STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **XLVI**, Number 2, 2001

PRINCIPAL COMPONENTS ANALYSIS BASED ON A FUZZY SETS APPROACH

HORIA F. POP

Abstract. As with any other multivariate statistical method, Principal Components Analysis is sensitive to outliers, missing data, and poor linear correlation between variables due to poorly distributed variables. As a result data transformations have a large impact upon PCA. This paper introduces a powerful approach to improve PCA: robust fuzzy PCA algorithm (FPCA). The matrix data is fuzzified, thus diminishing the influence of the outliers.

1. Introduction

Several statistical methods for the analysis of large quantities of data have been applied to scientific problems during the last decades. One of these methods, principal component analysis (PCA) showed special promise for furnishing new and unique insights into the data interactions.

PCA is designed to reduce the number of variables that need to be considered to a small number of indices (axes) called the principal components, that are linear combinations of the original variables. The new axes lie along the directions of maximum variance such that containing most of the information. PCA provides an objective way of finding indices of this type so that the variation in the data can be accounted for as concisely as possible. concisely= brève

In the case of an n-dimensional problem, often the number of components needed to describe, say 90% of the sample variance is less than n, so that PCA essentially affords one a technique whereby the dimensionality of the variable space can be reduced, i.e., it is a dimension reduction method. It may well turn out that usually two or three principal components provide a good summary of all the original variables. Moreover, PCA offers a second important tool for multidimensional analysis that derives, in fact, from its original application in the social sciences and from which it took its name. In other words, PCA can also reveal those underlying factors or combinations of the original variables that principally

2000 *Mathematics Subject Classification.* 68N30.

1998 *CR Categories and Descriptors.* I.5.1 [**Computing Methodologies**] : Pattern Recognition – *Models* – *Fuzzy set*; G.3 [**Mathematics of Computing**] : Probability and Statistics – *Data analysis* .

determine the structure of the data distribution and that not infrequently are related to some real influencing factors in the sample population. An important issue in PCA is the interpretation of components, to help determine after the reduction of the observation space, which initial variables has the greatest shares in the variance of particular principal components. This information can be obtained using coefficients of determination (loadings) established between the components and the initial variables.

2. Principal Components Analysis (PCA)

PCA is based on eigenanalysis of the covariance or correlation matrix. Let us consider a data set $X = \{x^1, \ldots, x^p\}$, and its covariance matrix M:

$$M_{ij} = \frac{1}{p-1} \sum_{k=1}^{p} (x_i^k - \bar{x}_i) \cdot (x_j^k - \bar{x}_j), \; i, j = 1, \ldots, n. \tag{1}$$

Let us also consider the ortonormal eigenvectors e^i of the matrix M, and the corresponding eigenvalues λ_i $(i = 1, \ldots, n)$.

The principal components of the data set X appear as linear combinations of the original variables in the form

$$\mathbf{PC_i} = e_1^i y^1 + e_2^i y^2 + \cdots + e_n^i y^n, \tag{2}$$

where y^i represents the i-th original variable $(y_j^i = x_i^j)$, and e_j^i represent the j-th element of the eigenvector e^i of the matrix M.

A constraint that $(e_1^i)^2 + (e_2^i)^2 + \cdots + (e_n^i)^2 = 1$ is imposed on all components. The constraint is introduced in order to ensure that $Var(\mathbf{PC}_i)$ cannot be increased by simply increasing any of the e_j^i values.

From the orthonormality of $e^1, e^2, \ldots, e^n$ it follows that

$$\begin{array}{ll} e^{iT} \cdot e^i = 1, & \text{for any } i \in \{1, \ldots, n\} \\ e^{iT} \cdot e^j = 0, & \text{for any } i, j \in \{1, \ldots, n\}, i \neq j \\ e^{iT} \cdot M \cdot e^i = 1, & \text{for any } i \in \{1, \ldots, n\} \\ e^{iT} \cdot M \cdot e^j = 0, & \text{for any } i, j \in \{1, \ldots, n\}, i \neq j, \end{array} \tag{3}$$

and

$$M = \lambda_1 e^1 e^{1T} + \lambda_2 e^2 e^{2T} + \cdots + \lambda_n e^n e^{nT}. \tag{4}$$

where ^{T} denotes the transposing operation.

The basic property of the new variables is their lack of correlation. We have that

$$\mathrm{Var}(e^i X) = \lambda_i, \; \text{for } i = 1, \ldots, n \tag{5}$$

and

$$\mathrm{Cov}(e^i X, e^j X) = 0, \; \text{for } i, j = 1, \ldots, n, i \neq j. \tag{6}$$

The first principal component $\mathbf{PC}_1$ is that linear combination of sample values for which the "scores" have maximum variation. The second component $\mathbf{PC}_2$ has scores that are uncorrelated with the scores for $\mathbf{PC}_1$. Among the many linear combinations with this property we select the one which has maximum variation among its scores. The third component $\mathbf{PC}_3$ is defined to be that linear combination which has the maximum variation among all those combinations whose scores are uncorrelated with the scores of the first two components. Subsequent components are defined analogously.

Principal component analysis as any other multivariate statistical methods are sensitive to outliers, missing data, and poor linear correlation between variables, due to poorly distributed variables. As a result, data transformations have a large impact upon PCA [3].

One of the most illuminating approach to robustify PCA appears to be the fuzzification of the matrix data by diminuishing in this way the influence of the outliers.

3. Fuzzy Principal Components Analysis (Fuzzy PCA)

Fuzzy clustering is an important tool to identify the structure in data. In general, a fuzzy clustering algorithm with objective function can be formulated as follows: let $X = \{x^1, \dots, x^n\} \subset \mathbb{R}^p$ be a finite set of feature vectors, where n is the number of objects (measurements) and p is the number of original variables, $x_k^j = [x_1^j, x_2^j, \dots, x_p^j]^T$ and $L = (L^1, L^2, \dots, L^s)$ be a s-tuple of prototypes (supports) each of which characterizes one of the s clusters; a partition of X into s fuzzy clusters will be performed by minimizing the objective function [2]:

$$J(P, L) = \sum_{i=1}^{s} \sum_{j=1}^{n} (A_i(x^j))^m d^2(x^j, L^i),$$

where $P = \{A_1, \dots, A_s\}$ is the fuzzy partition, $A_i(x^j) \in [0, 1]$ represents the membership degree of feature point x^j to fuzzy cluster A_i, $m > 1$ is the fuzziness index, and $d(x^j, L^i)$ is the distance from the feature point x^j to the prototype of the cluster A_i. If L_i are defined as points in the $\mathbb{R}^p$ Euclidean space, the distance d may be defined as the Euclidean distance.

According to the choice of prototypes and the definition of the distance measure, different fuzzy clustering algorithms are obtained. If the prototype of a cluster is a point — the cluster center — it will produce spherical clusters; if the prototype is a line it will produce tubular clusters, and so on. Also, elements with a high degree of membership in the i-th cluster (i.e. close to the cluster's center) will contribute significantly to this weighted average, while elements with a low degree of membership (far from the center) will contribute almost nothing.

Due to the problem at hand, we will consider that the fuzzy set is characterized by a linear prototype, denoted $L(u, v)$, where v is the center of the class and u, with

$\|u\| = 1$, is the main direction. This line is also called *the first principal component* of the set, and its direction is given by the unit eigenvector u associated with the largest eigenvalue λ_{max} of, for example, the covariance matrix $C = (C_{ij})$, formed by the elements

$$C_{ij} = \frac{\displaystyle\sum_{k=1}^{p} A(x^k)^m \cdot (x_i^k - \bar{x}_i) \cdot (x_j^k - \bar{x}_j)}{\displaystyle\sum_{k=1}^{p} A(x^k)^m}, \; i, j = 1, \ldots, n. \tag{7}$$

where $\bar{x}_i$ is the arithmetic mean of the i-th variable, $m > 1$ is the fuzziness index. The settings above mean that the fuzzy set A is characterized by the linear prototype $\mathbf{PC}_1$ produced considering the fuzzy covariance matrix C.

We wish to determine the particular membership degrees $A(x)$ such that the first principal component is best fitted along the items of the data set X. The algorithm proposed in this paper is a natural extension of the Fuzzy 1-Lines Algorithm [5].

Let us denote by α the membership degree corresponding to the farthest outlier. For the moment we consider that α is a value preset by the user. The membership degrees $A(x)$ will be produced using the following mechanism:

Algorithm Determine_Fuzzy_Memberships(α):

(1) Initialize $A(x) = 1$, for all $x \in X$;

(2) Determine the linear prototype $L(u, v)$: u is the eigenvector corresponding to the largest eigenvalue of the matrix C computed as in (7); v is the weighting center of the fuzzy cluster A, weighted by the m-th power of the membership degrees:

$$v = \frac{\displaystyle\sum_{j=1}^{n} A(x^j)^m \cdot x^j}{\displaystyle\sum_{j=1}^{n} A(x^j)^m};$$

(3) Determine the new fuzzy membership degrees $A(x^j)$:

$$A(x^j) = \frac{\dfrac{\alpha}{1-\alpha}}{\dfrac{\alpha}{1-\alpha} + \left(d^2(x^j, L)\right)^{\frac{1}{m-1}}};$$

(4) if the new fuzzy set is close enough to the old one, then Stop and return the new fuzzy set; else go to Step 2.

The algorithm suggested above depends on the input variable α. As opposed to the general case, we now do have a way to determine the best value for α. Of course, we are interested to find fuzzy membership degrees that contribute to producing a better fitted first principal component along the data set. But, since the eigenvalue associated to a principal component describes the scatter of data along that component, we are also interested in producing a first principal component characterised by an eigenvalue that is as large as possible. As a consequence, we will prefer that particular value of α that maximizes the eigenvalue associated to the first principal component.

Because of the fact that we are interested in real-world applications of this algorithm, an exact value of α is not required. Instead, we will simply work through a loop between 0 and 1, with a step to be chosen by the user, and select the value of α that maximizes our criterion. The produced algorithm follows:

```
Algorithm Determine_Best_Alpha():
(1) Initialize step as appropriate; initialize $\alpha_0 = 0$ and $\lambda_0 = 0$;
(2) Set $\alpha = step$, the first value to be considered;
(3) Call Determine_Fuzzy_Memberships($\alpha$) with the current value of $\alpha$,
    and determine the optimal fuzzy membership degrees $A(x)$;
(4) Using the fuzzy membership degrees determined above, compute the
    matrix $C$ as in (7), and compute the eigenvalue $\lambda$ corresponding to its
    largest eigenvector (i. e. the first principal component);
(5) If $\lambda > \lambda_0$ then set $\lambda_0 = \lambda$ and $\alpha_0 = \alpha$;
(6) Increment $\alpha$ by step; if $\alpha < 1$ then resume from Step 3; else stop, and
    return $\alpha_0$ as the optimal value for $\alpha$.
```

Now we have all the prerequisites for writing the algorithm. We will call this algorithm **Fuzzy (first component) Principal Component Analysis (FPCA)**:

```
Algorithm FPCA():
(1) Determine the optimal value of $\alpha$ by calling Determine_Best_Alpha();
(2) Call Determine_Fuzzy_Membership($\alpha$) with the value of $\alpha$ computed
    above, and determine the optimal value of the fuzzy membership degrees;
(3) Using the fuzzy membership degrees determined above, compute the
    matrix $C$ as in (7), and compute its eigenvalues and eigenvectors; these
    are the fuzzy principal components and the corresponding scatter values.
```

4. Experiments

We have selected for our experiments the set of 48 Roman pottery sherds presented in [1] and analysed in [4].

4.1. PCA on Roman pottery data.

The principal components produced using Classical PCA on Roman pottery data are depicted in Table 1, together with their associated eigenvalues.

Based on these values, we build reduction coefficients corresponding to different dimensionality reduction criteria. These reduction coefficients show the amount of original information explained by keeping only a limited number of variables or principal components, and are depicted in Table 2.

Eigenvalue	Eigenvector						
3.04969	-0.288946	-0.523259	0.352801	0.32056	-0.410301	-0.466635	0.171429
2.13202	-0.250928	-0.0121934	0.450562	-0.49078	0.306824	-0.301171	-0.555131
0.906438	0.787617	-0.104202	0.285086	-0.320179	0.0265746	-0.278653	0.326587
0.530287	-0.206816	-0.336004	0.218458	0.0857632	0.715959	0.266077	0.453712
0.193831	0.0999946	0.550648	0.651976	0.491662	0.00155423	0.136672	-0.0360862
0.135622	0.421228	-0.496785	-0.0255271	0.402459	0.144139	0.209852	-0.590197
0.0521156	-0.0547764	-0.228659	0.343225	-0.377891	-0.451031	0.693099	0.0171312

Table 1. Loadings of the principal components and their associated eigenvalues, for the classical PCA

Variables	Eigenvalue	Successive difference	Proportion	Cummulative proportion
1	3.04969	0.917665	0.435669	0.435669
2	2.13202	1.22558	0.304574	0.740244
3	0.906438	0.376152	0.129491	0.869735
4	0.530287	0.336456	0.0757552	0.94549
5	0.193831	0.0582092	0.0276901	0.97318
6	0.135622	0.083506	0.0193745	0.992555
7	0.0521156	0.0521156	0.00744509	1

Table 2. Reduction coefficients for the classical PCA

4.2. **Fuzzy PCA on Roman pottery data.** The principal components produced using Fuzzy PCA on Roman pottery data are depicted in Table 3, together with their associated eigenvalues. The optimal value of the α index has been determined to be 0.01.

Based on these values, we build reduction coefficients corresponding to different dimensionality reduction criteria. These reduction coefficients show the amount of original information explained by keeping only a limited number of variables or principal components, and are depicted in Table 4. The scores of the first two principal components are displayed in Figure 1.

By comparing Tables 1 and 3, we remark a larger value for the first eigenvalue as computed in the case of the Fuzzy PCA method. This shows an ability of the Fuzzy PCA method to get a better fit for the first principal direction among the data set.

A similar conclusion may be drawn by an analysis of Tables 2 and 4, with respect to the different reduction coefficients. For example, the cummulative proportion is 0.435669, 0.740244 and 0.869735 (for the first, the first two, and the first three variables, respectively) in the case of classical PCA, and 0.952995, 0.967357 and

Eigenvalue	Eigenvector						
5.10731	0.0282576	0.524877	-0.458777	-0.279638	0.299205	0.4783	-0.341669
0.076967	0.858295	0.0479117	0.181638	0.198671	0.0109479	0.128092	0.414782
0.0561507	-0.189394	-0.166621	0.0940642	0.121235	0.885688	0.0541047	0.354191
0.0439734	-0.396615	0.5612	0.0312381	0.267516	-0.267034	0.183698	0.591742
0.0311047	0.25096	0.592121	0.594966	-0.0673231	0.206804	-0.390974	-0.17963
0.0286412	-0.0314195	0.079254	-0.183425	0.877202	0.0802429	-0.0962796	-0.417007
0.0150737	0.0734132	-0.150373	0.599232	0.148506	-0.0734878	0.745667	-0.171598

TABLE 3. Loadings of the principal components and their associated eigenvalues, for FPCA

Variables	Eigenvalue	Successive difference	Proportion	Cummulative proportion
1	5.10731	5.03034	0.952995	0.952995
2	0.076967	0.0208163	0.0143616	0.967357
3	0.0561507	0.0121773	0.0104774	0.977834
4	0.0439734	0.0128687	0.00820519	0.986039
5	0.0311047	0.00246344	0.00580395	0.991843
6	0.0286412	0.0135676	0.00534429	0.997187
7	0.0150737	0.0150737	0.00281266	1

TABLE 4. Reduction coefficients for FPCA

0.977834 in the case of fuzzy PCA. This shows a better capability to concentrate the more information in less principal components, for the case of fuzzy PCA.

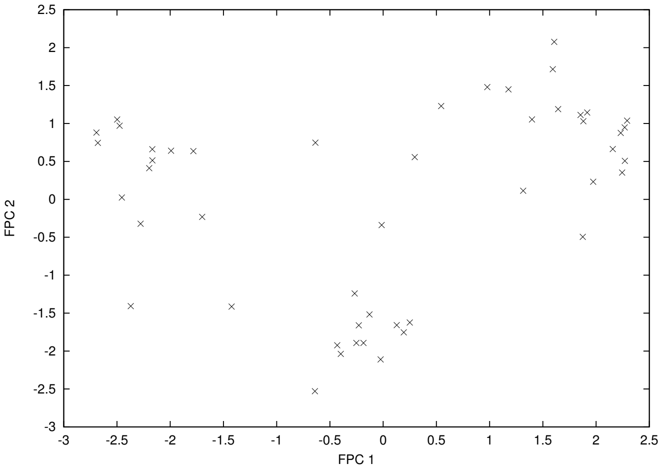

FIGURE 1. Scores of the first two principal components of Roman pottery data using FPCA

5. Conclusions

A fuzzy principal component analysis (FPCA) method for robust estimation of principal components has been described in this paper. The efficiency of the new algorithm was illustrated on a data set of 48 Roman pottery sherds. The FPCA method achieved better results mainly because it is more compressive than classical PCA. For the case of a two component model, FPCA accounts for 96.74% of the total variance, and PCA accounts only for 74.02%. Since much more classical principal components would be needed to account for the same total variance as two fuzzy principal components, the fuzzy PCA becomes a much more desirable data analysis tool.

This, together with a sharper data separation, encourages the further research on fuzzy principal components analysis, as well as the fuzzification of other important data analysis techniques.

References

[1] Aruga, R., Mirti, P., Casoli, A., Application of Multivariate Chemometric Techniques to the Study of Roman Pottery (Terra Sigillata), *Anal. Chim. Acta 276* (1993), 197–205.

[2] Dumitrescu, D., Sârbu, C., Pop, H. F., A Fuzzy Divisive Hierarchical Clustering Algorithm for the Optimal Choice of Sets of Solvent Systems, *Anal. Lett. 24* (1994), 1031–1054.

[3] Hubert, M., Rousseeuw, P. J., Verboven, S. A, Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemom. Intell. Lab. Syst. 60* (2002), 101–111.

[4] Pop, H. F., Dumitrescu, D., Sârbu, C., A Study of Roman Pottery (terra sigillata) Using Hierarchical Fuzzy Clustering, *Anal. Chim. Acta 310* (1995), 269–279.

[5] Pop, H. F., Sârbu, C., A New Fuzzy Regression Algorithm, *Anal. Chem. 68* (1996), 771–778.

Department of Computer Science, Babeş-Bolyai University, 1 M. Kogălniceanu St., RO-3400 Cluj-Napoca, Romania

E-mail address: hfpop@cs.ubbcluj.ro