Biometrical Letters
Vol. 44 (2007), No. 1, 33-49

A comparison between Principal Component Analysis and Nonlinear Principal Component Analysis

Paulo Canas Rodrigues

Department of Mathematics, Faculty of Science and Technology, Nova University of Lisbon, Monte da Caparica, 2829-516 Caparica, Portugal, e-mail: paulocanas@fct.unl.pt

SUMMARY

When dealing with large data sets, one of the main problems is how to extract the information. Principal component analysis (PCA) is the most used technique for reducing the data set while preserving significant features. However its availability is not immediate for categorical variables. PCA is applied to variables that are at least interval scaled. So if we have categorical variables (ordinal or nominal), we should use nonlinear principal component analysis (NLPCA).

We present a comparison between PCA and NLPCA from a practical point of view. With this comparison we intend to show how the results obtained from the inappropriate use of PCA (in detriment of NLPCA) may be misleading when we have variables with different measurement levels.

The analysis of two real data sets, one concerning characteristics of the countries of European Union and the other about some variables measured in people with heart failure, are presented and their interpretations in the context of our research will be discussed.

Key words: PCA, nonlinear PCA, CATPCA, European Union, Heart Failure.

1. Introduction

PCA is a technique of multivariate analysis used to reduce the dimension of a data set. PCA looks for a few linear combinations of the original variables that maximize the variance and may be used to summarize the data, losing as little information as possible. The linear combinations (or principal components) are also uncorrelated.

PCA assumes that all the variables are numerical and relations between them are linear. Performing a PCA to find relations that may be nonlinear, or to

variables with non-numerical measurement level, is controversial and may lead to non-significant results. Therefore, if we have ordinal or nominal variables we should use the NPCA.

NLPCA is a generalization of PCA which allows the inclusion of categorical variables in the analysis. To obtain this generalization, a nonlinear variable transformation, a homogeneity concept, as well as a minimization of the loss function of homogeneity using the alternating least squares algorithm are applied: Gifi (1990), Van Rijckevorsel (1983) and De Leeuw (2005).

We present a brief description of this methodology with emphasis on the logic of the process and the interest for its study in practical problems.

This work is divided in three parts. The first briefly presents some properties of categorical variables, specifically its codification and quantification. In the second and third parts we present the NLPCA (or categorical principal component analysis - CATPCA) and two applications of this method. One of the applications is about the countries of the European Union, and the other concerns heart failure.

2. Categorical data and codification

Categorical data analysis is an important part of the multivariate analysis which is widely used in disciplines such as Psychology, Social Sciences, Economy, Medicine and others.

Before applying a mathematical treatment to categorical/qualitative variables, we should quantify them. The quantification may be carried out through indicator matrices, which are presented in subsection 2.1. In subsection 2.2 we present a theoretical way of quantifying such variables.

2.1. The complete indicator matrix and its properties

We represent by n the number of objects and by m the number of variables $\mathbf{h}_j, j=1,...,m$ while the number of categories for the variables will be $k_j, j=1,...,m$.

A way of coding categorical variables data is, for each variable $\mathbf{h}_j$, defining an $n \times k_j$ binary matrix $\mathbf{G}_j = [g_{(j)ir}]$ by taking:

$$g_{(j)ir} = \begin{cases} 1, \text{ if the } i\text{th object is mapped in the } r\text{th category of } \mathbf{h}_j; \\ 0, \text{ if the } i\text{th object is not mapped in the } r\text{th category of } \mathbf{h}_j; \end{cases}$$

The matrix $\mathbf{G}_j$ is called the *indicator matrix* of $\mathbf{h}_j$. Such matrices may be collected in a partitioned matrix $\mathbf{G} = (\mathbf{G}_1, ..., \mathbf{G}_j, ..., \mathbf{G}_m)$ of dimension $n \times \Sigma k_j$, also called an *indicator matrix.*

The indicator matrix $\mathbf{G}_j$ is said to be *complete* if each row of $\mathbf{G}_j$ has only one element equal to one, and zeros elsewhere, so that the sum of the elements in each row of $\mathbf{G}_j$ is one. If all $\mathbf{G}_j$ are complete, the combined matrix $\mathbf{G}$ is also said to be complete, and then $\mathbf{Gu} = m\mathbf{u}$ where $\mathbf{u}$ is a vector of unit elements, so the elements in each row of $\mathbf{G}$ will add up to m, see Gifi (1990).

The columns of the matrices $\mathbf{G}_j$ are mutually orthogonal, so $\mathbf{D}_j = \mathbf{G}'_j \mathbf{G}_j$ is a diagonal matrix where the elements of the principal diagonal correspond to the marginal frequency of each category.

We can find some examples of this type of codification in Gifi (1990).

2.2. Quantification

When $\mathbf{h}_j$ is quantified its k_j categories will be given distinct numerical values which are the components of a vector $\mathbf{y}_j, j = 1, ..., m$. The values of the quantified variables will be the components of $\mathbf{q}_j = \mathbf{G}_j \mathbf{y}_j$, $j = 1, ..., m$.

We define x as the mean vector of all $\mathbf{q}_j$, i.e.

$$\mathbf{x} = \frac{\sum_{i=1}^{m} \mathbf{q}_j}{m}. \tag{1}$$

This vector contains the quantification of the objects. Besides this, we can consider the direct quantification of categories of the j-th variable using

$$\mathbf{y}_j = \mathbf{D}_j^{-1} \mathbf{G}'_j \mathbf{x}. \tag{2}$$

Quantification may be carried out to optimize the two equations (1) and (2) (or to minimize the loss function). When this is done we obtain optimal quantifications for the objects in the categories.

3. The CATPCA procedure

The NLPCA is useful when we intend to reduce the dimensionality and find patterns of variations in variables with different measurement levels. This procedure quantifies the categorical variables while reducing the dimensionality and is also known as CATPCA, see Meulman (2004).

Let $\phi_j(\mathbf{h}_j)$ be a nonlinear transformation of the variable $\mathbf{h}_j$. In order to obtain an optimal solution we might minimize the loss function

$$\sigma(\mathbf{x},\phi) = m^{-1}\sum_{j=1}^{m} SSQ(\mathbf{x}-\phi_j(\mathbf{h}_j)) \tag{3}$$

over the object scores $\mathbf{x}$ and the nonlinear transformations ϕ, where $SSQ(\mathbf{A})$ is the sum of squares of all elements of matrix $\mathbf{A}$. This minimization should be performed under the restriction $\mathbf{x}'\mathbf{x}=1$ or, alternatively under the condition $SSQ(\phi_j(\mathbf{h}_j))=1$, see Gifi (1990). The loss function quantifies the lost information arising from the substitution of the set of transformed variables, $\phi_j(h_j)$, by the variable $\mathbf{x}$. We can also say that the transformed (and original) variables are homogeneous with the loss given by the minimum of σ.

Suppose now that the matrix $\mathbf{Y}_j\,(k_j \times p)$ contains the coordinates of the scaled k_j categories in a p-dimensional space, and the matrix $\mathbf{X}\,(n \times p)$ contains the coordinates of the scaled observation units. $\mathbf{Y}_j$ and $\mathbf{X}$ are p-dimensional quantifications of the k_j categories of variable j with n observations. Then the CATPCA model can be written as

$$\mathbf{G}_j\,\mathbf{Y}_j \cong \mathbf{X},\quad j = 1,\ldots,m, \tag{4}$$

with the normalization $\mathbf{X}'\mathbf{X}=\mathbf{I}$ and $\mathbf{1}'\mathbf{X}=\mathbf{0}$. This model was obtained using the complete indicator matrix and the quantification in the previous section.

Equation (4) looks for some quantification $\mathbf{Y}_j$ and some quantification $\mathbf{X}$ in which the scores of the observations on the quantified variables correspond, as much as possible, to their own quantification. The perfect fit is defined as the coincidence of all observation points with their corresponding quantified scores on the m variables.

A solution to this problem can be given by minimizing the loss function. See, for example, Van Rijckevorsel (1983),

$$\sigma(\mathbf{X};\mathbf{Y}_j) = \sum_{j=1}^{m} tr(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j)'(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j). \tag{5}$$

This loss function is minimized using the Alternating Least Squares algorithm.

A particular case of the ϕ_j transformations in the loss function (3) relates to an application with splines. See, for example, Lavado (2004):

$$\phi_j(\mathbf{h}_j) = \sum_{i=1}^{z} \alpha_i I_i^{[k]}(\mathbf{h}_j), \tag{6}$$

where ϕ_j is a spline of degree k generated by z I-splines. The minimization of this loss function depends on: the degree of the spline, the number of nodes, the localization of the nodes, some restrictions on $\mathbf{x}$ and some restrictions on the coefficients α_i.

In the next two sections we present two applications in real data sets where we compare the results of NLPCA (or CATPCA), implemented in the SPSS software, with the results of PCA.

4. Application 1: European Union data set

This application refers to the 25 countries of the European Union, and the variables in Table 1 were considered. This data set of 20 variables was formed by a compilation of some reports presented by Eurostat and the variables were chosen without a preliminary study. The goal was only to compare PCA and CATPCA.

Since all the variables are continuous we performed a comparative analysis. Where possible, we present the advantages /disadvantages and the similarities/ differences which are more significant between the two methods.

Table 1. Variables used in the EU countries study.

Variable	Description
pop.1000	Population ($\times 10^3$)
area	Total area, in km^2 ($\times 10^3$)
g.d.p.	Gross domestic product per inhabitant in PPS. EU=100 (2003)
inf.rate	Inflation annual rate (in September of 2004)
trade	Intra-EU trade as a % of total trade (2003)
l.e.b.m.	Life expectancy at birth, years, in men (2003)
l.e.b.w.	Life expectancy at birth, years, in women (2003)
n.of.doc	Doctors per 100 000 inhabitants (2001)
inc.aids	AIDS incidence rate per million inhabitants (2003)
u.s.e.m.	% of 25-64 year-old men with at least upper secondary education (2003)
u.s.e.w.	% of 25-64 year-old women with at least upper secondary education (2003)
m.phone	Mobile phone subscriptions per 100 inhabitants (2003)
internet	Internet users per 100 inhabitants (2003)
motorway	Motorway density in km/1000 km^2 (2001)
railway	Railway density in km/1000 km^2 (2001)
r.energy	Renewable energy (electricity), % of total (2002)
agricult	% of employment in agriculture (2003)

industry	% of employment in industry (2003)
services	% of employment in services (2003)
civic.pa	% civic participation, at least once (2003)

4.1. Linear PCA

In Table 2 we present the component loadings of the first 6 principal components for all the 20 variables[1]. The last two lines represent the proportion of variance explained and the cumulative proportion of variance explained.

Table 2. Component loadings of the EU countries data set.

Variable	Component					
	1	2	3	4	5	6
pop.1000	0.35	-0.22	-0.41	-0.46	0.60	0.04
area	0.33	-0.22	-0.66	-0.03	0.55	0.10
g.d.p.	0.86	0.08	0.23	0.06	-0.07	0.12
inf.rate	-0.77	0.07	0.21	-0.10	0.29	-0.08
trade	-0.21	0.28	0.64	0.29	0.49	0.21
l.e.b.m.	0.86	-0.31	-0.07	0.07	-0.03	0.09
l.e.b.w.	0.80	-0.40	-0.17	0.15	0.13	0.09
n.of.doc	-0.05	-0.29	-0.36	-0.27	-0.33	0.52
inc.aids	0.03	-0.69	0.30	0.35	0.30	-0.06
u.s.e.m.	-0.22	0.89	-0.26	-0.10	0.10	0.16
u.s.e.w.	-0.29	0.85	-0.33	0.00	0.06	0.13
m.phone	0.63	-0.08	0.30	0.23	-0.20	0.58
internet	0.64	0.59	-0.17	0.29	0.07	-0.16
motorway	0.55	0.10	0.55	-0.20	0.09	-0.34
railway	0.22	0.39	0.54	-0.48	0.26	0.20
r.energy	0.09	0.20	-0.21	0.80	0.25	0.07
agricult	-0.69	-0.25	-0.17	0.28	-0.07	-0.28
industry	-0.67	-0.07	0.15	0.06	0.12	0.54
services	0.86	0.20	0.01	-0.22	-0.03	-0.17
civic.pa	0.48	0.74	-0.09	0.23	-0.14	-0.03
% of var. explained	30.82	18.58	11.79	8.86	7.23	6.62
cumulative % of var. explained	30.82	49.40	61.19	70.05	77.28	83.90

[1] The missing values were replaced by the mean of each variable. We considered this option in order to be able to continue the study without omitting variables.

We can see that the proportion of variance explained by the first principal components was almost uniform for the components after the third[2] (Table 2).

In this case we should choose more than three components to explain a significant proportion of variance. However we present the results for the first two components in Figure 1. With this representation we grouped the countries of the EU according to their proximity in the principal components.

In this case, if we chose only two principal components, the proportion of explained variance will be about 50%, and with the third we only explain an additional 12% of the original variability (Table 2). Although information is lost by choosing only two principal components, this is our best choice because we are able to establish a graphical interpretation of the data. Since this a real data set, there is often a dilemma whether to opt for one high cumulative proportion of variance explained without graphical interpretation or a possible graphical interpretation with low cumulative proportion of variance explained.

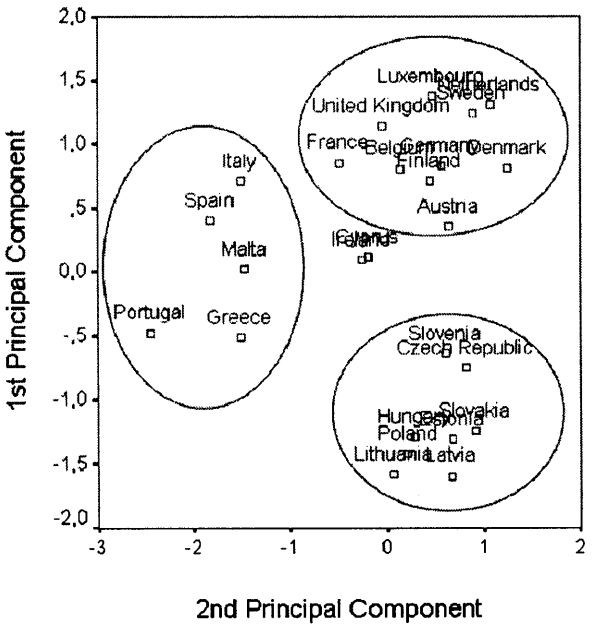

Figure 1. Plot of the first two principal components.

Figure 1 shows three groups of countries and two isolated countries. On the left side of Figure 1 are the Mediterranean countries, excluding France and

[2] The proportion of variance explained by the first two and three principal components was only 49.4% and 61.2%, respectively.

Cyprus. The inhabitants of these countries have the lower "proportion of 25-64 years old at least upper secondary education", a low "civic participation", "internet users per 100 inhabitants" below the average and high "AIDS rate". These conclusions were obtained from Table 2.

In the upper right corner of Figure 1 are the countries from northern Europe and the most developed EU countries. In the third group we have the latest countries to enter the EU, with the exception of Malta and Cyprus.

Finally, at the centre of the plot we have Ireland and Cyprus. Considering the latest trends, Ireland seems to be approaching the group of more developed countries. Since Cyprus has many missing values, and they were replaced by the mean of each variable, we consider that it is misplaced in Figure 1.

In the current study we tried to group variables so as to reduce the number of original variables and hence obtain higher proportion of explained variance. This was done by grouping similar variables (e.g. "% of 25-64 year-olds with at least upper secondary education" for men and women or "% of Agriculture employment", "% of Industry employment" and "% of Services employment") using PCA in order to obtain one first principal component that explains at least 95% of the variance of the referred group. Since this was not achieved, we do not present these results.

4.2. Nonlinear PCA (CATPCA)

We now present the results for the procedure CATPCA, as output by the SPSS software.

Since the data was constituted by continuous variables, we used the "Multiplying" discretization method, see Meulman (2004). The missing values were excluded from the correlations matrix calculus. This solution was chosen for comparison with the previous results because the other options tack off the variables or the objects with missing values. The chosen measurement level for all variables in the final solution was "spline ordinal" with degree two and two nodes[3].

As referred to in the previous subsection, the choice of the number of dimensions is controversial. As we did for PCA, we present the results of CATPCA with two dimensions/components. This enables us to make a comparison with the results from PCA and to achieve a graphical interpretation, which is useful in this kind of analysis.

[3] The same analysis was carried out for the measurement level "numerical" and the results were the same.

The proportions of variance explained by the first and second dimensions were 35.16% and 21.76%, respectively, where the cumulative proportion of variance explained is about 57%. This proportion is rather small when compared with the usual literature. Nevertheless, we will continue this study considering these results, not forgetting that we are working with a real data set.

Figures 2 and 3 show the countries' position in the two first dimensions and the variables' component loadings.

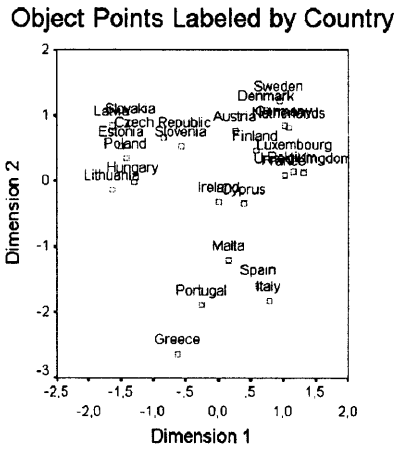

Figure 2. Position of the 25 EU countries in the first two dimensions.

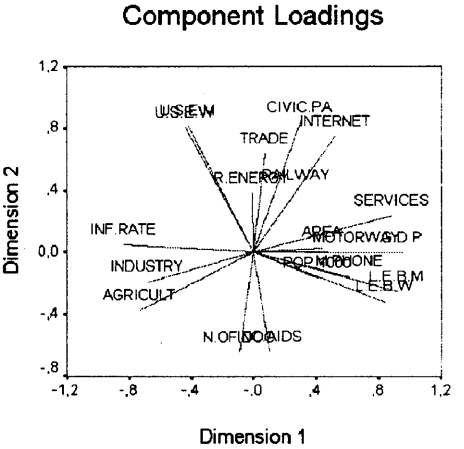

Figure 3. Component loadings for the first two dimensions of the EU data.

Figure 2 and Figure 1 are comparable, so the interpretation is identical to that presented in subsection 4.1. For analyzing Figure 3 we must have in mind that variables with a large component loading are more significant for the chosen model than variables with a short component loading. The direction (horizontal or vertical) of the component loadings reflects which variables have more importance in the associated dimension/component. Furthermore, relations between variables may be established by analyzing the angles made by them. Angles close to 0 or to 180 translate into a high direct or inverse correlation between variables, respectively. If the angle is close to 90, the variables are almost independent.

In this case, a two-dimensional model, having a short component loading for one variable, may correspond to a higher significance variable in a model with a different number of dimensions.

The variables "life expectancy at birth" and "% of 25-64 year-olds with at least upper secondary education" in men and women were highly related, as the

lines almost overlapped. The variables "total area" and "population" were also correlated, but with a shorter component loading, which means they had little importance for this model. The "gross domestic product" had a direct correlation with "life expectancy at birth" and "% of services employment" (acute angles), but with an inverse correlation with "inflation annual rate", "% of agriculture employment" and "% of industry employment" (obtuse angles).

For more accurate and less misleading interpretations, the correlation matrix should be taken into consideration. For example, in Figure 3, the variables "doctors per 100 000 inhabitants" and "AIDS incidence rate" seemed correlated. However, in the correlation matrix we could see that the correlation between them was only 0.158.

We stress that with the first dimension it is possible to distinguish between countries less developed within the EU (with a higher proportion of inhabitants employed in agriculture or industry, and with a high annual inflation rate) with countries with more developed features (higher proportion of employment in services, high gross domestic product per inhabitant and high life expectancy at birth). The second dimension makes it possible to define countries by educational/intellectual level. The most explained variables were "% of 25-64 year-olds with at least upper secondary education" for men and women, "% of civic participation" and "internet users per 100 inhabitants". So we have the most developed countries in EU on the right of Figure 2 and the most "intellectual" countries at the top.

With a three-dimensional solution we had the proportions 33.82%, 21.60% and 17.24% for the variance explained, resulting in a cumulative proportion of variance explained of 72.66%. The third dimension/component represented the size of the countries including the variables "area" and "population".

4.3. PCA *vs.* CATPCA

Comparing PCA and CATPCA (spline ordinal measurement level) procedures for this data set, with only continuous variables, similarities can be found. This is also similar to the results obtained using CATPCA with numerical measurement level for all variables. Generalizing, when we have a data set with only continuous variables it does not matter whether we use PCA or CATPCA (with numerical or spline ordinal measurement level).

5. Application 2: Heart Failure data set

Heart Failure (HF) is a serious public health problem in developed countries. It affects 4.3% of the Portuguese population over the age of 25 and is one of the main reasons for hospitalization, particularly in the case of patients over 65.

HF is a progressive disorder in which heart damage causes weakening of the cardiovascular system, and it is clinically manifested by fluid congestion or inadequate blood flow to tissues. HF may be the result of one or many causes. In the following subsections, we describe the population and the variables used in the application of PCA and CATPCA, trying to explain these causes. A comparative analysis will be carried out.

5.1. Population and variables

In this study, we considered a random sample of the population of patients with HF who used a Portuguese Central Hospital in the year 2001. This study included the period of internment and a follow-up of five years maximum. The aim of the study was to evaluate the influence of diabetes in sick people with cardiovascular illness. We considered cardiovascular illness, the presence of HF and/or acute coronary syndrome (ASC) in the users.

The group of patients with HF comprised 233 users, 98 men and 135 women, with a mean age of 68.1 (the 95% confidence interval for the mean age is [66.41; 69.78]), being recruited consecutively on reception into in-patient care at the Internal Medicine Department and at the Heart Failure Out-patient Clinic (medical infirmaries and coronary disease units).

The 17 variables in this study are presented in the first column of Table 3. The variables "type of ACS" and "number of diseased vasa" are categorical variables with four and five categories, respectively; the variable "age" and "time of the first internment" are numerical. All the others are binary variables.

5.2. Linear PCA

Table 3 presents the component loadings for the first 6 principal components of the 17 variables. The last two lines of the table show the proportion of variance explained by each variable and the cumulative proportion of variance explained.

Table 3. Component loadings and % of variance explained by each one of the first six components.

Variable	Component					
	1	2	3	4	5	6
Age	-0.55	-0.18	0.44	-0.11	-0.03	0.17
Sex	-0.30	-0.14	**0.59**	-0.08	0.28	0.13
Diabetes	0.21	**0.81**	0.35	-0.24	-0.13	-0.06
Acute coronary syndrome	**-0.69**	0.31	-0.42	-0.09	0.12	0.06
Type of ACS	**0.60**	-0.36	0.43	0.06	-0.06	-0.07
Heart failure	**0.67**	-0.08	0.23	0.17	-0.03	0.08
Event death	0.57	0.16	0.13	-0.20	0.07	-0.10
Event re-internment	-0.12	0.17	0.27	0.71	-0.16	-0.16
Time of internment	-0.16	-0.04	0.19	0.13	0.55	-0.62
Arterial hypertension	-0.16	**0.55**	-0.25	0.28	0.33	-0.23
Dyslipidemia	-0.42	0.30	0.16	0.46	-0.11	0.33
Auricular fibrillation	0.44	-0.19	-0.09	0.39	-0.14	-0.11
Chronic renal insufficiency	0.44	0.16	0.11	0.13	0.58	0.31
Anaemia	0.42	0.33	-0.25	0.20	0.13	0.49
Tobaccoism	-0.55	-0.09	**0.47**	0.01	0.19	0.23
Glycaemia	0.24	**0.81**	0.34	-0.22	-0.14	-0.09
Number of diseased vasa	0.36	-0.16	-0.19	-0.13	0.43	0.17
% of var. explained	19.67	13.21	10.27	7.31	6.94	6.41
cumulative % of var. explained	19.67	32.87	43.14	50.45	57.39	63.80

Table 3 shows a well-distributed proportion of variance explained along the first 6 components. Since the first two or three components do not have a significant proportion of variance explained, we should retain more than three components. However, with this option, we lose the advantage of the graphical representation.

Analyzing the component loadings, the most important variables for the first dimension were "acute coronary syndrome", "type of acute coronary syndrome" and "heart failure". For the second dimension we had "diabetes", "arterial hypertension" and "glycaemia", and for the third dimension we had "sex" and "tobaccoism".

5.3. Nonlinear PCA (CATPCA)

Among the 17 variables only "age" and "time of first internment" were continuous, thus we consider them as "ordinal spline" with degree 2 and 2 interior

knots[4]. The discretization method "grouping" with seven categories and the normal distribution were used[5].

In the following sub-subsections CATPCA was used considering two components (sub-subsection 5.3.1) and eight components (sub-subsection 5.3.2). For the first case we present a graphical interpretation with low variance explained. For the second case we obtained more than 70% cumulative proportion of variance explained without graphical interpretation.

5.3.1. Results for two dimensions

If we select two dimensions/components, the obtained proportions of variance explained were 22.05% and 13.11% for the first and second, respectively. The overall variance explained was about 35.16%. In this study, we present the plots of the patients labelled by the variables "diabetes", "ACS" and "HF" in Figures 4, 5 and 6, respectively. Theses plots were chosen randomly amongst the 17 variables in order to lighten the analysis.

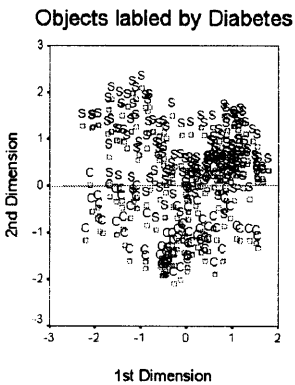

Figure 4. Objects labeled by the variable "Diabetes".

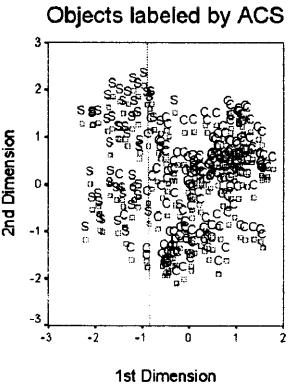

Figure 5. Objects labeled by the variable "ACS".

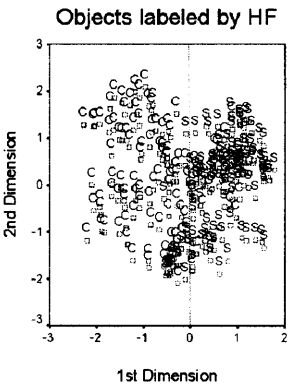

Figure 6. Objects labeled bythe variable "HF".

The second dimension of Figure 4 distinguishes between patients with "diabetes" ("C's") and without ("S's"). If we mark a horizontal line y = 0 we see that "C's" are below this line and "S's" above. In the Figures 5 and 6 we have the variables "ACS" and "HF", respectively. The distinction between the individuals

[4] These variables were also analyzed using the Numerical scale in SPSS. The results were similar to the ones presented.

[5] For more information about this procedure see SPSS Categories 13.0.

with ("C's") and without ("S's") the illness was carried out by the first dimension. The lines marked in Figures 5 and 6 were the best border.

We also obtained similar results for the remaining variables. The most important variables explaining the fist dimension were "ACS", "HF", "event death" and "auricular fibrillation", where the most important variables explaining the second dimension were "diabetes", "arterial hypertension" and "glycaemia". Adding the analysis of the centroid coordinates from SPSS, we have more three important variables for the first dimension ("age", "type of ACS" and "time of internment").

Figure 7 shows the component loadings for the variables in this study. The explanatory analysis of this graphic can be found in subsection 4.2.

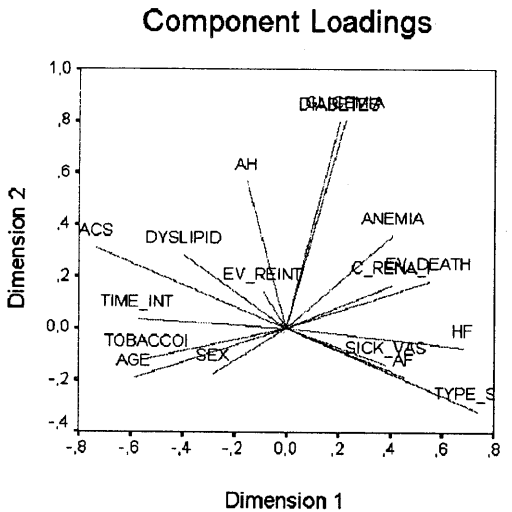

Figure 7. Component loadings for the variables of the HF study.

The strongest relation was between the variables "diabetes" and "glycaemia", because they have the closest component loadings and two of the longest lengths. This makes sense because of the similarity of the variables. The variables "ACS" and "type of ACS" are also related but with a correlation of -0.99. The sign of the correlation is not relevant since we are working with nominal variables. If we take "Glycaemia" and "Diabetes" we see that these variables form right angles with the variables "ACS", "Type of ACS" and "HF". This means that the correlation between them is close to zero.

5.3.2. Results for eight dimensions

Table 4 shows the component loadings and the proportion of variance explained for the first 8 components/dimensions obtained through the CATPCA procedure.

Table 4: Component loadings and proportion of variance explained by each one of the first eight components/dimensions through CATPCA.

Variable	Component loading							
	1	2	3	4	5	6	7	8
Age	**-0.57**	-0.11	**0.49**	-0.03	-0.01	-0.07	-0.37	-0.11
Sex	-0.29	-0.15	**0.60**	-0.22	0.15	0.36	0.24	0.20
Diabetes	0.20	**0.82**	0.33	-0.14	-0.21	-0.09	-0.08	0.20
Acute coronary syndrome	**-0.74**	0.28	**-0.47**	-0.09	-0.01	0.21	0.17	0.13
Type of ACS	**0.74**	-0.28	**0.48**	0.10	0.01	-0.20	-0.17	-0.13
Heart failure	**0.69**	-0.05	0.21	0.14	0.05	0.14	0.14	-0.13
Event death	**0.53**	0.20	0.04	-0.24	-0.12	0.25	**0.48**	-0.18
Event re-internment	-0.09	0.15	0.28	**0.65**	0.23	-0.08	0.33	-0.21
Time of internment	-0.29	0.00	0.19	-0.36	**0.48**	**-0.44**	0.29	0.01
Arterial hypertension	-0.17	**0.54**	-0.19	0.08	**0.46**	-0.29	0.10	-0.22
Dyslipidemia	-0.39	0.28	0.19	**0.52**	0.21	0.11	-0.24	0.26
Auricular fibrillation	0.47	-0.19	-0.10	0.32	0.10	-0.04	0.24	**0.64**
Chronic renal insufficiency	0.40	0.18	0.05	-0.22	**0.54**	0.38	-0.11	0.05
Anaemia	0.40	0.36	-0.33	0.17	0.17	**0.42**	-0.26	-0.22
Tobaccoism	**-0.54**	-0.10	0.42	-0.03	0.06	**0.43**	0.02	-0.08
Glycaemia	0.22	**0.83**	0.31	-0.12	-0.22	-0.11	-0.07	0.12
Number of diseased vasa	0.39	-0.17	-0.14	-0.27	0.43	-0.08	-0.33	0.18
Eigenvalue	3.62	2.22	1.80	1.26	1.19	1.13	1.02	0.85
% of var. explained	21.4	13.1	10.6	7.4	7.0	6.6	6.0	5.0
cumulative % of var. explained	21.4	34.5	45.1	52.5	59.5	66.1	72.1	77.1

If we choose eight dimensions for the final solution, we obtain about 77% cumulative proportion of variance explained. Graphical representations are too confused and an interpretation is impracticable. However, an interpretation is possible through the analysis of the component loadings from Table 4. Table 5 shows a resumé of the most significant variables to explain each dimension output in Table 4.

Table 5. Resumé of the most significant variables for each dimension.

Dimension	Most significant variables
1	Age; ACS; Type of ACS; HF; Event death; Tobaccoism
2	Diabetes; Arterial hypertension; Glycemia
3	Age; Sex; ACS; Type of ACS
4	Event re-internment; Dyslipidemia
5	Time of internment; Arterial hypertension; Chronic renal insufficiency
6	Time of internment; Anaemia; Tobaccoism
7	Event death; Event re-internment
8	Auricular Fibrillation

This analysis makes it possible to reduce the dimensionality of the original data set and find relations between the variables.

The obtained components/dimensions, besides enabling identification of an individual's profile, can also be used in other types of analysis, since the 8 components represent about 80% of all original variability.

6. Discussion: PCA *vs.* CATPCA

The most significant difference between linear PCA and nonlinear PCA, or CATPCA, is the preservation of the proportion of variance explained in PCA, while CATPCA does not preserve this proportion. That is to say, using CATPCA we have to perform the analysis every time we change the number of dimensions, because the component loadings and proportions of variance explained change with the number of dimensions. With PCA, the number of dimensions chosen can be easily reduced without recalculating the component loadings and the proportions of variance explained.

For a data set with continuous variables, the results from PCA or CATPCA, using spline ordinal or numerical measurement level, presents no difference.

For a data set of variables with different measurement levels the use of PCA is inadvisable and CATPCA should be used.

CATPCA produces a more extensive and complete output than PCA. A resumé of these differences is presented in Table 6.

The procedure CATPCA allows us to obtain better results when we have variables with different measurement levels. This method is the best one when we aim to reduce the dimensionality and to find patterns for this type of data.

Table 6. Differences between PCA and CATPCA.

Characteristic	PCA	CATPCA
Type of variables	Only continuous variables	All types of variables
Final % of variance explained	Preserved	Not Preserved
Output	Missing relations	Extensive and complete

Acknowledgements

The author wishes to thank A. T. Lima (Nova University of Lisbon) for her valuable help.

REFERENCES

De Leeuw J. (2005). Nonlinear Principal Component Analysis and Related Techniques, in Department of Statistics Papers, (University of California, Los Angeles).

Gifi A. (1990). Nonlinear Multivariate Analysis, in John Wiley & Sons.

Lavado N., Calapez T. (2005). Um enquadramento das variantes não-lineares da ACP via transformações spline, in Estatística Jubilar. Actas do XII Congresso Anual da Sociedade Portuguesa de Estatística. p 391-402.

Meulman J.J., Heiser W.J. (2004). SPSS Categories 13.0, in SPSS Inc. Chicago.

Van Rijckevorsel J., Walter J. (1993). An application of two generalizations of non-linear principal components analysis, in Elsevier Science Publishers B.V. (North-Holland).