CHUGOKUGAKUEN J. 2021
Vol. 20, pp. 1-6

CHUGOKUGAKUEN
Journal
http://www.cjc.ac.jp/

Original Article

A Study on the Placement of Variables in a Modified Constellation Graph

Mamoru Fukumori[1], Mika Fujiwara[1] and Shoji Kajinishi[2]

*[1] Department of Information Science and Business Management, Chugoku Junior College,
Niwase 83, Kita-ku, Okayama City, 701-0197, Japan*
*[2] Graduate School of Environmental and Life Science, Okayama University,
Tsushima-Naka 3-1-1, Kita-ku, Okayama City, 700-8530, Japan*

In this study, we proposed a method of variable placement in a modified constellation graph. In the graph, p variables are arranged at equal intervals on the circumference of a semicircle, and the vectors of length x_{ij} (i=1, 2, ..., n; j=1, 2, ..., p) are concatenated. This modified constellation graph has the advantage of providing an easier understanding of the object characteristics than is possible using the constellation graph. However, results of the modified constellation graph are affected by the placement of the variables; therefore, it is necessary to determine their optimum placement. In this study, we proposed to place the variables such that the covariance or horizontal distance of the objects is as large as possible. In addition, the proposed method was applied to an achievement survey.

Keywords: Constellation graph, Modified constellation graph, Principal component loadings, Maximum covariance, Maximum horizontal distance

Introduction

In statistical analyses, graphical representations are often used to facilitate the interpretation of the results. Graphical representation is a visualization method for intuitively grasping the summary of information and detailed features of data and is an effective means for the exploratory interpretation of quantitatively obtained results [1)].

Various graphs have been proposed as visualization methods for multivariate data, one of which is the constellation graph [2)]. In a constellation graph, variables are transformed into vectors, and those vectors are concatenated to form a star at the final point, i.e., a single line of connected vectors is drawn for one object. The line of these vectors is called a path, and the shape of the path makes it possible to analyze the characteristics of the variables. In addition, the total length of the vectors is adjusted to be one so that the final point is placed in the semicircle.

In a constellation graph, the position of the star at the final point is decided using the mean and variance. For example, in the case of grades of five subjects, the star shows the mean and variance of the five subjects for each student. Thus, it is not possible to identify the subjects in which a student performs well using the final position of the point. The path needs to be considered in detail to make such decisions as mentioned above. However, it is not easy to make an intuitive judgment at a glance because it is difficult to understand correspondence between

Corresponding author: Mamoru Fukumori
Department of Information Science and Business Management, Chugoku Junior College, Niwase 83, Kita-ku, Okayama City, 701-0197, Japan
Tel: +81 90 3378 4510
E-mail: mamo_fuku@cjc.ac.jp

vectors and variables.

Therefore, Fukumori, M. and Fujiwara, M. (2020) proposed an expanded constellation graph [3] that is an improved version of the constellation graph.

In the modified constellation graph, p variables are arranged at equal intervals on the circumference of a semicircle, and vectors of length x_{ij} (i=1, 2,..., n; j=1, 2, ..., p) are concatenated in order from right or left as per the following transformation.

$$\phi_j = \frac{j-1}{p-1}\pi;\ j=1, 2, ..., p$$

When drawing the i-th datapoint using ϕ_j, the coordinates (α, β) of the final point are as follows.

$$(\alpha, \beta) = \frac{1}{p}\left(\sum_{j=1}^{p}\frac{x_{ij}}{U_j}\cos\phi_j, \sum_{j=1}^{p}\frac{x_{ij}}{U_j}\sin\phi_j\right),$$

where x_{ij}>0, and U_j is defined as follows.

$$U_j = \max_{1\leq i\leq n} x_{ij}$$

In the modified constellation graph, the greater the variability in the variables, the further the final point is placed to the right or left. In other words, it is easier to capture the characteristics of the data when the final points are scattered in the left–right direction rather than being densely located near the center of the constellation graph. In this way, when analyzing variable characteristics, the modified constellation graph can be interpreted using the final point.

In the modified constellation graph, the direction is common to all variables, and vector length depends on the raw score. Therefore, the position of the final point is affected by the order of the variables placed on the semicircle. Order of the variables can be determined arbitrarily; however, variable order is an important point because it greatly changes the result. One of the methods of deciding the placement of variables is to use the loadings from the principal component analysis. This makes it easy to grasp the characteristics of each object.

In a modified constellation graph, the greater the lateral spread of variables, the more representative are the characteristics of the variables. Therefore, one method for variable placement is to arrange the variables so that they have the maximum dispersal in the horizontal direction. In addition, we propose a variable placement method that maximizes covariance between objects.

Then, the results of the proposed method are compared with the placement of variables using principal component analysis loadings.

Method of variable placement

For the modified constellation graph, let the data be represented as x_{ij} (i=1, 2, ..., n; j=1, 2, ..., p) with p variables, where, the possible number of variable placement patterns is p!.

The final distance between the points is calculated for each of these p! datasets. Then, the dataset with the largest distance in the left or right direction is selected, and a constellation graph using those datasets is drawn. The steps to find the best dataset are as follows:

1) Let M_k be the maximum value in the horizontal-axis direction of the dataset k(k=1, ..., p!), and m_k be the minimum value.

2) When the distance between two points is d_k (=M_k − m_k), let the dataset showing the maximum value D of all distances d be $D(x_{ij})$. Then, the maximum value D is as follows:

$$D = \max_{1\leq k\leq p!} d_k$$

3) A modified constellation graph is drawn using the arrangement of the variables at the determined D.

Another method of determining the order of the variables is to maximize the covariance between all the objects.

If all the final points of the constellation graph for one dataset are determined, we get n coordinates (α_1, β_1), (α_2, β_2), ..., and (α_n, β_n).

This data is then used to calculate the covariance $S_{\alpha\beta}$ using the following formula:

$$S_{\alpha\beta} = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \overline{\alpha})(\beta_i - \overline{\beta}),$$

where $\overline{\alpha}$ and $\overline{\beta}$ are the means of the respective coordinates.

If the covariance calculated using the i-th dataset is expressed as c_i, let the covariance calculated from all the datasets be expressed as

c_1, c_2, ..., and $c_{p!}$.

From these, consider the dataset that satisfies the following formula to be the optimum data set and create a constellation graph.

$$C = \max_{1 \leq i \leq p!} c_i$$

Application of a modified constellation graph

This study used the "The 4th Basic Study Survey (High School Student Edition)" conducted by the Benesse Institute for Educational Research from June to July 2006. The subjects of the survey were 4,464 second-year high school students (2,168 boys, 2,269 girls, and 27 gender unknown) attending ordinary high schools in Tokyo, local cities, and counties. In this study, responses from 809 people in Tokyo that included no missing data were used. The item asked respondents for their "favorite subject" out of the six listed subjects (Japanese, Mathematics, Geography and History, Science, Civics, and English). The responses were on a five-point scale such as "1) I like it very much," "2) I like it well," "3) I can't say either," "4) I hate it well," and "5) I hate it very much."

Placement of variables based on principal component loadings

Placement of variables in a modified constellation graph is determined based on principal component loadings.

The principal component loading $r(f_j, x_k)$ is the correlation coefficient between the principal component f_j and original variable x_k and is expressed as follows.

$$r(f_j, x_k) = \frac{V(f_j, x_k)}{\sqrt{V(f_j, f_j)V(x_k, x_k)}} = \frac{\sum_{l=1}^{p} w_{jl} V_{lk}}{\sqrt{\lambda_j V_{kk}}} = \frac{\lambda_j w_{jk}}{\sqrt{\lambda_j V_{kk}}} = \sqrt{\frac{\lambda_j}{V_{kk}}}\, w_{jk},$$

where f_j is the j-th principal component, x_k is the k-th variable, λ_j is the l-th eigenvalue, w_{jk} is the k-th element of the eigenvector corresponding to λ_j, and V is variance.

If the data is a variance matrix, the principal component loading is the value of the square root of the eigenvalues times the eigenvectors divided by the square root of the variance of the explanatory variables.

First, the six subjects were subjected to a principal component analysis. Two principal components were extracted based on the Gatman Kaiser's criterion (eigenvalue of 1.0 or more) as shown in Table 1. The contribution rates of the Factors 1 and 2 were 34.08% and 26.10%, respectively.

In the case of the first principal component (PCA1), the loadings of the science subjects, such as Science and Mathematics, were negative values, and the loadings of the liberal arts subjects, such as Geography and history, Civics, Japanese, and English, were positive values. Therefore, PCA1 was interpreted as "science subjects vs. liberal arts subjects." In addition, Mathematics and Science were placed at opposite ends.

In the case of the second principal component (PCA2), all variables had positive values. Therefore, PCA2 was interpreted as a comprehensive index. Furthermore, Science and Japanese were placed at opposite ends.

Fig. 1 shows a modified constellation graph, in which the variables are placed in the descending order of the first principal component loadings.

Table 1. Principal component loadings

Variables	First Loadings	Variables	Second Loadings
Geography and history	0.7362	Science	0.7950
Civics	0.6812	Mathematics	0.7372
Japanese	0.6538	Civics	0.4660
English	0.3686	Geography and history	0.3058
Science	-0.4231	English	0.2003
Mathematics	-0.5283	Japanese	0.1674

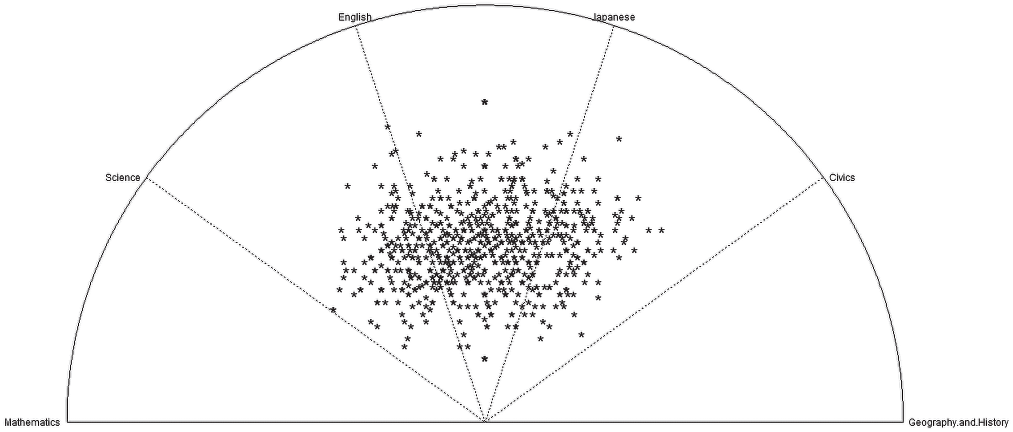

Fig. 1. Modified constellation graph using the first principal component

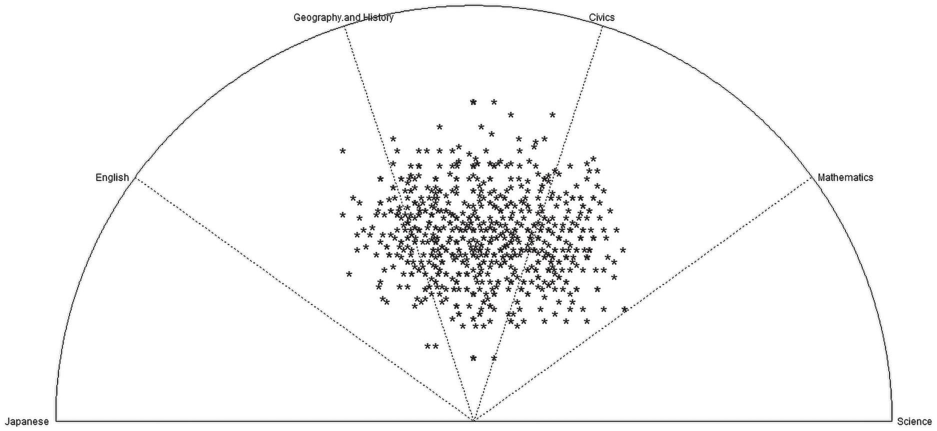

Fig. 2. Modified constellation graph using the second principal component

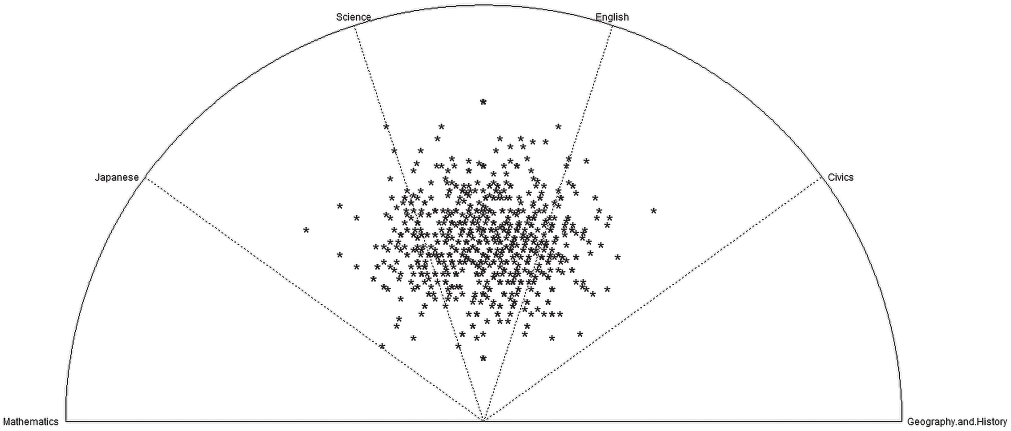

Fig. 3. Modified constellation graph with maximum covariance

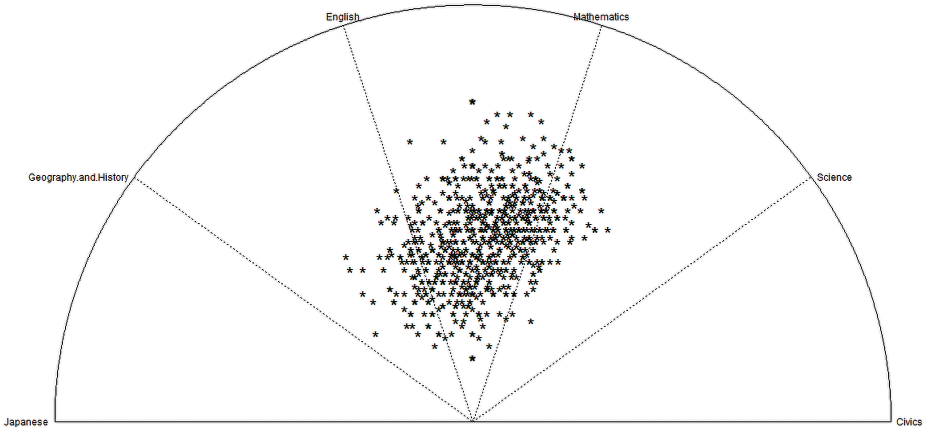

Fig. 4. Modified constellation graph with maximum horizontal distance

Fig. 2 shows a modified constellation graph, in which the variables are placed in the descending order of the second principal component loadings.

Both Figs. 1 and 2 are divided into science subjects and liberal arts subjects.

Fig. 3 shows a modified constellation graph based on the placement of the variables when covariance is maximized. In Fig. 3, the variables are arranged in the order of Civics, Science, Mathematics, English, Geography and history, and Japanese. Furthermore, Civics and Japanese are at opposite ends.

Fig. 4 shows a modified constellation graph based on the placement of variables when horizontal distance is maximized. In Fig. 4, the variables are arranged in the order of Geography and history, Civics, English, Science, Japanese, and Mathematics. In addition, "geography and history" and "mathematics" are at opposite ends.

Thus, the modified constellation graph with the maximum covariance or maximum horizontal distance gives an interpretation from a different perspective than the modified constellation graph based on the principal component loadings.

Conclusion

In this study, we proposed methods to order variables in a modified constellation graph based on maximizing covariance and horizontal distance. These methods were compared with the arrangement based on the principal component loadings.

As a result, it was shown that the proposed method produces a different arrangement from the principal component loadings and enables an analysis of the characteristics of the object from a new viewpoint.

The challenge for the future is to determine the best among the abovementioned methods. It is known that a modified constellation graph is affected by the arrangement of variables. Therefore, it remains an issue to determine the optimum method for obtaining the optimum results.

Acknowledgments

The data for this secondary analysis, "The 4th Basic Study Survey (High School Student Edition), Benesse Institute for Educational Research" was provided by the Social Science Japan Data Archive, Center for Social Research and Data Archives, Institute of Social Science, The University of Tokyo.

References

1) Wakimoto, K., Goto, M. and Matsubara, Y.: Multivariate graph analysis method. Asakura Shoten, 1979.

2) Wakimoto, K. and Taguri, M.: Constellation graphical method for representing multi dimensional data. *Annals of the Institute of Statistical Mathematics*, Vol. 30, Pt. A, pp. 77-84, 1978.

3) Fukumori, M. and Fujiwara, M.: A modification of constellation graph and its application to cluster analysis, *Journal of Chugokugakuen*, Vol.19, pp. 45-51, 2020.

Accepted March 31, 2021.