Available online at www.sciencedirect.com
SCIENCE DIRECT®
Expert Systems with Applications xxx (2005) xxx–xxx
Expert Systems with Applications
www.elsevier.com/locate/eswa

Comparison of dimension reduction methods using patient satisfaction data

Mevlut Ture [a,*], Imran Kurt [a], Zekeriya Akturk [b]

[a] *Trakya University, Medical Faculty, Department of Biostatistics, 22030 Edirne, Turkey*
[b] *Trakya University, Medical Faculty, Department of Family Medicine, 22030 Edirne, Turkey*

Abstract

In this study, we compared classical principal components analysis (PCA), generalized principal components analysis (GPCA), linear principal components analysis using neural networks (PCA-NN), and non-linear principal components analysis using neural networks (NLPCA-NN). Data were extracted from the patient satisfaction query with regard to the satisfaction of patients from hospital staff, which was applied in 2005 at the outpatient clinics of Trakya University Medical Faculty. We found that percentages of explained variance of principal components from PCA-NN and NLPCA-NN were highest for doctor, nurse, radiology technician, laboratory technician, and other staff using a patient satisfaction data set. Results show that methods using NN which have higher percentages of explained variances than classical methods could be used for dimension reduction.

Keywords: Principal components analysis; Artificial neural networks; Generalized principal components analysis; Dimension reduction; Patient satisfaction

1. Introduction

High number of variables and interactions between variables makes it harder to interpret and summarize the results as well as to apply multivariate statistics. Principle Components Analysis (PCA) is a method developed to remove the dependence between the variables and to reduce p variables to m variables ($p > m$) with longitudinal components (Ozdamar, 2004; Tatlıdil, 1996). PCA only examines continuous relationships between ordinal or continuous variables. Generalized Principal Components Analysis (GPCA), which is developed as an alternative for PCA, is the optimal scaling method for data sets with continuous, ordinal, and nominal variables.

Another method currently used for dimension reduction is the neural networks (NN). Inspired by biological neural networks, the NN imitates functions of the human brain such as knowledge transfer and storage. This method can determine the basic variables by evaluating linear and non-linear relationships without any restriction to the type of the variables.

Some researchers have investigated the performance of these methods with different data sets. Dong and MacAvoy (1996) have compared PCA and non-linear principal components analysis using neural networks (NLPCA-NN) in image compression. Monahan (2000) has compared the dimension reduction performances of PCA and NLPCA-NN using a data set related with climate. Albanis and Batchelor (1999) have compared PCA, PCA-NN, and NLPCA-NN in dimension reduction of financial rates in evaluating a data set with long term credit continuity. Hsieh (2001) has compared PCA, rotated PCA, and NLPCA-NN using surface temperature data of the Pacific Sea.

This study aimed to compare the PCA, GPCA, PCA-NN, and NLPCA-NN methods in the dimension reduction of questions examining satisfaction of the 294 patients applying at Trakya University Medical Faculty in 2005.

* Corresponding author. Tel.: +90 284 2357641x1631; fax: +90 284 2357652.
E-mail address: ture@trakya.edu.tr (M. Ture).

0957-4174/$ - see front matter
doi:10.1016/j.eswa.2005.12.003

2. Material and methods

2.1. Data

Our study included 294 consecutive patients (46.3% males, 53.7% females) admitted to the outpatient clinics of Trakya University Medical Faculty on 12 January 2005. A patient satisfaction questionnaire consisting of 31 items with a 5-point Likert scale (1—very bad, 2—bad, 3—average, 4—good, 5—excellent) evaluating the satisfaction of patients from the hospital staff was applied to the patients.

The reliability coefficient of the questionnaire (Cronbach α) was 0.96. The 31 items were subdivided as satisfaction from the doctor, nurse, radiology technician, laboratory technician, and other staff. The reliability coefficients for doctor, nurse, radiology technician, laboratory technician, and other staff were 0.96, 0.93, 0.91, 0.93, and 0.84 respectively (Table 1).

2.2. Principal components analysis

Essentially, PCA maximizes the correlation between the original variables to form new variables that are mutually orthogonal, or uncorrelated (A project funded by the Tsunami Initiative, 1999; Ozdamar, 2004). PCA was used to describe the variance in data sets of n observations on p variables (Jolliffee, 1986; Manly & Bryan, 1986). PCA is a statistical technique that linearly transforms the original set of variables into a substantially smaller set of uncorrelated variables that represents the maximum amount of information in the original set of variables. A small set of uncorrelated variables is much easier to understand and use in further analyses than a larger set of correlated variables.

Principal components are able to describe different dimensions of a given $n \times p$ data set of n observations on p variables. The first principal component (Y_1) can be defined as a linear combination of the elements of the data matrix, **X**:

Table 1
Reliability coefficients of the different groups

Question	Cronbach α
Items related with satisfaction from the doctor	
1. Giving you enough time during the consultation	0.96
2. Facilitating you to explain you problems	
3. Involving you in clinical decisions related with your care	
4. Listening to you	
5. Having fast relief for your complaints	
6. Examining you	
7. Explaining the aims of tests ant treatments	
8. Responding to your information requests	
9. Helping you to deal with your anxieties	
10. Helping to become aware of the importance of his suggestions	
11. Remembering what he/she said and did during the previous encounters	
12. His/her respect to you	
13. The interest/listening of the doctor	
14. Cheerfulness of the doctor	
15. Giving you information	
Items related with the satisfaction from the nurse	
1. Respect to you	0.93
2. Interest/listening to you	
3. Cheerfulness of the nurse	
4. Giving you information	
Items related with satisfaction from the radiology technician	
1. Respect of the radiology technician	0.91
2. Interest/listening to you	
3. Cheerfulness of the radiology technician	
4. Giving you information	
Items related with satisfaction from the laboratory technician	
1. Respect of the laboratory technician	0.93
2. Interest/listening of the laboratory technician	
3. Cheerfulness of the laboratory technician	
4. Giving you information	
Items related with satisfaction from other staff	
1. Interest and listening of the registration desk staff to you	0.84
2. Cheerfulness of the registration staff	
3. Respect of the registration staff to you	
4. Attitudes of the security staff	

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$

where coefficients that chosen so as to maximize the variance represented by the first principal component are simply the eigenvectors of the symmetric covariance matrix.

The eigenvalues of the covariance matrix represent the variation of each principal component, where

$$\mathrm{Var}(Y_i) = \lambda_i$$

Ideally, a principal component analysis will yield several components that describe the majority of the total variation of the data set. Geometrically, the first PC is the line of closest fit to the n observations. It minimizes the sum of the squared distances of the n observations from the line where the distance is defined in a direction perpendicular to the line. The second PC a line of closest fit to the residuals from the first PC, the third PC is a line of closest fit to the residuals from the second PC, and so on (Ozdamar, 2004; Sharma, 1996; Tatlıdil, 1996).

2.3. Generalized principal components analysis

The GPCA procedure quantifies categorical variables using optimal scaling, resulting in optimal principal components for the transformed variables. The variables can be given mixed optimal scaling levels and no distributional assumptions about the variables are made (Gifi, 1990; Michailidis & De Leeuw, 1998, 2000; SPSS Inc., 1999).

In GPCA, dimensions correspond to components, and object scores correspond to component scores (Gifi, 1990; SPSS Inc., 1999).

The GPCA objective is to find object scores X and a set of $\underline{\mathbf{Y}}_j$ (for $j = 1, 2, \ldots, m$) so that the function

$$\boldsymbol{\sigma}(\mathbf{X}; \underline{\mathbf{Y}}) = n_{\mathrm{w}}^{-1} \sum_j c^{-1} \mathrm{tr}((\mathbf{X} - \mathbf{G}_j\underline{\mathbf{Y}}_j)' \mathbf{M}_j \mathbf{W} (\mathbf{X} - \mathbf{G}_j\underline{\mathbf{Y}}_j))$$

with c is p if $j \in J$,

is minimal, under the normalization restriction $\mathbf{X}'\mathbf{M}_*\mathbf{W}\mathbf{X} = n_{\mathrm{w}} m_{\mathrm{w}} \mathbf{I}$ ($\mathbf{I}$ is the $p \times p$ identity matrix). The inclusion of $\mathbf{M}_j$ in $\boldsymbol{\sigma}(\mathbf{X}; \underline{\mathbf{Y}})$ ensures that there is no influence of passive missing values. $\mathbf{M}_*$ contains the number of active data values for each object. The object scores are also centered. $\underline{\mathbf{Y}}$ is that collection of category quantifications for variables with multiple nominal scaling level, and vector coordinates for non-multiple scaling level. $\mathbf{G}_j$ is indicator matrix for variable j, of order $n_{\mathrm{total}} \times k_j$. n_{w} is the weighted number of analysis cases, m_{w} is the weighted number of analysis variables, and $\mathbf{W}$ is diagonal $n_{\mathrm{total}} \times n_{\mathrm{total}}$ matrix, with w_i on the diagonal (Gifi, 1990; Michailidis & De Leeuw, 1998, 2000).

2.4. Linear principal components analysis using neural networks

PCA-NN is mainly used for classification and feature extraction. The goal of PCA is to find a set of orthogonal components that minimize the error in the reconstructed data. An equivalent formulation of PCA is to find an orthogonal set of vectors that maximize the variance of the projected data (Diamantras & Kung, 1996).

Sanger proved that one-layered linear neural network is equivalent to the linear standard PCA. And the neural networks which implement this learning algorithm is called PCA-NN. We are assuming that the network has m outputs, each given by

$$y_j(n) = \sum_{i=1}^{p} w_{ij}(n)x_i(n), \quad j = 1, 2, \ldots, m$$

and p inputs ($m < p$). To apply Sanger's rule the weights ($w_{ij}(n)$) are updated according to

$$\Delta w_{ji}(n) = \eta[y_j(n)x_i(n) - y_j(n)\sum_{k=1}^{j} w_{ki}(n)y_k(n)], \quad i = 1, 2, \ldots, p$$

where η is the step size. In this rule, the input to each neuron is modified by subtracting the product of the outputs from the preceding neurons and the respective weights. This implements the deflation method after the system converges. That is, after convergence of the first neuron weights will the second neuron weights converge completely to the eigenvector that corresponds to the second largest eigenvalue (Albanis & Batchelor, 1999; Hassoun, 1995; Principe, Euliano, & Lefebvre, 2000).

2.5. Non-linear principal components analysis using neural networks

The NLPCA-NN is a general purpose feature extraction algorithm producing features that retain the maximum possible amount of information from the original data set. If non-linear correlations between variables exist and sufficient data to support the formulation between more complex mapping functions are available, then NLPCA will describe the data with greater accuracy than PCA (Albanis & Batchelor, 1999).

Kramer presented a NLPCA-NN method based on autoassociative neural networks that are trained by backpropagation (Krammer, 1991). NLPCA-NN uses five-layer feed-forward network with a bottleneck layer of nodes to reduce the dimension of the input variables and each layer fully connected to the next (Krammer, 1991; Oja, 1992). The second and fourth layers of the network have sigmoidal activation functions, so layers 1–3 and layers 3–5 model non-linear functions. The activation functions of the third and fifth layers are linear. The input (first) and output (fifth) layers have p units (the number of variables in the data set). The third layer has fewer nodes ($m < p$) than the first or fifth. The values of the output nodes in layer 5 are trained to approximate the inputs. After the network has been trained, bottleneck node activation values in layer 3 give a lower dimensional representation of the inputs (Fotheringhame & Baddeley, 1997; Monahan, 2000).

The goal of the network is to minimize the error term (e). The network is then trained to try and reproduce the input pattern at the output layer by using an error term which is simply the squared difference between the network prediction (X_i) and input pattern (X'_i);

$$e = \sum_{i=1}^{p} (X_i - X'_i)^2 \quad (i = 1, 2, \ldots, p)$$

NLPCA-NN reduces the dimension of the inputs by fitting a curve through the data. The first three layers of the network project the original data onto the curve and the activation values of the bottleneck layer, called scores, give the location of the projection. The last three layers define the curve (Albanis & Batchelor, 1999; Daszykowski, Walczak, & Massart, 2003; Hsieh, 2001; Michailidis & De Leeuw, 2000).

2.6. Package programs

In this study, data were analyzed using SPSS 10.5 (PCA, GPCA, Hierarchical Cluster Analysis) and NeuroSolutions 5.0 (PCA-NN, NLPCA-NN).

3. Results

The whole data set was used for PCA and GPCA. Before building NN, the data set was randomly split into two parts: 80% ($n = 235$) of the data for a training set and 20% ($n = 59$) for a cross validation set.

In Table 2, we present the percentages of variance explained by dimension reduction methods for a principal component which represented items of each group. The minimum–maximum levels for percentage of variance explained with PCA, GPCA*, GPCA+, PCA-NN and NLPCA-NN were 62.2–82.1%, 60.2–82.1%, 63.4–84.0%, 84.9–91.7%, and 84.9–96.1% respectively (Table 2).

As it can be seen from Table 2 and Fig. 1, PCA-NN and NLPCA-NN had the highest percentages of variance explained for doctor, nurse, radiology technician, laboratory technician, and other staff. Ordinal GPCA performed better than numeric GPCA and PCA.

Percentages of variance explained were used as input variables in the Hierarchical Cluster Analysis (HCA) (Ture, Kurt, Kurum, & Ozdamar, 2005). HCA was done to identify homogenous groups of dimensionality reduction techniques based on percentages of variance explained. The dendrogram from centroid clustering method that was obtained is shown in Fig. 2. In the dendrogram, the data points appear to cluster in two groups. The first cluster includes PCA, GPCA*, and GPCA+. The second cluster includes PCA-NN and NLPCA-NN. We found that both PCA-NN and NLPCA-NN explain a higher amount of variation in the original set of variables than other techniques.

4. Discussion

Frequently the investigated issues are under the effect of multiple variables and thus it is mandatory to evaluate the effecting variables together in order to ascertain reliability and validity. However, having multiple variables and relationship between variables makes data analysis more difficult. Dimension reduction methods are used frequently to reduce the data into m dimensions instead of working in a p dimension environment ($p > m$). Since PCA is a method developed for determining linear relationships, it does not consider non-linear relationships. Therefore it remains insufficient in determining representative variables in the data set in case of non-linear relationships.

Albanis and Batchelor (1999) have demonstrated that the PCA-NN and NLPCA-NN are superior to PCA in dimension reduction using the long term credit continuity data set. Dong and MacAvoy (1996) have shown that NLPCA-NN is better than PCA in image compression. Monahan (2000) has used a data set related with climate and compared the dimension reduction performances of PCA and NLPCA-NN where he has shown that NLPCA-NN is a more superior method in doing this. Hsieh (2001) compared PCA, rotated PCA, and NLPCA-NN using surface temperature data set of the Pacific Sea where he demonstrated a better performance of NLPCA-NN. In our study using the patient satisfaction data set, the highest variance explanation rates for variables determined for doctor, nurse, radiology technician, laboratory technician, and other staff was with the PCA-NN and NLPCA-NN methods. Ordinal scaled GPCA method showed a slightly higher performance than the proportionally scaled GPCA and PCA methods.

Table 2
Percentage of variance explained after dimensionality reduction

Principal component	Variable	Percentage of variance explained (%)				
		PCA	GPCA*	GPCA+	PCA-NN	NLPCA-NN
Doctor	15	62.2	60.2	63.4	84.9	84.9
Nurse	4	82.0	80.5	82.7	88.4	88.7
Radiology technician	4	79.4	79.2	84.0	91.7	96.1
Laboratory technician	4	82.1	82.1	83.8	90.9	90.9
Other staff	4	67.6	66.5	69.8	85.9	86.1

* Variables were determined as continuous.
+ Variables were determined as ordinal.

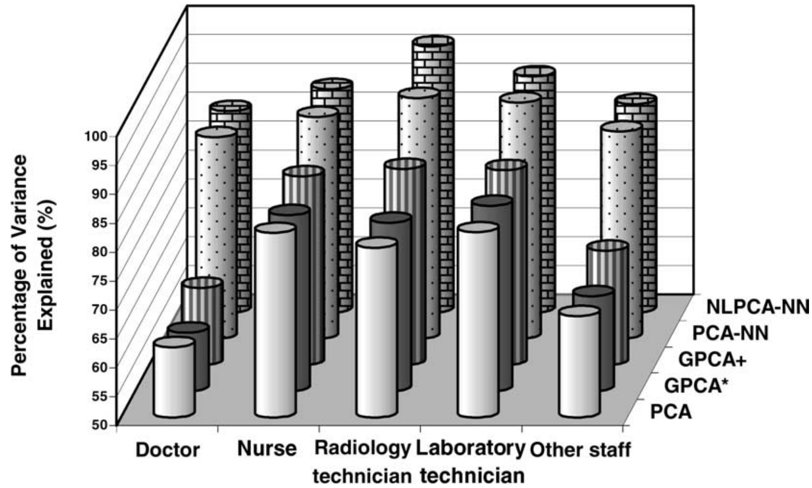

Fig. 1. Percentages of variance explained according to principal component of methods.

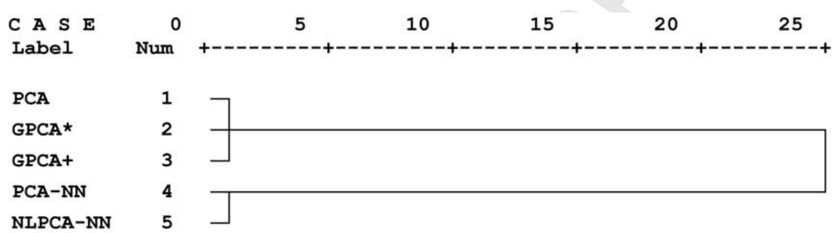

Fig. 2. Dendrogram showing relationship among dimension reduction methods.

It can be concluded from this study that instead of working with variables of lower explanatory rates, methods containing NN should be implemented in dimension reduction due to its advantage of containing alternative methods in contrast to many classical methods. It should be kept in mind that NN has the ability to determine the best variables due to its advantage of considering the non-linear relationships along with the linear relationships.

References

Albanis, G. T., & Batchelor, R. A. (1999). Assessing the long-term credit standing using dimensionality reduction techniques based on neural networks—an alternative to overfitting. In *The proceedings of the SCI 99/ISAS 99 conference, Orlando, US*.

A project funded by the Tsunami Initiative (1999). Investigation of the risk from climate variability and change over Northern Europe. Available from http://www.nerc-bas.ac.uk/public/tsunami/.

Daszykowski, M., Walczak, B., & Massart, D. L. (2003). A journey into low-dimensional spaces with autoassociative neural networks. *Talanta, 59*, 1095–1105.

Diamantras, K. I., & Kung, S. Y. (1996). *Principal component neural networks*. New York: John Wiley and Sons.

Dong, D., & MacAvoy, T. J. (1996). Batch tracking via nonlinear principal component analysis. *AIChE Journal, 42*(8), 2199–2208.

Fotheringhame, D., & Baddeley, R. (1997). Nonlinear principal components analysis of neuronal spike train data. *Biological Cybernetics, 77*(4), 283–288.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*. Cambridge, MA: MIT Press.

Hsieh, W. W. (2001). Nonlinear principal component analysis by neural networks. *Tellus, 53A*, 599–615.

Jolliffee, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Krammer, A. M. (1991). Non-linear principal component analysis using autoassociative neural networks. *AIChE Journal, 37*(21), 223–243.

Manly, B. F., & Bryan, F. J. (1986). *Multivariate statistical methods: A primer*. London: Chapman and Hall.

Michailidis, G., & De Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science, 13*(4), 307–336.

Michailidis, G., & De Leeuw, J. (2000). Multilevel homogeneity analysis with differential weighting. *Computational Statistics & Data Analysis, 32*(3–4), 411–442.

Monahan, A. H. (2000). Nonlinear principal component analysis by neural networks: Theory and applications to the Lorenz system. *Journal of Climate, 13*, 821–835.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks, 5*, 927–935.

Ozdamar, K. (2004). *Paket programlar ile istatistiksel veri analizi-2*. Eskişehir: Kaan Kitabevi.

Principe, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). *Neural and adaptive systems: Fundamentals through systems*. New York: John Wiley & Sons.

SPSS Inc. (1999). *SPSS Categories 10.0*. Chicago: SPSS Inc.

Tatlıdil, H. (1996). *Uygulamalı çok değişkenli istatistiksel analiz*. Ankara: Akademi Matbaası.

Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications, 29*(3), 583–588.