TWO DIMENSIONAL QUANTIFICATION BASED ON THE MEASURE OF DISSIMILARITY AMONG THREE ELEMENTS

CHIKIO HAYASHI

(Received Sept. 27, 1971)

We have presented the method of quantification based on the measure of similarity between two elements, which we call e_{ij}-type quantification [1]. This e_{ij}-type quantification is found to be closely related to Guttman's SSA or POSA and Hayashi's MDA [2], [3], [4], [5], in the sense that this type of quantification is applied in these methods.

In the e_{ij}-type quantification, e_{ij} means similarity. Even if e_{ij} means dissimilarity, we can apply the same formula by using $(-e_{ij})$ instead of e_{ij}. Here, as a generalization of e_{ij}-type quantification, we consider two dimensional quantification based on the measure of dissimilarity $e_{ijk} \geqq 0$ among three elements i, j, k, where $i, j, k=1, 2, \cdots, R$, R being the size of the elements. We assume that e_{ijk}'s are represented by numerical values for all combinations of i, j, k, $i, j, k=1, 2, \cdots, R$, and the value of e_{ijk} is independent of the order of i, j, k, i.e. uniquely determined for the combination of three elements i, j, k. And furthermore we assume that, if $e_{ijk} > e_{lmn}$, the degree of dissimilarity for i, j, k is higher than that for l, m, n.

This problem arises from the following essential examples; construction of the color space from the information of color harmony data among (more than) three colors, finding of the spatial configuration of persons from the productivity data among (more than) three persons, etc., these data being not essentially derived from the relations between the two.

Based on this information, we consider the problem of finding the configuration of the R elements in two dimensional space, which preserves the dissimilarity information with respect to e_{ijk}'s $(i, j, k=1, 2, \cdots, R)$ as much as we can. This is a generalization of the e_{ij}-type quantification.

Model of e_{ijk}-type quantification

We take two dimensional Euclidean space and summarize the information with respect to the relations e_{ijk}'s $(i, j, k=1, 2, \cdots, R)$ as the configuration of R elements in that space. We reduce this process to the one of finding a correspondence of e_{ijk} to a dispersion measure among points i, j, k in two dimensional Euclidean space, i.e. a representation of e_{ijk} by a dispersion measure among i, j, k. It is natural that we define the dispersion measure among i, j, k as square of the area of triangle (i, j, k). Let (x_i, y_i), (x_j, y_j) and (x_k, y_k) be the coordinates of points i, j, k in the two dimensional Euclidean space. Thus we take the corresponding measure $\varphi(i, j, k)$ among i, j, k instead of the corresponding measure $(x_i - x_j)^2$ between i and j in the e_{ij}-type quantification, where

$$\varphi(i, j, k)=\begin{vmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_k & y_k & 1 \end{vmatrix}^2$$

and this is equal to the square of the area of the triangle except a constant multiplier. Let us require that the above-mentioned configuration of R elements should be determined so that $\varphi(i, j, k)$ becomes a monotone increasing function of e_{ijk} as far as possible. As an analogy with the e_{ij}-type quantification, we take

$$Q=\sum_i^R \sum_j^R \sum_k^R \frac{e_{ijk}\varphi(i, j, k)}{\sigma_x^2 \sigma_y^2},$$

where $\sigma_x^2(\sigma_y^2)$ is the variance of $x(y)$ and we may take $\sigma_x^2(\sigma_y^2)=1$ without loss of generality. Of course, $Q \geqq 0$ holds.

We want to require (x_i, y_i) $i=1, 2, \cdots, R$ to maximize Q. For simplicity, we take $(\bar{x}, \bar{y})$ as the origin where $\bar{x}(\bar{y})$ is the mean value of $x(y)$.

Quantification of the elements

From $\partial Q/\partial x_u=0$, $u=1, 2, \cdots, R$ and $\partial Q/\partial y_v=0$, $v=1, 2, \cdots, R$, we obtain

(1.1) $$A_{uu}(y)x_u+\sum_{i \neq u} A_{ui}(y)x_i=\lambda^2 x_u, \qquad u=1, 2, \cdots, R$$

(1.2) $$A_{vv}(x)y_v+\sum_{j \neq v} A_{vj}(x)y_j=\lambda^2 y_v, \qquad v=1, 2, \cdots, R$$

where

$$e_{uij}+e_{iuj}+e_{iju}=E_{uij} \quad (=3e_{uij}=3e_{iuj}=3e_{iju})$$

$$A_{uu}(y)=\sum_{\substack{i,j\neq u\\ i\neq j}}\sum \begin{vmatrix} y_i & 1 \\ y_j & 1 \end{vmatrix} \begin{vmatrix} y_i & 1 \\ y_j & 1 \end{vmatrix} E_{uij}$$

$$A_{ui}(y)=-2\sum_{\substack{j\neq u\\ j\neq i}} \begin{vmatrix} y_u & 1 \\ y_j & 1 \end{vmatrix} \begin{vmatrix} y_i & 1 \\ y_j & 1 \end{vmatrix} E_{uij}=A_{iu}(y)$$

$$A_{vv}(x)=\sum_{\substack{i,j\neq v\\ i\neq j}}\sum \begin{vmatrix} x_i & 1 \\ x_j & 1 \end{vmatrix} \begin{vmatrix} x_i & 1 \\ x_j & 1 \end{vmatrix} E_{vij}$$

$$A_{vj}(x)=-2\sum_{\substack{i\neq v\\ i\neq j}} \begin{vmatrix} x_v & 1 \\ x_i & 1 \end{vmatrix} \begin{vmatrix} x_j & 1 \\ x_i & 1 \end{vmatrix} E_{vij}=A_{jv}(x)$$

$$\lambda^2=Q/R \quad \text{and of course} \quad \lambda^2\geqq 0\,.$$

We require the maximum characteristic root λ^2 ($\lambda^2\neq 0$) and the corresponding characteristic vector.

(1.1) is symmetric and a characteristic equation with respect to x if y's are fixed. (1.2) is symmetric and a characteristic equation with respect to y if x's are fixed. The mean value of x's which satisfy (1.1) is always equal to zero, because $A_{uu}(y)+\sum_{i\neq u} A_{ui}(y)=0$, $A_{ui}(y)=A_{iu}(y)$, and $\lambda^2\neq 0$. The mean value of y's which satisfy (1.2) is always equal to zero, because $A_{vv}(x)+\sum_{j\neq v} A_{vj}(x)=0$, $A_{vj}(x)=A_{jv}(x)$ and $\lambda^2\neq 0$.

From the above relations, formally speaking, $\lambda=0$ is a characteristic root and constant vector is a solution in (1.1) and (1.2). However, the form of Q does not admit it.

In the e_{ij}-type quantification, the solution is to be found invariant except the constant multiplier under the transformation $\alpha e_{ij}+\beta$ (α, β are constants). However, in the present quantification, the solution is not generally invariant under the transformation $e_{ijk}+\beta$, though the solution is invariant except the constant multiplier under the condition that the vector x and y are to be orthogonal i.e. $\sum_i^R x_i y_i=0$. Of course we can solve the equations adding this condition.

We must solve (1). To solve (1), we use successive approximation method.

[First Step] We take the first approximation of y's and then we solve the characteristic equation with respect to x and require the characteristic vector corresponding to the maximum characteristic root. We normalize the vector x (as $\bar{x}=0$ holds, we take the variance of x equal to 1).

[Second Step] Using the obtained normalized x's and calculating $A_{vv}(x)$ and $A_{vj}(x)$, $v=1, 2, \cdots, R$, $j=1, 2, \cdots, R$, we solve the characteristic equation with respect to y and require the characteristic vector corresponding to the maximum characteristic root. We normalize the vector y (as $\bar{y}=0$ holds, we take the variance of y equal to 1).

[Third Step] Using the obtained normalized y's and calculating $A_{uu}(y)$ and $A_{ui}(y)$, $u=1, 2, \cdots, R$, $i=1, 2, \cdots, R$, we solve the characteristic equation with respect to x and determine the characteristic vector corresponding to the maximum characteristic root. We normalize the vector x.

[Repeat this process] We generally confirm the convergence from the forms of (1) by finding a good first approximation. We can judge the convergence by the followings.

a. The maximum characteristic root to satisfy (1.1) is approximately equal to the maximum characteristic root to satisfy (1.2).

b. The obtained vectors x and y are the approximately equal to the vectors x and y obtained previously.

Goodness of fit

The goodness of fit is given by the measure $\eta_P'^2$ defined as follows:

$$\eta_P'^2=1-\frac{S^2}{\sigma_e^2},$$

where σ_e^2 is the variance of e's (e_{ijk}, $i, j, k=1, 2, \cdots, R$) and

$$S^2=\min_P \sum_i \sum_j \sum_k [e_{ijk}-P(\varphi)]^2,$$

where $P(\varphi)$ is a monotone increasing function. Conventionally we take polynomial as $P(\varphi)$ and, in the simplest case, we take $P(\varphi)=a\varphi+b$, a and b being constants. Occasionally we take $P(\varphi)=a\varphi^2+b\varphi+c$, where a, b and c are constants.

S^2/σ_e^2 means how much the variance of e reduces by using the model φ mentioned above, comparing with the variance of e. $\eta_P'^2$ is considered to represent the effectiveness of the model. We have $1\geqq \eta_P'^2\geqq 0$.

First approximation

We find the first approximation of y by the e_{ij}-type quantification. Now let us define $b_{ij}=\sum_{k\neq i,j}^{R} e_{ijk}$ $(i\neq j)$ and quantify the elements using b_{ij}. As the first approximation vector y, take the vector corresponding to the maximum characteristic root μ of the following characteristic equation.

$$\Big(\sum_{\substack{l=1\\ l\neq w}}^{R} a_{wl}\Big) y_w - \sum_{\substack{l=1\\ l\neq w}}^{R} a_{wl} y_l = \mu y_w\,, \qquad w=1, 2, \cdots, R$$

where $a_{kl}=b_{kl}+b_{lk}$.

Example

Let e_{ijk} be the unproductivity (dissimilarity) measure ($0\sim10$) in i-j-k-cooperation team, i.e. $10-e_{ijk}$ be the productivity (similarity) measure in i-j-k-cooperation team. This example was calculated by Fumi Hayashi, using C-F-Hayashi-EQCP computer program.

$R=6$

$e_{123}=1.0$	$e_{234}=8.0$
$e_{124}=7.0$	$e_{235}=7.0$
$e_{125}=6.0$	$e_{236}=9.0$
$e_{126}=9.0$	$e_{245}=6.0$
$e_{134}=7.0$	$e_{246}=8.0$
$e_{135}=6.0$	$e_{256}=7.0$
$e_{136}=9.0$	$e_{345}=3.0$
$e_{145}=4.0$	$e_{346}=5.0$
$e_{146}=9.0$	$e_{356}=3.0$
$e_{156}=6.0$	$e_{456}=1.0$

We take the following as the conditions which are to be used to judge the convergence:

1) Let the maximum characteristic root λ_x^2 satisfy (1.1) and the maximum characteristic root λ_y^2 satisfy (1.2). $|\lambda_y/\lambda_x-1|\leqq0.0001$. We neglect the absolute value of the ratio minus 1 smaller than 0.0001 and regard $\lambda_x(\lambda_y)$ as being equal to $\lambda_y(\lambda_x)$.
2) We solve the equations (1.1) and (1.2) by Jacobi-method. We take 0.005 as the convergence condition in Jacobi-method.

Thus we have

$x_1=-0.1342$	$y_1=0.6230$
$x_2=0.6403$	$y_2=-0.5256$
$x_3=0.1437$	$y_3=0.2579$
$x_4=0.0915$	$y_4=0.0830$
$x_5=-0.0041$	$y_5=0.0691$
$x_6=-0.7369$	$y_6=-0.5074$.

See Fig. 1 (with scale of 10 times), where $1\rightarrow$A, $2\rightarrow$B, $3\rightarrow$C, $4\rightarrow$D, $5\rightarrow$E, $6\rightarrow$F.

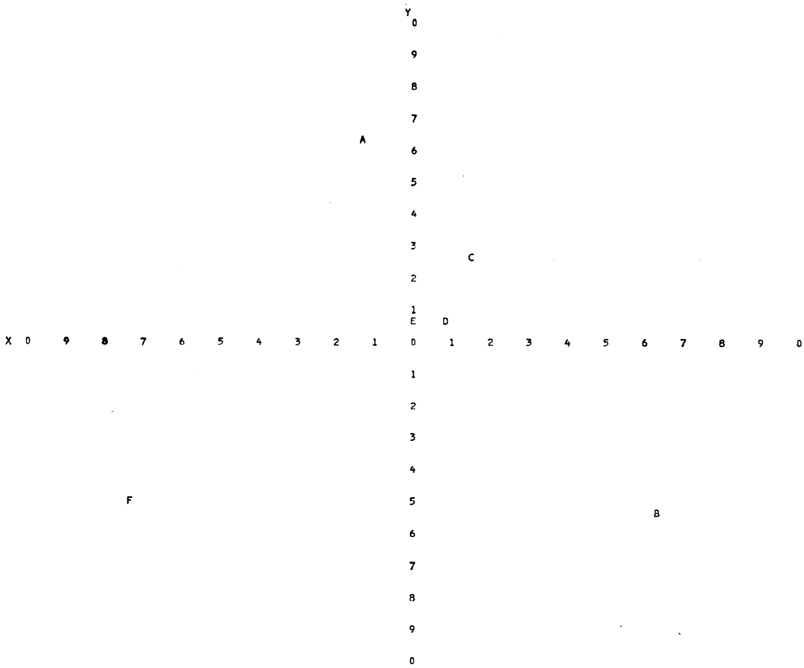

Fig. 1. Configuration of the elements (last solution)

3) Pseudo-Correlation ratio

$$\eta_P'^2 = 1 - \frac{\sum_i \sum_j \sum_k (e_{ijk} - \boldsymbol{P}(\varphi))^2}{\sigma_e^2}$$

(i) When $P(\varphi)$ is linear $a\varphi + b$, $\eta_L'^2$ is 0.2579 $(P=L)$.
(ii) When $P(\varphi)$ is quadratic $a\varphi^2 + b\varphi + c$, $\eta_Q'^2$ is 0.3649 $(P=Q)$.

Computer program

The computer program for this quantification is called C-F-Hayashi-E3QCP designed by Fumi Hayashi in our Institute which will be published in Proceedings of the Institute of Statistical Mathematics Vol. 20, No. 1. The communications relating to this program should be addressed to the author.

Acknowledgement

The author would like to acknowledge the considerable assistance of Miss F. Hayashi and the valuable advices of Dr. T. Uematu in the Institute.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematic-statistical point of view, *Ann. Inst. Statist. Math.*, **2**, 93-96; (1967), Note on multidimensional quantifiation of data obtained by paired comparison, *Ann. Inst. Statist. Math.*, **19**, 363-365.

[2] Guttman, L. (1968). A general non metric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika*, **33**, 469-505.

[3] Guttman, L. (1970). Some examples of partial order scalogram analysis, *Israel Institute of Applied Social Research*, (stencilled).

[4] Hayashi, C. (1969). Distance and dimension from a statistical point of view, *Market Research* (Published from Yoron Kagaku Kyokai), **1**, 20-34.

[5] Hayashi, C. (1971). MSA, POSA and Theory of Quantification, *Bulletin of the Research Institute of Public Opinion* (Yoron Kagaku Kyokai).