ZASTOSOWANIA MATEMATYKI
V (1960)

J. PERKAL (Wrocław)

ON THE ANALYSIS OF A SET OF CHARACTERISTICS

In natural sciences individuals are characterized by a set of characteristics. Thus for instance a psychotechnician characterizes an individual under examination by the estimates of several tests, a physician characterizes the patient by the results of several auxiliary examinations. The investigator is very often confronted with the problem of characterizing the individual in question by a number of numbers (indices) that is smaller than the number of characteristics in the set. This can be done for instance by deleting less important characteristics. In general it should aim at to obtaining by means of a small number of indices the largest possible part of the information resulting from the full set of characteristics.

The above problem absorbs the attention of scientists dealing with the so called analysis of factors. Several methods of solving this and similar problems have been developed. None of them, however, is entirely satisfactory and each has serious drawbacks pointed out in world publications. In this paper I shall describe the principles and disadvantages of a few of the more important methods of analysing a set of characteristics and I shall present my own method, which seems to me better than any of the previous ones.

Let the random variables $x_1, x_2, \ldots, x_k$ denote k characteristics of the set in question. We investigate the population of individuals or a representative sample P and denote by x_{ij} a number, constituting the value of the i-th characteristics of the j-th individual ($j = 1, 2, \ldots, n$, where n is the size of the sample or of the population).

Hotelling's method (see [2]) and a similar method which I have published (see [5]) consist in finding k linear combinations of the initial features

$$y_p = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pk}x_k \quad \text{where} \quad p = 1, 2, \ldots, k,$$

the coefficients a_{pi} being chosen so that the combinations y_p are uncorrelated (characteristics x_i are generally correlated). We find the variances ω_p

of combinations y_p and we replace combinations of small variance (as compared with the variances of other combinations) by constants, i. e. by the expected values of those combinations. In this way we obtain a number (less than k) of combinations (indices) containing the greatest part of information given by the initial set of characteristics. Those mathematically correct and neat methods have two drawbacks. First, the numerical difficulties make them practically useless in analysing sets containing more than a few characteristics. E. g., the calculations for a set of ten characteristics must last several months. Adding one more characteristics at least doubles the number of calculations. The second drawback is the difficulty of a natural interpretation of the combinations obtained. E. g., the students of the AWF (Academy of Physical Training) have been characterized by three performances: x_1 — one-hundred metre race, x_2 — high jump, x_3 — ball throw. We have obtained three linear combinations, of which the first (with the greatest variance) increases during the studies (and training), the second is constant and the third (with the least variance) decreases. The first of those combinations can easily be interpreted as the physical capacity of the students. The interpretation of the second and the third components involves considerable difficulties and causes differences of opinion among experts. In general the whole interpretation is highly subjective and rather stirs the fancy of the interpreter. Those difficulties are well-known to psychotechnicians, who, taking psychotechnical tests as a starting point, would like to regard the suitable combinations as indices of ability, diligence and other features of the character and the mind of the person tested.

The method of Spearman (see [7]) can only be used for a set of characteristics of which every one is non-negatively correlated with every other one. Spearman says that there exists then a certain common factor contained in the individual characteristics. When that factor is eliminated from the characteristics x_i there remains a residual z_i. Spearman determines the common factor g as a linear combination of all characteristics x_i in such a way that it is not correlated with the residuals z_i; those residuals cannot be correlated too. This is also a mathematically correct method. It is a disadvantage that it replaces a whole set of characteristics by a single number — the common factor, which necessarily involves the loss of a too large part of the information contained in the initial set.

The method of Thurstone (see [10]) may be regarded as a generalization of the preceding one. Unlike Spearman, Thurstone sees in the set under consideration a few orthogonal factors, $g_1, g_2, \ldots$, not correlated with the residuals z_i. That method makes it possible to enclose an

arbitrary part of the total information in the factors (their number must of course be sufficient). The number of factors is defined rather subjectively, according to what part of the information can be disregarded in a given investigation. But even if the number of factors has been settled, Thurstone's method will not be unique. For the factors determine a certain linear space and instead of the given system of factors we can choose any other basis of that linear space. Thurstone suggests that the system of factors should be turned over and over until we obtain factors with a clear natural interpretation. This advice, however, can justifiably be regarded as insufficient. The method has thus two serious drawbacks, namely it is not unique and requires a subjective interpretation of factors (I discussed those difficulties earlier while describing Hotelling's method and my own). Moreover, it involves considerable numerical difficulties.

As can be seen, all the methods described above—the most important methods of analysing a set of characteristics—have serious drawbacks, such as cumbersome calculations, difficulty of natural interpretation and (in Thurstone's method) lack of unicity.

In 1953 I published (see [6]) a new method of analysing a set of characteristics, similar to that of Spearman. I call two individuals *naturally similar* if the differences between the values of the characteristics of those individuals are proportional to the standard deviations of those characteristics. The invariants of that similarity, i. e. the natural indices, are obtained in the following way. Denote by $\bar{x}_i$ the mean value and by σ_i the standard deviation of a characteristics x_i. The characteristics x_i are normalized so that their mean value is equal to 0 and their standard deviation is equal to 1. Instead of the initial characteristics x_i we obtain new random variables

$$t_i = \frac{x_i - \bar{x}_i}{\sigma_i},$$

which will be termed *normalized characteristics*. The value of the i-th normalized characteristics of the j-th individual will be

$$t_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i}.$$

The arithmetic mean of normalized characteristics

$$m = \frac{1}{k} \sum_{i=1}^{k} t_i$$

I term the *overall index of magnitude*. The value of that index for the j-th individual

$$m_j = \frac{1}{k}\sum_{i=1}^{k} t_{ij} = \frac{1}{k}\sum_{i=1}^{k}\frac{x_{ij}-\bar{x}_i}{\sigma_i} \tag{1}$$

characterizes the overall magnitude of the j-th individual in population P with respect to the set of characteristics, i. e. it orders the individuals of population P with respect to the overall magnitude. This index plays an analogous part to that of the common factor in Spearman's method.

I apply the term *natural indices* to the residuals

$$w_i = t_i - m \qquad (i = 1, 2, \ldots, k).$$

The values of these indices for the j-th individual,

$$w_{ij} = t_{ij} - m_j,$$

characterize the magnitude of the individual characteristics in relation to all the characteristics of the j-th individual that have been investigated. In paper [6] I prove that a necessary and sufficient condition of natural similarity between the individuals numbered h and j is a system of k equations,

$$w_{ih} = w_{ij} \qquad (i = 1, 2, \ldots, k),$$

of which every one results from the remaining equations since natural indices are bound by one linear equation $\sum_{i=1}^{k} w_i = 0$. Besides, they play a similar role to that of the residuals z_i in Spearman's analysis. The index m is usually correlated with the indices w_i, and also the indices w_i are correlated with one another (with Spearman they are not). However, the advantage of my method over Spearman's and other methods of analysing a set of characteristics lies in the fact that it is unique, objective, easy in calculation and clear as regards natural interpretation.

Teissier (see [9]) investigated the index m as a common factor of a set of characteristics. He pointed to the additional advantage resulting from the treatment of that index, namely to the fact, that m as the arithmetic mean is charged with a smaller error than the individual features x_i. It is particularly important in those cases where measurement errors play a considerable part. The stronger the correlation of the features of the initial set the better the results obtained by the method of analysis here described (see e. g. H. Milicer's paper [4]).

In the sequel I shall be concerned with the above method as regards the index m, which I shall term the *common vector* (or briefly — *vector*) of a given set of characteristics. It should be imagined as a vector of length m lying in a k-dimensional space with axes $x_1, x_2, \ldots, x_k$ on the axis perpendicular to a hyperplane with the equation $\sum_{i=1}^{k} w_i = 0$.

The method proposed in this paper will be similar to Thurstone's one in the same sense as that in which the method of natural indices described above is similar to Spearman's method. I shall divide a set of characteristics into systems, if necessary smaller and smaller, and I shall find the vector m for each system. Those vectors will be analogues of Thurstone's factors. Although they will not be orthogonal and will be correlated with the residuals of the initial features, in practice those correlations will not be very great. Instead the method is almost objective, easy in calculation, and the vectors have a clear natural interpretation.

H. Milicer, in her works on the somatic features of children, divides the set of characteristics which she is investigating into systems concerning four properties of a child, namely the characteristics of his flesh tissue, fat tissue, skeleton width and skeleton length. For each of those four systems she has found the index m and obtained interesting results. This idea has suggested to me the notion of dividing sets of characteristics into systems, and then dividing systems into subsystems in such a way as to have the characteristics in the systems and the residuals in the subsystems positively correlated.

Denote by r_{pq} the correlation coefficient between the characteristics x_p and x_q. By $r_{pp} = 1$ we shall understand the correlation of the p-th characteristics with itself. The matrix R of those correlation coefficients is known in classical statistics. It is a square matrix with k lines and k columns, symmetric ($r_{pq} = r_{qp}$) and having unities on the main diagonal. Frisch's theorem (see Cramer's manual [1], pp. 297 and 298) states that the order of that matrix is equal to the dimension of the hyperplane containing the whole mass of the distribution. Thus the order of that matrix is equal to the number of linearly independent indices which are combinations of the initial characteristics x_i. But the probability that the matrix order is less than the number of characteristics is equal to zero. In practice we often come across sets of characteristics such that all correlations r_{pq} are non-negative. A *set* of that kind will be termed a *system* of characteristics. If that is not the case, then we can divide the set into systems of characteristics in such a way that the characteristics in one system will be non-negatively correlated. It is true that the division into systems may happen to be non-unique, i. e. a certain characteristics may be arbitrarily assigned to one or to another system: this

difficulty, however, is rather of natural character. I shall write about it, and also about the method of dividing a set of characteristics into systems, in another paper. We can also come across an exceptional situation, namely such that all correlation coefficients r_{pq} are negative. Then every characteristic constitutes a separate system. It will be observed, however, that by changing the sign of some, or even of one characteristic x_i, we shall obtain a change of sign of some correlations, making the grouping of characteristics possible.

Take for example a set of 8 anthropological characteristics:

x_1 — length of head (g-op), x_2 — width of head (eu-eu),
x_3 — width of face (zy-zy), x_4 — length of face (n-gn),
x_5 — length of nose (n-sn), x_6 — width of nose (al-al),
x_7 — colour of eyes, x_8 — colour of hair,

which have been investigated by Krauze and Łozińska in paper [3]. The matrix of the correlation coefficients

$$R = \left[\begin{array}{cccccc|cc} 1 & 0{,}28 & 0{,}26 & 0{,}19 & 0{,}12 & 0{,}20 & 0{,}02 & -0{,}02 \\ 0{,}28 & 1 & 0{,}41 & 0{,}14 & 0{,}10 & 0{,}11 & 0{,}01 & 0{,}02 \\ 0{,}26 & 0{,}41 & 1 & 0{,}20 & 0{,}12 & 0{,}18 & 0{,}02 & -0{,}01 \\ 0{,}19 & 0{,}14 & 0{,}20 & 1 & 0{,}41 & 0{,}06 & 0{,}02 & -0{,}00 \\ 0{,}12 & 0{,}10 & 0{,}12 & 0{,}41 & 1 & 0{,}10 & -0{,}00 & -0{,}04 \\ 0{,}20 & 0{,}11 & 0{,}18 & 0{,}06 & 0{,}10 & 1 & 0{,}00 & -0{,}03 \\ \hline 0{,}02 & 0{,}01 & 0{,}02 & 0{,}02 & -0{,}00 & 0{,}00 & 1 & 0{,}35 \\ -0{,}02 & 0{,}02 & -0{,}01 & -0{,}00 & -0{,}04 & -0{,}03 & 0{,}35 & 1 \end{array}\right]$$

makes it possible to divide that set of characteristics into two systems: W_1 — of the characteristics x_1 to x_6, i. e. all the metric characteristics, and W_2 — of the pigmentation characteristics x_7 and x_8. The vectors of those systems, m_1 and m_2, have a clear natural sense. The first defines the „mean" value of the head with respect to its 6 metric characteristics of the head, and the second defines the „mean" pigmentation.

We should now consider what part of the information contained in the initial set of characteristics will be found in the vectors and what part of that information will be lost, i. e. left in the residuals.

Formula (1) implies that the expected value of the quantity m is

$$E(m) = \frac{1}{n}(m_1 + m_2 + \ldots + m_n) = 0,$$

and

$$E(m^2) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{k}\sum_{p=1}^{k}\frac{x_{pj}-\bar{x}_p}{\sigma_p}\right)^2 = \frac{1}{k^2}\sum_{p=1}^{k}\sum_{q=1}^{k} r_{pq} = \bar{\bar{r}},$$

where $\bar{\bar{r}}$ denotes, as follows from the above, the mean correlation coefficients between the features of the system. Similarly, it is easy to find that $E(w_i) = 0$ and $E(w_i^2) = 1+\bar{\bar{r}}-2\bar{r}_i$, where we denote by $\bar{r}_i$ the mean correlation coefficient in the i-th column of the matrix of correlation coefficients. The greater the variance of the vector m the greater the amount of information contained in that vector,

$$\sigma_m^2 = E(m^2) = \bar{\bar{r}}.$$

The amount of information remaining in the residuals w_i increases with rhe increase of the mean variance of the residual w_i, i. e.

$$\bar{\sigma}_{wi}^2 = \frac{1}{k}\sum_{p=1}^{k} E(w_i^2) = 1-\bar{\bar{r}}.$$

The amounts of information retained and lost by replacing a system of characteristics by one vector m are given by the numbers $\bar{\bar{r}}$ and $1-\bar{\bar{r}}$.

Correlations will be found in a similar manner as we have found the expected values and variances. The correlation coefficient between the vector m and the i-th residual is equal to

$$r_{mw_i} = (\bar{r}_i-\bar{\bar{r}})/\sqrt{\bar{\bar{r}}(1+\bar{\bar{r}}-2\bar{r}_i)}.$$

Those correlations are not very great and of different signs. The correlation coefficient between two residuals

$$r_{w_pw_q} = (r_{pq}+\bar{\bar{r}}-\bar{r}_p-\bar{r}_q)/\sqrt{(1+\bar{\bar{r}}-2\bar{r}_p)(1+\bar{\bar{r}}-2\bar{r}_q)}. \tag{2}$$

The sign of that coefficient is the same as the sign of the numerator on the right side, $r_{pq}+\bar{\bar{r}}-\bar{r}_p-\bar{r}_q$. Summing these expressions for $p = 1, 2, \ldots, k$ and $q = 1, 2, \ldots, k$ we obtain 0. Thus if any coefficient $r_{w_pw_q}$ is different from 0, it can never occur that all the residuals should be positively correlated. Formula (2) will be useful below. Finally, if a set of characteristics has been divided into a few systems, and m_1 denotes the vector of one and m_2 the vector of another system, then the coefficient of the correlation between these two vectors is

$$r_{m_1m_2} = \frac{1}{n_1n_2}\sum_{p,q} r_{pq},$$

where p runs over the n_1 numbers of the features of the first system and q tuns over the n_2 numbers of the features of the second system.

In the example of the matrix of correlation coefficients calculated by Krauze and Łozińska I have divided the set of 8 characteristics into a system of metric characteristics and a system of pigmentation characteristics. The vector m_1 of the first system contains comparatively little

information since its amount is characterized by the ratio 0,33:0,67 relatively to the information contained in the residuals $w_1, \ldots, w_6$. The vector m_2 of the second system contains thus much more information regarding the characteristics of its system since the corresponding ratio is 0,68:0,32. The vector m_1 is weakly correlated with the characteristics of its system. The corresponding correlation coefficients are 0,01, 0,01, 0,07, 0,00, −0,04 and −0,10. The vector m_2 is uncorrelated with the residuals of the characteristics of its system. Finally the coefficient of the correlation between the vector m_1 and m_2 is −0,001.

The amount of information contained in the vector m of the system under consideration may prove to be too small, as for instance in the first system of the above example. Let us explain how to divide a system of characteristics into subsystems and how to describe a set of characteristics by further vectors, namely the vectors of those subsystems.

On subtracting the system vector from the normalized characteristics we are left with the residuals w_i. The correlations between those residuals are given by formula (2). Some of those coefficients must be negative. If they were all negative, the method proposed here would not permit a further division of features into subsystems. However, if some correlations prove to be positive, we shall group the features of the system in question—as before — into subsystems in such a manner that the residuals in each subsystem will be non-negatively correlated. The vector of the whole system will be called a *vector of the first order* and the vectors of the residuals in the subsystems will be called *vectors of the second order*. If necessary we can repeat the procedure, obtaining vectors of the higher order.

In the example investigated above we shall calculate according to formula (2) the coefficients of the correlation between the residuals of the features of the first system. They are given in the following matrix:

$$\begin{bmatrix} 1 & -0{,}11 & -0{,}18 & -0{,}23 & -0{,}29 & -0{,}13 \\ -0{,}11 & 1 & 0{,}06 & -0{,}30 & -0{,}31 & -0{,}15 \\ -0{,}18 & 0{,}06 & 1 & -0{,}25 & -0{,}33 & -0{,}19 \\ -0{,}23 & -0{,}30 & -0{,}25 & 1 & 0{,}14 & -0{,}31 \\ -0{,}29 & -0{,}31 & -0{,}33 & 0{,}14 & 1 & -0{,}22 \\ -0{,}13 & -0{,}15 & -0{,}19 & -0{,}31 & -0{,}22 & 1 \end{bmatrix}$$

As we see, two subsystems of residuals have been formed: for the characteristics 2 and 3, i. e. for the widths of head and face, and for the characteristics 4 and 5, i. e. for the lengths of face and nose. The vectors of the second order will contain ample information. The corresponding ratios are 0,53:0,47 and 0,57:0,43. Thus to define an individual from

the population examined in the work of Krauze and Łozińska (the material is taken from the anthropological survey of the district of Rybnik [8]) we can use, instead of the 8 anthropological characteristics, the vector m_1 of the metric characteristics, the vector m_2 of pigmentation, the vector of the second order m_3 of the indices of width (of head and face) and the vector of the second order m_4 of the indices of length (of face and nose). The significance of the individual vectors is evidenced by their standard deviations 0,58, 0,82, 0,73 and 0,75. If we divide by these numbers the quantities of the vectors m_1, m_2, m_3 and m_4 respectively, we shall obtain results normalized with respect to the mean 0 and variance 1. It permits a better orientation as regards the magnitudes of the individual vectors for definite individuals and a comparison of the individual vectors of the same individual. Here is an example of the first 5 individuals from paper [8].

The table of the 8 initial characteristics

No.	g-op	eu-eu	zy-zy	n-gn	n-sn	al-al	eyes	hair
1	181	149	128	114	55	38	6	*S*
2	181	152	131	118	48	39	9	*V*
3	183	157	130	118	50	38	7	*Q*
4	190	159	151	131	55	42	11	*V*
5	190	163	151	119	50	41	12	*R*

The table of the 4 normalized vectors

No.	vector of 6 metric characteristics m_1	vector of pigmentation m_2	vector of width m_3	vector of length m_4
1	−0,69	−0,84	−1,41	0,25
2	−0,53	−0,84	−0,85	−0,37
3	−0,31	−0,40	−0,52	−0,28
4	2,18	−0,55	−0,04	−0,29
5	1,14	0,15	1,27	−1,30

For some problems, particularly psychotechnical ones, a slightly different analysis of a set of characteristics is needed. Namely, after the elimination of the vectors of the first order interesting correlations (positive or negative) may be found between the residuals of the characteristics belonging

to different systems. It is then necessary to represent by means of vectors of the second order all that is common to all the residuals correlated. In an analysis of that kind a characteristic may appear first in one system of characteristics and then in another subsystem of the set of characteristics in question. The analysis of Hotelling and Thurstone provides for such contingencies. They can be met by my method also. I shall devote another paper to that problem.

References

[1] H. Cramér, *Mathematical methods of statistics*, Princeton 1946.

[2] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology 24 (1933), p. 417-441 and 498-520.

[3] M. Krauze and W. Łozińska, *Korelacje obszarowe ośmiu cech antropologicznych* (*Regional correlations of eight anthropological features*), Przegląd Antropologiczny (Anthropological Review) 23 (1957), p. 446-454.

[4] H. Milicerowa, *Zastosowanie wskaźników Perkala do charakterystyki budowy ciała bokserów* (*Application of Perkal indices to the characteristics of the build of boxers' bodies*), Materiały i Prace Antropologiczne (Anthropological Materials and Works) 20 (1956).

[5] J. Perkal, *O pewnych korelacjach obszarowych*, Časopis pro pěstování matematiku a fyziku Roč 75 (1949).

[6] — *O wskaźnikach antropologicznych* (*On anthropological indices*), Przegląd Antropologiczny (Anthropological Review) 19 (1953), p. 210-221.

[7] Ch. Spearman, *The abilities of man*, New York 1927.

[8] K. Stołyhwo, B. Jasicki and P. Sikora, *Zdjęcie antropologiczne Śląska, powiat rybnicki* (*Anthropological survey of Silesia, the district of Rybnik*), Materiały i Prace Antropologiczne (Anthropological Materials and Works) 13 (1956).

[9] G. Teissier, *Allométrie de taille et variabilité chez Maïa squinado*, Zoologie Expérimentale et Générale 92 (1955), p. 221-264.

[10] L. L. Thurstone, *Multiple-factor analysis*, Chicago 1947.

UNIWERSYTET WROCŁAWSKI, INSTYTUT MATEMATYCZNY (DZIAŁ ZASTOSOWAŃ)
UNIVERSITY OF WROCŁAW, INSTITUTE OF MATHEMATICS (SECTION OF APPLICATIONS)

Praca wpłynęła 30. 10. 1958

J. PERKAL (Wrocław)

O ANALIZIE ZESPOŁU CECH

STRESZCZENIE

W pracy tej rozpatruje się zespół cech $x_1, x_2, \ldots, x_k$ indywiduów pewnej populacji P. Znane są metody składowych (Hotellinga) i faktorów (Spearmana i Thurstone'a) analizy takiego zespołu. Wadami tych metod są trudności rachunkowe i interpretacyjne, a także niejednoznaczność (metody Thurstone'a). W pracy opisano dwie nowe metody analizy zespołu cech: jednowektorową i wielowektorową.

Pierwsza z nich pozwala zastąpić zespół cech jednym wskaźnikiem sumarycznym m, określającym sumaryczną wielkość indywiduum ze względu na zespół cech (pozwalającym też na uporządkowanie indywiduów populacji P według wielkości). Po wytrąceniu wskaźnika m z i-tej cechy pozostaje reszta w_i, tzw. *wskaźnik przyrodniczy* i-tej cechy, określający wielkość tej cechy w porównaniu z wszystkimi cechami zespołu.

Wielowektorowa metoda pozwala zastąpić zespół cech kilkoma wskaźnikami sumarycznymi $m_1, m_2, \ldots, m_r$, $r < k$. Każdy z nich jest wskaźnikiem sumarycznym pewnego podzespołu cech albo reszt. Metody te są rachunkowo łatwe i nie przedstawiają trudności interpretacyjnych. Natomiast wektory m_i nie są ortogonalne i mogą być skorelowane z resztami.

Ю. ПЕРКАЛЬ (Вроцлав)

ОБ АНАЛИЗЕ СОВОКУПНОСТИ КАЧЕСТВ

РЕЗЮМЕ

В настоящей работе автор предлагает два метода анализа совокупности качеств: одновекторный и многовекторный. Они аналогичны однофакторному и многофакторному методам анализа факторов. Вместо ортогональных факторов, некоррелированных с оставшимися, автор предлагает неортогональные векторы и в общем случае коррелированные с оставшимися. Зато векторы получаются из начальных качеств при помощи линейных операций. Этот метод легок арифметически и в интерпретации, чего нельзя утверждать об анализе факторов.