PLISKA
STUDIA MATHEMATICA
BULGARICA

ПЛИСКА
БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

AN EXTENTION TO SERIATION BASED ON INCIDENCE MATRICES

LILIANA I. BONEVA

This paper extends Kendall's mathematical model of seriation from incidence, or (0, 1) matrices. Given an $n\times k$ matrix A, it was established up to now that $A'A$ gives a partial information about the possibility of rearranging A into a P-matrix (i. e. in each column the 1's are bunched together in a single run). However, according to Kendall, it is AA' which contains sufficient information to decide if A is P-convertible and to construct the row-permutation, which converts A into a P-matrix.

The extention considered here is based on both $A'A$ and AA'. A special 3×3 matrix, with pattern called B^*, turned out to be very relevant for seriation as well. If there is no submatrix B^* in A then both $A'A$ and AA' give sufficient information about the row- and column-permutations, if any, which convert A into an L-matrix, i. e. all 1's are bunched together in a single block.

The main reason to attach a great importance to the seriation problem is that one is able to obtain quantitative information from a qualitative one. That is why seriation proved recently to be a great success when applied not only in archaeology, where it was actually born, but also in sociology, philology, control theory, genetics, history, geology, geography, etc.

What seriation aims at? Suppose we have numerous statistical observations on n objects characterized by k features. We are looking for some reasonably "true" chronological (temporal, spatial or any other) order of these n objects, or k features or both, using only the information about the degree of similarity between the pairs of objects, or features, or both. For instance, if we are looking for a reconstruction of the chronological order, if any, of a set of n objects, it will be one of their $n!$ possible rearrangements. But to try all of them, especially when n and k are comparatively large, is an extremely difficult task. That is why there have been developed various, more or less effective methods for solving the ordering problem. For $k=1$, the ordering could be obtained quite trivially, but the situation is more complicated when $k>1$. There are methods in which the task is to find some summarized function of k, depending on what aim one is after. This function, called sometimes ordering function, is perceived as a generalized feature. The trouble in this case is that the ordering purposes might be different, so as to be difficult to decide which one of the possible functions will be the optimal one. For similar purposes Hotelling's component analysis [1] could be used, though connected with laborious calculations, or the discriminant analysis of Fisher [2] or Mahalanobis [3]. There exists also the ordering method of the Polish anthropologist Chekanowski [4], used for arranging (with the help of diagrams) various groups of people according to their anthropometrical features, and also the more reliable graph-theoretic method, called "taxonomy of Wroclaw" [5], giving as a result a linear dendrite or a dendrite of a higher degree. But may be most similar to seriation is the travelling salesman

PLISKA Studia mathematica bulgarica. Vol. 7, 1984, p. 3—9.

problem for finding the shortest route (the hamiltonian line) between the objects taken into cosideration.

With that end of view the seriation theory and its application were created recently by Shepard [6], Kruskal [7], Kendall [8] etc. In fact, the seriation theory consists not only of algorithms for solving practical "undetermined" problems but also of mathematical models, describing ideal "determined" situations excluding, of course, the cases of missing data, nonsymmetry, ties, etc. An extended approach to the mathematical model is given bellow.

Mathematical approach. Suppose we have n objects characterized by k features disposed in a matrix $A=(a_{ij})_{n,k}$, $(i=1,\dots,n;\ j=1,\dots,k)$, with as many rows as there are objects and as many columns as there are features.

Let us give first several basic definitions:

D.1. A is called an incidence matrix when $a_{ij}=1$ if the i-th object owns the j-th feature, and $a_{ij}=0$ if the i-th object does not own the j-th feature.

D.2. A is called a P matrix (is in P form) if it has the "consecutive 1's property" for columns, i. e. in every column the 1's are bunched together in a single run.

D.3. A is called a P convertible if there is a row-permutation matrix π which converts A into P form.

D.4. A is called an L matrix (is in L form) if it has the "consecutive 1's property" for both rows and columns, i. e. all 1's in A are buched together in a single block.

D.5. A is called L convertible if there are such row and column permutation matrices π_1 and π_2, for which $\pi_1 A\pi_2$ is in L form.

D.6. A square symmetric matrix is called an R matrix (is in R form) if, when going away from the main diagonal, whether along a row or a column, the elements never increase.

D.7. The square symmetric matrices $A'A$ and AA' are called R convertible if there are such row and column permutation matrices π_1 and π_2, exactly the same which convert A into L, for which $(\pi_1 A\pi_2)'(\pi_1 A\pi_2)$ and $(\pi_1 A\pi_2)(\pi_1 A\pi_2)'$ are in R form.

The mathematical background of the seriation theory developed by Kendall is a theorem due to Fulkerson and Gross [9]. A natural question underlying it is: do one needs to know A itself so as to be able to decide whether it is P convertible or not, prcvided A is P convertible in principle?

Theorem (F and G). *If the incidence matrices A and B satisfy $AA'=B'B$ then they both are either P convertible or not. Moreover, if they have equal number of rows then there is a permutation π such that $B=\pi A$.*

In fact, this theorem gives a partial answer to the question mentioned above, i. e. it tells that, when A is P convertible, it is enough to know only $A'A$ for deciding whether seriation is possible or not. But is does not answer how to do the seriation. Here comes in help the effective Kendall's theorem for incidence matrices [10].

Theorem (K). *If A is a P convertible incidence matrix and π is any permutation matrix, then πA will be in P form if and only if $\pi(AA')\pi'$ is in R form.*

The most important conclusion which follows from Kendall's theorem is that AA' contains enough information about A, i. e. retains enough of the structure of A, so as to enable us to find a row-permutation matrix, provided

A is P convertible, which converts A into a P matrix. Hence, $A'A$ tells us whether there is a seriation solution, while AA' tells us how to find it.

Let us now consider the problem from the dual point of view, i. e. whether there is a solution with respect to the features. In such a case one could use the following

Corollary. *If the transpose of an incidence matrix A is P convertible and π is any column-permutation matrix, then $\pi' A'$ will be in P form if and only if $\pi'(A'A)\pi$ is in R form.*

The proof is the same as in Kendall's theorem.

Meanwhile, a simple theorem concerning both AA' and $A'A$ can be stated.

Theorem 1. *If A is an incidence matrix then, for every row and column permutation matrices π_1 and π_2 respectively,*

1) $A'A=(\pi_1 A)'(\pi_1 A)$

and

2) $AA'=(A\pi_2)(A\pi_2)'$

are fulfilled.

The proof is evident taking into account that $\pi_1\pi_1'=\pi_2\pi_2'=I$, i. e. that the square matrices π_1 and π_2 are orthogonal.

Let us turn our attention now to the simultaneous seriation of objects and features. By this we are coming to the case of L convertible incidence matrices. First of all let us introduce for 3×3 incidence matrices the following

Pattern B^*: there is exactly one 0 in each row and in each column.

Let us recall also that

1. Each element of $G=AA'=(g_{ij})_{n,n}$ gives the number (or the sum) of the 1's common to the i-th and j-th rows of A, that is

$$g_{ij}=\sum_{s=1}^{k} a_{is}a_{js}, \quad i, j=1, \dots, n.$$

2. Each element of $V=A'A=(v_{ij})_{k,k}$ gives the number (or the sum) of the 1's common to the i-th and j-th columns of A, that is

$$v_{ij}=\sum_{m=1}^{n} a_{mi}a_{mj}, \quad i, j=1, \dots, k.$$

3. The elements of the main diagonal of $G=AA'$ are equal to the sums of the 1's in the corresponding rows of A, i. e.

$$g_{ii}=\sum_{s=1}^{k} a_{is}, \quad i=1, \dots, n.$$

4. The elements of the main diagonal of $V=A'A$ are equal to the sums of the 1's in the corresponding columns of A, i. e.

$$v_{jj}=\sum_{m=1}^{n} a_{mj}, \quad j=1, \dots, k.$$

5. The sum of the elements of the main diagonal of each one of $G=AA'$ and $V=A'A$ gives the total number of 1's in A, i. e.

$$\sum_{i=1}^{n} g_{ii}=\sum_{j=1}^{k} v_{jj}=\sum_{i,j} a_{ij}, \quad i=1,\dots,n, \quad j=1,\dots,k.$$

6. For each incidence matrix A the main diagonals of $G=AA'$ and $V=A'A$ are weakly dominant, i. e. each element of the main diagonals of both G and V is greater or equal than each of the remining elements respectivelly.

We shall prove now several lemmas.

Lemma 1. *If an incidence matrix A is in L form, then both G and V are in R form, i. e. for each triplet $i<s<p$ and $j>q>r$ the elements of the upper triangles of both G and V satisfy*

$$g_{ii}\geqq g_{is}\geqq g_{ip}, \tag{1}$$

$$g_{jj}\geqq g_{qj}\geqq g_{rj}, \tag{1'}$$

and

$$v_{ii}\geqq v_{is}\geqq v_{ip}, \tag{2}$$

$$v_{jj}\geqq v_{qj}\geqq v_{rj}. \tag{2'}$$

Proof. If $A=(a_{ij})_{n,k}$ is an L matrix, then for each triplet $i<s<p$ (i, s, $p=1,\dots,n$) the sum of the 1's in the i-th row of A will be greater or equal than the sum of those 1's in the s-th row common to the i-th row, and than the sum of those 1's in the p-th row common to the i-th row.

At the same time, for each triplet $j>q>r$ (j, q, $r=1,\dots,k$), the sum of the 1's in the j-th column of A will be greater or equal than the sum of those 1's in the q-th column common to the j-th column, and than the sum of those 1's of the r-th column common to the j-th column.

Should it be otherwise A would have a "window" either in a row $(\dots 1\ 0\ 1 \dots)$ or in a column

$$\left\| \begin{matrix} \dots & 1 & \dots \\ \dots & 0 & \dots \\ \dots & 1 & \dots \end{matrix} \right\|,$$

or both, i. e. at least one submatrix will be of the type, say, B^*.

These arguments lead to the inequalities

$$\sum_{j=1}^{k} a_{ij}\geqq \sum_{j:\,a_{ij}=1} a_{sj}\geqq \sum_{j:\,a_{ij}=1} a_{pj} \tag{3}$$

and

$$\sum_{i=1}^{n} a_{ij}\geqq \sum_{i:\,a_{ij}=1} a_{iq}\geqq \sum_{i:\,a_{ij}=1} a_{ir}. \tag{3'}$$

If we multiply A elementwise by a_{ij}, then

$$\sum_{j=1}^{k} a_{ij}a_{ij}\geqq \sum_{j:\,a_{ij}=1} a_{ij}a_{sj}\geqq \sum_{j:\,a_{ij}=1} a_{ij}a_{pj} \tag{4}$$

and

(4′) $$\sum_{i=1}^{n} a_{ij}a_{ij} \geqq \sum_{i:a_{ij}=1} a_{iq}a_{ij} \geqq \sum_{i:a_{ij}=1} a_{ir}a_{ij}.$$

Thus, bearing in mind the six points recalled above, we have

(5) $$g_{ii} \geqq g_{is} \geqq g_{ip}$$

and

(5′) $$g_{jj} \geqq g_{qj} \geqq g_{rj},$$

and similarly for A' (because of the row-column duality)

(6) $$v_{ii} \geqq v_{is} \geqq v_{ip}$$

and

(6′) $$v_{jj} \geqq v_{qj} \geqq v_{rj}.$$

Evidently, the inequalities (5), (5′) and (6), (6′) are equivalent to (1), (1′) and (2), (2′) respectively. With this the lemma is proved.

Lemma 2. *A will be L if and only if every submatrix of A is L.*

Proof. Suppose A is L. Then the suppressions of rows and columns of A does not distroy the L pattern of the obtained submatrices.

Let now every submatrix of A be L. Then A will be also L because any matrix is a submatrix of itself.

Corollary. *A is L convertible if and only if every submatrix of A is L convertible.*

Notice that further we shall denote by B^* all 3×3 incidence matrices which are combinatorically equivalent [11] to pattern B^*.

Lemma 3. *Let B be any* 3×3 *incidence matrix. If* $G=BB'$ *and* $V=B'B$ *are simultaneously R convertible, then every B, with exception of the* B^**'s, is L convertible.*

Proof. 1. Let

$$B^* = \left\| \begin{matrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{matrix} \right\|.$$

Denote $G^*=B^*B^{*\prime}$ and $V^*=B^{*\prime}B^*$. We shall show that B^* is not L convertible though its G^* and V^* are R.

As B^* is only a 3×3 matrix it is easy to check that it is not L convertible, i. e. that there are neither row nor column permutation matrices which could remove the row and the column windows in B^*. From the other hand, it could be seen (by direct multiplication) that G^* and V^* are R, and even that $G^*=V^*$ because $B^*=B^{*\prime}$.

2. Suppose now that for any B its $G=BB'$ and $V=B'B$ are simultaneously R convertible. We shall prove that all possible B, except B^*, are L convertible.

Consider all possible triplets of 0's and 1's, i. e. $2^3=8$ triplets, taken for instance in columns, namely

$$\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 1 & 0
\end{array}$$

From all 3×3 matrices formed from these triplets, we have to consider only those without repetitions and not containing the 1-st and the 8-th triplets. In fact, repetitions and columns 1 and 8 are not to be taken into account because they lead to 3×2 matrices which are always L convertible. Therefore, only six from the eight triplets are essential. Denote their 3×6 matrix with

$$C=\left\|\begin{array}{cccccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right\|.$$

According to Lemma 2, C is not L convertible because it contains a submatrix B^*, which, as shown above, is not L convertible. Analogously, each submatrix of C which contains B^* is not L convertible. Let us see then what happens with the 3×5 submatrices not containing B^*. It turns out that they are enough for our purposes. There are three such submatrices, and they are actually combinatorically equivalent. Hence, we have to check only one 3×5 submatrix of C, which could be easily done directly. For instance, let us take columns 1, 3, 4, 5, 6. This matrix is L convertible and its G and V are R convertible. The same is true for its submatrices. Thus, all possible 3×3 submatrices of C, whose G and V are R convertible, are L convertible, except B^*. That proves the lemma.

Lemma 4. *If A is an incidence matrix, whose G and V are R convertible, then A will be L convertible if and only if it does not contain a 3×3 submatrix with pattern B^*.*

Proof. Following out the proof of Lemma 3, it could be seen that every $3\times k$ matrix ($k\geqq3$), whose G and V are R convertible, is L convertible if it does not contain a submatrix with pattern B^*. The same is true for each $n\times3$ matrix ($n\geqq3$), because of the duality. Hence, each $n\times k$ matrix A, whose G and V are R convertible, will be L convertible if it does not contain a 3×3 submatrix with pattern B^*. From the other side, in accordance with the corollary of Lemma 2, if A does not contain a submatrix with pattern B^* it is L convertible. This finishes the proof.

We are ready now to state the following

Theorem 2. *Let A be an incidence matrix, which does not contain a submatrix with pattern B^*, and let π_1 and π_2 be a row and a column permutation matrices. Then $\pi_1 A\pi_2$ is L if and only if both $\pi_1 AA'\pi_1'$ and $\pi_2' A' A\pi_2$ are R.*

Proof. Suppose that A does not contain a 3×3 submatrix with pattern B^* and that a row and a column permutation matrices π_1 and π_2 exist. In such a case we prove easily the following:

1. If $\pi_1 A\pi_2$ is L, then, according to Lemmas 2 and 4, its $G=\pi_1 AA'\pi_1'$ and $\pi_2' A' A\pi_2$ will be R.

2. In accordance with Lemma 4, if $G=\pi_1 AA'\pi_1'$ and $V=\pi_2' A' A\pi_2$ are R, then $\pi_1 A\pi_2$ will be L, and that finishes the proof.

Some final notes.

1. The square symmetric matrices G and V are very useful because: (i) whatever column permutations to apply to A, its G does not change (see Theorem 1). Moreover, if A does not contain a submatrix with pattern B^*, then if G is R we conclude that A is P; (ii) whatever row permutation to apply to A, its V does not change (see T. 1). Moreover, when A does not contain any B^* and when its V is R, then we may conclude that A' is P.

2. If A does not contain any B^*, and if both G and V are R (or R convertible), then, even if A is lost, we may assert that A is L (or L convertible, i. e. permits L sorting).

3. It seems quite senseful to work out a computer algorithm which will tell us whether A contains B^* or not. In that case we may apply some of the known sorting algorithms without any fear that we may seriate things which should not be seriated. An algorithm for recognizing the $B^{*\prime}s$ in A will appear elsewhere soon.

4. At last, it seems very useful to be able to find a significant "approximate seriation solution" when working with real data, given in a rather big A matrix. In other words, it will be useful to find a criterion which gives us the critical value above which we shall have to reject the hypothesis that A is arrangeable. Such criterion is in preparation.

REFERENCES

1. H. H. Harman. Modern Factor Analysis. Chicago, 1962.
2. R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.*, 7, 1963, 179—188.
3. P. C. Mahalanobis. Historical Note on the D^2-statistics. *Sankhya*, 9, 1948, 237.
4. J. Czekanowski. Zarys antropologii Polski. Lwów, 1930.
5. J. Perkal. Taksonomia wrocławska. *Przegląd Antrop.*, 19, 1953, 82—96.
6. R. N. Shepard. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. *Psychometrika* (I, II), 27, 1962, 125—139, 219—246.
7. J. B. Kruskal. Multidimensional Scaling. *Psychometrika* (I, II), 29, 1964, 1—27, 28—42
8. D. G. Kendall. Incidence Matrices, Interval Graphs and Seriation in Archaeology. *Pacif. J. Math.*, 28, 1969, 565—570.
9. D. R. Fulkerson, O. A. Gross. Incidence Matrices and Interval Graphs. *Pacif. J. Math.*, 15, 1965, 835—855.
10. D. G. Kendall. Seriation from Abundance Matrices. — In: Mathematics in the Archaeological and Historical Sciences (F. R. Hodson, D. G. Kendall and P. Tautu, Eds.). Edinburgh, 1971, 215—252.
11. D. T. Finkerbeiner. Introduction to Matrices and Linear Transformations. San Francisco and London, 1966.

Centre for Mathematics and Mechanics
1090 Sofia P. O. Box 373

Received 15. 2. 1979