Journ. Japan Statist. Soc.
6, 2, 1976, 17~25

GRAPHICAL REPRESENTATION OF CORRELATION ANALYSIS OF ORDERED DATA BY LINKED VECTOR PATTERN

Masaaki Taguri*, Makoto Hiramatsu**,
Tomoyoshi Kittaka** and Kazumasa Wakimoto**

(Received Aug. 12, 1975, Revised May 30, 1976)

In this paper, we propose a new method of simultaneous graphical representation of certain correlations between objective variable and several explanatory variables which is done through visual processing of linked vector patterns corresponding to the objective variable and each of the explanatory variables. Merits of the method lie in the following points: (1) It is very efficient when correlation changes in part. (2) Besides whole correlation can be grasped, it makes possible to examine individual data in detail. (3) It is possible to express correlations of several parts of variables on the same graph. Moreover, two measures are defined to represent the degree of correlation relation quantitatively, which highly correlate with the traditional correlation coefficients.

1. Introduction

Correlation between objective variable and explanatory variable has been discussed by using the dispersion graph, Spearman's rank correlation coefficient and Kendall's one (cf. [2]). By the dispersion graph, it is possible to grasp the correlation between the two kinds of variables, but it is extremely difficult to grasp simultaneously the correlations between objective variable and several explanatory variables. The dispersion graph may make clear correlation between the two variables as a whole, however it is almost impossible to find out information related with closely individual data. On the other hand, the rank correlation coefficient given by Spearman or Kendall is convenient for representing quantitatively the correlation between two variables. These measures are, however, not appropriate if positive and negative correlations are mixed or there exist correlations in parts. Taking into account of the features of these traditional method, we introduce a new method of graphical representations, through which the unsatisfactory points described above are partly dissolved.

Now, let us denote objective variable by p and k kinds of explanatory variables by $r_1, r_2, \cdots, r_k$, where the variables p and r_j $(j=1, 2, \cdots, k)$ represent ranks from 1 to n. The proposed method makes it possible to show the correlation between p and r_j $(j=1, 2, \cdots, k)$ graphically on a 2-dimensional plane by means of linked vector patterns (see [1], [3]), which are drawn by linking certain vectors whose arguments being related to p and r_j. One novel characteristic of our method is to be able to make clear underlying correlation relation on the data through visual processing of the graphical pattern rather than using a numerical method. A

* Chiba University.
** Okayama University.

graphical display may be utilized effectively in most cases. Advantages of this method may be summarized as follows:

(a) k kinds of relations between p and r_j $(j=1, 2, \cdots, k)$ can be represented on a sheet of graphic paper simultaneously.

(b) Besides the information in case of grouping n sets of data, it is possible to obtain informations in detail for individual data or a sub-group at the same time.

(c) It is possible to define a quantitative measure which is highly correlated with the Spearman's rank correlation coefficient and Kendall's one.

It should be noted that the discussion on this paper is restricted to the descriptive statistics.

2. Graphical representation of correlation relation

In this section, a method of representation of correlation relation using linked vector pattern is described. Let us denote the number of explanatory variables by k and the number of given data for each explanatory variable by n.

(i) As shown in Table 1, let p_i and r_{ji} $(j=1, 2, \cdots, k;\ i=1, 2, \cdots, n)$ be ranks of the i-th observations of objective variable and the j-th explanatory variable, respectively. When ties are present among observations of objective variable, the ranks of these are decided equally by a certain random mechanism, and when ties are present among observations of the j-th explanatory variable, the ranks are given in accordance with those of objective variable. The n values of the objective variable are transformed into radian values from 0 to π by the following formula:

$$\beta_i=\frac{p_i-1}{n-1}\pi,\quad (i=1, 2, \cdots, n). \tag{2-1}$$

Similarly, for k kinds of explanatory variables r_j the following transformations are made:

$$\alpha_{ji}=\frac{r_{ji}-1}{n-1}\pi,\quad (j=1, 2, \cdots, k;\ i=1, 2, \cdots, n). \tag{2-2}$$

(ii) Rearrange n $(k+1)$-vectors $(\beta_i, \alpha_{1i}, \cdots, \alpha_{ki})$, $(i=1, 2, \cdots, n)$, in order of the magnitudes of β_i's:

$$0=\beta_{(1)}<\beta_{(2)}<\cdots<\beta_{(n)}=\pi. \tag{2-3}$$

Let $\alpha_{j(i)}$ be the value of α_j corresponding to $\beta_{(i)}$ and put

Table 1. Original ordered data

Variables / Data No.	p	r_1	r_2	$\cdots$	r_j	$\cdots$	r_k
1	p_1	r_{11}	r_{21}	$\cdots$	r_{j1}	$\cdots$	r_{k1}
2	p_2	r_{12}	r_{22}	$\cdots$	r_{j2}	$\cdots$	r_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	p_i	r_{1i}	r_{2i}	$\cdots$	r_{ji}	$\cdots$	r_{ki}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	p_n	r_{1n}	r_{2n}	$\cdots$	r_{jn}	$\cdots$	r_{kn}

Table 2. The data transformed by formulas (2-1)~(2-4)

Variables / Data No.	η	ξ_1	ξ_2	$\cdots$	ξ_j	$\cdots$	ξ_k
(1)	η_1	ξ_{11}	ξ_{21}	$\cdots$	ξ_{j1}	$\cdots$	ξ_{k1}
(2)	η_2	ξ_{12}	ξ_{22}	$\cdots$	ξ_{j2}	$\cdots$	ξ_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(i)	η_i	ξ_{1i}	ξ_{2i}	$\cdots$	ξ_{ji}	$\cdots$	ξ_{ki}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(n)	η_n	ξ_{1n}	ξ_{2n}	$\cdots$	ξ_{jn}	$\cdots$	ξ_{kn}

$$\eta_i=\beta_{(i)},\ \xi_{ji}=\alpha_{j(i)},\quad (j=1, 2, \cdots, k;\ i=1, 2, \cdots, n). \tag{2-4}$$

Thus we have Table 2 obtained from Table 1.

(iii) Let us associate the vectors $\vec{\eta}_i$, $\vec{\xi}_{ji}$ with the variable η_i, ξ_{ji} in such a manner as

$$\begin{cases}|\vec{\eta}_i|=1,\\ \arg(\vec{\eta}_i)=\eta_i,\end{cases}\quad (i=1, 2, \cdots, n). \tag{2-5}$$

$$\begin{cases}|\vec{\xi}_{ji}|=1,\\ \arg(\vec{\xi}_{ji})=\xi_{ji},\end{cases}\quad (j=1, 2, \cdots, k;\ i=1, 2, \cdots, n). \tag{2-6}$$

Here, $\vec{\eta}_i$ and $\vec{\xi}_{ji}$ are called the vector of objective variable and the vector of explanatory variable, respectively.

(iv) By linking successively the vectors in each of the sets $\{\vec{\eta}_1, \vec{\eta}_2, \cdots, \vec{\eta}_n\}$ and $\{\vec{\xi}_{j1}, \vec{\xi}_{j2}, \cdots, \vec{\xi}_{jn}\}$ $(j=1, 2, \cdots, k)$ starting from the origin, linked graphical patterns on a 2-dimensional plane are obtained (see Fig. 1), which are called the linked vector pattern of objective variable (P-pattern) and the linked vector pattern of the j-th explanatory variable (R_j-pattern) $(j=1, 2, \cdots, k)$, respectively. A set of patterns thus obtained is considered to represent a kind of degree of correlation between objective and explanatory variables. Note that the end points of the P-pattern and R_j-patterns always coincide on the ON-axis (cf. Fig. 1), and that the P-pattern displaies itself along a semicircle.

3. Method to read correlation relation from the linked vector patterns

In this section, method of reading the correlation relation between objective variable and explanatory variable from the linked vector patterns is stated. Generally speaking, the degree of correlation is determined by comparing such factors as proximity, smoothness and shape of P-pattern and of R_j-patterns, where the smoothness of P-pattern and R_j-patterns are defined as $|\arg(\vec{\eta}_{i+1})-\arg(\vec{\eta}_i)|$ and $|\arg(\vec{\xi}_{j,i+1})-\arg(\vec{\xi}_{j,i})|$ $(i=1, 2, \cdots, n-1)$, respectively.

The correlation relation between objective variable and explanatory variable can be classified by the feature of the R_j-pattern as follows:

(a) no correlation

(b) the case that there exist some correlations

(b-1) monotonic correlation

(b-1-1) positive correlation

(b-1-2) negative correlation

(b-2) local correlation

(b-2-1) positive correlation in parts

(b-2-2) negative correlation in parts

(b-2-3) positive and negative correlations in parts.

In case (b-2), the linked vector patterns can be interpreted by applying the characteristics in cases (a) and (b-1). Therefore, the linked vector patterns in cases (a), (b-1-1) and (b-1-2) are exemplified below as three basic patterns.

Example 1

For the data in Table 3, the linked vector patterns are obtained by the procedure explained in the preceeding section. The result is shown in Fig. 1, which presents three basic patterns R_1, R_2 and R_3.

1° The R_1-pattern

This pattern will appear when there is no correlation between the objective and the explanatory variables. Thus, in case (a), R_J-pattern may resemble to the R_1-pattern in Fig. 1, and it goes up along the ON-axis in zigzag apart from the P-pattern.

2° The R_2-pattern

This pattern suggests us the case (b-1-1), and it indicates that positive correlation between the two variables is strong. Then in case (b-1-1), R_J-pattern and P-pattern are similar in shape, and they appear close to each other.

3° The R_3-pattern

This pattern suggests us that there is strong negative correlation (b-1-2) between the two variables. In this case R_J-pattern and P-pattern are nearly symmetric with respect to the ON-axis in shape.

For case (b-2-1), (b-2-2) or (b-2-3), the linked vector patterns can be interpreted

Table 3. The data in Example 1

p	r_1	r_2	r_3	p	r_1	r_2	r_3	p	r_1	r_2	r_3
1	11	2	19	8	1	9	13	15	12	16	5
2	8	3	20	9	10	10	12	16	3	17	7
3	2	4	18	10	4	7	11	17	19	11	4
4	18	5	17	11	6	12	10	18	14	18	3
5	9	1	16	12	15	13	8	19	13	19	2
6	20	6	15	13	5	15	6	20	7	20	1
7	16	8	14	14	17	14	9				

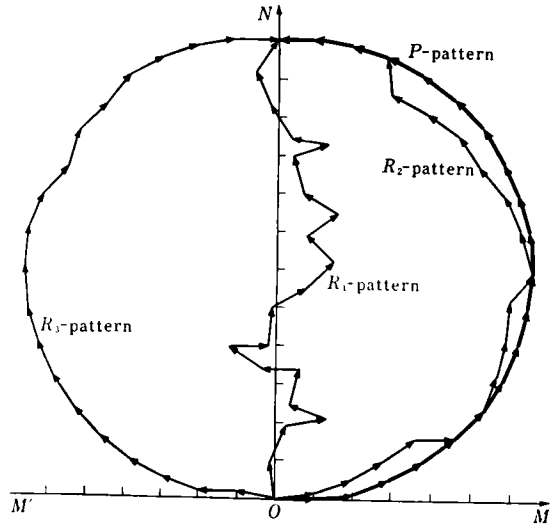

Fig. 1. Linked vector patterns for the data in Example 1

Table 4. The data in Example 2

p	r_4	r_5	r_6	p	r_4	r_5	r_6	p	r_4	r_5	r_6
1	1	9	20	8	10	13	13	15	19	8	6
2	2	16	19	9	11	15	12	16	12	5	7
3	4	11	18	10	20	19	1	17	17	4	8
4	6	17	16	11	5	20	2	18	13	2	9
5	7	14	17	12	18	12	3	19	16	3	10
6	8	18	15	13	3	7	5	20	15	1	11
7	9	10	14	14	14	6	4				

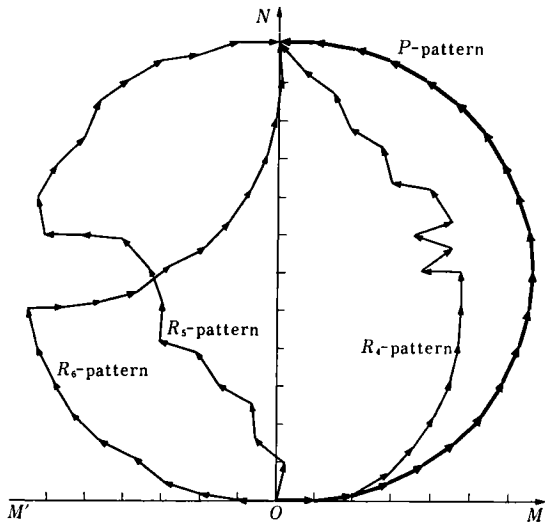

Fig. 2. Linked vector patterns for the data in Example 2

as a combination of some of these three patterns. The following example is for these cases.

Example 2

The linked vector patterns composed from the data in Table 4 are P, R_4, R_5 and R_6-pattern in Fig. 2.

1° The R_4-pattern

This pattern suggests us the case (b-2-1) where there are positive correlations in parts between the two variables. The local pattern from $\vec{\xi}_{4,10}$ to $\vec{\xi}_{4,15}$ in Fig. 2 corresponds to case (a) of no correlation, while that from $\vec{\xi}_{4,1}$ to $\vec{\xi}_{4,9}$ corresponds to case (b-1-1) of positive correlation. The linked vector pattern from $\vec{\xi}_{4,16}$ to $\vec{\xi}_{4,20}$ suggests us that there is weak positive correlation.

2° The R_5-pattern

This pattern is an example for case (b-2-2). The local pattern from $\vec{\xi}_{5,12}$ to $\vec{\xi}_{5,20}$ in Fig. 2 suggests us the case (b-1-2) of negative correlation, while that from $\vec{\xi}_{5,1}$ to $\vec{\xi}_{5,6}$ suggests us that there is weak negative correlation. The middle part of the pattern from $\vec{\xi}_{5,7}$ to $\vec{\xi}_{5,11}$ shows that there is weak positive correlation.

3° The R_6-pattern

This pattern gives an example for case (b-2-3). Judging from the shape of the

pattern for case (b-1-1) or (b-1-2), it may be concluded that the linked vector pattern from $\vec{\xi}_{6,1}$ to $\vec{\xi}_{6,9}$ in Fig. 2 suggests us negative correlation and the local pattern from $\vec{\xi}_{6,10}$ to $\vec{\xi}_{6,20}$ does positive correlation.

As shown in Example 2, it is possible to read correlation relation from a given linked vector pattern as a combination of some of the three basic patterns described in Example 1.

4. Two measures representing the degree of correlation

In the present section, certain measures to grasp the degree of correlation quantitatively are considered.

(i) Area ratio correlation coefficient

The following symbols are defined;

A_0: area of the domain enclosed by P-pattern and the ON-axis.

A_j: area of the domain enclosed by R_j-pattern and the ON-axis. The domain on the right-side of the ON-axis is defined to have positive sign area and the domain on the left-side of the ON-axis is defined to have negative sign area. If R_j-pattern exists on both sides of the ON-axis, its area is defined to be the sum of these areas.

Now, a measure of degree of correlation a_j, between the objective variable p and the j-th explanatory variable r_j, is defined by the following formula and it is called the area ratio correlation coefficient:

$$a_j=\frac{A_j}{A_0}, \quad (j=1, 2, \cdots, k). \tag{4-1}$$

Since $\max_{1\leq j\leq k}|A_j|\leq A_0$, we have $|a_j|\leq 1$. If there exists strong positive correlation between p and r_j, then A_j may be close to A_0, so a_j will have a value closed to 1. If there exists strong negative correlation, A_j close to $-A_0$, so a_j will have a value closed to -1.

The values of a_j have been computed for the data in Example 1 and Example 2, and are shown in Table 5 in which ρ_j represents the value of Spearman's rank correlation coefficient. Note that the values of a_j and ρ_j are fairly close to each other in most case of the six patterns.

(ii) Smoothness coefficient

Another kind of measure concerning with the linked vector pattern is the smooth-

Table 5. The values of ρ_j, a_j and s_j for the data in Example 1 and Example 2

	ρ_j	a_j	s_j
R_1	0.062	0.081	0.211
R_2	0.943	0.907	0.842
R_3	−0.988	−0.980	0.936
R_4	0.708	0.639	0.538
R_5	−0.701	−0.582	0.713
R_6	−0.669	−0.513	0.918

ness coefficient s_j, which is defined by using the sum of the absolute values of the differences of successive ranks:

$$s_j=1-\frac{2\sum_{i=1}^{n-1}|r_{j,i+1}-r_{j,i}|-2(n-1)}{(n-1)(n-2)}. \tag{4-2}$$

Since

$$n-1\leqq\sum_{i=1}^{n-1}|r_{j,i+1}-r_{j,i}|\leqq\frac{1}{2}n(n-1),$$

we have $0\leqq s_j\leqq 1$. The values of s_j have been computed for the data in Table 3 and 4, and the results are shown in Table 5. If p and r_j have strong positive correlation, then the R_j-pattern may be smooth and the value $\sum_{i=1}^{n-1}|r_{j,i+1}-r_{j,i}|$ will be close to $n-1$, and the value of s_j is expected to be close to 1. Moreover, if p and r_j have strong positive correlation in one part and have strong negative correlation in the other part, the value of s_j will get near of 1, as is seen from the value of R_6 in Table 5. It may be said empirically that the value of s_j is close to 0.3 when data are arranged randomly in the rank.

5. An application for analysis of scholarly attainment in university

This graphical method of linked vector pattern stated above is applied to evaluate the scholarly attainment in college education. The objective group analyzed consists of 40 science majored students. Their comprehensive scholastic results of natural science, including college physics, general chemistry and mathematics, at the end of the first academic year are considered as the elements of the objective

Table 6. The data with four variables for 40 science majored students

Natural science	Foreign language	Entrance examination	Scholarly attainment in high school	Natural science	Foreign language	Entrance examination	Scholarly attainment in high school
1	1	35	6	21	8	37	14
2	2	13	8	22	13	16	25
3	3	21	1	23	17	20	24
4	4	10	7	24	26	33	22
5	6	3	12	25	29	23	36
6	10	28	16	26	18	4	5
7	7	29	9	27	30	31	26
8	37	30	31	28	27	36	11
9	23	1	10	29	34	9	2
10	11	5	3	30	39	27	23
11	9	6	27	31	32	19	28
12	22	7	4	32	31	26	35
13	24	11	19	33	12	2	15
14	16	8	17	34	35	34	39
15	19	18	13	35	15	17	29
16	14	14	20	36	28	38	40
17	36	12	38	37	21	32	18
18	33	39	21	38	38	24	30
19	25	22	32	39	40	40	33
20	5	15	34	40	20	25	37

variable.

On the other hand, the comprehensive scholastic results of foreign language at the same educational phase with the above are treated as those of the explanatory variable. Two other educational factors are also introduced as explanatory variables in this study. They are the score of the entrance examination to the university and the comprehensive scholarly attainment in senior high school. For the last factor no revision has been done if there is any scholastic disparity among schools.

Data of all the four variables for the 40 students are shown in Table 6. The linked vector patterns for these data are shown in Fig. 3. The followings are some educational information derived from the patterns in Fig. 3.

(i) Informations pertaining to the scholastic results of natural science

a) The correlation between the comprehensive scholastic results of natural science

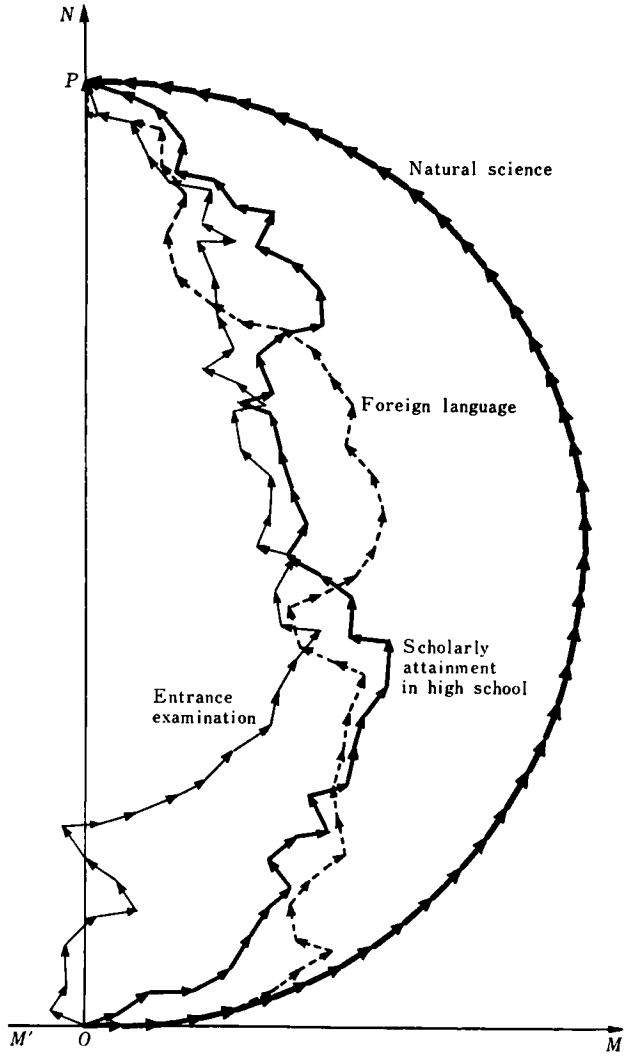

Fig. 3. Linked vector patterns for the data with four variables for 40 science majored students

and that of foreign language is rather bigger than that between the former and score of the entrance examination.

b) For the students who belong to the poorest seven in the outcome in natural science, there seems to exist strong correlations among all the variables considered except one, namely that of entrance examination. In other words, those students mentioned above seem not to study seriously in classes in spite of getting good scores at the entrance examination. There may be some reasons. They may only devote themselves to club activities because that lectures in general education may not give any satisfaction to them. Or, they may have lost their energy to study more when they hardly passed the narrow gate, so-called the entrance examination. In any reason, the lowest seven persons in the patterns in Fig. 3 may be interpreted as that about 20% of the students whose scholastic results occupy there lost their volition of study in the first year of college life.

It may be pointed out from the above applicational example that this new method has two advantages. First, the linked vector patterns can be make possible to express the correlations between objective variable and several explanatory variables in a figure and to evaluate the correlations simultaneously. Next, the method by the linked vector patterns provides possibility to find out and to discriminate features and characteristics of sub-groups which may exist in analyzing group.

(ii) Area ratio correlation coefficient a_j and smoothness coefficient s_j

Spearman's rank correlation coefficient ρ_j, the area ratio correlation coefficient a_j and the smoothness coefficient s_j were calculated for the data shown in Table 6. These values are given in Table 7. From this table, it is found that very strong positive correlations exist between a_j and ρ_j, and also between s_j and $|\rho_j|$.

Table 7. The values of Spearman's rank correlation coefficient ρ_j, area ratio correlation coefficient a_j, and smoothness coefficient s_j, for 40 science majored students.

	ρ_j	a_j	s_j
Foreign language	0.606	0.516	0.587
Score of entrance examination	0.338	0.285	0.372
Scholarly attainment in high school	0.570	0.493	0.422

Acknowledgement

The authors would like to express their grateful appreciation to the referees for their valuable comments concerning the manuscript.

References

[1] Hiramatsu, M., Kittaka, T. and Wakimoto, K. (1975). Correlation analysis of the learning outcomes by linked vector patterns (in Japanese), *J. Physics Education Soc. Japan*, **23**, No. 1, 25-29.

[2] Kendall, M.G. and Stuart, A. (1961). *The Advanced Theory of Statistics, Vol. 2*, 474-509, Charles Griffin & Company Ltd., London.

[3] Wakimoto, K. and Taguri, M. (1974). On the representation method of multiple correlation by pattern of connected vectors (in Japanese), *J. Japan Statist. Soc.*, 5, No. 1, 9-24.