論文

経営・情報研究（多摩大学研究紀要） **15**, 65-75, 2011

A Hybrid Model for Preference Data †

今泉　忠 [*1]

多摩大学経営情報学部 [*1]

Preference scores to n objects of N individuals is a popular data collected in Marketing, Behavior Science, etc. A vector model or an unfolding distance model have been used to analyze these type of data matrix. However,it is difficult to understand what attributes contribute on preference evaluation using these continuous mapping models as the decomposition of data is not unique. The overlapping cluster models and methods such as ADCLUS (Shepard and Arabie, 1979) have interesting features to find the attributes in similarity data. So we propose a modified model of overlapping model, a hybrid model, to discover the hidden attributes of objects by putting a decomposition constraints. And we also show an application to real data set.

キーワード: ADCLUS, 選好度，ベクトルモデル，クラスター

Key Words: ADCLUS,vector model,clusters

1. Introduction

A number of different models have been proposed to account for individual differences in preference data. Let s_{ij} be a preference score to object $j, j = 1, 2, \cdots, n$ of individual $i, i = 1, 2, \cdots, N$. And we assume a set of t dimensions of factors to be common to all objects and all individuals. Two models are well-known to analyze the preference data, one is the vector model and the other is the unfolding model.

$$s_{ij} = \hat{s}_{ij} + e_{ij}, i = 1, 2, \cdots, N; j = 1, 2, \cdots, n, \tag{1}$$

where $\hat{s}_{ij}$ is a obtained preference score by the model, and e_{ij} is error,respectively. We obtain a t-dimensional score vector $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \cdots, y_{iT}]$ of individual i, and a T-dimensional score

2011 年 2 月 17 日受理

† Tadashi Imaizumi[*1] : A Hybrid Model for Preference Data

School of Management & Information Sciences[*1] , Tama University, 4-1-1 Hijirigaoka, Tama-shi, Tokyo, 206-0022 Japan

vector $\boldsymbol{x}_i[x_{i1}, x_{i2}, \cdots, x_{iT}]$ such as

$$\hat{s}_{ij} = \sum_{t=1}^{T} y_{it} x_{jt}, \tag{2}$$

in the vector model, where e_{ij} denote the error term. And

$$\hat{s}_{ij} = -\sqrt{\sum_{t=1}^{T} (y_{it} - x_{jt})^2}, \tag{3}$$

in the unfolding model. These N vectors of T-dimensionality $\boldsymbol{y}_i, i = 1, 2, \cdots, N$ represent the individual differences in common space. As these models give us the geometric representation of N individuals and n objects, we can understand an overall relation among N individuals and n objects. However it is difficult to understand what attributes contributes on preference evaluation when we assume these continuous mapping models. On the other hand, the overlapping cluster models and methods such as ADCLUS (Shepard and Arabie, 1979;Arabie and Carroll, 1980) have interesting features to find attributes in similarity data. Chaturvedi and Carroll(1994) discussed on a generalzied INDCLUS model. Their model is also applicable to analyzing two-mode preference scores as

$$s_{ij} = \sum_{t=1}^{T} w_p p_{it} q_{jt}, i = 1, 2, \cdots, N; j = 1, 2, \cdots, n \tag{4}$$

where p_{it} and q_{jt} take one of $\{0, 1\}$,

$$p_{it} = \begin{cases} 1 & \text{if individual } i \text{ employ the attribute } t \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

$$q_{jt} = \begin{cases} 1 & \text{if object } j \text{ has the attribute } t \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

This ADCLUS type models suggest us what attributes contribute on the similarity evaluation process. ten Berg and Kier(2005) proposed an algorithm to obtain the SINDCLUS model parameters. Baier, Gaul and Schader(1996) discussed on two-mode overlapping clustering. And Krolar-Schwerdt and Wiedenbeck and (2006) investigated the properties of two-mode ADCLUS type model. These models represent objects and individuals on qualitative dimensions. However it has some contradiction when similarity values or preference scores are quantitative. And Chaturvedi and Carroll(2006) proposed a hybrid ADCLUS model,CLUSCAL, in which they assume a quantitative dimensions and qualitative dimensions simultaneously.

$$s_{ijk} \approx \sum_{r=1}^{R} w_{kr} x_{ir} x_{jr} + \sum_{t=1}^{T} u_{kt} p_{it} q_{jt} + c_k, \tag{7}$$

whete k denote the k-th source or individuals. This model is an interesting one as the attributes will be determinated uniquely, but, we need to check on the unique decomposition of data matrix. By the way,it may natural for us to assume object having a quantitive property rather than a qualitative since observed preference score are measured on a quantitive scale in general. This suggests us each object would be represented as having quantitive scores rather than having qualitative attributes. On the hand, Each individual will rate his or her preference to objects independently. And individuals may be classified into one of several groups for simplicity. So, we propose other hybrid model of ADCLUS model. We assume

- Each object is measured on quantitive dimensions.
- Preference scores are represented as a weighed sum of objects scores on these dimensions which are common to all individuals
- Weight of each individual on these dimension is +1,0, or −1

2. A Hybrid Model

We want to analyze the preference score to n objects of N individuals. Let s_{ij} be an observed preference score to object $j, j = 1, 2, \cdots, n$ of individual $i, i = 1, 2, \cdots, N$. We propose a hybrid vector model in which objects are represented as points in t dimensional space,$\boldsymbol{x}_j, j = 1, 2, \cdots, n$. Let $y_{it} (i = 1, 2, \cdots, N)$ be one of $\{+1, 0, -1\}$. Then We assume s_{ij} are represented by

$$s_{ij} = \sum_{t=1}^{t} p_{it} x_{jt} + e_{ij}, i = 1, 2, \cdots, N; j = 1, 2, \cdots, n, \tag{8}$$

where e_{ij} is error term. This is a modified ADCLUS model includes the original ADCLUS model as the special case with

$$w_t q_{jt} = x_{jt}, t = 1, 2, \cdots, T. \tag{9}$$

This modeling assumes that N individuals share T common dimensions, and that $s_{ij}, j = 1, 2, \cdots, n$ for some individual i is embedded into subspace of T dimensional space. Some individuals dislike some attribute property of object which some individuals like. So we assume that dimensional weight of individual i to the dimension t,p_{it} takes negative value,

$$p_{it} = \begin{cases} 1 & \text{if individual } i \text{ weights the attribute } t \text{ positively} \\ 0 & \text{does not concern this attribute} \\ -1 & \text{if individual } i \text{ weights the attribute } t \text{ negatively} \end{cases}, \tag{10}$$

When $p_{i3} = 0, thepreferencescoresoftheindividualiareembeddedinsubspaceof\mathrm{R}^t$. For example, in the case of $p_{i1} = 1, p_{i2} = -1, p_{i3} = 0$, then

$$s_{ij} = x_{j1} - x_{j2} + e_{ij}, \tag{11}$$

And N individuals are grouped into one of 3^t groups by assumption on $P = [p_{it}]$.

2.1 Metric Scaling

We must obtain $\boldsymbol{X} = [x_{jt}], \boldsymbol{P} = [p_{it}]$, and t from observed preference scores, $S = [s_{ij}]$ Let $\tilde{s}_{ij}$ denote the computed preference score to the object j of the individual i.

$$\hat{s}_{ij} = \sum_{t=1}^{T} t_{it}x_{jt}, i = 1, 2, \cdots, N; j = 1, 2, \cdots, n. \tag{12}$$

As the degree of fitness of our model to the data, we adopt a least square criterion for given dimensionality T

$$LSQ(\boldsymbol{X}, \boldsymbol{P}|T) = \sum_{i=1}^{N}\sum_{j=1}^{n}(s_{ij} - \hat{s}_{ij})^2 / \sum_{i=1}^{N}\sum_{j=1}^{n} s_{ij}^2, \tag{13}$$

As the dimensional weights $\boldsymbol{P}$ take only one of three values $+1, 0, -1$, we must use two step minimization procedure for $LSQ(\boldsymbol{X}, \boldsymbol{P}|T)$

2.2 Obtaining X matrix for given $\boldsymbol{P}$ and T

the conditional LSE of $\boldsymbol{X}$ will be given by

$$\boldsymbol{X} = \boldsymbol{S}'\boldsymbol{P}(\boldsymbol{P}'\boldsymbol{P})^{-1} \tag{14}$$

2.3 Obtaning $\boldsymbol{P}$ by a Heuristic Optimization

We must update the individual weights $\boldsymbol{p}_i, i = 1, 2, \cdots, N$. One convenient method is a discretization method which discretize a continuous $\boldsymbol{P}$ as being adopted by Shepard and Arabie(1979). We will adopt another method, a heuristics method instead of a discretization method.

(1) compute three vectors for the dimension t^* in which

$$\tilde{p}_{iq}^{0} = \begin{cases} p_{iq} & \text{if } q \neq t^* \\ 0 & \text{if } q = t^* \end{cases}, \tag{15}$$

of individual i.

$$\tilde{p}^{+}_{iq} = \begin{cases} p_{iq} & \text{if } q \neq t^{*} \\ 1 & \text{if } q = t^{*} \end{cases}, \tag{16}$$

$$\tilde{p}^{-}_{iq} = \begin{cases} p_{iq} & \text{if } q \neq t^{*} \\ -1 & \text{if } q = t^{*} \end{cases}, \tag{17}$$

(2) And compute three sum of squares for each $\{\tilde{p}^{0}_{iq}, \tilde{p}^{+}_{iq}, \tilde{p}^{-}_{iq}\}$,

$$ssq^{0}_{it} = \sum_{j=1}^{n}(s_{ij} - \sum_{t=1}^{T}\tilde{p}^{0}_{it}x_{jt})^2 \tag{18}$$

$$ssq^{+}_{it} = \sum_{j=1}^{n}(s_{ij} - \sum_{t=1}^{T}\tilde{p}^{+}_{it}x_{jt})^2 \tag{19}$$

$$ssq^{-}_{it} = \sum_{j=1}^{n}(s_{ij} - \sum_{t=1}^{T}\tilde{p}^{-}_{it}x_{jt})^2 \tag{20}$$

$$\tag{21}$$

(3) compare the above three sum of squares, and adopt one of $\tilde{p}^{0}_{it}, \tilde{p}^{+}_{it}$ and $\tilde{p}^{-}_{it}$ which minimizes sum of squares as new p_{it}

3. Computational Procedure

For a given data matrix $\boldsymbol{S}$, we obtain $\boldsymbol{X}$ and $\boldsymbol{P}$ which minimize $LSQ(\boldsymbol{X}, \boldsymbol{P}|T)$ iteratively.

3.1 Initial Matrix of $\boldsymbol{X}$ and $\boldsymbol{P}$

An initial matrix of $\boldsymbol{P}$ and $\boldsymbol{X}$, $\boldsymbol{P}^{(0)}$,$\boldsymbol{X}^{(0)}$ are obtained by SVD of Preference score matrix $\boldsymbol{S}$ where (0) indicates iteration number. $\boldsymbol{S}$ will be decomposed by using SVD,

$$\boldsymbol{S} = \boldsymbol{U}(\boldsymbol{\Lambda}\boldsymbol{V})', \tag{22}$$

and the initial matrix of $\boldsymbol{X}^{(0)}$ by

$$\boldsymbol{X}^{(0)} = (\boldsymbol{\Lambda}\boldsymbol{V}). \tag{23}$$

$$\tag{24}$$

And $p_{it}^{(0)}$ will be obtained by discretizing $\boldsymbol{U}$,

$$p_{it}^{0} = 1 \text{ if } u_{ip} > 0.334 \tag{25}$$

$$p_{it}^{(0)} = 0 \text{ if } |u_{ip}| \leq 0.334 \tag{26}$$

$$p_{it}^{(0)} = -1 \text{ if } u_{ip} < -0.334 \tag{27}$$

$$\tag{28}$$

3.2 Updating matrix $\boldsymbol{X}$

Matrix $\boldsymbol{X}$ will be updated by

$$\boldsymbol{X}^{(l+1)} = \boldsymbol{S}'\boldsymbol{P}^{(l)}(\boldsymbol{P}^{(l)\prime}\boldsymbol{P}^{(l)})^{-1} \tag{29}$$

3.3 Updating Individual Scores

Individual scores P will be updated the procedure in **2.3**. if all values of individual i were 0, i.e. $p_{it} = 0, t = 1, 2, \cdots, T$, then $ssq_{it}^{0} = \sum_{j=1}^{T} s_{ij}^{2}$. And this supports one value of $p_{pt}, t = 1, 2, \cdots, T$ is not 0.

3.4 Normalization of Preference Score

We assume to fit the proposed data to the collected data directly. And we do not include the constant term c in our model. the supplemental dimension whose p_{is} is 1 or -1 for all individuals may be obtained. To avoid such situation, some pre-processing on data may be useful.

- First one is that the mean of individual scores is set to 0,

$$\bar{s}_i = \sum_{j=1}^{n} s_{ij} = 0. \tag{30}$$

- Second one is that sum of squares of individual scores is set to n

$$s_i^2 = \sum_{j=1}^{n} s_{ij}^2/n = 1. \tag{31}$$

4. Application

4.1 Application to Green & Rao Food Items Data

Green and Rao collected the preference ranking to 15 food items of 42 individuals. We normalize this data to the mean of individual scores to 0. We show the joint plotting of Object configuration and Individual configuration in Figure 1. We calculated a $LSQ(\boldsymbol{X}, \boldsymbol{U}|2)$

for the results of MDPREF, and it's value was 0.125 for $T = 2$. And we also applied the proposed model to same data by setting $T = 2$ for the comparison. the obtained the value of $LSQ(\boldsymbol{X}, \boldsymbol{Y}|2) = 0.518$. this value was 4 times of that of MDPREF. We show the joint configuration in Figure 2. We added small jitter to the individual weights as we could understand the number of individuals at each corner and the origin. We done another analysis to the original data, and computed correlation coefficient between the each mean of 15 food items and the obtained scores of objects, and it was 0.993. This suggests our pre-processing on that data was suitable.

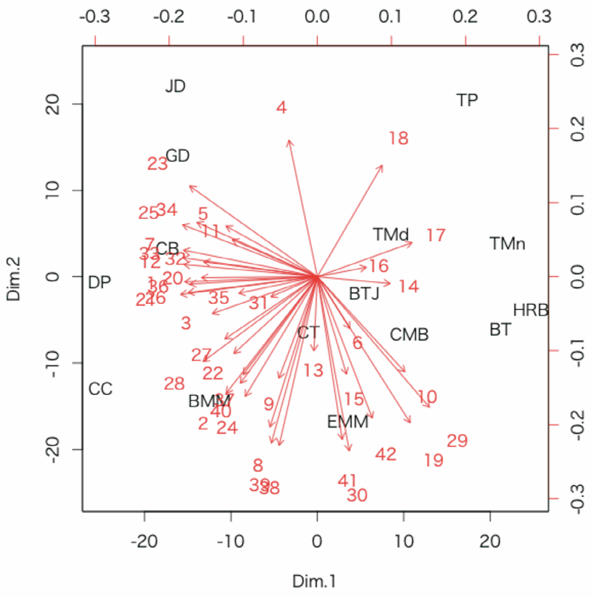

図 1 Joint Configuration of Green and Rao's Preference Data to Food Items

Object Configuration looks to same to that by MDPREF. And We can classify individuals from Figure 2 very easily.

4.2 Application to the Number of Children Data by Delbeke

Delbeke(1968) constructed a set of stimuli by systematically varying the number of boys and the number of girls in a family. By factorially combining four levels (0 to 3) each of the two variables, 20 combinations were constructed. 1D80 students responded his or her preference to each number of children. We applied MDPREF and the proposed model to the data matrix of deviates from mean of each individuals. The $LSQ(\boldsymbol{X}, \boldsymbol{V}|2)$ by MDPREF

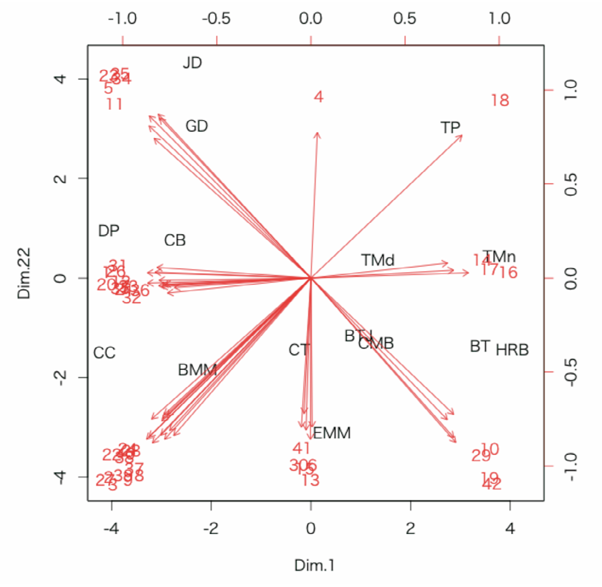

図 2　Joint Configuration of Green and Rao’s Preference Data to Food Items

表 1　Combination of Boys and Girls

Number of Boys	No Daughter	One Daughter	Two Daughter	Three Daughter	Four Daughter	Five Daughter
No Boy		1D	2D	3D	4D	5D
One Son	1S	1S1D	1S2D	1S3D	1S4D	
Two Sons	2S	2S1D	2S2D	2S3D		
Three Sons	3S	3S1D	3S2D			
Four Sons	4S	4S1D				
Five Sons	5S					

was 0.035 and $LSQ(\boldsymbol{X}, \boldsymbol{U}|2)$ by the proposed model was 0.164. The joint configuration by MDPREF was shown in Figure 3. And Figure 4 shows the configuration obtained by the proposed model.

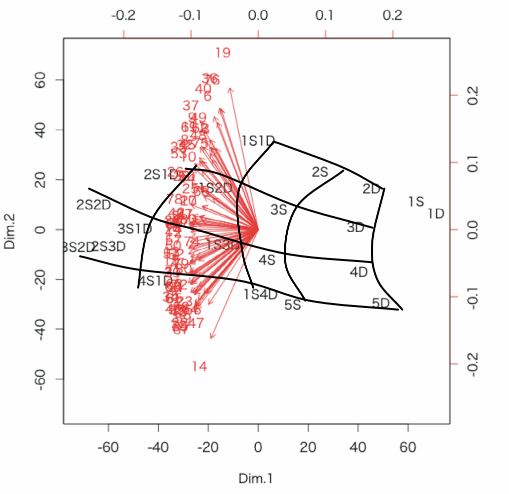

図 3　Joint Configuration by MDPREF

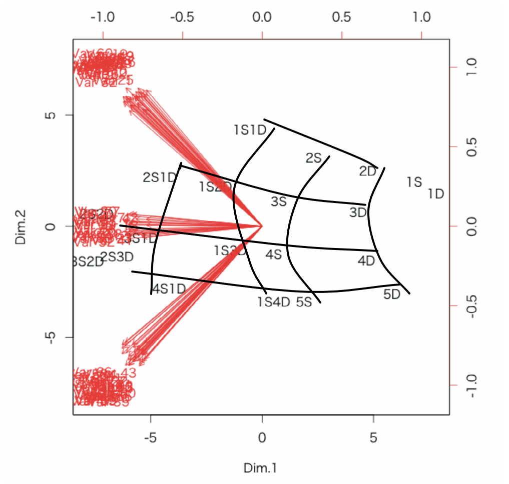

図 4　Joint Configuration by MDPREF

The lines added are connected between two points whose sex are same or the number of children are same. Both figures look very similar though the factorial lattice structure will be more recovered in Figure 4 and we can guess the student preference to the number of children easily.

5. Conclusion

We proposed a hybrid model and method for analyzing a preference data matrix. This model assumes that N individuals share the common space in preference scoring and some individuals ignore some dimensions in his or her preference scoring. However, we can assume the different decomposition of data matrix. We assume the qualitative attributes of objects, and individuals differently weight to these attributes. Then the model

$$s_{ij} = \sum_{t=1}^{T} y_{it} q_{jt} + e_{ij} \tag{32}$$

will be more reasonable. The external analysis approach as in PREFMAP(Carroll,1972) will be suitable when we assume this modeling.

参考文献

[1] Arabile, P., and Carroll, J.D. (1980):MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model, Psychometrika, 45, 211—235.

[2] BAaier, D., Gaul, W. and Schader, M. (1996):Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring, in *Classification and Knowledge Organisation*, Proceedings of the 20th Annual Conference of the Gesellschaft fuer Klassikation, Eds., F. Klar and O. Opitz, Berlin et. al. Springer, Heidelberg,Germany.

[3] Carroll, J.D. (1972): Individual Differences and Multidimensional Scaling, in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, (Vol. 1, pp. 105—155), Eds. R.N. Shepard, A.K. Romney and S.B. Nerlove, New York: Seminar Press. [Reprinted (1984) in Key Texts on Multidimensional Scaling (pp. 267— 301), Eds. P. Davies and A.P.M. Coxon, Portsmouth, NH: Heinemann.]

[4] Chaturvedi, A.D., and Carroll, J.D. (1994):An Alternating Combinatorial Approach to Fitting the INDCLUS and Generalized INDCLUS Models, *Journal of Classification*, 11, 155—170.

[5] Chaturvedi, A.D., and Carroll, J.D.(2006):CLUSCALE (”CLUstering and multidimensional SCAL[E]ing”): A Three-Way Hybrid Model Incorporating Overlapping Clustering and Multidimensional Scaling Structure,*Journal of Classification*,23 (2) pp. 269-299.

[6] Krolar-Schwerdt, K and Wiedenbeck, M.(2006). :The Recovery Performance

of Two-mode Clustering Methods: Monte Carlo Experiment, *From Data and Information Analysis to Knowledge Engineering*, pp. 190-197,Springer, Heidelberg,Germany.

[7] Shepard, R.N., and Aravie, P. (1979) : Additive Clustering: Representation of Similarities as Combinations of Discrete Overlapping Properties, *Psychological Review*, 86, 87—123.

[8] ten Berge, J.M.F., and Kiers, H.A.L. (2005), A Comparison of Two Methods for Fitting the INDCLUS Model, *Journal of Classification*, 22, 273-286.