CHAPTER 16

Simultaneous Clustering and Scaling

16.1 INTRODUCTION

In the usual case when the data consist of a number of variables measured in different scales, it is necessary to express the variables in a common scale before distances between cases may be computed. A typical *ad hoc* rescaling requires all variables to have variance one or, more generally, requires every variable to make the same average contribution to the distance.

If the variables are $V(1), V(2), \ldots, V(N)$, then a scale V will be such that $V(I) = T(V, I)$, where $T(V, I)$ is a transformation of the common scale V to the variable $V(I)$. The transformation T will be linear for interval scale variables and monotonic for ordered variables. The variance standardizing transformation would be

$$V(I) = A(I)V + B(I),$$

where $A(I)$ is the standard deviation of $V(I)$ and $B(I)$ is the mean. In Table 16.1, relationships between the votes on various questions in the U.N. General Assembly (1969–1970) are tabulated. These show the necessity of various monotonic transformations to represent the responses on a common scale. For example, the relationship between $V1$ and $V3$ is essentially that the large yes vote on $V1$ has fragmented into yes, abstain, and no votes, in about equal proportions, for $V3$. A suitable common scale would take five values, 1, 2, 3, 4, 5, with

$$T(1, 1) = T(2, 1) = T(3, 1) = 1, \qquad T(4, 1) = 2, \qquad T(5, 1) = 3$$

and

$$T(1, 2) = 1, \qquad T(2, 2) = 2, \qquad T(3, 2) = T(4, 2) = T(5, 2) = 3.$$

In other words, the five values on the common scale correspond to values of $(V1, V3)$, successively: (1, 1), (1, 2), (1, 3), (2, 3), (3, 3).

Returning to the case of interval variables, there are serious defects with the method of equalizing variances. The principal one is that the variance calculation is very much affected by the presence of outliers or other clusters in the data. What is necessary is to continue rescaling as cluster information is exposed and to use standardizing techniques, such as equalizing interquartile ranges, that are not too sensitive to outliers or other clusters.

There follows a number of algorithms, of the joining type, for simultaneously clustering cases and variables while rescaling variables. These algorithms are different according to the type of variable being rescaled. The more difficult and intricate procedures necessary for combining different types of variables have been neglected.

Table 16.1 Relationship Between Votes on Various U.N. Questions (1969–1970)

		V1		
		1	2	3
V3	1	27	2	1
	2	23	3	1
	3	23	0	44

SHIFT OF V1 YES TOWARD V4 NO .

		V2		
		1	2	3
V3	1	2	6	22
	2	7	10	10
	3	54	13	0

REVERSAL V2 YES TO V4 NO .

		V4		
		1	2	3
V5	1	11	0	0
	2	0	26	2
	3	0	4	69

IDENTICAL QUESTIONS.

		V4		
		1	2	3
V1	1	2	22	41
	2	0	1	3
	3	9	8	26

WEAK RELATIONSHIPS.

1, yes; 2, abstain; 3, no. $V1$, declare the China admission question an important question; $V2$, to make the study commission on China admission "important"; $V3$, to form a study commission on China admission; $V4$, replace last paragraphs of preamble, on South Africa expulsion from UNCTAD, by Hungarian amendment; $V5$, adopt the Hungarian amendment of paragraph 1 and 2 on South Africa expulsion.

16.2 SCALING ORDERED VARIABLES

Preliminaries. Given two ordered variables X and Y taking values $\{X(I), Y(I), 1 \leq I \leq M\}$ on M cases, it is desired to find a scale Z, an ordered variable taking values $\{Z(I), 1 \leq I \leq M\}$, and monotonic transformations $T(Z, 1)$ and $T(Z, 2)$ of Z, such that $X(I) = T[Z(I), 1]$ and $Y(I) = T[Z(I), 2]$ with maximum frequency.

Let the values taken by X be the integers $1, 2, \ldots, N1$, let the values taken by Y be the integers $1, 2, \ldots, N2$, and let $N(I, J)$ denote the number of cases with values $X = I$ and $Y = J$. The variable Z will take values $[I(1), J(1)], \ldots, [I(K), J(K)]$, where $I(1) \leq I(2) \leq \cdots \leq I(K)$ and $J(1) \leq J(2) \leq \cdots \leq J(K)$ or $J(1) \geq J(2) \geq \cdots \geq J(K)$, and $N[I(1), J(1)] + N[I(2), J(2)] + \cdots + N[I(K), J(K)]$ is a maximum. [The transformations are $T[I(L), J(L), 1] = I(L)$, and $T[I(L), J(L), 2] = J(L)$.] The algorithm uses a maximization technique similar to dynamic programming. Let $\mathrm{NMAX}(I, J)$ denote the maximum value of $N[I(1), J(1)] + N[I(2), J(2)] + \cdots + N[I(K), J(K)]$ subject to the constraints $1 \leq I(1) \leq I(2) \leq \cdots \leq I(K) \leq I$ and $1 \leq J(1) \leq J(2) \leq \cdots \leq J$. Then $\mathrm{NMAX}(I, J) = N(I, J) + \max [\mathrm{NMAX}(I, J - 1), \mathrm{NMAX}(I - 1, J)]$. In this way, an optimal sequence increasing in I and J, connecting $(1, 1)$ to (N_1, N_2), is discovered. Similarly, discover an optimal sequence increasing in I but decreasing in J.

STEP 1. Compute $N(I, J)$, the number of times variable X takes value I and variable Y takes value J. Set

$$\mathrm{NMAX}(0, J) = \mathrm{NMAX}(I, 0) = 0 \qquad (1 \leq J \leq N1, 1 \leq I \leq N2).$$

STEP 2. For each $J(1 \leq J \leq N2)$, compute for each $I(1 \leq I \leq N1)$ $\text{NMAX}(I, J) = N(I, J) + \max [\text{NMAX}(I, J-1), \text{NMAX}(I-1, J)]$.

STEP 3. Set $L = N1 + N2 - 1$, $I(L) = N1$, $J(L) = N2$.

STEP 4. By definition,

$$\text{NMAX}[I(L), J(L)] = N[I(L), J(L)] + \text{NMAX}[I(L), J(L) - 1]$$

or

$$\text{NMAX}[I(L), J(L)] = N[I(L), J(L)] + \text{NMAX}[I(L) - 1, J(L)].$$

In the first case $I(L-1) = I(L), J(L-1) = J(L) - 1$, and in the second case $I(L-1) = I(L) - 1, J(L-1) = J(L)$. If $L = 2$, go to step 5. Otherwise, decrease L by 1 and repeat this step.

STEP 5. Define a new variable U by $U = N2 - J + 1$ when $Y = J$. Discover the optimal monotonic relationship between X and U, following Steps 1–4. If

Table 16.2 Scaling Components of Mammal's Milk

					Reversed Protein	
	ANIMAL	WATER %	PROTEIN %	NMAX	JMAX	NMAX	JMAX
1.	Dolphin-----	44.9	10.6	1	-	1	-
2.	Seal--------	46.4	9.7	1	-	2	1
3.	Reindeer----	64.8	10.7	2	2	2	4
4.	Whale-------	64.8	11.1	3	3	1	-
5.	Deer--------	65.9	10.4	2	2	3	3
6.	Elephant----	70.7	3.6	1	-	4	5
7.	Rabbit------	71.3	12.3	4	4	1	-
8.	Rat---------	72.5	9.2	2	6	4	5
9.	Dog---------	76.3	9.3	3	8	4	5
10.	Cat---------	81.6	10.1	4	9	4	5
11.	Fox---------	81.6	6.6	2	6	5	9
12.	Guinea Pig--	81.9	7.4	3	11	5	9
13.	Sheep-------	82.0	5.6	2	6	6	12
14.	Buffalo-----	82.1	5.9	3	13	6	12
15.	Pig---------	82.8	7.1	4	14	6	12
16.	Zebra-------	86.2	3.0	1	-	7	15
17.	Llama-------	86.5	3.9	2	16	7	15
18.	Bison-------	86.9	4.8	3	17	7	15
19.	Camel-------	87.7	3.5	2	16	8	18
20.	Monkey------	88.4	2.2	1	-	9	19
21.	Orangutan---	88.5	1.4	1	-	10	20
22.	Mule--------	90.0	2.0	2	21	10	20
23.	Horse-------	90.1	2.6	3	22	9	19
24.	Donkey------	90.3	1.7	2	21	11	22

Data ordered by water percentage.

NMAX($N1$, $N2$) is larger for U and X than for Y and X, the optimal relationship overall is increasing in X and decreasing in Y.

NOTE If the ordered variables X and Y take a very large number of different values, the contingency table $N(I, J)$ will mostly consist of 0's and 1's and will be rather expensive to store and manipulate. Suppose that the variables X and Y take the values $\{X(I), Y(I)\}$ $N(I)$ times. Assume the data ordered so that $X(I) \leq X(J)$ if $I \leq J$, and $Y(I) < Y(J)$ if $X(I) = X(J)$ and $I < J$. The quantity NMAX(I) is the maximum value of $N[I(1)] + \cdots + N[I(K)]$ subject to

$$X[I(1)] \leq X[I(2)] \leq \cdots \leq X[I(K)] = X(I)$$

and

$$Y[I(1)] \leq Y[I(2)] \leq \cdots \leq Y[I(K)] = Y(I).$$

Compute NMAX(I) iteratively ($1 \leq I \leq M$), setting

$$\text{NMAX}(I) = \max_J [\text{NMAX}(J)] + N(I),$$

where $X(J) \leq X(I)$ and $Y(J) \leq Y(I)$ and $I \neq J$. The quantity JMAX(I) is the value of J which maximizes NMAX(J) under the above constraint. The sequence of Z values is $I(1)$, maximizing NMAX(I), then $I(2)$ = JMAX[$I(1)$], $I(3)$ = JMAX[$I(2)$], and so on.

This algorithm is applied to mammal's milk components in Table 16.2. A scale is computed with water and protein both increasing and also with water increasing and protein decreasing. The second relationship is preferred since 11 of 24 points are covered in the fitting curve. The curve is graphed in Figure 16.1.

16.3 SCALING ORDERED VARIABLES APPLIED TO U.N. QUESTIONS

The questions to be scaled are $V2$ and $V3$ as given in Table 16.1, two questions on a study commission on the China admission question.

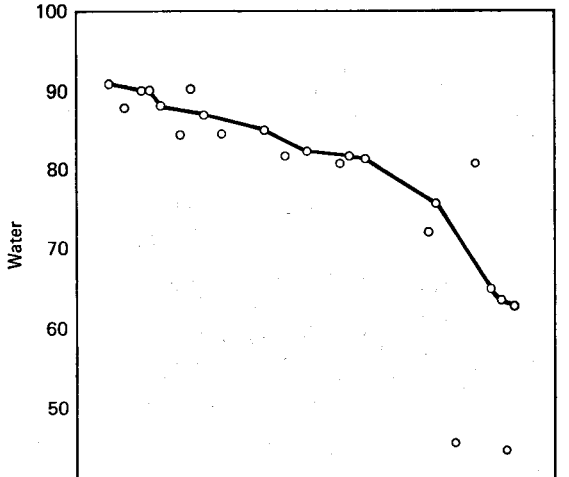


Figure 16.1 Monotonic scale for mammals' milk.

STEP 1. In the terminology of the algorithm, $X = V3$ and $Y = V2$. Then $N(1, 1) = 2$, $N(1, 2) = 6$, $N(1, 3) = 22$, $N(2, 1) = 7$, $N(2, 2) = 10$, $N(2, 3) = 10$, $N(3, 1) = 54$, $N(3, 2) = 13$, $N(3, 3) = 0$. $\text{NMAX}(0, 1) = \text{NMAX}(0, 2) = \text{NMAX}(0, 3) = 0$, $\text{NMAX}(1, 0) = \text{NMAX}(2, 0) = \text{NMAX}(3, 0) = 0$.

STEP 2 First $\text{NMAX}(1, 1) = 2$, then

$$\text{NMAX}(2, 1) = N(2, 1) + \max [\text{NMAX}(1, 1), \text{NMAX}(2, 0)]$$
$$= 7 + 2 = 9.$$

$\text{NMAX}(3, 1) = 63$, $\text{NMAX}(1, 2) = 8$, $\text{NMAX}(2, 2) = 19$, $\text{NMAX}(3, 2) = 76$, $\text{NMAX}(1, 3) = 30$, $\text{NMAX}(2, 3) = 40$, $\text{NMAX}(3, 3) = 76$.

STEP 3. Set $L = 5$, $I(5) = 3$, $J(5) = 3$.

STEP 4. Since $\text{NMAX}(3, 3) = 0 + \text{NMAX}(3, 2)$, $I(4) = 3$, $J(4) = 2$. Decrease L to 4, since $\text{NMAX}(3, 2) = N(3, 2) + \text{NMAX}(3, 1)$. Therefore $I(3) = 3$, $J(3) = 1$. Similarly, $I(2) = 2$, $J(2) = 1$ and $I(1) = 1$, $J(1) = 1$. The final optimal increasing sequence is thus (1, 1), (2, 1), (3, 1), (3, 2), (3, 3).

STEP 5. Define the variable U: $U = 1$ if $V2 = 3$, $U = 2$ if $V2 = 2$, $U = 3$ if $V2 = 1$. Repeating Steps 1–4, discover the sequence (1, 3), (1, 2), (2, 2), (3, 2), (3, 1), which covers 109 points. This sequence is thus preferred to the sequence increasing in I and J. The final scale is $Z = (1, 3), (1, 2), (2, 2), (3, 2), (3, 1)$ with $T[(I, J), 2] = J$. The function $T(Z, 1)$ is increasing; the function $T(Z, 2)$ is decreasing.

A number of such scales, which can be computed very quickly by hand for ordered variables taking just a few values, are given in Table 16.4.

16.4 JOINER SCALER

Preliminaries. The data matrix $\{A(I, J), 1 \leq J \leq M, 1 \leq J \leq N\}$ is a collection of N ordered variables measured on different scales. During the algorithm's execution, pairs of the variables are joined to form new variables, pairs of cases are joined to form new cases, and a common scale for all variables is constructed.

The output consists of data clusters within which all values are equal when expressed in the common underlying scale. The data clusters $1, 2, \ldots,$ KD are determined by the corresponding row and column clusters $\text{IR}(I)$, $\text{IC}(I)$ for the Ith clusters. The tree structure of the row clusters $1, 2, \ldots,$ KR is determined by the function $\text{JR}(I)$ which is the smallest row cluster properly including cluster I. The tree structure of the column clusters $1, 2, \ldots,$ KC is determined by the function $\text{JC}(I)$, which is the smallest column cluster properly including cluster I.

STEP 1. Set $\text{KR} = M$, $\text{KC} = N$, $\text{JR}(I) = 0$ $(1 \leq I \leq M)$, and $\text{JC}(I) = 0$ $(1 \leq I \leq N)$.

STEP 2. Compute distances between all pairs of row clusters I, J $[1 \leq I, J \leq \text{KR}$, $\text{JR}(I) = \text{JR}(J) = 0]$ as the proportion of columns in which $A(I, K) \neq A(J, K)$, among columns for which $A(I, K)$ and $A(J, K)$ are both defined and in which $\text{JC}(K) = 0$. Let the smallest distance be DROW and the corresponding rows be IROW, JROW.

Table 16.3 Mammal's Milk

	WATER	PROTEIN	FAT	LACTOSE	ASH
Horse	90.1	2.6	1.0	6.9	0.35
Orangutan	88.5	1.4	3.5	6.0	0.24
Monkey	88.4	2.2	2.7	6.4	0.18
Donkey	90.3	1.7	1.4	6.2	0.40
Hippo	90.4	0.6	4.5	4.4	0.10
Camel	87.7	3.5	3.4	4.8	0.71
Bison	86.9	4.8	1.7	5.7	0.90
Buffalo	82.1	5.9	7.9	4.7	0.78
Guinea Pig	81.9	7.4	7.2	2.7	0.85
Cat	81.6	10.1	6.3	4.4	0.75
Fox	81.6	6.6	5.9	4.9	0.93
Llama	86.5	3.9	3.2	5.6	0.80
Mule	90.0	2.0	1.8	5.5	0.47
Pig	82.8	7.1	5.1	3.7	1.10
Zebra	86.2	3.0	4.8	5.3	0.70
Sheep	82.0	5.6	6.4	4.7	0.91
Dog	76.3	9.3	9.5	3.0	1.20
Elephant	70.7	3.6	17.6	5.6	0.63
Rabbit	71.3	12.3	13.1	1.9	2.30
Rat	72.5	9.2	12.6	3.3	1.40
Deer	65.9	10.4	19.7	2.6	1.40
Reindeer	64.8	10.7	20.3	2.5	1.40
Whale	64.8	11.1	21.2	1.6	1.70
Seal	46.4	9.7	42.0	-	0.85
Dolphin	44.9	10.6	34.9	0.9	0.53

From *Handbook of Biological Data* (1956), William S. Spector, ed., Saunders.

STEP 3. Compute distances between each pair of columns by finding the monotonic scale which covers most rows, as in the algorithm for scaling ordered variables. (Look only at columns I, J for which $\mathrm{JC}(I) = \mathrm{JC}(J) = 0$, and look only at rows K for which $\mathrm{JR}(K) = 0$ and for which $A(I, K)$ and $A(J, K)$ are both defined.) The distance between I and J is the number of rows not covered by the monotonic scale, divided by the total number of rows considered less 2. (The reason for the 2 is that two rows will always be covered.) Let the smallest distance be DCOL and the corresponding columns be ICOL, JCOL.

STEP 4. (If the minimum of DCOL and DROW is 1, go to Step 6.) If DCOL $<$ DROW, go to Step 5. Otherwise increase KR by 1, $\mathrm{JR}(\mathrm{IROW}) = \mathrm{JR}(\mathrm{JROW}) = \mathrm{KR}$, $\mathrm{JR}(\mathrm{KR}) = 0$. For each column K $[1 \leq K \leq \mathrm{KC},\ \mathrm{JC}(K) = 0]$, set $A(\mathrm{KR}, K) = A(\mathrm{IROW}, K)$ if $A(\mathrm{IROW}, K) = A(\mathrm{JROW}, K)$. If $A(\mathrm{IROW}, K)$ is undefined, set $A(\mathrm{KR}, K) = A(\mathrm{JROW}, K)$. If $A(\mathrm{IROW}, K) \neq A(\mathrm{JROW}, K)$, leave $A(\mathrm{KR}, K)$ undefined and define data clusters KD + 1 and KD + 2 by $\mathrm{IR}(\mathrm{KD} + 1) = \mathrm{IROW}$, $\mathrm{IC}(\mathrm{KD} + 1) = K$, $\mathrm{IR}(\mathrm{KD} + 2) = \mathrm{JROW}$, $\mathrm{IC}(\mathrm{KD} + 2) = K$. Increase KD by 2

and go to the next column cluster K. If all column clusters have been adjusted, return to Step 2.

STEP 5. Increase KC by 1, define JC(ICOL) = JC(JCOL) = JC, JC(KC) = 0. For each row cluster K [$1 \leq K \leq$ KR, JR(KR) = 0] define $A(K, \mathrm{KC})$ to be the value in the new scale corresponding to $A(K, \mathrm{JCOL})$ and $A(\mathrm{K}, \mathrm{JCOL})$] if this value is uniquely defined. Otherwise, define data clusters KD + 1 and KD + 2 by IR(KD + 1) = K, IC(KD + 1) = ICOL, IR(KC + 2) = K, JC(KD + 2) = JCOL, and increase KD by 2. Return to Step 2.

STEP 6. A single underlying scale has been constructed with monotonic functions from this scale to each original variable. Within each data cluster, consider the data values that are not included in some smaller cluster. Each such data value corresponds to a range of scale values. The intersection of these ranges is always nonempty, and this intersection range is recorded for each data cluster.

Beginning with the largest clusters and moving toward the smaller, eliminate a cluster I if the smallest cluster containing it has an intersection range which includes the intersection range for I. Otherwise, change the intersection range for I to be the smallest value in the range.

Table 16.4 Scaling U.N. Questions

V1

V3	1	2	3
1	27	2	1
2	23	3	1
3	23	0	44

V2

V3	1	2	3
1	2	6	22
2	7	10	10
3	54	13	0

V4

V5	1	2	3
1	11	0	0
2	0	26	2
3	0	4	69

V4

V1	1	2	3
1	2	22	41
2	0	1	3
3	9	8	26

Blocks are different values of constructed scale.

16.5 APPLICATION OF JOINER-SCALER ALGORITHM TO U.N. VOTES

It is natural to apply a two-way clustering algorithm to the U.N. votes (Table 16.5) because there are blocs of countries such as Bulgaria, Romania, and the USSR that vote similarly, and blocs of questions that arise from the same issue, such as "China admission," "importance of China admission," "study China admission," "importance of studying China admission."

Table 16.5 Selected Votes in the United Nations (1969–1970)

Y = YES N = NO A = ABSTAIN

	1	2	3	4	5	6	7	8	9	10
1. CANADA	N	A	Y	A	N	A	A	Y	Y	Y
2. CUBA	Y	A	N	Y	Y	N	Y	A	N	N
3. MEXICO	N	Y	Y	N	N	Y	Y	A	A	Y
4. UNITED KINGDOM	N	N	Y	Y	N	A	N	A	Y	Y
5. NETHERLANDS	N	N	Y	A	N	Y	A	A	Y	Y
6. FRANCE	N	A	N	Y	A	N	A	A	Y	Y
7. SPAIN	N	A	Y	N	Y	Y	A	A	A	Y
8. PORTUGAL	A	N	A	A	A	A	N	N	Y	Y
9. POLAND	Y	Y	N	Y	A	N	Y	Y	A	A
10. AUSTRIA	N	A	A	A	A	A	A	Y	Y	Y
11. HUNGARY	Y	Y	N	Y	Y	N	Y	Y	A	A
12. CZECHOSLOVAKIA	Y	Y	N	Y	A	N	Y	Y	A	A
13. ITALY	N	A	Y	N	N	Y	A	A	Y	Y
14. BULGARIA	Y	Y	N	Y	Y	N	Y	Y	A	A
15. ROMANIA	Y	Y	N	Y	Y	N	Y	Y	A	A
16. USSR	Y	Y	N	Y	A	N	Y	Y	A	A
17. FINLAND	A	A	N	Y	A	N	A	Y	Y	Y
18. GAMBIA	N	A	Y	N	A	N	A	A	A	A
19. MALI	A	Y	N	Y	Y	N	A	Y	N	N
20. SENEGAL	A	Y	Y	A	A	A	Y	Y	N	N
21. DAHOMEY	A	Y	Y	N	Y	N	Y	Y	N	N
22. NIGERIA	N	Y	Y	N	Y	N	Y	Y	N	N
23. IVORY COAST	N	Y	Y	N	Y	N	Y	Y	A	A
Y/N/A	7/11/5	12/3/8	11/10/2	11/7/5	9/5/9	4/14/5	12/2/9	14/1/8	8/5/10	10/5/8

Columns: 1, to adopt USSR proposal to delete item on Korea unification; 2, to call upon the UK to use force against Rhodesia; 3, declare the China admission question an important question; 4, recognize mainland China and expel Formosa; 5, to make study commission on China admission important; 6, to form a study commission on China admission; 7, convention on no statutory limits on war crimes; 8, condemn Portuguese colonialism; 9, defer consideration of South Africa expulsion; 10, South Africa expulsion is important question.

Also, the questions must be rescaled. For example, "importance of China admission" and "study China admission" are similar questions translated on an underlying scale, so that some of the yes votes on "importance" become abstains on "study." The other two China questions are very similar in producing opposite votes from almost every country.

STEP 1. To initialize, set KR = 23, KC = 10, JR(I) = 0 ($1 \leq I \leq 23$), and JC(I) = 0 ($1 \leq I \leq 10$).

STEP 2. Compute the distance between all pairs of rows. For example, row 1 and row 2 match in just one vote, so the distance between Canada and Cuba is $\frac{9}{10}$. The smallest row distance (there are several, and one is chosen arbitrarily) is DROW = 0, IROW = 12 (Czechoslovakia), JROW = 16 (USSR).

STEP 3. Find the monotonic scale for columns 1 and 2 that covers most rows. This is done by using the previous algorithm of Section 16.2. The optimal scale has five values, YY, AY, NY, YA, NN, which cover 18 of 23 rows. The distance between columns 1 and 2 is therefore $1 - \frac{5}{21}$. (Note that 2 is subtracted from the 23, because a monotonic scale always covers two points for free. This becomes important in the later stages of the algorithm when just a few rows and columns remain.) Examining all pairs of columns, discover DCOL = 0 for ICOL = 9, JCOL = 10.

STEP 4. Increase KR to 24, define JR(12) = 24, JR(16) = 24, JR(24) = 0. Since rows 12 and 16 are identical, row 24 is the same as row 12. Return to Step 2, and amalgamate rows 24 and 9 to be row 25, rows 14 and 15 to be row 26, and rows 11 and 26 to be row 27. On the next return to Step 2, columns 9 and 10 are closer than any pair of rows, and Step 5 is taken.

STEP 5. Increase KC to 11, JC(9) = JC(10) = 11, JC(11) = 0. The monotonic scale takes values 1, 2, 3, 4, corresponding to the pairs (Y, Y), (A, Y), (A, A), (N, N). Note that this sequence is monotonic in both variables. All pairs of votes fall in one of these four categories, so no data clusters are formed.

STEP 4 REPEATED. The next closest pair of row or column clusters are rows 21 and 22. Set KR = 28, JR(21) = JR(22) = 28, JR(28) = 0. Define $A(28, K) = A(21, K)$ except for $K = 1$, since $A(21, 1) \neq A(22, 1)$. Define two data clusters by IR(1) = 21, IR(2) = 22, IC(1) = 1, IC(2) = 1, and increase KD to 2. The algorithm continues in this way until a single column remains and several row clusters which are a distance of 1 from each other. The data clusters at this stage are given in Table 16.6. Also all original variables are monotonic functions of the scale of the column cluster which replaced them. These column clusters are joined, pairwise, with other column clusters till a single column cluster remains. All original variables will be monotonic functions of the scale of this final column cluster, given in Table 16.7.

STEP 6. Each data cluster generates a range of scale values, the intersection of the ranges of scale values over all values in the cluster. Consider, for example, the data cluster corresponding to rows 1 and 17 and columns 3, 4, 6, 1, 2, 5. The data values which are not included in smaller clusters are A, N, A, N for row 17 and columns 6, 1, 2, 5. From Table 16.7, these correspond to ranges of final scale values, 5–8, 6-E, 4-7, 7-E. The intersection of these ranges is the value 7. Such a value is associated with every data cluster.

For some data clusters, the intersection range includes that of the next largest cluster, and the data cluster is deleted. For example, the data cluster rows 21, 22, 23 by columns 7, 8, 9, 10 has intersection range C-E which includes that of the next largest cluster, rows 3–23 by columns 3–10, intersection range *C*. This data cluster is deleted.

Beginning with the largest clusters, every cluster is either deleted or has its intersection range replaced by the smallest value in it. A single scale value is thus associated with each remaining cluster, as in Table 16.8. The original data is recoverable from this representation in 41 data clusters, using the scale-to-variable transformations in Table 16.7.

The clusters of countries are {Senegal}, {African bloc}, {Netherlands, Italy}, {Soviet bloc}, {fringe neutrals}, {Portugal}, and {United Kingdom}. The clusters of questions are {China questions}, {African questions}.

Table 16.6 Preliminary Data Clusters in Applying Joiner-Scaler Algorithm to U.N. Data

	3	4	6	1	2	5	7	8	9	10
20	Y	A	A	A	Y	A	Y	Y	N	N
3	Y	N	Y	N	Y	N	Y	A	A	Y
7	Y	N	Y	N	A	Y	A	A	A	Y
6	Y	N	N	N	A	A	A	A	A	A
18	Y	N	N	N	A	A	A	A	Y	Y
21	Y	N	N	A	Y	Y	Y	Y	N	N
22	Y	N	N	N	Y	Y	Y	Y	N	N
23	Y	N	N	N	Y	Y	Y	Y	A	A
5	Y	A	Y	N	N	N	A	A	Y	Y
13	N	Y	Y	N	A	N	A	A	Y	Y
19	N	Y	N	A	Y	Y	A	Y	N	N
2	N	Y	A	Y	A	Y	Y	A	N	N
11	N	Y	A	Y	Y	Y	Y	Y	A	A
14	N	Y	A	Y	Y	Y	Y	Y	A	A
15	N	Y	A	Y	Y	Y	Y	Y	A	A
9	N	Y	A	Y	Y	A	Y	Y	A	A
12	N	Y	A	Y	Y	A	Y	Y	A	A
16	N	Y	A	Y	Y	A	Y	Y	A	A
10	A	A	A	N	A	A	A	Y	Y	Y
17	Y	N	N	A	A	A	A	Y	Y	Y
1	Y	A	A	N	A	N	A	Y	Y	Y
8	A	A	A	A	N	A	N	N	Y	Y
4	Y	Y	A	N	N	N	N	A	Y	Y

16.6 THINGS TO DO

16.6.1 Running the Joiner Scaler

The algorithm assumes that given variables are obtained by monotonic transformation from some underlying scale to be discovered in the course of the algorithm. It thus produces results invariant under monotonic transformation of the variables.

Table 16.7 Common Scale for All U.N. Questions, Output of Joiner-Scaler

SCALE.	1	2	3	4	5	6	7	8	9	B	C	D	E
QUESTION 1.	Y	A	A	A	A	N	N	N	N	N	N	N	N
2.	Y	Y	Y	A	A	A	A	N	N	N	N	N	N
3.	N	N	A	A	A	A	Y	Y	Y	Y	Y	Y	Y
4.	Y	Y	A	A	A	A	A	N	N	N	N	N	N
5.	Y	Y	A	A	A	A	N	N	N	N	N	N	N
6.	N	N	N	N	A	A	A	A	Y	Y	Y	Y	Y
7.	N	N	N	N	N	N	N	N	A	A	Y	Y	Y
8.	N	N	N	N	N	A	A	A	A	Y	Y	Y	Y
9.	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	A	A	N
10.	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	A	N

It is expensive to use if each variable takes many different values. The time is proportional to $M^2K^2N^2$, where K is the number of different values taken by each variable, averaged over different variables. In using it with continuous variables, it is suggested that you reduce the number of different values taken by each variable to between 5 and 10.

16.6.2 Monotonic Subsequences

For any sequence of length n, show that there is an increasing subsequence of length r and a decreasing subsequence of length s, such that $rs \geq n$. Thus for any n points in two dimensions, there is a monotone curve passing through at least $\sqrt{n}$.

16.6.3 Category Data

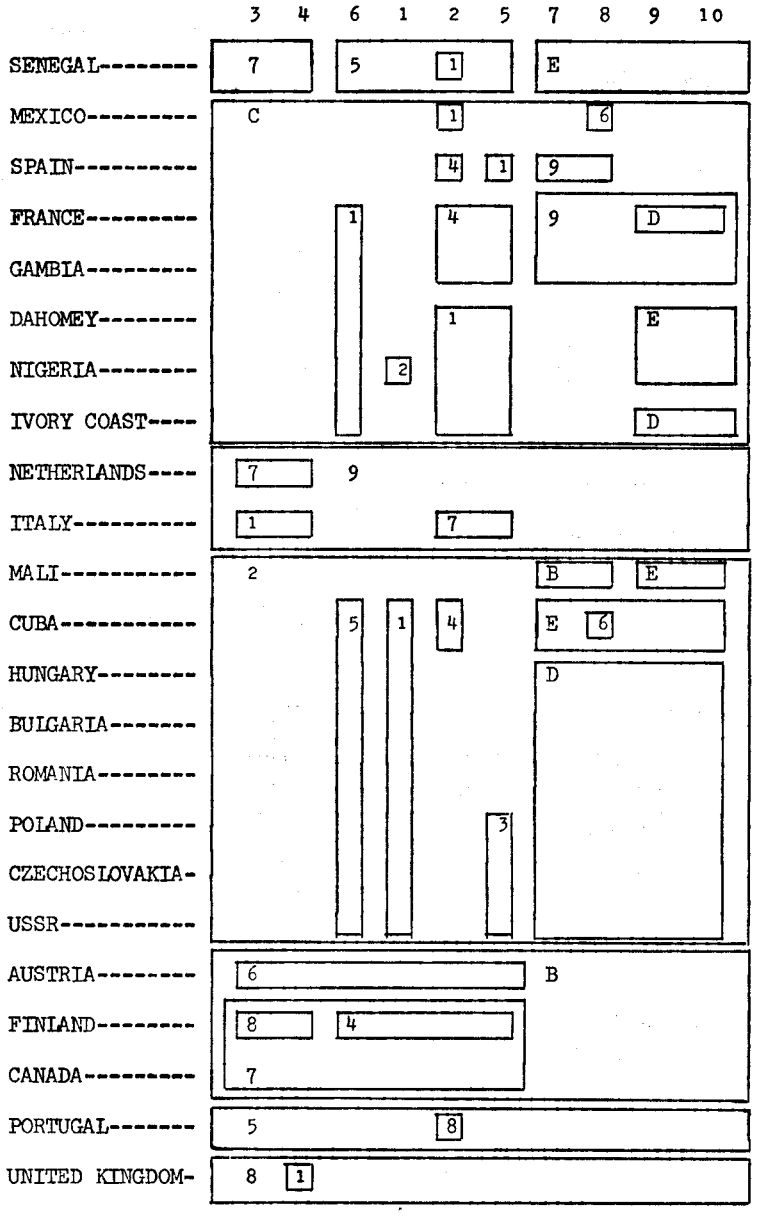
If each variable is a category variable, the results should be invariant under arbitrary one-to-one transformations of each variable. Therefore there will be an underlying scale of block values, a category scale, from which the given variables must be obtained by transformation.

In the joining algorithm, the basic problem is always the distance, the amalgamation rule, and block construction for pairs of rows or columns. The rows will be handled as usual by using matching distances and constructing blocks at the mismatches. The variables require new treatment. One simple procedure measures the distance between two variables as $\sum K(I,J)[K(I,J) - 1]/M(M-1)$, where $K(I,J)$ counts the number of times, in M cases, that the first variable takes the value I and the second variable takes the value J. The new variable just takes the set of values (I,J) which actually occur, and no blocks are constructed when variables are joined.

16.6.4* Continuous Data

In both the monotonic data and category data approaches, the final blocks have the property that every variable within a block is constant over cases within a block. This property is not realistic in the continuous data case. It is plausible to consider either monotonic transformation or linear transformations from the block-value scale, but it is necessary that a threshold be given for each variable, such that the variable ranges within the threshold over the cases in a block.

Table 16.8 Final Data Clusters in Applying Joiner Scaler to U.N. Votes



To translate this table, look at Mexico on Question 3, taking the value C . From 16.7, the value C on Question 3 is Y . Thus Mexico votes Yes on Question 3.

Original data are recovered by relating scale values to questions (Table 16.7).

In considering the linear case, pairs of cases are treated as in the range algorithm in Chapter 15, but pairs of variables must be considered freshly. Suppose that X and Y are variables taking values $[X(I), Y(I)]$ with thresholds TX and TY. A new variable Z will be constructed, connected to X and Y by

$$X = A(1)Z + A(2)$$

and

$$Y = B(1)Z + B(2).$$

There will be a threshold TZ for Z that is the minimum of TX/$A(1)$ and TY/$B(1)$. For each case I, there is a difference between the Z values $D(I) = |[X(I) - A(2)]/A(1) - [Y(I) - B(2)/B(1)]|$. Define

$$\mathrm{DD}(I) = D(I)/\mathrm{TZ} \quad \text{if} \quad D(I) \leq \mathrm{TZ},$$
$$\mathrm{DD}(I) = 1 \quad \text{if} \quad D(I) \geq \mathrm{TZ}.$$

Then $\sum \{1 \leq I \leq M\} \mathrm{DD}(I)$ measures the distance between X and Y for the particular choice of scale parameters $A(1)$ and $A(2)$, $B(1)$ and $B(2)$. Of course, these must be chosen to minimize $\mathrm{DD}(I)$. You see instantly that $B(1) = 1$, $B(2) = 0$ without loss. It is true also that the optimal choice of $A(1)$ and $A(2)$ is such that $D(I) = 0$ for two cases I. Thus the optimal values of $A(1)$ and $A(2)$ are obtained by searching over all the lines through pairs of points. (The time for a complete join of all rows and columns is thus proportional to M^3N^2.) Blocks are constructed, as in the homogeneous case, whenever a value is out of threshold with the value it is being joined to and is out of threshold with the value it is likely to be joined to next.

Complications arise later on in the algorithm, when each value becomes a range of values. For a pair of variables, the range is a rectangle with four corners. The optimal scale choice passes through corners for two cases, and so the same search procedure finds the optimal scaling.

16.6.5 Greater Generality

To handle data in which different variables are on entirely different scales, such as continuous, ordered, or category scales, it is supposed that there is an underlying block scale. All values in a block take a single block value z. For a variable I, there is a transformation $T(I, z)$ which specifies the value of variable I when the block value is z.

Thus $T(I, z)$ might be a linear transformation of z, or a monotonic transformation, or an arbitrary transformation, according to the type of variable. The problem of combining different types of variables to produce such a scale remains to be solved.

16.6.6 Median Regression

If X, Y are variables taking values $X(I)$, $Y(I)$, a median regression line of Y on X is the line $y = a + bx$, where a, b are chosen to minimize

$$\sum \{1 \leq I \leq M\} \, |Y(I) - a - bX(I)|.$$

Show that there is a median regression line for which $Y(I) = a + bX(I)$ for two values of I. Suppose that cases I, J are such that $Y(I) = a + bX(I)$, $Y(J) = a + bX(J)$. Suppose that for every K the lines through I, K and J, K have larger sums of absolute deviations than the line through I, J. Then the line through I, J is a median regression line.

REFERENCES

HAMMERSLEY, J. M. (1972). "A few seedlings of research." *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 345–393. U. of California, Berkeley. Considers various properties of the longest increasing subsequence of a random permutation of n integers. For example, the length L of the subsequence exceeds $\sqrt{n} - 1$, a result proved by Erdos and Szekeres in 1935. Hammersley shows that $L/\sqrt{n} \to c$ as $n \to \infty$, where $\frac{1}{2}\pi \leq c \leq e$. Hammersley presents a variety of nonrigorous arguments suggesting that c is close to 2.