS. The data were based on co-citation frequencies (cf. Part III). Co-citation data as a measure of similarity are also used by McCain (1986a,b) to plot (via MDS) authors, showing disciplines.

3. Engineering journals have been studied (also using citation data and MDS) by Miyamoto and Nakayama (1983).

MDS is frequently combined with cluster analysis, which is the topic of the next section.

I.5.4. Cluster analysis

Cluster analysis is one of the most popular multivariate techniques. Here as well, the starting point is the matrix C of raw data. The aim is again to obtain a two-dimensional representation of the k-dimensional scatterplot of n points. Hence cluster analysis falls into the category of dimensionality-reduction techniques (Kinnucan et al., 1987). Nevertheless, cluster analysis is mainly concerned with the recognition of natural groups (clusterings), rather than with the k-dimensional configuration itself.

The result of cluster analysis is a tree-like structure called a '*dendrogram*'. One is frequently interested in both configuration and clustering. Fortunately, many examples abound in the literature in which cluster analysis is combined with PCA, MDS or factor analysis (e.g. Small (1986), Small and Garfield (1985), Leydesdorff (1986), Leydesdorff and Zaal (1988)). An introduction to cluster analysis can be found in Hartigan (1975).

I.5.4.1. General principles

Although scientists have developed several different clustering techniques, many of these methods also have features in common. These general principles will be described in this subsection.

Let C be a matrix of raw data :

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & & & \\ \vdots & & & \\ c_{n1} & & & c_{nk} \end{pmatrix} . \qquad [I.5.13]$$

This matrix is then transformed into a standardised distance or dissimilarity matrix $D_1 = (d_{ij})$. For this matrix $d_{ii} = 0$ for $i = 1,\ldots,n$ and $d_{ij} = d_{ji}$ (D_1 is symmetric). At this point we consider every point as a separate cluster. In that case, larger clusters are formed, one by one, until we have obtained one cluster, containing all points.

Points i and j such that d_{ij} is the smallest non-zero entry in D_1 are combined into one cluster, denoted by (i,j). This is represented in Fig.I.5.9, where i = 1 and j = 2.

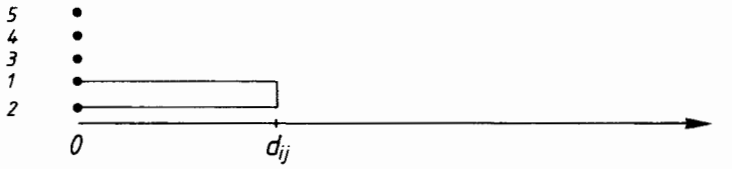

Fig.I.5.9 Dendrogram after the first clustering

A new matrix $D_2 = (d_{ij}^{(2)})$ is calculated : for points k,ℓ different from i and j, $d_{k\ell}^{(2)} = d_{k\ell}$. New d-values are computed for the distances from a point to the new cluster (i,j). Several different ways of doing this will be explained in the next subsection. Assume that D_2 looks like the following :

$$D_2 = \begin{array}{c} \\ (12) \\ (3) \\ \vdots \\ (n) \end{array} \begin{array}{c} \begin{array}{cccc} (12) & (3) & \ldots & (n) \end{array} \\ \begin{pmatrix} 0 & & & \\ d_{3,(12)} & 0 & & \\ \vdots & \vdots & & \\ d_{n,(12)} & d_{n3} & \ldots & 0 \end{pmatrix} \end{array} . \qquad [I.5.14]$$

In D_2 we take the smallest non-zero value, giving rise to the next cluster. There are two types of alternatives here : either a new cluster, say (45), is formed next to (12) or a new point, say 4, joins the cluster which has already been formed, producing cluster (124). Both alternatives are illustrated in Figs.I.5.10 and I.5.11.

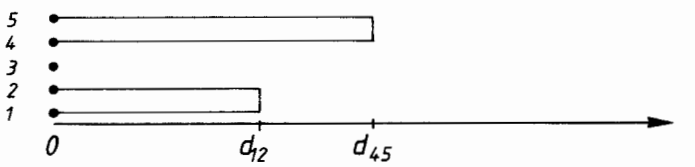

Fig.I.5.10 Dendrogram

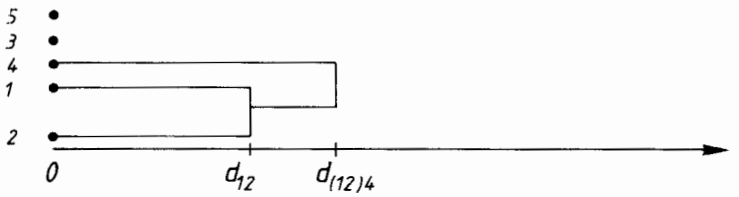

Fig.I.5.11 Dendrogram

This procedure continues until only one cluster remains, as illustrated in Fig.I.5.12 (based on Fig.I.5.10).

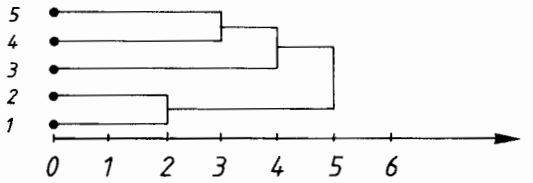

Fig.I.5.12 Dendrogram of a totally clustered set

Finally, the dendrogram is analysed to find natural clusters, preferably those that can be interpreted : this means trying to find a relatively long interval in which no clusters are formed. The dendrogram is cut at that point and natural clusters appear. For Fig.I.5.12 one can say that (12), (3) and (45) are three natural clusters. A procedure for cutting a dendrogram is called a '*stopping rule*'.

I.5.4.2. Cluster techniques that follow the general outline

The cluster techniques discusses in this section differ only in the way they define the distance between clusters.

I.5.4.2.1. Single link method (nearest neighbour method)

We will explain this method by means of a simple (unstandardised) example. Let $D = D_1$ be a dissimilarity (often a distance) matrix derived from a data

matrix C :

$$D_1 = \begin{array}{c} \\ (1) \\ (2) \\ (3) \\ (4) \\ (5) \end{array} \begin{array}{c} \begin{array}{ccccc} (1) & (2) & (3) & (4) & (5) \end{array} \\ \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array} .$$

Here $d_{12} = 2$ is the smallest value different from zero. Hence (1) and (2) are combined into one cluster. The distance between clusters is then defined as the smallest of all distances between elements of the first cluster and elements of the second cluster. (A point that has not yet been joined with another one is considered as a cluster consisting of one element.) This rule results in this case in :

$$d_{(12)3} = \min \{d_{13}, d_{23}\} = d_{23} = 5 ,$$
$$d_{(12)4} = \min \{d_{14}, d_{24}\} = d_{24} = 9 ,$$
$$d_{(12)5} = \min \{d_{15}, d_{25}\} = d_{25} = 8 ,$$

leading to the following new matrix :

$$D_2 = \begin{array}{c} \\ (12) \\ (3) \\ (4) \\ (5) \end{array} \begin{array}{c} \begin{array}{cccc} (12) & (3) & (4) & (5) \end{array} \\ \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 9 & 4 & 0 & \\ 8 & 5 & 3 & 0 \end{pmatrix} \end{array} .$$

In this matrix $d_{45} = 3$ is the smallest strictly positive value. This yields cluster (45). Then $d_{(12)(45)} = 8$ and $d_{3(45)} = 4$. The new D-matrix becomes :

$$D_3 = \begin{array}{c} \\ (12) \\ (3) \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (3) & (45) \end{array} \\ \begin{pmatrix} 0 & 0 & \\ 5 & 0 & \\ 8 & 4 & 0 \end{pmatrix} \end{array} .$$

Here $d_{3(45)} = 4$ is the smallest, yielding cluster (345), and $d_{(345)(12)} = 5$. Next,

$$D_4 = \begin{matrix} \\ (12) \\ (345) \end{matrix} \begin{matrix} \begin{matrix} (12) & (345) \end{matrix} \\ \begin{pmatrix} 0 & \\ 5 & 0 \end{pmatrix} \end{matrix} .$$

Lastly, (12) and (345) are clustered. This clustering procedure is illustrated as a dendrogram in Fig.I.5.12.

The main disadvantage of the *single link method* is that clusterings sometimes occur too soon, as illustrated in Fig.I.5.13. This tendency to form loosely bound clusters with little internal cohesion is called '*chaining*'.

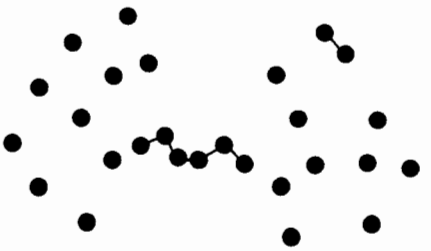

Fig.I.5.13 Chaining

I.5.4.2.2. Complete link method (furthest neighbour method)

This method differs from the preceding one in that distances between clusters are now defined as the maximum of all distances between elements of the first cluster and elements of the second cluster. Clusters themselves are still formed on the basis of the shortest 'distance' between clusters, just as in the single link method.

The next series of matrices illustrates the *complete link method* for matrix D in I.5.4.2.1. Fig. I.5.14 shows the corresponding dendrogram.

$$D_1 = \begin{array}{c} \\ (1) \\ (2) \\ (3) \\ (4) \\ (5) \end{array} \begin{array}{c} \begin{array}{ccccc} (1) & (2) & (3) & (4) & (5) \end{array} \\ \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array},$$

$$D_2 = \begin{array}{c} \\ (12) \\ (3) \\ (4) \\ (5) \end{array} \begin{array}{c} \begin{array}{cccc} (12) & (3) & (4) & (5) \end{array} \\ \begin{pmatrix} 0 & & & \\ 6 & 0 & & \\ 10 & 4 & 0 & \\ 9 & 5 & 3 & 0 \end{pmatrix} \end{array},$$

$$D_3 = \begin{array}{c} \\ (12) \\ (3) \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (3) & (45) \end{array} \\ \begin{pmatrix} 0 & & \\ 6 & 0 & \\ 10 & 5 & 0 \end{pmatrix} \end{array},$$

$$D_4 = \begin{array}{c} \\ (12) \\ (345) \end{array} \begin{array}{c} \begin{array}{cc} (12) & (345) \end{array} \\ \begin{pmatrix} 0 & \\ 10 & 0 \end{pmatrix} \end{array}.$$

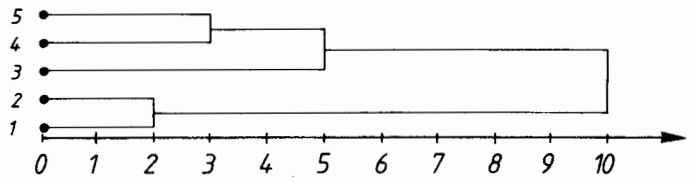

Fig.I.5.14 Dendrogram illustrating the furthest neighbour method

I.5.4.2.3. Group average clustering (average linking)

This is an intermediate method : the distance between clusterings is defined as an average. If cluster A is merged with cluster B, then the distance from cluster C to the new cluster (AB) is defined as the average of

all distances between all points from C and all points from (AB). Applied to our example, this *group average cluster method* results in the following matrices :

$$D_2 = \begin{array}{c} \\ (12) \\ (3) \\ (4) \\ (5) \end{array} \begin{array}{c} \begin{array}{cccc} (12) & (3) & (4) & (5) \end{array} \\ \begin{pmatrix} 0 & & & \\ 5.5 & 0 & & \\ 9.5 & 4 & 0 & \\ 8.5 & 5 & 3 & 0 \end{pmatrix} \end{array} ,$$

$$D_3 = \begin{array}{c} \\ (12) \\ (3) \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (3) & (45) \end{array} \\ \begin{pmatrix} 0 & & \\ 5.5 & 0 & \\ 9 & 4.5 & 0 \end{pmatrix} \end{array} ,$$

$$D_5 = \begin{array}{c} \\ (12) \\ (345) \end{array} \begin{array}{c} \begin{array}{cc} (12) & (345) \end{array} \\ \begin{pmatrix} 0 & \\ 7.83 & 0 \end{pmatrix} \end{array} .$$

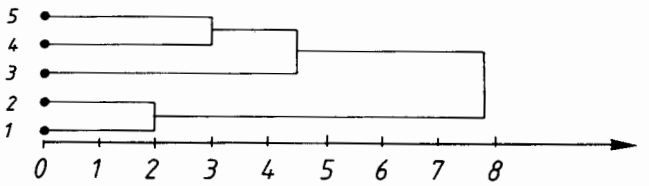

Fig.I.5.15 Dendrogram illustraging group average clustering

I.5.4.3. Ward's error sum of squares method (Ward (1963))

I.5.4.3.1. The method

This method differs slightly from those outlined in the above section. The general principle of reducing the number of clusters step by step remains, but distances (or dissimilarities) between points are defined rather than differences between clusters.

We will describe how to apply Ward's algorithm to go from K clusters to K-1 clusters.

Step 1. Take any two clusters and merge them into one cluster, yielding K-1 clusters, say $C_1, C_2, \ldots, C_{K-1}$.

Step 2. For cluster C_i, consisting of points $x_{i1}, \ldots, x_{in_2}$ (where $n_i = \#C_i$), define

$$X_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad . \qquad [I.5.15]$$

Point X_i is then the barycentre of this cluster (usually not a point of C_i).

Step 3. For i = 1 to K-1 take

$$ESS_i = \sum_{j=1}^{n_i} d(X_{ij}, X_j)^2 \ , \qquad [I.5.16]$$

where d is defined the same as for the determination of the distances between points and ESS stands for 'Error Sum of Squares'.

Step 4. Take

$$E = \sum_{i=1}^{K-1} ESS_i \quad . \qquad [I.5.17]$$

Step 5. Repeat this for every possible clustering of two K-clusters.

Step 6. Retain that clustering which minimises E.

Note that there are K(K-1)/2 possible combinations to consider in this one step!

I.5.4.3.2. An intuitive explanation : 'the trees and the wood'

In the perfectly unclustered situation one can say that all 'the trees' are completely visible, but 'the wood' is totally unknown. In this situation E = 0.

Constructing clusters brings 'the wood' in sight but 'the trees' fade away. However, Ward's method minimises the fading of the trees (E is minimal), leaving a maximum of information. Although the method finally clusters everything (we see only 'the wood'), the real clusters that remain will certainly have good properties. This explains - intuitively - why Ward's method is often felt to be the best one.

I.5.4.4. General properties of clustering methods

a. The clustering techniques discussed here all have the property that the linking distance at level j-1 is smaller than the linking distance at

level j. This is a good property, enhancing the evaluation possibilities of dendrograms.

b. The clustering techniques encountered here are all *agglomerative* in the sense that for every j, the partition of the n points on level j-1 is finer than the partition on level j.

c. Furthermore, these techniques are *hierarchical*, meaning that once level j has been passed, the partial dendrogram up to this level is no longer altered anymore by further internal clustering activities. Note that there are also nonhierarchical clustering methods (see e.g. the 'single-pass' iterative clustering methods described in Salton and McGill (1984; p.137)), but these seem to be generally less effective than hierarchical methods.

d. If the distance between points in a dendrogram is defined as the first level (on the axis) on which these two points appear in the same cluster, this distance, denoted by d', satisfies all axioms for a metric. Moreover, for every i,j :

$$d'(i,j) \leq \max_k (d'(i,k),d'(k,j)) \ . \qquad [I.5.18]$$

This type of metric is called an '*ultrametric*'. In this sense a clustering technique can be thought of as a transformation from a metric space into an ultrametric space. For a review on the history and the use of ultrametricity (and its recent introduction in physics) we refer the reader to Rammal et al. (1986).

I.5.4.5. Evaluation of cluster techniques

The techniques discussed above always yield clusters. If there is a large interval in which no clusters are formed, the evaluation is easy : see Fig. I.5.16. However, one is usually not so lucky!

Main problems in the evaluating of results of cluster techniques concern the occurrence of artefacts (clusters are mainly the result of the applied technique), the stability of cluster structures and the interpretation of the results (Braam et al. (1988)).

Shaw (1985) and Logan and Shaw (1987) have investigated the validity of clusters in co-citation and co-author graphs. They use as a null hypothesis the random-graph hypothesis. This hypothesis states that lines of a graph are selected randomly from the set of all possible lines. If lines are random, there is strong evidence that no clustering structure will exist in the data. Other tests are considered, such as these outlined in Strauss (1975). Dubes (1987) reported the results of a Monte-Carlo (simulation) study on estimating

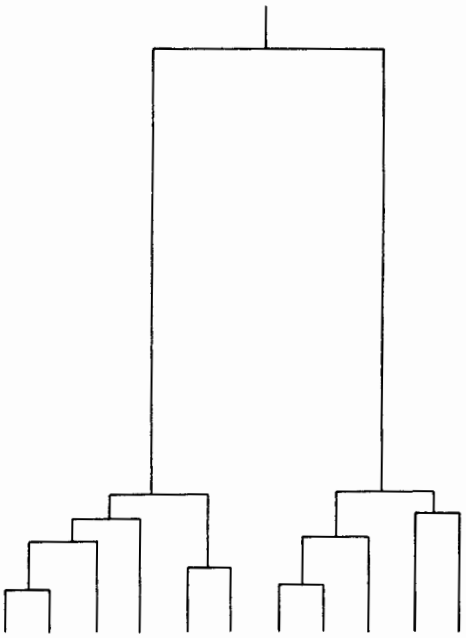

Fig.I.5.16 Dendrogram yielding two obvious clusters

the number of clusters using some specially constructed indexes. Among other things, he found that the complete link clustering method recognises the true number of clusters consistently better than the single link method.

An extensive critical review of clustering methods focusing mainly on their use in document retrieval systems has been written by Peter Willett (1988).

I.5.4.6. Examples

1. Fig.I.5.17 shows a dendrogram of botanics journals based on Ward's method (Keteleer (1986)). Botanics is not an isolated field, and subfields are not always clearly defined. When we cut the dendrogram along the dotted line, we find the clusters indicated in Fig.I.5.5.

2. Arms and Arms (1978) cluster journals in social science using citations. They conclude that cluster analysis on the basis of citations is not a practical method of designing secondary services in the social sciences. The group average method is used in Todorov and Vlachý (1986) to find groups of countries with a similar publication behaviour in physics. Pharmacology journals are clustered in Rousseau (1989b) on the basis of their 'importance'

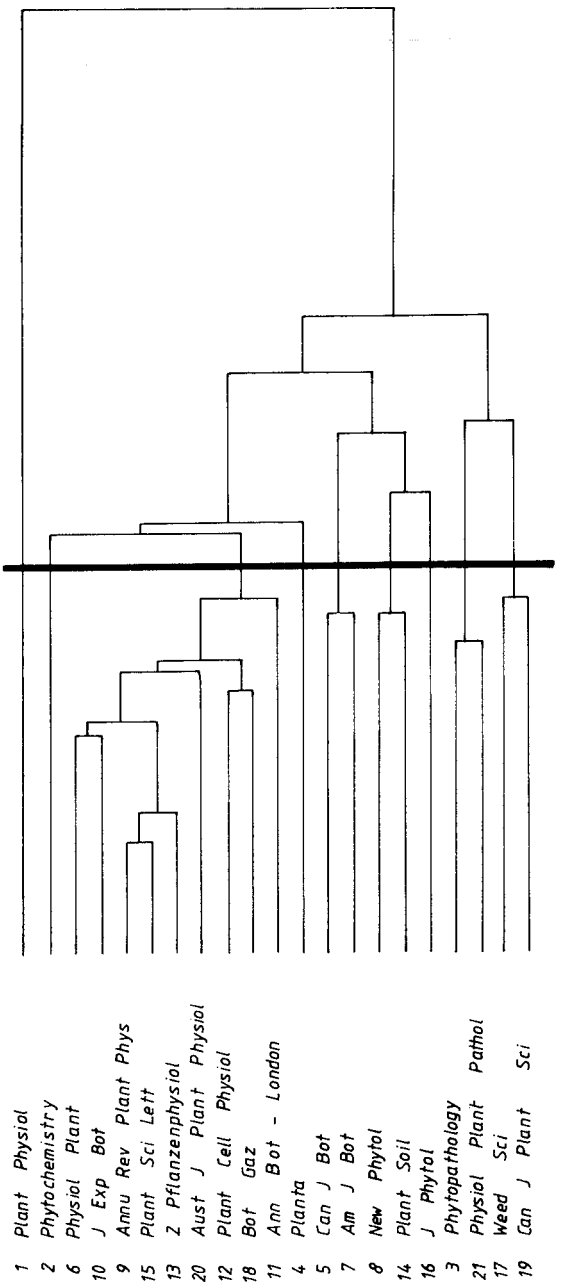

Fig.I.5.17 Dendrogram of botanical journals

(single link method).

3. An application concerning the automatic classification of documents can be found in Griffiths et al. (1984). In this case Ward's method performs best. In the same field but applied to retrieval we mention Jardine and Van Rijsbergen (1971), Croft (1977), Willett (1984), Voorhees (1986). These authors mainly use the single link method, but group average and complete link techniques are also discussed.

4. ISI's citation and co-citation studies mainly use the single link method (Small (1986), Small and Garfield (1985)). This approach, particularly the use of the single link method, is criticised by, for example, Leydesdorff (1987).

5. An evaluation of clustering methods can also be found in Mojena (1977).

I.5.4.7. Combination of MDS and cluster techniques

When these techniques are combined, multidimensional scaling or principal components analysis is used to obtain a two-dimensional image of an n-dimensional configuration. A cluster technique is applied independently, resulting in

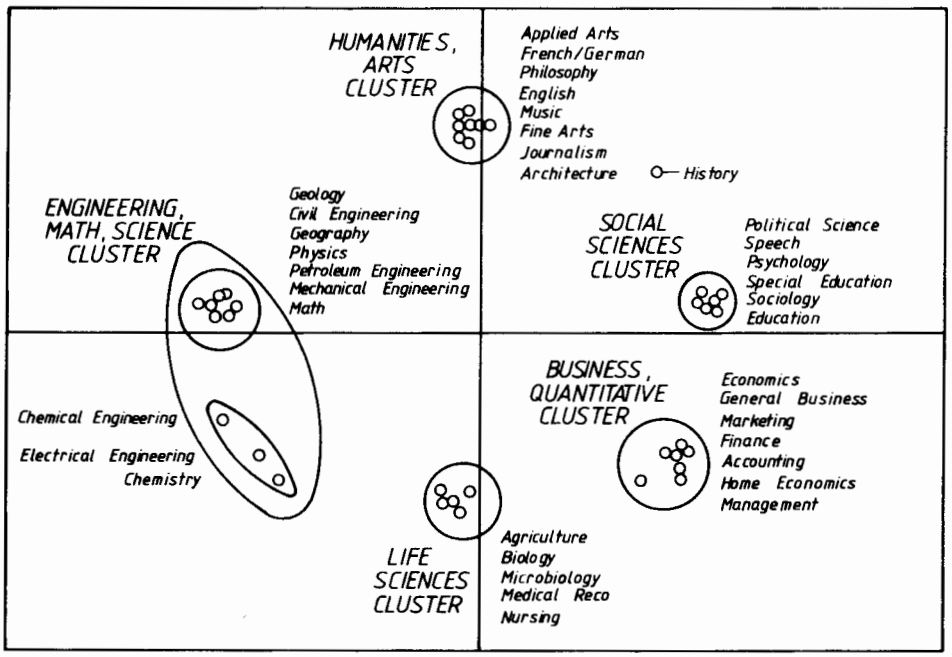

Fig.I.5.18 Library clusters (taken from J Acad Libr)

natural groupings. Instead of showing dendrograms, some programs draw Venn diagrams, in the two-dimensional image, around points belonging to the same cluster. This creates an optimal representation of the data (cf. Fig.I.5.5).

This combined technique is used, for example, in Keteleer (1986), Small (1986), Miyamoto and Nakayama (1983). An interesting example of this combined technique can be found in McGrath (1986). He studies the following problem : allocate libraries on a campus in such a way that every department is situated as close as possible to that library which contains the most books devoted to its field of investigation. An important constraint is, of course, the fact that there have to result the least possible number of libraries. MDS and cluster analysis based on circulation data of 37 academic disciplines yield five meaningful clusters, shown in Fig.I.5.18.