Environmental Health Perspectives
Vol. 32, pp. 125-129, 1979

Discussion of Tanaka's Paper

by Chikio Hayashi*

The theory of quantification is a method of statistical data analysis of categorical data. In other words, this is a kind of data theory and is closely related to optimum scaling method. This method has been mainly developed by Guttman in Israel and Hayashi in Japan. The multidimensional scaling method, which has been recently developed, is considered to be a continuation of the theory of quantification. Tanaka discussed mathematically some of Hayashi's methods of quantification. The present paper, gives an overview of the methods developed by him and other closely related methods and gives the orientation of those methods introduced by Tanaka. Then, as an illustration of exploratory categorical data analysis, the experimental data of Grizzle are analyzed by using the second method of quantification. The data structure is shown heuristically as a spatial configuration of factors in two-dimensional Euclidean space.

Tanaka discussed my early methods of quantification from the stand point of mathematical statistics and added his newly developed method in the case of ordered categories with some asymptotic theories (*1*).

The terminology and notations in his paper are somewhat different from those in my papers. It is only remarked here that external criterion or criterion variable is used for outside criterion or outside variable in my original papers. As for the notations, readers must carefully follow. The method of quantification is considered to be a kind of scaling method of categorical data. The most important problem of quantification, both in fundamental idea and in methodology, is assigning numerical vectors to categorical data from the point of view of optimization for our purpose under some minimum assumptions. The idea is briefly described in a previous paper (*2*). From this idea, many methods including the four methods detailed in Tanaka's paper (*1*), have been developed; these are shown in Table 1.

This list contains the methods of quantification published by Hayashi with some closely related important methods to orientate his methods. Of course, this is not exhaustive; besides these, interesting methods have been developed by Hayashi's colleagues inside or outside Japan. The methods of quantification are frequently used in data analysis because the computer programs are now available in some of them as Tanaka mentioned.

*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan.

The leading ideas in the methods of quantification play a productive role in data analysis and in developing new statistical methods necessary for detective analysis of data. For example, the fourth method gives one similar realization of the aim of nonmetric multidimensional scaling (MDS) (*17-23*) and naturally proceeds to MDS.

Some comments are added here. Usually the responses are given in the form,†

$$\delta_i(j,k_j) = 1, \text{ the } i\text{-th element responses in the } k_j \text{ category in the } j\text{-th item}$$
$$= 0, \quad \text{otherwise}$$

However, we often meet the situation that the responses are not always determinative but may be expressed as a probabilistic event. In this case

$$\delta_i(j,k_j) = p_h(j,k_j)$$

if $i \in h$; $h = 1, 2, \ldots, H$; $H << N$ (size of sample),

where

$$1 \geqq p_h(j,k_j) \geqq 0$$

and

$$\sum_{k_j} p_h(j,k_j) = 1 \text{ for all } j$$

†In Tanaka's paper,

$$x_\alpha(i,j) = \begin{cases} 1, \text{ if subject } \alpha \text{ belongs to category } j \text{ of the } i\text{-th attribute item} \\ 0, \text{ otherwise} \end{cases}$$

is used for the 1st method, while other notations are used for the same event, in other methods.

Table 1. Method of quantification (or scaling) as one type of analysis of categorical data.[a]

I. Existence of outside variable
(quantification or scaling of factors for estimating outside variable)
- A. Numerical outside variable
  - 1. Unidimensional
    1st method (one type of regression analysis)
  - 2. Multidimensional
    1st method (by means of a vector correlation coefficient)
- B. Categorical outside variable
  - 1. Classification into two groups
    - a. Absolute inference
      - (1) Discrimination based on a measure of correlation ratio; second method (one type of discriminant analysis)
      - (2) Discrimination based on a measure of success rate of estimation (or prediction)
    - b. Relative inference
      Guttman's quantification method of categorical factors in case of paired comparison
  - 2. Classification into more than three groups
    - a. Absolute inference
      - (1) Unidimensional or ordered classification 2nd method (scaling by assignment of unidimensional numerical value based on correlation ratio
      - (2) Multidimensional or unordered classification
        - (a) Scaling by assignment of multidimensional numerical values based on generalized correlation ratio (2nd method)
        - (b) Unidimensional scaling of multifactors (multidimensional metrical space construction by multifactors) based on generalized variance
    - b. Relative inference
      - (1) By paired comparison (application of 2nd method or Guttman's method generalized)
      - (2) By simultaneous many objects comparison (e.g., ordering of N objects); application of 2nd method

II. Nonexistence of outside variable
(quantification or scaling of factors for understanding their data structure)
- A. Data based on response pattern of elements
  - 1. Representation of a degree of mutual dependence between two variables; quantification of categorical variable by maximization of correlation coefficient
  - 2. Construction of spatial configuration of data based on relations among more than three variables; third method (in the case of those variables being numerical, factor analysis or principal component analysis may be used under some strict conditions)
- B. Data based on relations between (among) elements
  - 1. Numerical representation of similarity or dissimilarity
    - a. Between two elements
      - (1) Nonmetrical treatment if valid; 4th method (e_{ij}-type quantification by use of information of those relations with flexibility)
      - (2) Metrical treatment if valid; K-L type quantification and Torgerson-Gower method
    - b. Among more than three elements
      - (1) Nonmetrical treatment if valid; e_{ijk}-type quantification; generalization of 4th method
      - (2) Metrical treatment if valid; Torgerson's metrical multidimensional scaling or MDS
  - 2. Nonmetrical representation
    - a. Representation of relations between two elements by an absolute judgement or criterion
      - (1) Rank-ordered representation of similarity or dissimilarity; nonmetric MDS
        - (a) Shepard method
        - (b) Kruskal method
        - (c) Smallest space analysis, SSA (Guttman, Lingoes)
        - (d) Individual difference model (Carroll)
        - (e) Asymmetric model (Young, Hayashi)
      - (2) Belonging representation of similarity or dissimilarity to rank-ordered group : nonmetric MDS (Minimum Dimension Analysis MDA or MDA-OR)
      - (3) Nominal classification; MDA-UO
    - b. Representation of relations by a relative judgement
      - (1) By paired comparison (Hayashi's multidimensional unfolding method)
      - (2) By simultaneous many objects comparison; Coomb's multidimensional unfolding method

[a]Hayashi's papers on the theoretical aspects of the method of quantification are listed in the references (*3-16*).

$p_h(j,k_j)$ denotes the probability that i element responses in category k_j of the j-th item when i belongs to h class. Even in such cases, the calculation in methods of quantification is done in quite the same way as in the dichotomous 1, 0 responses. The information of $i \in h$ and p's must be given in the data. For example, $H = 3., h = +, \pm, -$ in item 1 which has of course three categories $+, \pm, -$. It is supposed that

$$p_+(1,+) = 0.80, p_+(1, \pm) = 0.15, p_+(1, -) = 0.05$$

if there are i responses in + in item 1,

$$p_\pm(1, +) = 0.20, p_\pm(1, \pm) = 0.60, p_\pm(1, -) = 0.20$$

if there are i responses in $\pm$, and

$$p_-(1, +) = 0.00, p_-(1, \pm) = 0.10, p_-(1, -) = 0.90$$

if there are i responses in $-$.

This model must be verified; also the values of probability must be estimated by some fundamental research before the present analysis. This idea may be crucial in some medical data. From our experience, fluctuation of measurement data which is due to bioactivity and measurement error is not usually neglected and fairly large even if the conditions of

Table 2. External criterion (type of lesion).

Microcardial scar	Infarct	π_i
−	−	π_1
−	+	π_2
+	−	π_3
+	+	π_4

subject
j

Table 3. Factors x . . (k,l).[a]

$k = 1$	Location and race	
	New Orleans,	white
		$(l = 1)$ s_{11}
	Oslo	$(l = 2)$ s_{12}
	New Orleans,	Negro
		$(l = 3)$ s_{13}
$k = 2$	Age	
	35-44	$(l = 1)$ s_{21}
	45-54	$(l = 2)$ s_{22}
	55-64	$(l = 3)$ s_{23}
	65-69	$(l = 4)$ s_{24}

[a] s_{kl} means numerical vector given to category l in the k-th item.

measurements are strictly regulated.

As an illustration of the second method of quantification, the data of Table 3 in Grizzle's paper (*24*) on cases of coronary heart disease classified by type of lesion, age, location and race are used. These data are reproduced in Table 2. However this application may not be satisfied because of the properties of the data; this analysis will be done for the understanding of the second method.

According to Tanaka's notation, we have the factors listed in Table 3.

Subject j belongs to one of the π and has a response in one category in item 1, i.e., location and race ($k = 1$), New Orleans white, Oslo, New Orleans Negro and a response in one category in item 2, i.e. age ($k = 2$), 35-44, 45-54, 55-64, 65-69. Grizzles data are rewritten as convenient for understanding of the second method in Table 4.

This analysis gives an information for an individual element (subject) of a group while Grizzle's result gives an information for in-group relations. Note that the meaning is rather different. Apart from this point, the calculation is shown as below. The numerical vectors given to item-categories and the calculated mean values of external criteria $Y_{(c)i}$ are

Table 4. Data of Grizzle as presented without weighting.

Myocardial scar	Infarct	Location and race 11	12	13	Age 21	22	23	24	Total
−	−	40	28	18	20	26	32	8	86
−	+	94	49	32	18	42	82	33	175
+	−	61	75	33	10	37	81	41	169
+	+	46	71	11	14	29	62	23	128
11 New Orleans White		241	0	0	29	66	114	32	
12 Oslo			223	0	17	32	110	64	
13 New Orleans Negro				94	16	36	33	9	$N =$
21 Age 35-44			(Symmetry)		62	0	0	0	558
22 45-54						134	0	0	
23 55-64							257	0	
24 65-69								105	

Table 5. Numerical vectors given to S_{kl}.

k			1st dimension	2nd dimension	3rd dimension	Size of sample
1	New Orleans White	1	−0.63	0.43	0.84	241
	Oslo	2	0.89	−0.89	−0.13	223
	New Orleans Negro	3	−0.49	1.02	−1.84	94
2	Age 35-44	1	−1.46	−1.87	−0.40	62
	45-54	2	−0.21	−0.54	−0.16	134
	55-64	3	0.25	0.25	0.30	257
	65-69	4	0.51	1.18	−1.18	−0.28
105						
						558 (total)

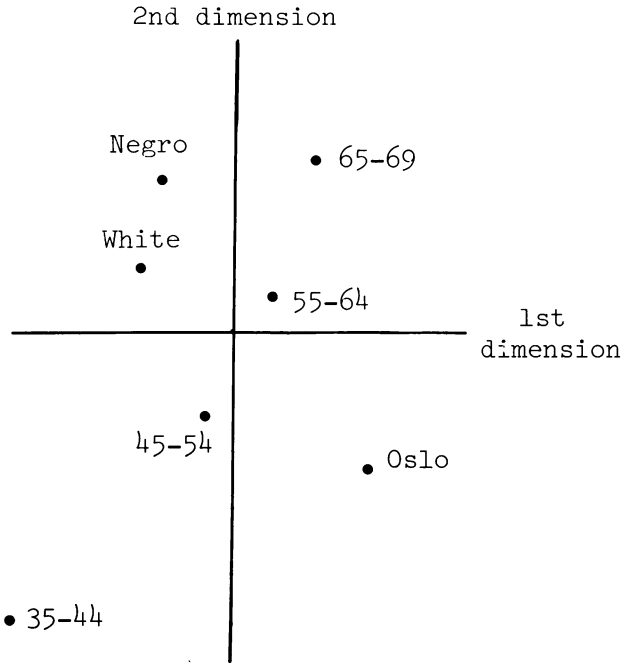

FIGURE 1. Vector s_{kl}.

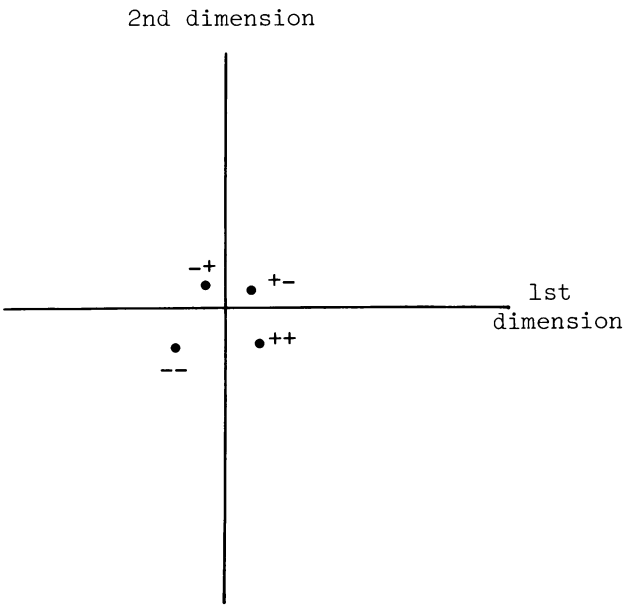

FIGURE 2. Mean value $Y_{(c)i\cdot}$.

Table 6. Mean value $Y_{(c)i\cdot}$ of external criteria.

		1st dimension	2nd dimension	3rd dimension	Size of sample
−	−	−0.37	−0.27	−0.10	86
−	+	−0.17	0.18	0.08	175
+	−	0.18	0.14	−0.10	169
+	+	0.23	−0.25	0.08	128
					558 (total)

shown in Tables 5 and 6 when the total variance is taken to be equal to 1.

The square root of correlation ratios which are obtained as latent roots in the latent equation are 0.23, 0.20, and 0.09 respectively. The first dimension and second dimension are adopted corresponding to the maximum and second maximum latent root. To make clear the features, s_{kl} and $Y_{(c)i\cdot}$ are shown in Figures 1 and 2. However, the discrimination power may be weak, as expected from the properties of the data; the configuration is interesting and the data structure is well shown.

In Figure 1, age has a linear structure and, New Orleans white and New Orleans Negro have similar features, quite different from Oslo. It is observed in Figure 2 that the values in the first dimension give the discrimination between myocardinal scar existence and nonexistence, whereas the values in the second dimension give the discrimination between positive relation (++, −−) and negative relation (+−, −+). The correspondence of Figure 1 to Figure 2 reveals the meaning of items.

REFERENCES

1. Tanaka, Y. Review of methods of quantification. Environ. Health Perspect. 32: 111 (1979).
2. Hayashi, C. Methodological problems in mass communications research, — from a statistico-mathematical standpoint. Studies of Broadcasting, No. 9, Nippon Hoso Kyokai, (1973, pp. 121-151.
3. Hayashi, C. On the quantification of qualitative data from the mathematicostatiscal point of view, Ann. Inst. Statist. Math. 2: 35 (1950).
4. Hayashi, C. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematicostatistical point of view. Ann. Inst. Statist. Math 3: 69 (1952).
5. Hayashi, C. Multidimensional quantification I, II. Proc. Japan. Acad. 30: 61, 165 (1954).
6. Hayashi, C. Multidimensional quantification with the application to analysis of social phenomena. Ann. Inst. Statist. Math. 5: 121 (1955).
7. Hayashi, C. Sample survey and theory of quantification. Bull. ISI, 38: 505 (1961).
8. Hayashi, C. Multidimensional quantification of the data obtained by the method of paired comparison. Ann. Inst. Statist. Math. 16: 231 (1964).
9. Hayashi, C. Note on multidimensional quantification of data obtained by paired comparison, Ann. Inst. Statist. Math. 19: 363 (1967).
10. Hayashi, C. One-dimensional quantification and multidimensional quantification. Ann. Japan Assoc. Phil. Sci. 3: 29 (1968).

11. Hayashi, C. Response error and biased information. Ann. Inst. Statist. Math. 20: 211 (1968).
12. Hayashi, C. Two-dimensional quantification based on the measure of dissimilarity among three elements. Ann. Inst. Statist. Math. 24: 251 (1972).
13. Hayashi, C. Minimum dimensional analysis MDA. Behaviormetrika 1: 1 (1974).
14. Hayashi, C. Minimum dimension analysis, MDA-OR and MDA-UO. In: Essays in Probability and Statistics. S. Ikeda, et al. Ed., Shinko Tsusho Co. Tokyo, 1976, pp. 395-412.
15. Hayashi, C., and Suzuki, T. Quantitative approach to a cross-societical research I, II. a comparative study of Japanese national character, Ann. Inst. Statist. Math. 26: 455 (1974); Ibid., 27: 1 (1975).
16. Hayashi, C., and Hayashi, F. Comparison of two types of multidimensional scaling methods: minimum dimension analysis MDA-OR and MDA-UO. Ann. Inst. Statist. Math 30: 199 (1978).
17. Shepard, R. N. The analysis of proximites: Multidimensional scaling with an unknown distance function I, II. Psychometrika 27: 125, 219 (1962).
18. Kruskal, J. B. Multidimensional scaling: a numerical method. Psychometrika 29: 1 (1964).
19. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika 29: 115 (1964).
20. Guttman, L. A general non-metric technique for finding the smallest coordinate space for a configuration of points. Psychometrika 33: 469 (1968).
21. Carroll, J. D., and Chang, J. J. Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart-Young" decomposition. Psychometrika 35: 283 (1970).
22. Takane, Y., Young, F. W., and de Leeuw, J. Nonmetric individual difference multidimensional scaling, The L. L. Thurstone Psychometric Laboratory, University of North Carolina, No. 147, (1975). of nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features, Psychometrika 42: 7 (1976).
23. Carroll, J. D. Spatial, non-spatial and hybrid models for scaling. Psychometrika 41: 439 (1976).
24. Grizzle, J. E., and Koch, G. G. Some applications of categorical data analysis to epidemiological studies. Environ. Health Perspect. 32: 000 (1979).