12. *Multidimensional Quantification. I*

By Chikio HAYASHI
Institute of Statistical Mathematics, Tokyo
(Comm. by Z. SUETUNA, M.J.A., Feb. 12, 1954)

This is a continued report from the papers 1) and 2) (see references) previously published, in which are treated some methods of quantification of qualitative data in multidimensional analysis and especially how to quantify qualitative patterns to secure the maximum success rate of prediction of phenomena from the statistical point of view. The important problem in multidimensional analysis is to devise the methods of the quantification of complex phenomena (intercorrelated behaviour patterns of units in dynamic environments) and then the methods of classification. Quantification means that the patterns are categorized and given numerical values in order that the patterns may be able to be treated as several indices, and classification means prediction of phenomena. The aim of multidimensional quantification is to make numerical representation of intercorrelated patterns synthetically to maximize the efficiency of classification, i. e. the success rate of prediction. Quantification does not mean finding numerical values but giving them patterns on the operational point of view in a proper sense. In this sense, quantification has not absolute meaning but relative meaning to our purpose. To achieve this purpose, we should always be aware of rational behaviour that is to pursue the so-called "optimum".

Thus we can give numerical values (weights) and distances to qualitative patterns (data) by analysing complex phenomena, quantifying and synthesizing. For example, "Kan" (efficient subjective judgement of experts) will be able to be analyzed and treated quantitatively and so to become a common property to us. In the present paper, the methods of quantification of qualitative patterns are considered in the case where an outside variable (realized by the outside criterion) is given in the form of qualitative classification. In this case it is most important that we must devise the methods to fulfil the property of validity. Let us take a universe of n elements, each of which has, as a label, behaviour patterns categorized by a survey method and is classified into only one class by the definite outside criterion (this is an outside variable). Here the outside criterion must be on an absolute scale and not change relatively to what elements of universe are classified (judged). That

is to say, $J(O_i)=J(O_j)=$constant independent of i, j; $i \neq j$, $i, j=1, 2, \ldots, n$ where $J(O_j)$ represents symbolically the frame of criterion when O_j is classified (judged), and O_i is the i-th element. It is our aim to predict to which class an element will belong in future which has a definite behaviour pattern at present, by the method of quantification using the past data.

The case where elements are classified into $S(\geqq 3)$ strata by an outside criterion which is unidimensional; that is to say, it means that the so-called law of transitivity holds in the field where elements are judged by only one norm. Whenever $S=2$, this method is applicable. Each element has a response pattern in R items which have several sub-categories respectively and the label of the stratum classified into. Response patterns are represented by sub-categories in items which an element checks in. The essential point of this method is same as in the paper 2) §3 (see references). Let $\{C_{11}, C_{12}, \ldots, C_{1K_1}\}$, $\{C_{21}, C_{22}, \ldots, C_{2K_2}\}, \ldots,$ $\{C_{R1}, C_{R2}, \ldots, C_{RK_R}\}$ be sub-categories in items. Let us consider that we give a numerical value x_{lm} to the m-th sub-category in the l-th item, C_{lm}, from the mathematico-statistical point of view. Response patterns of n elements are, for example, as below.

Response patterns (behaviour patterns) of elements

By outside criterion, outside variable ↓ Number of stratum	Item / Sub category / Element	1				.....	R			
		C_{11}	C_{12}	⋯	C_{1K_1}		C_{R1}	C_{R2}	⋯	C_{RK_R}
1	1	V				.....	V			
	2		V			.....		V		
	⋮					⋮				
	n_1		V			.....		V		
⋮	⋮					⋮				
S	1				V	.....		V		
	⋮					⋮				
	n_S				V	.....				V

V sign means the check in response of elements.

n_t is the number of elements belonging to the t-th stratum, where $n=\sum_{t=1}^{S} n_t$. Let $\{X_{1(i)}, X_{2(i)}, \ldots, X_{R(i)}\}$ be the response pattern of i-element, where $X_{j(i)}$ means the sub-category of the j-th item that i-element checks in. Now we use the score $a_i=x_{1(i)}+x_{2(i)}+\cdots x_{R(i)}$ as the score of i-element, where $x_{j(i)}$ is the numberical value given to the sub-category in the j-th item i-element checks in. The linear form is considered to be appropriate, according to the outside criterion being unidimensional and the idea of the first approximation. This linear form is not so restricted, considering from the point of view of making the data (having not linear relations) linear by quantifying qualitative patterns.

We have $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(a_i-\bar{a})$ as the toatal variance with respect to elements, where $\bar{a}=\frac{1}{n}\sum_{i=1}^{n} a_i$. Now we should like to quantify the sub-categories (items) so as to maximize the effect of stratification, that is, so as to maximize the correlation ratio $\eta^2=\frac{\sigma_b^2}{\sigma^2}$ where σ_b^2 is the variance between strata. This is a reasonable method of quantification, because η^2 is a measure of discriminative power of items, i.e. it is a measure of efficiency of classification (success rate of prediction).

If η^2 is large in the result of quantification, we can treat quantitatively the behaviour patterns by using x_{lm} (or a). Thus we can introduce a metric into qualitative patterns and define the distances between qualitative patterns (data) using the obtained values x_{lm} and so to speak, obtain the functional form of them which is valid in the above sense. In some cases, the degree of efficiency of items in prediction of phenomena will be given as quantitative functional relations in qualitative data. So as to require x_{lm} to maximize η^2, let us introduce the following definition.

Let

$$\delta_i(jk)\begin{cases}=1, & \text{if } i\text{-element checks in the } k\text{-th sub-category in the } j\text{-th item,}\\ =0 & \text{otherwise.}\end{cases}$$

Then

$$\sum_{k=1}^{K_j}\delta_i(jk)=1, \qquad \sum_{j=1}^{R}\sum_{k=1}^{K_j}\delta_i(jk)=R, \qquad a_i=\sum_{j=1}^{R}\sum_{k=1}^{K_j}\delta_i(jk)x_{jk}.$$

So

$$\bar{a}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{R}\sum_{k=1}^{K_j}\delta_i(jk)x_{jk},$$

$$\sigma^2=\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{R}\sum_{k=1}^{K_j}\delta_i(jk)x_{jk}\right)^2-\bar{a}^2=\frac{1}{n}\left(\sum_{j=1}^{R}\sum_{k=1}^{K_j}n_{jk}x_{jk}^2+\sum_{l}{}'\sum_{m}{}'\sum_{j=1}^{R}\sum_{k=1}^{K_j}f_{jk}(lm)x_{jk}x_{lm}\right)-\bar{a}^2,\right.$$

where

$$n_{jk}=\sum_{i=1}^{n}\delta_i(jk), \qquad f_{jk}(lm)=\sum_{i=1}^{n}\delta_i(jk)\delta_i(lm)$$

which represents correlation pattern between responses in items of each element, $\sum_{l}'\sum_{m}'\sum_{j=1}^{R}\sum_{k=1}^{K_j}$ covers all ranges of l, m, j, k, except $l=j$, $m=k$ holds simultaneously.

$$\sigma_b^2=\sum_{t=1}^{S}(\bar{\alpha}_t-\bar{\alpha})^2\frac{n_t}{n},$$

where

$$\bar{\alpha}_t=\frac{1}{n_t}\sum_{j=1}^{R}\sum_{k=1}^{K_j}g^t(jk)x_{jk},$$

$$g^t(jk)=\sum_{i(+)=1}^{m_t}\delta_{i(t)}(jk), \qquad n_{jk}=\sum_{t=1}^{S}g^t(jk), \qquad n_t=\sum_{k=1}^{K_j}g^t(jk),$$

$\delta_{i(t)}(jk)$ means $\delta_i(jk)$ which the i-element belonging to the t-th stratum has. Thus we have

$$\eta^2=\frac{\sigma_b^2}{\sigma^2}.$$

To maximize η^2 with respect to x_{uv} $(u=1, 2, \ldots, R,\ v=1, 2, \ldots, K_u)$, we take $\frac{\partial\eta^2}{\partial x_{uv}}=0$, that is, $\frac{\partial\sigma_b^2}{\partial x_{uv}}=\eta^2\frac{\partial\sigma^2}{\partial x_{uv}}$. Calculating this, using $\bar{\alpha}=0$ without loss of generality as easily be seen,

$$\frac{\partial\sigma_b^2}{\partial x_{uv}}=\frac{2}{n}\sum_{j=1}^{R}\sum_{k=1}^{K_j}\left(\sum_{t=1}^{S}\frac{g^t(jk)g^t(uv)}{n_t}\right)x_{jk}=\frac{2}{n}\sum_{j=1}^{R}\sum_{k=1}^{K_j}h_{uv}(jk)x_{jk},$$

$$\frac{\partial\sigma^2}{\partial x_{uv}}=\frac{2}{n}\left(\sum_{l=1}^{R}\sum_{m=1}^{K_l}f_{uv}(lm)x_{lm}\right),$$

where

$$h_{uv}(jk)=\sum_{t=1}^{S}\frac{g^t(jk)g^t(uv)}{n_t}.$$

Then

$$\sum_{j=1}^{R}\sum_{k=1}^{K_j}h_{uv}(jk)x_{jk}=\eta^2\sum_{l=1}^{R}\sum_{m=1}^{K_l}x_{lm}f_{uv}(lm) \quad (u=1, 2, \ldots, R,\ v=1, 2, \ldots, K_u).$$

Let the matrix $(h_{uv}(jk))$ be H, the matrix $(f_{uv}(lm))$ be F, vector (x_{jk}) be X. The above equation is written as follows:

$$HX=\eta^2FX.$$

It is our problem to solve this under the conditions $\sum_{k=1}^{K_j}n_{jk}x_{jk}=0$, $(j=1, 2, \ldots, R)$, and to require the largest maximum value (this is the largest value) of η^2 being not equal to 1 and the corresponding vector X to it. It is easily proved that $\eta^2(0\leqq\eta^2\leqq1)$ satisfying the above equations exists and the values we require exist. We can obtain the required η^2 and x_{lm} by the successive approximation methods.

The method of classification (prediction) by quantified behaviour

patterns and its efficiency are described in the paper 2) §3 (listed below), because it is considered that the stratum means the label of outcome of elements in future.

The applications of this method to analyse social phenomena will be shown in papers 3) and 4) (listed below).

References

1) C. Hayashi: On the quantification of qualitative data from the mathematico-statistical point of view, Annals of the Institute of Statistical Mathematics, **2**, No. 1 (1950).

2) C. Hayashi: On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, Annals of the Institute of Statistical Mathematics, **3**, No. 2 (1952).

3) C. Hayashi: Multidimensional quantification, with applications to the analysis of social phenomena, Annals of the Institute of Statistical Mathematics, **4**, No. 2 (1954) (in press).

4) C. Hayashi: Measurement and quantification of social attitude (in Jap.), Proceedings of the Institute of Statistical Mathematics, **1**, No. 2 (1954) (in press).