

Introduction of Nonlinear Principal Component Analysis with an Application in Health Science Data

Canan Demir^{1*}, Siddik Keskin²

¹Vocational School of Health Services, Van Yuzuncu Yil University, Van, Turkey

²Department of Biostatistics, Faculty of Medicine, Van Yuzuncu Yil University, Van, Turkey

ABSTRACT

Nonlinear Principal Component Analysis is one of the explanatory dimension reducing technique and presents numerical and graphical results for variable set included linear or nonlinear relationships. In this study, Nonlinear Principal Component Analysis was introduced and the relationship between students' sexual and physical trauma stories and demographic characteristics was examined with this method. In the study, the relationship between trauma and 9 variables obtained by questionnaire from 548 students was evaluated by Nonlinear Principal Component Analysis. The eigenvalue of first dimension has been found to be 1.766 and the eigenvalue of second dimension has been found to be 1.504. The variance explanation rate of these eigenvalues are 17.656% and 15.044% respectively. The total explained variance is seen as 28.550%. With Nonlinear Principal Component Analysis, categorical variables are scaled to the desired size in the most appropriate way, and thus, nonlinear relationships can be modeled as well as linear relationships between variables. With this analysis, gender, age, marital status and suicide variables were found to be effective on trauma.

Keywords: Component loading; Dimension reduction; Nonlinear Principal Component Analysis; Optimal Scaling

Introduction

Principal Component Analysis was put forward by Karl Pearson in the early 1900s for the first time. Then application fields were developed by Hotelling in 1933 and by Rao in 1964 (1). Principal Component Analysis which is used to reduce the size of the structure of dependence between variable is used as a data preparation technique as well as being an analysis technique used alone. Principal Components Analysis is a multivariate analysis technique which benefits from linear relationship between numerous original variable that include continuous data and is composed of linear combination of linear variates with minimal information loss and it includes the process of acquiring fewer variables called principal components and interpreting the results. The Nonlinear Principal Component Analysis is an explanatory dimension reduction technique which gives numerical and visual data for datasets between which there is linear or nonlinear relationship and which include continuous, categorical and discrete variables. The solution technique of Nonlinear Principal Component Analysis was developed first by Van Bekker and Leeuw (1988). It is considered that Nonlinear Principal Component Analysis have not

found an application area because of both difficulties in analysis process and complicity of interpretation of the results obtained. De Leeuw and Rijkevorsel (1980), in their study, showed applicability of Nonlinear Principal Component Analysis in biological data. In Nonlinear Principal Component Analysis, Leeuw (1982) and Stuetzl Hastie (1989) presented a general structure using a nonmetric measure which forms nonlinear optimization problem (2-5).

In this study, Nonlinear Principal Component Analysis was introduced and the theoretical basis, process steps, and the relationship of it with other techniques was examined and besides; after making an application of it in healthcare field, the results obtained have been commented on.

Material and Method

In this study, as application data, some of the data obtained from a survey conducted on 548 students by Boysan and et al. (2009) was used (6). Variables and properties used in this study are given in Table 1. For data analysis; applying Nonlinear Principal Component Analysis in SPSS (ver: 13.0) statistical software package, trauma histories of

*Corresponding Author: Canan Demir, Van Yuzuncu Yil University, Vocational School of Health Services, Van, Turkey

E-mail: canandemir@yyu.edu.tr, Gsm:+90 (505) 521 35 10, Fax: 0 (432) 225 13 03

ORCID ID: Canan Demir: 0000-0002-4204-9756, Siddik Keskin: 0000-0001-9355-6558

Received: 20.10.2021, Accepted: 25.03.2022

students and the relationship of them with these variates have been examined.

The Optimization of Nonlinear Principal Component Analysis

Start: If initial configuration has not been determined, initial value with random numbers is given for X object scores. Then X is orthonormalized. Thus, $u'M_*WX=0$ and $X'M_*WX=n_w m_w I$ and X^+_{w} is obtained. Initial Component Loads are calculated through reciprocal multiplication of X^+_{w} matrix. In addition, the original variables $(I-M_j uu'W/(u'M_j W))h_j$ is centralized and is rescaled to unit length.

Updating the Category Scale: X^+_{w} of the constant value the updated by Y_j with average. And $\tilde{Y}_j = D_j^{-1} G'_j X^+_{w}$. $Y_j^+ = \tilde{Y}_j$ is for multiple nominal. For those which have no Multi-scale level, update process is calculated as $\tilde{Y}_j = D_j^{-1} G'_j X^+_{w}$ in the same way. The alternating least squares (ALS) algorithm is used to calculate the rank decomposition of \tilde{Y}_j . $\tilde{y}_j = \tilde{Y}_j a_j$.

And $\tilde{y}_j = \tilde{Y}_j a_j$ $y_j^* = \tilde{y}_j$ in nominal variables. If the j. variable is taken as an extra category for the next four optimal scaling levels, it is included in the $y_j^* k_j$ category in first stage and removed in the final stage. $y_j^* = d_j + S_j b_j$ for spline nominal spline ordinal variables. Spline transformation \tilde{y}_j is calculated as the weighted regression and weighted with D_j diagonal elements. The elements of b_j is restricted for spline ordinal scaling level in such a way that it will not be negative and that makes y_j^* monotone increasing.

$y_j^* \leftarrow WMON(\tilde{y}_j)$ for ordinal variables. The $WMON()$ function is used to define the weighted monotone regression process and this makes y_j^* monotonically increasing. The weights used are the diagonal elements of D_j . $y_j^* \leftarrow WLIN(\tilde{y}_j)$ is for the numeric variables. $WLIN()$ is used to define the weighted linear regression process. The weights are D_j diagonal elements. Then y_j^* is

normalized (it is included in y_j^* If j. variable is included as an extra category). $y_j^+ = n_w^{1/2} y_j^* (y_j^* D_j y_j^*)^{-1/2}$ after normalization.

Here from the component loads are obtained as $a_j^+ = n_w^{-1} \tilde{Y}' D_j y_j^+$ and lastly $Y_j^+ = y_j^+ a_j^{++}$.

Updating the Object Scores: Firstly Z $Z \leftarrow \sum_j M_j G_j Y_j^+$ which is the auxiliary matrix is calculated. it is centralized by W and M_* . From here, $X_* = (I - M_* uu'W / (u'M_* Wu))Z$.

Orthonormalization: The least squares approach and Procrustes rotation is used to find out M_* i and orthonormal X^+ . $n_w^{1/2} m_w^{1/2} M_*^{-1/2} W^{1/2} KL'$ is obtained from the equality $m_w^{1/2} M_*^{-1/2} W^{1/2} X^* = K \Lambda^{1/2} L'$ with singular value decomposition. From here, orthonormal weighted object scores are calculated as $X^+_{w} \leftarrow n_w^{1/2} m_w M_*^{-1} W X^* \Lambda^{-1/2} L'$ and $X^+ = W^{-1} X^+_{w}$. L and λ is based on the householder transformation following QL algorithm and the triple diagonalization.

Convergence Test (repeat 2-4): Convergence test is done by the equality $TFIT = (pn_w)^{-1} \sum_{j \in J} V_j tr(Y'_j D_j Y_j) + \sum_{j \notin J} V_j a'_j a_j$. After convergence, the value $TFIT$ is $tr(\Lambda^{1/2})$. Λ is calculated in the step orthonormalization in the process of the last iteration.

Turning (notation): The matrix X^+ is converted the matrix L to obtain the principal axis in the conversion result. In addition, s. column of the matrix X^+ reflects squares average which belongs to s-dimensional loadings. The negative signs indicate that the average of quadratic loadings is higher.

Maximum Number of Rank: Maximum number of rank " ρ_{max} " points out the maximum number of dimensions which can be calculated for any data set and generally P_{max} is calculated as

$$\rho_{max} = \begin{cases} \min \{ (n-1), r_1, r_2 \} & K = 2 \\ \min \{ (n-1), \max r_k \} & K > 2 \end{cases}$$

The value r_k is defined as in the equality $r_k = \sum_{j \in JM(k)} k_j + m_{k1} - m_{k2}$, m_{k1} is the number of multiple variables with non-missing observation in the set of variable k. m_{k2} is the number of single variables in the set of variable k. $JM(k)$ is index set and the number of variables in multiple sets.

Overall is stopped if the following conditions are not fulfilled.

- 1) $r_k < n_k - 1$
- 2) $n_k > 2$
- 3) $\sum_k r_k \leq \sum_k (n_k - 1) - (n_{\max} - 1)$

Here n_k is the number of object with non-missing observation in k. set. $r_{n_{\max}}$ indicates the number of objects in n_k .

Compliance and Loss Measurements: Total fit is the equality TFIT in convergence in previous process. Total loss function is multiple and single loss function and is written as follows.

$$TMLOSS = (m_{w1} + pm_{w2}) - \left((n_w p)^{-1} \sum_{j \in J} v_j \text{tr}(Y'_j D_j Y_j) + n_w^{-1} \sum_{j \notin J} v_j \text{tr}(Y'_j D_j Y_j) \right)$$

$$SLOOS = n_w^{-1} \sum_{j \in J} v_j \text{tr}(Y'_j D_j Y_j) - \sum_{j \in J} v_j a'_j a_j$$

Cronbach's Alpha: Cronbach's Alpha is calculated with the equality $a_s = m_w (\lambda_s^{1/2} - 1) / (\lambda_s^{1/2} (m_w - 1))$ and Total Cronbach's Alpha is calculated with the equality $a = m_w (\sum_j \lambda_s^{1/2} - 1) / \sum_j \lambda_s^{1/2} (m_w - 1)$ for each dimension. Here λ_s is s. diagonal elements of Λ and it is calculated in the orthonormalization step in the most recent iteration process.

Variance Explanations Rates: Variance explanations rates are calculated for multiple nominal variables for each dimensions to ($s=1, \dots, p$):

$$VAF1_s = n_w^{-1} \sum_{j \in J} v_j \text{tr}(y'_{(j)s} D_j y_{(j)s}), \quad (\% \text{ variance } VAF1_{s \times 100 / m_{w1}})$$

and for non-multiple nominal variables;

$$VAF2_2 = \sum_{j \in J} v_j a_{js}^2, \quad (\% \text{ variance } VAF2_{s \times 100 / m_{w2}}).$$

Eigenvalues is $\lambda_s^{1/2} = VAF1_s + VAF2_s$, for each dimension. λ_s is s. diagonal elements of Λ and it is calculated in the orthonormalization step in the most recent iteration process.

Total variance is calculated by the equality $\text{tr}(\Lambda^{1/2}) = p^{-1} \sum_s VAF1_s + \sum_s VAF2_s$, for multiple nominal and non-multiple nominal variables on the average of dimensions (7). This equality is known as total eigenvalues. $\Lambda^{1/2}$ are

eigenvalues of correlation matrix weighted by variable weights and is calculated as follows if there is no passive missing observation.

$$r_{jj}^w = v_j r_{jj}, \quad \text{ve} \quad r_{j1}^w = r_{1j}^w = v_j^{1/2} r_{j1}$$

Eigenvalues is $m_w Q'_c M_*^{-1} Q_c$, and $Q_c = n_w^{-1/2} (I - M_* u u' W / (u' M_* W u)) Q$ if there is passive missing observation.

Centroid Coordinates: Centroid coordinates is calculated by the equality $VAF_{js} = v_j \text{tr}(Y'_{js} D_j Y_{js})$.

Vector Coordinates: Vector coordinates for each

$$j \notin J: VAF_{js} = v_j a_{js}^2$$

Correlations and Eigenvalues

Before Transformation: $R = n_w^{-1} H'_c W H_c$ is weighted and normalized with the H_c matrix. 1st row and 1st column is removed if the variable j is complementary variable for eigenvalue decomposition of R matrix. Then r_{ij} is multiplied by $(v_i v_j)^{1/2}$.

After Transformation: Correlation matrix is $q_j = G_j y_j$ and $R = n_w^{-1} Q' W Q$ if analysis is done for non-multiple variables and if there is non-missing observation or it is defined as passive. The 1st p Eigenvalue of R is equal to $\Lambda^{1/2}$. p correlation matrix is calculated by the equality $R_{(s)} = n_w^{-1} Q'_{(s)} W Q_{(s)}$ to ($s=1, \dots, p$) if there is multiple nominal variables in analysis. It is $q_{(s)j} = G_j y_j$ for non-multiple variables and it is $q_{(s)j} = n_w^{1/2} G_j Y_{(j)s} (Y'_{(j)s} D_j Y_{(j)s})^{-1/2}$ for multiple nominal variables. Generally 1st eigenvalue of $R_{(s)}$ is higher. Lower values of $\Lambda^{1/2}$ is mostly second or subsequent eigenvalues of $R_{(s)}$ matrix. Numeric variables go in model without calculating correlations if missing observation is defined as passive. In this case, eigenvalues of correlation matrix are not equal to Scaling for extra categories of multiple nominal variables is calculated by

$$Y_{(j)(k+1)s} = \left(\sum_{i \in I} W_i \right)^{-1} \sum_{i \in I} W_i X_{is}$$

it is calculated for non-multiple variables as follows.

$$y_{(k+1)j} = n_w^{1/2} \left(\sum_s a_{js}^2 \right)^{-1} \sum_s a_{js} Y_{(j)(k+1)s}$$

1st row and column of R matrix is removed to calculate eigenvalues of R (eigenvalue decomposition).

Object Scores and Loadings: The normalization decomposition of the first p. singular values of $n_w^{-1/2}W^{1/2}QV^{1/2}$ is obtained by dividing by m_w and A , then it is obtained by multiplying by v_j which are diagonal elements of Q and V from object scores of X. According to this, the singular value decomposition of $n_w^{-1/2}W^{1/2}QV^{1/2}$ is $SVD(n_w^{-1/2}W^{1/2}QV^{1/2}) = K\Phi^{1/2}L'$.

$X=K_p$ (here p index indicates the first p. columns of K) and $A = (L\Phi^{1/2})_p$, XA' gives the best approach with p dimension of $n_w^{-1/2}W^{1/2}QV^{1/2}$.

The first p. singular values of $\Phi_p^{1/2}$ are equal to $\Lambda^{1/4}$. Λ is calculated in the orthonormalization step in the latest iteration process.

It is written as $(K\Phi^{1/2}L')_p = K_p \Phi_p^{a/2} \Phi_p^{b/2} L'_p = K_p \Lambda^{a/4} \Lambda^{b/4} L'_{p,(a+b=1)}$

for the decomposition of the first p. singular values. Principal variable normalization is used in the optimization process and after convergence, it is $X = n_w^{1/2}W^{-1/2}K_p$ and $A = V^{-1/2}L_p \Lambda^{1/4}$. It is equal to $n_w^{-1/2}W^{1/2}Q_{(s)}V^{1/2} \Lambda_s^{1/4}$ if multiple nominal variables are analyzed. The normalized factors is showed for also centroids if variables are in the multiple nominal scaling level. And it is $Y_j^n = Y_j \Lambda^{1/4(b-1)}$.

Measurement: Measurements in the multiple scaling level are showed as y_j . vector coordinates are one of other normalization options $y_j(a_j^n)$ centroid coordinates y_j principal normalization and are $D_j^{-1}G_j'WX^n$. Measurements for multiple nominal variables are centroid coordinates and Y_j^n . Only centroid coordinates is showed for this category if any category are used as complementary objects and it is calculated as $y_{(j)r} = n_w^{1/2}n_{jr}^{-1} \sum_{j \in J} x_j^n$. It is calculated as $y_{(j)r} = n_w^{1/2}n_{jr}^{-1} \sum_{j \in J} x_j \Lambda^{1/4(b-1)}$ for variables which is not in multiple scaling level. Here $y_{(j)r}$ is r. row of Y_j , n_{jr} is the number of objects in r. category and I is the record index value for objects in r. category.

Remains: Remains are given as $j(G_j y_j)$ variable of Xa_j and scatter graph for non-multiple variables approximately. Scatter graph for each dimension in multiple nominal variables is obtained by drawing $G_j y_{(j)s}^n$ against to x_s^n .

Planned Centroids: The centroid of i. variable which is planned on j. variable is $Y_j a_j (a_j a_j)^{-1/2}$ if $j \notin J$.

Graph of Scaled and Load Factor of Two or Three: Centroids and object scores are scaled again by using the following scaling factors in the graphs of both of centroids and loadings or with object scores (8).

$$\text{Scaling Factor} = \frac{2 \sum_{s=1}^p \max(a_{1s}^n, \dots, a_{ms}^n)}{\sum_{s=1}^p I \min(x_{1s}^n, \dots, x_{ns}^n) I + (\max(x_{1s}^n, \dots, s_{ns}^n))}$$

Results

Eigenvalues of each dimension is seen when Table 2 is examined. These eigenvalues are identical with the ones in standard Principal Component Analysis and they indicate how much of the total variation is explained by each dimension. Eigenvalues also indicate how many dimensions are required. However, generally, the process of reducing to two dimensions is performed and the results are shown in two dimensions. In the study, the eigenvalue of first dimension has been found to be 1.766 and the eigenvalue of second dimension has been found to be 1.504 the variance explanation rate of these eigenvalues are 17,656% and 15,044% respectively. The total explained variance is found as 28,550%. As a general rule for multiple nominal variables, the number of suitable dimensions can be found with the equality "The total category numbers of variables - the number of variables - 1".

When Table 3 is examined, the highest correlation value is found to be 0.251 between suicide and mental status variables. This is followed by the correlations between age group and gender, address and gender and age group and marital status variable. The lowest correlation is recorded to be between economy and marital status, trauma and marital status, suicide and address, suicide and residence, with a value of 0.002.

When Table 4 is examined; the highest correlation

Table 1. The Considered Variables and Properties

Variable name	Categories	Type
Sex	(1) Girl (2) Male	Nominal
Marital status	(1) Married (2) Bachelor	Nominal
Age group	(1) 17- 19 (2) 20- 25 (3) 26- 30 (4) 31- 35	Ordinal
Mental status	(1) Yes (2) No	Nominal
Address	(1) Next of Family (2) Dormitory (3) Pension (4) Single at Home (5) Friends at Home	Multiple Nominal
Residing	(1) Metropolitan (2) City (3) Town (4) Village	Multiple Nominal
Economic	(1) Low (2) Medium (3) Hight	Ordinal
Number brother group	(1) 0- 5 (2) 6- 10 (3) 11- 15	Ordinal
Suicide	(1) Yes (2) No	Nominal
Trauma	(1) Sexual Trauma (2) Pysical Trauma (3) Sexual and Pysical Trauma (4) Absent	Nominal

in the first dimension is seen to be between gender and address variables with the value of 0.359. The highest correlation in the second dimension is seen to be between the age and gender with the value of 0.271.

When Table 5 is examined, while in the first-dimension gender variable has the highest load value with 0.771, in the second dimension, suicide variable has the highest load value with 0.557 (Figure 1).

Discussion

Nonlinear Principal Component Analysis is known as nonlinear generalized form of Principal Component Analysis. Nonlinear Principal Component Analysis includes multiple concepts. Similarly, with Principal Components Analysis it is the technique to explain the nonlinear relationships (9). Nonlinear Principal Component Analysis not only makes reduction in data set which is composed of original variables but also provides presentation of these relationships visually. Bekker and de Leeuw (1988), Burgers and

Table 2. The Number Size and Proportion of Variance Explanation

Dimension	Cronbach's Alpha	Variance explanations rates					
		Multiple nominal variables		No multiple variables		Total (Eigenvalues)	% Variance
		Total	% Variance	Total	% Variance		
1	0.482	0.442	22.079	1.324	16.550	1.766	17.656
2	0.373	0.388	19.420	1.116	13.950	1.504	15.044
Total	0.722	0.415	20.750	2.440	30.500	2.855	28.550

Table 3. Correlations Between The Original Variables

	Sex	Marital status	Address	Residing	Economic	Mental status	Suicide	Age Gr.	Brother Gr.	Trauma
Sex	1									
Marital status	0	1								
Address	0.218	-0.12	1							
Residing	0.062	-0.024	0.042	1						
Economic	0.156	0.002	-0.109	-0.036	1					
Mental status	0.068	0.021	-0.006	0.046	-0.064	1				
Suicide	0.102	0.05	-0.002	-0.002	-0.014	0.251	1			
Age Gr.	0.229	0.213	0.084	-0.022	0.008	0.064	0.065	1		
Brother Gr.	0.171	-0.057	0.121	0.184	-0.06	0.044	0.012	0.052	1	
Trauma	0.157	0.002	-0.088	0.018	0.03	0.038	0.031	0.051	-0.093	1
Dimension	1	2	3	4	5	6	7	8	9	10
Eigenvalues	1.647	1.328	1.193	1.05	0.962	0.914	0.781	0.752	0.733	0.639

de Leeuw (1988), in their studies, investigated the relationship between the analysis of variance and Nonlinear Principal Component Analysis (10,11). Türe et al (2002), applied Nonlinear Principal Component Analysis and cluster analysis to the data obtained from a satisfaction survey about hospital staff conducted on 294 patients who applied to Trakya University Medical Faculty Family Practice. Researchers have indicated that according to their method %89.4 of total variation have been explained with the aid of the predicted first component (12). Al-Shaban et al. (2010), in their study, examined complex systems like human brain which is composed of thousands of neurons with Nonlinear Principal Component Analysis (13). Gower and Blasius (2005), in their study, examined the Nonlinear Principal Component Analysis and mentioned on important issues about the subject (14). DeMers and Cottrell (1993), in a study, applied Kramer's methods about Nonlinear

Principal Component Analysis to many different problems successfully (15).

In applications, nonlinear relation structure between variables is a very common case. In addition, variables that are subject to the study can generally be in different structures. These two cases limit the use of several statistical methods commonly used in practice. Researchers solves this problem either making some transformations to the variables or leaving variables in different forms out of analysis. In both cases, the relation structure between original variables can be destroyed and the results obtained can be deceptive. Therefore, Nonlinear Principal Component Analysis come into prominence as a suitable analysis method. Analysis explains as big part of the variation in original variation cluster as possible and aims to get fewer new variables between which there is no correlation and which

Table 4. Correlations Between Transformed Variables

	Sex	Marital status	Address	Residing	Economic	Mental status	Suicide	Age Gr.	Brother Gr.	Trauma
Sex	1	0	0.359	0.072	-0.156	0.068	0.102	0.271	0.169	-0.184
Marital status	0	1	0.035	-0.014	0.012	0.021	0.05	0.109	-0.057	0.02
Address	0.033	0.214	1	0.047	-0.021	0.021	0.047	0.197	0.096	-0.081
Residing	0.023	0.015	0.116	1	-0.065	0.042	0.007	0.038	0.2	0.022
Economic	0.156	0.012	0.128	0.04	1	-0.071	-0.016	0.015	-0.056	0.026
Mental status	0.068	0.021	0.033	0.039	-0.071	1	0.251	0.067	0.042	0.175
Suicide	0.102	0.05	0.099	0.023	-0.016	0.251	1	0.1	0.01	0.136
Age Gr.	0.271	0.109	-0.002	-0.003	-0.015	0.067	0.1	1	0.054	-0.069
Brother Gr	0.169	-0.057	-0.091	-0.064	-0.056	0.042	0.01	0.054	1	-0.109
Trauma	0.184	0.02	0.101	0.012	0.026	0.175	0.136	0.069	-0.109	1
Eigenvalues 1	1.766	1.396	1.217	0.988	0.965	0.856	0.774	0.746	0.719	0.572
Eigenvalues 2	1.533	1.504	1.222	0.992	0.959	0.897	0.814	0.764	0.718	0.596

Table 5. Component Loadings

	Dimension			Dimension	
	1	2		1	2
Sex	0.771	-0.056	Mental status	0.218	0.500
Marital status	0.083	0.404	Suicide	0.256	0.557
Address			Age Gr.	0.547	0.126
Residing			Brother Gr.	0.411	-0.265
Economic	-0.268	0.156	Trauma	-0.265	0.528

are composed of combination of original variables. This approach has the same objective with the standard Principal Component Analysis. However, variables such as nonlinear relationship between the variables, this approach will be measured at different levels. In other words, in Nonlinear Principal Component Analysis, along with numeric variables, classifier and ordinal variables can be analyzed at the same time. Nonlinear Principal Component Analysis is a suitable method used for dimension reduction particularly for serial variables. Because the order of the categories for individuals are considered simultaneously and the possible multi-dimensional cognitive values are obtained.

Particularly in social sciences; some of the assumptions of the standard Principal Component Analysis are not usually provided. One of these is related to the measurement levels of the variables, and measuring levels of categorical variables are generally categorical or ordinal. Likert-type scales can also be assumed as interval scaling. However, these scales can be taken as ordinal scaling, and Nonlinear Principal Component Analysis can be applied. Ordinal number is given categorically. However, they do not need to be equally weighted. Thus, Nonlinear Principal Component Analysis can also be taken as ordinal scale can be put to analysis. Although Likert-type scale data can be taken as equally spaced and can be applied

to standard Principal Component Analysis, in case gender, marital status, age group and region variables are added to these

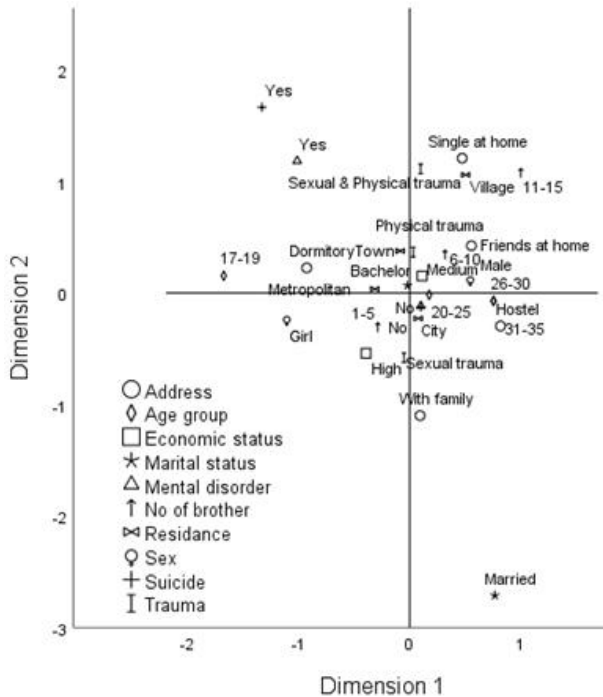


Fig 1. Configuration map for the categories of the variables

variables, standard Principal Component Analysis is not suitable while Nonlinear Principal Component Analysis can be a more suitable method. Nonlinear Principal Component Analysis (NLPCA) model is identical to the standard Principal Component Analysis model. However, it is applied to data made nonlinear transformation. Thus, the mathematical properties of the Principal Component Analysis are also discussed here. Variables are converted to score values assigned to with optimal scaling to category. Thus, transformed variables take numerical values. Nonlinear Principal Component Analysis makes scaling of categories for transformed variables with optimal scaling in such a way that the general variation will be maximum.

Advantages of Nonlinear Principal Component Analysis can be summarized as follows.

1. It does not require assumptions like normality and linearity.
2. Taking into account the different types of variables and variable structures, it provides the opportunity to examine the relationship between them.
3. The Principal Component Analysis does not use correlation matrix like standard Principal Component Analysis but reach a solution through data itself.

4. The variables in mixed structure can also be included in the model simultaneously.

5. Moreover, this analysis method, allows the graphical display of the variables considered in the study in two-dimensional map.

On the other hand; it can be said that there are two important disadvantages. The first is that continuous variables are included in the analysis by being categorized. Thus, this situation can be considered as the loss of some information. Secondly, any statistical significance (test of significance) test is not done. Therefore, Nonlinear Principal Component Analysis is seen as one of the investigation analysis techniques rather than confirmative analysis techniques.

References

1. Timm NH. Applied Multivariate Analysis (2 nd ed). USA: 2002.
2. Leeuw JD. Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. Computational Statistics and Data Analysis 2006; 50(1): 21-39.
3. Leeuw JD, Rijkevorsel JV. Homals and princials: Some generalizations of principal components analysis. Data Analysis and Informatics 1980; 89: 231-242.
4. Leeuw JD. Nonlinear principal component analysis. Proceedings in Computational Statistics COMPSTAT 1982; October 30 -September 3, Toulouse, 77-86.
5. Hastie T, Stuetzle W. Principal curves. Journal of the American Statistical Association 1989; 84(406): 502-516.
6. Boysan M, Goldsmith RE, Çavuş H, Kayri M, Keskin S. Relations among anxiety, depression and dissociative symptoms, the influence of abuse subtype. Journal of Trauma and Dissociation 2009; 10: 83-101.
7. Demir Y, Keskin S, Çavuşoğlu Ş. Doğrusal olmayan temel bileşenler analizinin tanıtımı ve uygulanabilirliği. KSÜ Tarım ve Doğa Derg 2021; 24(2): 442-450.
8. SPSS Base 13.0 for Windows. Chicago, IL, SPSS Inc. 2004.
9. Scholz M, Kaplan F, Guy CL, Kopka J, Selbig J. Non-linear PCA: a missing data approach. Bioinformatics 2005; 21(20): 3887-3895.
10. Bekker P, Leeuw JD. Component and Correspondence Analysis (1 nd ed). NewYork: 1988.
11. Burgers D, Leeuw JD. Homogeneity analysis with k sets of variables. An alternating least squares method with optimal scaling features. Psychometrika 1988; 53(2): 177-197.

12. Türe M, Süt N, Yavuz E, Kurt İ, Türe H. Akademik personelin tercihlerin belirlenmesinde, Doğrusal Olmay haritaları”, T Araştırmalar Dergisi 2002; 2(1): 72-76.
13. Al-Shaban SA, Al-Faysale MS, Al- Neami AQH. Nonlinear principal component analysis neural network for blind source separation of EEG signals. Research Journal of Applied Sciences, Engineering and Technology 2010; 2(2): 180-190.
14. Türe H, Süt N, Yavuz E, Kurt İ, Türe H. İyile prediction with its analysis. Theory, Quality & Quantity 2005; 39: 359-372.
15. DeMers D, Cottrell G. Nonlinear dimensionality reduction. Advances in Neural Information Processing Systems 1993; 5: 580-587.