

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,800

Open access books available

142,000

International authors and editors

180M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Variable Selection in Nonlinear Principal Component Analysis

Hiroko Katayama, Yuichi Mori and Masahiro Kuroda

Abstract

Principal components analysis (PCA) is a popular dimension reduction method and is applied to analyze quantitative data. For PCA to qualitative data, nonlinear PCA can be applied, where the data are quantified by using optimal scaling that nonlinearly transforms qualitative data into quantitative data. Then nonlinear PCA reveals nonlinear relationships among variables with different measurement levels. Using this quantification, we can consider variable selection in the context of PCA for qualitative data. In PCA for quantitative data, modified PCA (M.PCA) of Tanaka and Mori derives principal components which are computed as a linear combination of a subset of variables but can reproduce all the variables very well. This means that M.PCA can select a reasonable subset of variables with different measurement levels if it is extended so as to deal with qualitative data by using the idea of nonlinear PCA. A nonlinear M.PCA is therefore proposed for variable selection in nonlinear PCA. The method, in this chapter, is based on the idea in “Nonlinear Principal Component Analysis and its Applications” by Mori et al. (Springer). The performance of the method is evaluated in a numerical example.

Keywords: quantification, categorical data, modified PCA, stepwise selection, cumulative proportion, RV-coefficient

1. Introduction

Principal components analysis (PCA) is a popular dimension reduction method and is applied to analyze quantitative data. For PCA to qualitative data, the data are quantified by using optimal scaling that nonlinearly transforms qualitative data into quantitative data. The PCA with optimal scaling is called nonlinear PCA. Nonlinear PCA reveals all qualitative variables uniformly as numerical variables by using optimal scaling quantifiers in the analysis, that is, it can deal with nonlinear relationships among variables with different measurement levels.

Using this quantification, we can consider variable selection in the context of PCA for qualitative data. In PCA for quantitative data, Tanaka and Mori discussed a method called modified PCA (M.PCA) that can be used to compute principal components (PCs) using only a selected subset of variables that represents all of the variables, including those not selected [1]. Since M.PCA includes variable selection procedures in the analysis, if we quantify all the qualitative variables by using the

optimal scaling and then apply M.PCA to the quantified data, we can select a reasonable subset of variables from the qualitative data.

In this chapter, we refer to Mori et al. [2] to revisit a variable selection problem in PCA for qualitative data. The proposed method here (we call it nonlinear M.PCA or NL.M.PCA) is an extension of M.PCA so as to deal with a mixture of quantitative and qualitative data. In Section 2 we provide the overview of NL.M.PCA (optimization, the original M.PCA and NL.M.PCA for qualitative data) based on studies by Mori et al. [2], and in Section 3, we apply this method to the customer engagement data [3] to show how it works in the real data and how you use the output from the method for variable selection, and to evaluate the performance of the method.

2. Modified PCA for mixed measurement level data

2.1 Quantification of qualitative data

We must use a suitable quantification method in the context of PCA because we here wish to consider a variable selection problem in PCA. One of the best methods is the optimal scaling in nonlinear PCA. Nonlinear PCA is a method to deal with qualitative data, which estimates the parameters of PCA and quantifies qualitative variables simultaneously by alternating between estimation and quantification. PRINCIPALS of Young et al. [4] and PRINCIPALS of Gifi [5] are algorithms for nonlinear PCA. Here we use PRINCIPALS.

PRINCIPALS is an algorithm using the alternating least squares (ALS) algorithm as follows: Let $\mathbf{Y} = (\mathbf{y}_1 \mathbf{y}_2 \dots \mathbf{y}_p)$ be a data matrix of n objects by p categorical variables and let \mathbf{y}_j of \mathbf{Y} be a qualitative vector with K_j categories labeled $1, \dots, K_j$. PRINCIPALS minimizes the loss function

$$\sigma_L(\mathbf{Z}, \mathbf{A}, \mathbf{Y}^*) = \text{tr}(\mathbf{Y}^* - \hat{\mathbf{Y}})^\top (\mathbf{Y}^* - \hat{\mathbf{Y}}) = \text{tr}(\mathbf{Y}^* - \mathbf{Z}\mathbf{A}^\top)^\top (\mathbf{Y}^* - \mathbf{Z}\mathbf{A}^\top), \quad (1)$$

where \mathbf{Y}^* is an optimally scaled matrix form \mathbf{Y} , \mathbf{Z} is an $n \times r$ matrix of n component scores on r ($1 \leq r \leq p$) components, and $\mathbf{A} = \{\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r\}$ is a $p \times r$ weight matrix that gives the coefficients of the linear combinations. PRINCIPALS alternately makes two estimations: the model parameters \mathbf{Z} and \mathbf{A} for ordinary PCA, and the data parameter for optimally scaled data \mathbf{Y}^* .

In the computation of PRINCIPALS, \mathbf{Y}^* are standardized for each variable such as to satisfy restrictions $\mathbf{Y}^{*\top} \mathbf{1}_n = \mathbf{0}_p$ and $\text{diag} \left[\frac{\mathbf{Y}^{*\top} \mathbf{Y}^*}{n} \right] = \mathbf{I}_p$. We denote the value θ estimated the t -th iteration by $\theta^{(t)}$. Given the initial data $\mathbf{Y}^{*(0)}$ (the observed data \mathbf{Y} may be used as $\mathbf{Y}^{*(0)}$ after the above standardization), PRINCIPALS iterates the following two steps:

- *Model estimation step*: By solving the eigenvalue problem (EVP) of the covariance matrix of $\mathbf{Y}^{*(t)}$ ($= \mathbf{S}$)

$$[\mathbf{S} - \lambda \mathbf{I}] \mathbf{a} = 0, \quad (2)$$

where λ is the eigenvalues, obtain $\mathbf{A}^{(t+1)}$ and compute $\mathbf{Z}^{(t+1)} = \mathbf{Y}^{*(t)} \mathbf{A}^{(t+1)}$. Update $\hat{\mathbf{Y}}^{(t+1)} = \mathbf{Z}^{(t+1)} \mathbf{A}^{(t+1)\top}$.

- *Optimal scaling step*: Obtain $\mathbf{Y}^{*(t+1)}$ such that

$$\mathbf{Y}^{*(t+1)} = \arg \min_{\mathbf{Y}^{*(t)}} \text{tr} \left(\mathbf{Y}^{*(t)} - \hat{\mathbf{Y}}^{(t+1)} \right)^\top \left(\mathbf{Y}^{*(t)} - \hat{\mathbf{Y}}^{(t+1)} \right) \quad (3)$$

for fixed $\hat{\mathbf{Y}}^{(t+1)}$ by separately estimating \mathbf{y}_j^* for each variable j under the measurement restrictions on each of the variables. That is, compute $\mathbf{q}_j^{(t+1)}$ for nominal variables as

$$\mathbf{q}_j^{(t+1)} = \left(\mathbf{G}_j^\top \mathbf{G}_j \right)^{-1} \mathbf{G}_j^\top \hat{\mathbf{y}}_j^{(t+1)}, \quad (4)$$

where \mathbf{q}_j is a $K_j \times 1$ category score vector for \mathbf{y}_j^* and \mathbf{G}_j is an $n \times K_j$ indicator matrix

$$\mathbf{G}_j = \left(g_{jik} \right) = \begin{pmatrix} g_{j11} & \cdots & g_{j1K_j} \\ \vdots & \vdots & \vdots \\ g_{jn1} & \cdots & g_{jnK_j} \end{pmatrix} = \left(\mathbf{g}_{j1} \cdots \mathbf{g}_{jK_j} \right), \quad (5)$$

where

$$g_{jik} = \begin{cases} 1 & \text{if object } i \text{ belongs to category } k \\ 0 & \text{if object } i \text{ belongs to some other category } k' (\neq k), \end{cases} \quad (6)$$

and then the optimally scaled vector \mathbf{y}_j^* is obtained by $\mathbf{y}_j^* = \mathbf{G}_j \mathbf{q}_j$.

Re-compute $\mathbf{q}_j^{(t+1)}$ for ordinal variables using the monotone regression [6]. For nominal and ordinal variables, update $\mathbf{y}_j^{*(t+1)} = \mathbf{G}_j \mathbf{q}_j^{(t+1)}$ and standardize $\mathbf{y}_j^{*(t+1)}$. For numerical variables, standardize the observed vector \mathbf{y}_j and set $\mathbf{y}_j^{*(t+1)} = \mathbf{y}_j$.

These two steps alternately iterate until convergence, and \mathbf{y}_j^* obtained at convergence is the quantified variable while \mathbf{A} and \mathbf{Z} are the solutions of PCA for qualitative data.

2.2 Modified PCA

M.PCA of Tanaka and Mori [1] derives PCs that are computed using only a selected subset but represent all of the variables, including those not selected. This means that M.PCA naturally includes variable selection procedures in its estimation process. Although there are several variable selection methods in PCA, we use M.PCA, because a subset of variables selected by M.PCA can represent all the variables very well and it is easy to incorporate the quantification method in Section 2.1 into M.PCA, which will be described in Section 2.3.

Suppose we obtain an $n \times p$ data matrix \mathbf{Y} that consists of numerical variables or optimally quantified variables. Let \mathbf{Y} be decomposed into an $n \times q$ submatrix \mathbf{Y}_1 and an $n \times (p - q)$ submatrix \mathbf{Y}_2 ($1 \leq q \leq p$). \mathbf{Y} is represented by r PCs, which is a linear combination of a submatrix \mathbf{Y}_1 , that is, $\mathbf{Z} = \mathbf{Y}_1 \mathbf{A}$, where r is the number of PCs ($1 \leq r \leq q$). To derive $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_r)$, the following Criterion 1 based on Rao [7] and Criterion 2 based on Robert and Escoufier [8] can be used:

(Criterion 1) The prediction efficiency \mathbf{Y} is maximized using a linear predictor in terms of \mathbf{Z} .

(Criterion 2) The closeness of configurations between \mathbf{Y} and \mathbf{Z} is maximized using the RV -coefficient.

We denote the covariance matrix of $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ as $\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}$, where the subscript i of \mathbf{S} corresponds to \mathbf{Y}_i . The maximization criteria for the above Criterion 1 and Criterion 2 are given by the proportion P

$$P = \sum_{j=1}^r \lambda_j / \text{tr}(\mathbf{S}), \quad (7)$$

and the RV -coefficient

$$RV = \left\{ \sum_{j=1}^r \lambda_j^2 / \text{tr}(\mathbf{S}^2) \right\}^{1/2}, \quad (8)$$

respectively, where λ_j is the j -th eigenvalue with the order of magnitude of the EVP

$$[(\mathbf{S}_{11}^2 + \mathbf{S}_{12}\mathbf{S}_{21}) - \lambda\mathbf{S}_{11}]\mathbf{a} = 0. \quad (9)$$

The solution is obtained as a matrix \mathbf{A} , the columns of which consist of the eigenvectors associated with the largest r eigenvalues of EVP (9), and \mathbf{Y}_1 that provides the largest value of P or RV is the best subset of q variables among all possible subsets of size q . Thus, to obtain a reasonable subset of variables with size q in PCA, you apply M.PCA to the data and find the subset of size q , \mathbf{Y}_1 , that has the largest P or RV . The selected subset \mathbf{Y}_1 is reasonable in the sense of PCA because it contains information that includes not only the selected variables \mathbf{Y}_1 but also the deleted ones \mathbf{Y}_2 .

2.3 Modified PCA for mixed measurement level data

M.PCA is a good method to find a reasonable subset of numerical variables as described in the previous section. To select variables from mixed measurement level data by using a criterion in M.PCA, qualitative/categorical variables in the data should be quantified in an appropriate manner. Based on the original idea in ref. [9], considering PRINCIPALS in Section 2.1 and M.PCA in Section 2.2, it is easy to incorporate the quantification (PRINCIPALS) into M.PCA, because we can formulate M.PCA for qualitative data only by replacing the EVP (2) in the *Model estimation step* of PRINCIPALS by the EVP (9) to get the model parameters \mathbf{A} and \mathbf{Z} for M.PCA. Thus, M.PCA and optimal scaling are alternately executed until $\theta^* = \text{tr}(\mathbf{Y}^* - \hat{\mathbf{Y}})^\top (\mathbf{Y}^* - \hat{\mathbf{Y}}) = \text{tr}(\mathbf{Y}^* - \mathbf{Z}\mathbf{A}^\top)^\top (\mathbf{Y}^* - \mathbf{Z}\mathbf{A}^\top)$ is minimized. This is nonlinear M.PCA or NL.M.PCA.

Here, we rewrite the ALS algorithm of PRINCIPALS as follows—for given initial data $\mathbf{Y}^{*(0)} = (\mathbf{Y}_1^{*(0)}, \mathbf{Y}_2^{*(0)})$ from the original data \mathbf{Y} , the following two steps are iterated until convergence:

- *Model estimation step*: From $\mathbf{Y}^{*(t)} = (\mathbf{Y}_1^{*(t)}, \mathbf{Y}_2^{*(t)})$, obtained $\mathbf{A}^{(t)}$ by solving the EVP (9).

Compute $\mathbf{Z}^{(t)}$ from $\mathbf{Z}^{(t)} = \mathbf{Y}_1^{*(t)} \mathbf{A}^{(t)}$. Update $\hat{\mathbf{Y}}^{(t+1)} = \mathbf{Z}^{(t)} \mathbf{A}^{(t)}$.

- *Optimal scaling step*: Obtain $\mathbf{Y}^{*(t+1)}$ for fixed $\hat{\mathbf{Y}}^{(t+1)}$ by separately estimating \mathbf{y}_j^* ($=\mathbf{G}_j \mathbf{q}_j$) for each variable j under the measurement restrictions. Re-compute $\mathbf{Y}_j^{*(t+1)}$ by an additional transformation to keep the monotonicity restriction for ordinal variables and skip this computation for numerical variables.

$\mathbf{Y}^* = (\mathbf{Y}_1^*, \mathbf{Y}_2^*)$ obtained after convergence is an optimally scaled (quantified) matrix of \mathbf{Y} , and \mathbf{Y}_1 corresponding to \mathbf{Y}_1^* is a subset to be selected and \mathbf{Y}_2 to \mathbf{Y}_2^* is one to be deleted.

NL.M.PCA procedure for fixed q is as described above, but since the variable selection performs M.PCA calculation for $q = p, \dots, r$ and ${}_p C_q$ times to find the best \mathbf{Y}_1 , there are three possible *types* of selection according to where the quantification is implemented in the computation flow (see Fig. 4.1 in [2]).

The first type (*Type 1*) is that the quantification is performed only once at first, that is, nonlinear PCA is applied to the data \mathbf{Y} to obtain the quantified data \mathbf{Y}^* , and ordinary M.PCA selection is applied to \mathbf{Y}^* . No more quantification is carried out in the selection stage. The second type (*Type 2*) is that the quantification is carried out every time after the best subset of size q is found in the selection stage. That is, the quantified $(\mathbf{Y}_1^*, \mathbf{Y}_2^*)$ based on the best subset of the size q found in the previous selection is used to find the best subset of size $q - 1$ or $q + 1$ in the next selection. The third type (*Type 3*) is that the quantification is carried out for every temporary $(\mathbf{Y}_1, \mathbf{Y}_2)$ in the section stage, that is, NL.M.PCA is performed whenever temporary $(\mathbf{Y}_1, \mathbf{Y}_2)$ is given to compute its criterion value.

A reasonable subset of size q is given as \mathbf{Y}_1 corresponding to the best subset \mathbf{Y}_1^* which is finally found at q when the selection procedure is terminated.

3. A numerical example

3.1 Data

The data we analyze here was gathered in the survey about the relationship among customer engagement on “fashion,” “brand,” and “shop staff” [3]. The questions (variables) are divided into three groups based on the purposes for consumption: “Involvement” (16 variables), “Expectations” (35 variables), and “Values” (34 variables). The total number of questions is 85 on a five-level scale and 825 responses are obtained. Ohyaabu et al. [3] analyzed this data to find the structure of the customer consciousness, but we use this data simply as sample data for variable selection in PCA without considering the original purpose in ref. [3]. Here we apply NL.M.PCA to the second question group “Expectation” (35 variables) to show the performance of the proposed method. The questions asked in the survey are indicated in the “Question” column of **Table 1** and answers (responses) are shown in **Table 2**.

Group	Item	Question	$q = 25$
About the fashion	Q	We would like to ask you about your thoughts and behaviors about fashion.	
	Q1	I think about fashion by putting on clothes or choosing the clothes.	×
	Q2	I think about fashion by putting on clothes or choosing the clothes.	×
	Q3	I want to know about fashion by putting on clothes or choosing clothes.	×
	Q4	I'm enthusiastic when I think about fashion.	×
	Q5	I'm happy with thinking about fashion.	×
	Q6	I feel relaxed when I think about fashion.	×
	Q7	I'm proud of my fashions when I think about fashion.	×
	Q8	I spend a lot of one's time when I think about the fashions.	
	Q9	I talk about fashion with my friends.	
	Q10	I'm checking about SNS or writing comment for fashion.	×
Q11	I'm posting about a fashion to SNS.	×	
About the brand	Q	We would like to ask you about your thoughts and behaviors about fashion brands.	
	Q12	I think about the brand by putting on clothes or choosing the clothes.	
	Q13	I think about the brand by putting on clothes or choosing the clothes.	×
	Q14	I want to know about the brand by putting on clothes or choosing the clothes.	×
	Q15	I'm enthusiastic when I think about the brand.	×
	Q16	I'm happy when I think about the brand.	
	Q17	I feel relaxed good when I think about the brand.	×
	Q18	I'm proud when I think about the brand.	×
	Q19	I spend a lot of one's time when I think about the brand.	×
	Q20	I always use a specific brand when I wear or choose clothes.	
	Q21	I always use the brand when I clothes or choice of clothes.	×
Q22	I'm checking about SNS or writing comment for the brand.	×	
Q23	I'm posting about a brand to an SNS of mine.	×	
About the shop staff	Q	We would like to ask you about your thoughts and behaviors about the staff member	
	Q24	I think about the staff member by talking to other staff	×
	Q25	I think about staff members when I speak to other shop staff.	
	Q26	I want to know more about shop staff by speaking.	×
	Q27	I'm enthusiastic when I'm talking with staff members.	
	Q28	I'm happy when I'm talking with staff members.	×
	Q29	I feel relaxed when I'm talking with staff members.	×
	Q30	I'm proud when I'm talking with shop staff.	
	Q31	I spend a lot of time talking with shop staff.	×
	Q32	I always talk to the specific staff member when choosing clothes or putting on clothes.	

Group	Item	Question	$q = 25$
	Q33	I always talk the specific staff member.	×
	Q34	I'm checking about the specific staff member of SNS or writing comment for the brand.	
	Q35	I'm posting about the specific staff member to my SNS.	×

Table 1. 35 questions in “expectation” and 25 selected ones (marked by × in the right column).

Respondent	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35			
1	1	1	1	1	1	1	3	1	1	1	1	2	2	2	2	2	2	3	2	2	3	3	4	2	2	3	3	3	2	4	2	3	2	5	5			
2	1	1	1	1	1	1	1	1	3	3	3	1	1	1	1	1	1	1	1	1	2	4	3	1	1	2	1	1	1	1	1	1	1	3	3			
3	2	2	2	2	2	4	3	2	3	4	2	2	2	2	2	2	4	3	3	3	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5			
4	1	1	2	1	1	2	1	1	1	2	3	1	2	2	2	2	1	2	1	1	1	3	4	1	1	1	1	1	1	1	2	2	2	2	3			
5	2	2	3	2	2	3	3	3	3	5	3	3	3	3	3	3	3	3	3	5	3	5	3	5	5	5	5	5	5	5	5	5	5	5	4			
6	4	4	4	4	4	5	4	4	5	5	5	4	5	4	4	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4	5	4	4	4	3			
7	2	3	1	1	1	1	2	1	2	1	1	2	2	2	2	2	1	1	1	1	2	1	2	2	3	3	3	1	3	2	1	1	1	4	1			
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	1	2	1	5	5	5	5	5	5	5	5	5	5	5	5			
9	2	2	2	3	2	2	3	4	3	5	5	2	2	3	3	3	3	3	2	2	2	4	5	4	4	4	5	3	3	4	4	3	3	5	5			
10	1	1	1	1	1	1	2	1	4	3	5	2	3	1	1	1	1	3	1	1	1	3	5	4	4	4	3	2	4	5	5	4	4	4	5			
11	2	2	2	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	5	5		
12	1	1	2	1	1	1	4	3	2	4	4	2	2	1	1	1	3	2	2	1	4	4	2	4	4	2	2	2	4	3	3	2	4	4	4			
13	2	2	2	2	1	2	2	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	3	2	1	2	1	2	2	1	2	2	2	2	4			
14	1	1	1	2	2	2	4	2	1	1	4	2	2	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	5	2	2		
15	2	2	3	2	1	2	4	3	3	2	5	2	2	2	2	2	1	2	2	2	2	3	3	2	1	4	2	2	2	2	1	2	2	5	5	5		
16	1	1	2	2	2	2	3	3	2	2	4	2	2	3	2	2	1	2	2	2	1	4	4	5	5	5	5	2	3	3	3	3	5	2	5	5		
17	1	1	1	1	1	1	2	2	1	2	3	2	2	1	1	1	1	2	3	2	3	4	4	3	4	4	4	4	4	4	4	4	3	4	4	5		
18	2	2	3	2	2	2	4	4	3	5	5	2	2	3	2	2	2	2	3	3	2	5	5	5	5	1	5	5	5	5	5	5	5	5	5	5		
19	2	2	2	2	2	2	4	3	3	3	3	3	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	3	3	4	4	4	3	4	4	4		
20	2	2	2	2	2	2	2	2	2	4	2	2	2	2	2	2	2	2	2	2	2	2	2	3	4	2	2	2	2	2	2	2	2	2	4	5		
21	1	1	1	1	1	1	3	2	1	1	5	1	1	1	1	1	1	1	1	1	1	1	5	1	1	1	1	1	1	1	1	1	1	1	5	5		
22	1	1	1	1	1	1	3	2	3	2	3	1	1	1	1	1	1	1	2	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	4	5	
23	2	1	2	1	1	1	1	3	4	2	5	2	1	3	1	1	1	1	2	3	1	2	5	3	4	5	2	2	2	4	3	3	2	5	5	5		
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
25	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
26	1	1	1	2	2	1	3	3	2	2	3	1	2	2	1	1	1	2	3	2	1	2	4	2	3	3	2	3	3	4	3	2	2	5	5	5		
27	1	1	1	2	1	1	2	2	2	2	3	1	1	1	2	1	1	2	1	2	1	2	3	1	1	1	1	1	1	1	2	3	2	3	2	3		
28	2	1	3	3	3	2	1	4	3	2	2	1	2	2	2	1	2	2	1	2	2	1	3	4	2	2	1	1	1	2	1	1	1	4	2	2		
29	1	1	1	1	1	1	1	1	1	2	2	2	2	1	2	1	1	1	2	1	1	1	2	2	1	2	2	1	1	2	2	2	2	2	2	3	4	
30	2	2	2	2	2	2	2	2	3	5	3	3	3	3	2	3	2	3	2	3	3	2	2	3	3	3	3	3	3	3	3	3	3	3	2	5	5	
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
824	3	1	1	1	2	4	3	1	2	4	4	1	2	2	5	3	3	1	1	2	1	1	3	2	1	1	4	3	2	4	1	1	1	1	1	1		
825	3	3	1	1	1	5	2	1	4	3	1	1	5	1	4	2	2	2	2	2	2	2	2	2	3	2	2	2	2	3	3	3	3	2	2	3	3	

Table 2. Expectation data (825 responses on 35 variables).

3.2 Output from NL.M.PCA

Table 3 shows the output of NL.M.PCA when NL.M.PCA is applied to Expectation data with $r = 5$ of the number of PCs, proportion P as a criterion, and forward-backward stepwise selection and *type 3* quantifications as selection procedures. The number q is the number of selected variables and the value P is the criterion value. $\mathbf{Y}_1|\mathbf{Y}_2$ shows that the left side of each row is the question numbers to be selected (\mathbf{Y}_1) and the right side to be deleted (\mathbf{Y}_2). If you have a specific number q for variables to be used, such as 20, 10, or $2/3 = 24$, $1/2 = 18$, you can use variables whose numbers are displayed in \mathbf{Y}_1 at that q . If the number of variables to be used is not determined, the proportion P can be used. For example, since the proportion P is 66.95% with all 35 variables, if you want to keep P up to 65%, looking at the row of $P = 0.6512$ (i.e., $q = 20$), you can use 20 variables in \mathbf{Y}_1 . Alternatively, if the difference between the proportion with all 35 variables and that with selected variables should be less than 1%, 25 variables can be used because $0.6695 - 0.01 = 0.6595$, which is the P value at

Principal Component Analysis

q	$Y_1 Y_2$																																			P	
35	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	0.6695	
34	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	26	27	28	29	30	31	32	33	34	35	25	0.6689	
33	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	26	27	28	29	30	31	33	34	35	25	32	0.6682	
32	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	26	28	29	30	31	33	34	35	8	25	27	32	0.6675
31	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	26	28	29	30	31	33	34	35	8	16	25	27	32	0.6666
30	1	2	3	4	5	6	7	9	10	11	12	13	14	15	17	18	19	20	21	22	23	24	26	28	29	30	31	33	34	35	8	16	25	27	32	0.6656	
29	1	2	3	4	5	6	7	9	10	11	12	13	14	15	17	18	19	20	21	22	23	24	26	28	29	31	33	34	35	8	16	25	27	30	32	0.6646	
28	1	2	3	4	5	6	7	9	10	11	13	14	15	17	18	19	20	21	22	23	24	26	28	29	31	33	34	35	8	12	16	25	27	30	32	0.6636	
27	1	2	3	4	5	6	7	9	10	11	13	14	15	17	18	19	20	21	22	23	24	26	28	29	31	33	35	8	12	16	25	27	30	32	34	0.6624	
26	1	2	3	4	5	6	7	10	11	13	14	15	17	18	19	20	21	22	23	24	26	28	29	31	33	35	8	9	12	16	25	27	30	32	34	0.6611	
25	1	2	3	4	5	6	7	10	11	13	14	15	17	18	19	21	22	23	24	26	28	29	31	33	35	8	9	12	16	20	25	27	30	32	34	0.6598	
24	1	2	3	4	6	7	10	11	13	14	15	17	18	19	21	22	23	24	26	28	29	31	33	35	5	8	9	12	16	20	25	27	30	32	34	0.6583	
23	1	2	3	4	6	7	10	11	13	14	15	17	18	19	21	22	23	24	26	29	31	33	35	5	8	9	12	16	20	25	27	28	30	32	34	0.6568	
22	1	2	3	4	6	7	10	11	13	14	15	17	18	19	21	22	24	26	29	31	33	35	5	8	9	12	16	20	23	25	27	28	30	32	34	0.6552	
21	2	3	4	6	7	10	11	13	14	15	17	18	19	21	22	24	26	29	31	33	35	1	5	8	9	12	16	20	23	25	27	28	30	32	34	0.6533	
20	2	3	4	6	7	10	11	13	14	17	18	19	21	22	24	26	29	31	33	35	1	5	8	9	12	15	16	20	23	25	27	28	30	32	34	0.6512	
19	2	3	4	6	7	10	11	13	14	17	18	19	21	22	26	29	31	33	35	1	5	8	9	12	15	16	20	23	24	25	27	28	30	32	34	0.6492	
18	2	3	4	6	7	10	11	13	14	17	19	21	22	26	29	31	33	35	1	5	8	9	12	15	16	18	20	23	24	25	27	28	30	32	34	0.6467	
17	2	4	6	7	10	11	13	14	17	19	21	22	26	29	31	33	35	1	3	5	8	9	12	15	16	18	20	23	24	25	27	28	30	32	34	0.6435	
16	1	4	6	7	10	11	13	14	15	19	21	22	25	27	32	35	2	3	5	8	9	12	16	17	18	20	23	24	26	28	29	30	31	33	34	0.6403	
15	1	4	6	7	10	11	14	15	19	21	22	25	27	32	35	2	3	5	8	9	12	13	16	17	18	20	23	24	26	28	29	30	31	33	34	0.6370	
14	1	6	7	10	11	14	15	19	21	22	25	27	32	35	2	3	4	5	8	9	12	13	16	17	18	20	23	24	26	28	29	30	31	33	34	0.6326	
13	1	6	7	10	11	14	15	19	21	25	27	32	35	2	3	4	5	8	9	12	13	16	17	18	20	22	23	24	26	28	29	30	31	33	34	0.6278	
12	1	6	10	11	14	15	19	21	25	27	32	35	2	3	4	5	7	8	9	12	13	16	17	18	20	22	23	24	26	28	29	30	31	33	34	0.6224	
11	1	6	10	11	14	15	19	25	27	32	35	2	3	4	5	7	8	9	12	13	16	17	18	20	21	22	23	24	26	28	29	30	31	33	34	0.6154	
10	1	6	10	11	14	19	25	27	32	35	2	3	4	5	7	8	9	12	13	15	16	17	18	20	21	22	23	24	26	28	29	30	31	33	34	0.6068	
9	1	6	10	11	14	19	25	27	35	2	3	4	5	7	8	9	12	13	15	16	17	18	20	21	22	23	24	26	28	29	30	31	32	33	34	0.5968	
8	1	6	11	14	19	25	27	35	2	3	4	5	7	8	9	10	12	13	15	16	17	18	20	21	22	23	24	26	28	29	30	31	32	33	34	0.5870	
7	1	6	11	14	19	25	32	2	3	4	5	7	8	9	10	12	13	15	16	17	18	20	21	22	23	24	26	27	28	29	30	31	33	34	35	0.5714	
6	1	6	11	14	25	32	2	3	4	5	7	8	9	10	12	13	15	16	17	18	19	20	21	22	23	24	26	27	28	29	30	31	33	34	35	0.5400	
5	6	11	14	21	27	1	2	3	4	5	7	8	9	10	12	13	15	16	17	18	19	20	22	23	24	25	26	28	29	30	31	32	33	34	35	0.5319	

Table 3. Selection results (expectations, $r = 5$, proportion P , forward-backward stepwise selection, Type 3).

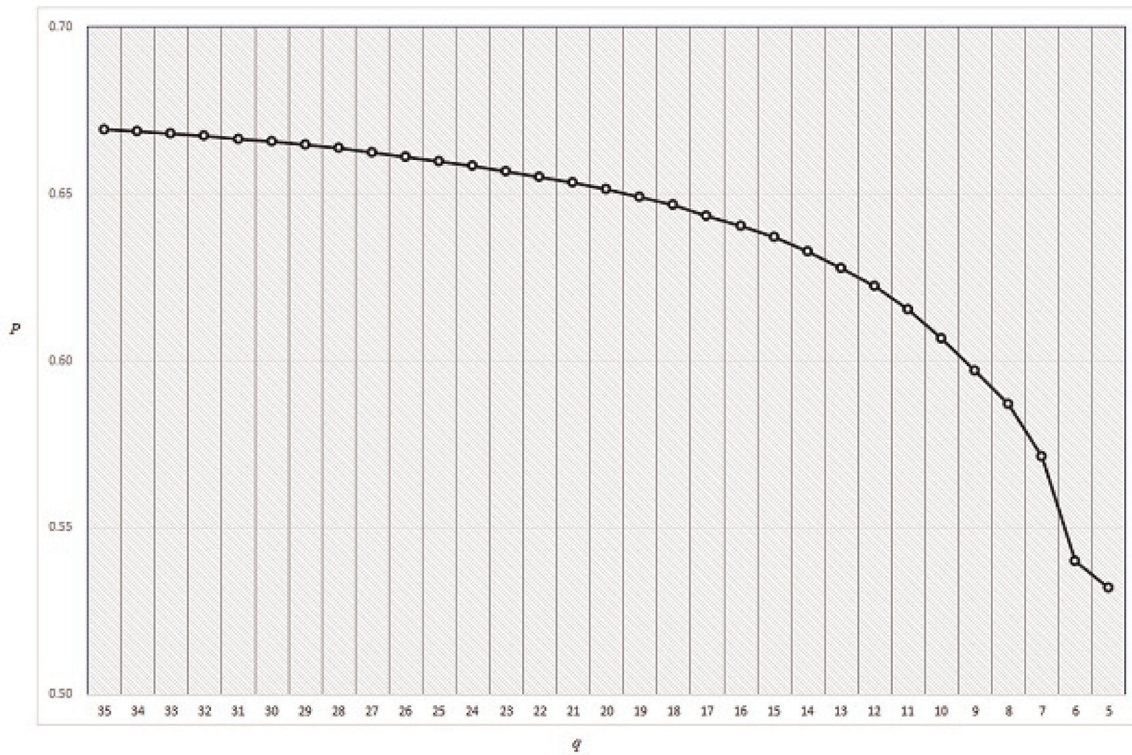


Figure 1. Change of the proportion P for every q (from 35 to 5).

$q = 25$. **Figure 1** shows the change of P for every q . This graph can be used to obtain guidance on the determination of the number of variables. Looking at this graph, if there is a large drop in P , the number of variables just before that point can be used (for this data, no particular drop is observed).

When using RV , the same considerations are applied, and scatter plots are also considered to see how close the configurations are.

3.3 Results of variable selection

Here we select a subset of variables from 35 variables of Expectation data focusing on the loss of proportion P . Suppose we want to keep it under 1%, $q = 25$ which is assigned from **Table 3** and **Figure 1**. The selected variables are marked by \times in the right column of **Table 1**. As far as looking at the variables deleted from each block, two variables {8, 9} from 11 variables in “fashion” block, three variables {12, 16, 20} from 12 variables in “brand” block, and five variables {25, 27, 30, 32, 34} from 12 variables in “shop staff” block are deleted. That is, nine variables are selected from the first two blocks and seven from the third block. It can be stated that the proposed method selects a reasonable subset of variables. Comparing the number of deleted variables in the three blocks, a slightly larger number of variables are removed from the third block, so it is thought that questions on “shop staff” have little information rather than those in the other two blocks and some of them have less significance on the prediction efficiency. From this point of view, we can evaluate the usefulness of each question in the questionnaire.

To evaluate the significance of variables, we observe how many times each variable is selected through the selection for $q = 35, \dots, 5$. Extracting the variables selected over $2/3$ times (24 or more), for example, in the “fashion” block, variables {1, 6, 10, 11} were selected. Given the fact that the close-up questions are located close to each other (1 to 3, recognition on fashion, 4 to 7—consciousness on fashion, 8 to 11—activity on fashion), it is generally clear that NL.M.PCA using the proportion P selects variables well-balanced from the close-up questions. Similarly, if the most frequently selected variables (such as the above four items) are considered as the most important questions, they should be involved in future surveys. If variables are selected a few times, they should not be involved in the future. In such a way, there is a possibility to use the selection results to evaluate the questionnaire itself.

4. Concluding remarks

We reconsider a variable selection problem in PCA for qualitative data based on the idea of Mori et al. [2]. For the problem of how to deal with qualitative data, we apply optimal scaling with the ALS algorithm [4] to the qualitative data. For the variable selection in PCA, we use the criteria in M.PCA of Tanaka and Mori [1] for optimally quantified data. That is, the proposed method is an extension of M.PCA by implementing optimal scaling into M.PCA so as to select a subset of qualitative variables. Using this method, since the quantification is done separately for each variable, we can select a subset of variables from mixed measurement level data.

We apply this method to real data from a customer engagement study [3] to select a subset of qualitative variables by using a criterion that maximizes the prediction efficiency. For a case where there is no preassigned number of variables to be selected, it can be suggested to specify the number in such a way that the maximum loss of the efficiency is not over a certain percentage.

As a result, variables are selected in a well-balanced manner from questions asking similar contents, and the selected subset, therefore, provides as much information as possible. It is expected that the nonlinear M.PCA works well for any mixed measurement level data.

IntechOpen

Author details

Hiroko Katayama*†, Yuichi Mori^{2†} and Masahiro Kuroda²


1 Graduate School of Informatics, Okayama University of Science, Okayama, Japan

2 Department of Management, Okayama University of Science, Okayama, Japan

*Address all correspondence to: hk22@pub.ous.ac.jp

† These authors contributed equally.

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Tanaka Y, Mori Y. Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematics and Management Sciences*. 1997;17(1 & 2):61-89
- [2] Mori Y, Kuroda M, Makino N. *Nonlinear Principal Component Analysis and its Applications (JSS Research Series in Statistics)*. Singapore: Springer; 2017
- [3] Ohyaabu R, Kuroda M, Seino S, Zhang Z. Exploring interplay among customer engagements with multiple objects, the 6th Naples forum on service. *Service Dominant Logic, Network & Systems Theory and Service Science: Integrating three Perspectives for a New Service Agenda*. 2019:103
- [4] Young FW, Takane Y, de Leeuw J. Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*. 1978;43:279-281
- [5] Gifi A. *Nonlinear Multivariate Analysis*. Chichester: Wiley; 1990
- [6] Kruskal JB. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*. 1964;29:115-129
- [7] Rao CR. The use and interpretation of principal component analysis in applied research. *Sankhya*. 1964;A26:329-358
- [8] Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics*. 1976;A25:257-265
- [9] Mori Y, Tanaka T, Tarumi T. Principal component analysis based on a subset of qualitative variables. In: Hayashi C, editor. *Data Science, Classification and Related Methods*. Springer. 1997:547-554