

Principal Component Analysis, A Powerful Un-Supervised Learning Technique

George C. J. Fernandez
Department of Applied Economics and Statistics / 204
University of Nevada - Reno
Reno NV 89557

ABSTRACT

Data mining is a collection of analytical techniques to uncover new trends and patterns in massive databases. These data mining techniques stress visualization to thoroughly study the structure of data and to check the validity of the statistical model fit which leads to proactive decision making. Principal component analysis (PCA) is one of the unsupervised data mining tools used to reduce dimensionality in multivariate data. In PCA, dimensionality of multivariate data is reduced by transforming the correlated variables into linearly transformed uncorrelated variables. PCA summarizes the variation in a correlated multi-attribute to a set of uncorrelated components, each of which is a particular linear combination of the original variables. The extracted uncorrelated components are called principal components (PC) and are estimated from the eigenvectors of the covariance or correlation matrix of the original variables. Therefore, the objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number components that account for most of the variation in the original multivariate data and to summarize the data with little loss of information. A user-friendly SAS application utilizing the latest capabilities of SAS macro to perform PCA is presented here. Previously published diabetes screening data containing multi-attributes is used to reduce multi attributes into few dimensions.

Key Words: Data mining, SAS, Data exploration, Bi-Plot, Multivariate data, SAS macro,

Introduction

Data mining is the process of selecting, exploring, and modeling large amounts of data to uncover new trends and patterns in massive databases. These analyses lead to proactive decision making by stressing data exploration to thoroughly study the structure of data and to check the validity of statistical models that fit. Data mining techniques can be broadly

classified into unsupervised and supervised learning methods. The main difference between supervised and unsupervised learning methods is the underlying model structure. In supervised learning, relationships between input and the target variables are being established. But, in unsupervised learning, no variable is defined as a target or response variable. In fact, for most types of unsupervised learning, the targets are same as the inputs. All the variables are assumed to be influenced by a few components in unsupervised learning. Because of this feature, it is possible to study large complex models with unsupervised learning than with supervised learning.

Unsupervised learning methods are used in many fields including medicine and in microarray studies [1] under a wide variety of names. Analysis of multivariate data plays a key role in data mining and knowledge discovery. Multivariate data consists of many different attributes or variables recorded for each observation. If there are p variables in a database, each variable could be regarded as constituting a different dimension, in a p -dimensional hyperspace. This multi-dimensional hyperspace is often difficult to visualize, and thus the main objectives of unsupervised learning methods are to reduce dimensionality and scoring all observations based on a composite index. Also, summarizing multivariate attributes by, two or three that can be displayed graphically with minimal loss of information is useful in knowledge discovery.

The most commonly practiced unsupervised methods are latent variable models (principal component and factor analyses). In principal component analysis (PCA), dimensionality of multivariate data is reduced by transforming the correlated variables into linearly transformed uncorrelated variables. In factor analysis, a few uncorrelated hidden factors that explain the maximum amount of common variance and are responsible for the observed correlation among the multivariate data are extracted. The relationship

between the multi-attributes and the extracted hidden factors are then investigated.

Because it is hard to visualize multi-dimensional space, principal components analysis (PCA), a popular multivariate technique, is mainly used to reduce the dimensionality of p multi-attributes to two or three dimensions. A brief account on non-mathematical description and application of PCA are discussed in this paper. For a mathematical account of principal component analysis, the readers are encouraged to refer the multivariate statistics book [2]

PCA summarizes the variation in a correlated multi-attribute to a set of uncorrelated components, each of which is a particular linear combination of the original variables. The extracted uncorrelated components are called principal components (PC) and are estimated from the eigenvectors of the covariance or correlation matrix of the original variables. Therefore, the objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number components that account for most of the variation in the original multivariate data and to summarize the data with little loss of information.

In PCA, uncorrelated PC's are extracted by linear transformations of the original variables so that the first few PC's contain most of the variations in the original dataset. These PCs are extracted in decreasing order of importance so that the first PC accounts for as much of the variation as possible and each successive component accounts for a little less. Following PCA, analyst tries to interpret the first few principal components in terms of the original variables, and thereby have a greater understanding of the data. To reproduce the total system variability of the original p variables, we need all p PCs. However, if the first few PCs account for a large proportion of the variability (80-90%), we have achieved our objective of dimension reduction. Because the first principal component accounts for the co-variation shared by all attributes, this may be a better estimate than simple or weighted averages of the original variables. Thus, PCA can be useful when there is a severe high-degree of correlation present in the multi-attributes.

In PCA, the extractions of PC can be made using either original multivariate data sets or using the covariance or the correlation matrix if the original data set is not available. In deriving PC, the correlation matrix is commonly used when different variables in the dataset are measured using different units (annual income, educational level, numbers of cars owned per family) or if different variables have different variances. Using the correlation matrix is equivalent to standardizing the variables to zero mean and unit

standard deviation. The statistical theory, methods, and the computation aspects of PCA are presented in details elsewhere [2].

While the PCA and Exploratory factor analysis (EFA) analyses are functionally very similar and are used for data reduction and summarization, they are quite different in terms of the underlying assumptions. In EFA, the variance of a single variable can be partitioned into common and unique variances. The common variance is considered shared by other variables included in the model and the unique variance that includes the error component is unique to that particular variable. Thus, EFA analyzes only the common variance of the observed variables whereas PCA summarizes the total variance and makes no distinction between common and unique variance.

The selection of PCA over the EFA is dependent upon the objective of the analysis and the assumptions about the variation in the original variables. EFA and PCA are considered similar since the objectives of both analyses are to reduce the original variables into fewer components, called factors or principal components. However, they are also different since the extracted components serve different purposes. In EFA, a small number of factors are extracted to account for the inter-correlations among the observed variables and to identify the latent factors that explain why the variables are correlated with each other. But, the objective of PCA is to account for the maximum portion of the variance present in the original variables with a minimum number of PC.

If the observed variables are measured relatively error free, or if it is assumed that the error and unique variance represent a small portion of the total variance in the original set of the variables, then PCA is appropriate. But if the observed variables are only indicators of the latent factor, or if the error variance represents a significant portion of the total variance, then the appropriate technique is EFA.

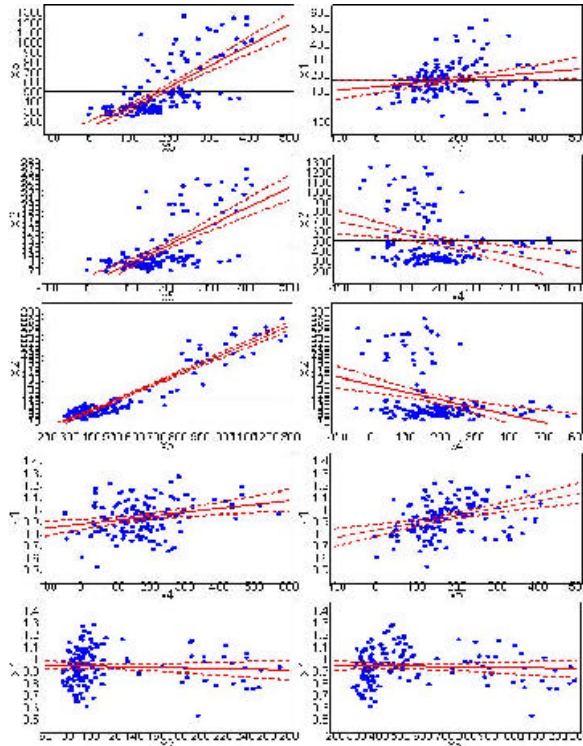
Previously published diabetes screening data [3] containing multi-attributes from glucose tolerance tests, relative weight (x1: RELWT), fasting plasma glucose levels (x2: GLUFAST), test plasma glucose levels (x3: GLUTEST), plasma insulin level during test (x4: INSTEST), and steady state plasma glucose levels (x5: SSPG) are used here to demonstrate the features of PCA. A user-friendly SAS application developed by the author utilizing the latest capabilities of SAS macros to perform PCA and EFA analyses with data exploration is presented here. Instructions are also provided in the Appendix to download the SAS

macro-call file and perform the PCA analysis reported in this paper.

Data exploration

The descriptive simple statistics of all multi-attributes generated by the SAS PROC CORR [4] are presented in Table 1. The number of observations (N) per variable is useful in checking for missing values for any given attribute and providing information on the size of the $n \times p$ coordinate data. The estimates of central tendency (mean) and the variability (standard deviation) provide information on the nature of multi-attributes that can be used to decide whether to use standardized or un-standardized data in the PCA analysis. The minimum and the maximum values describe the range of variation in each attribute and help to check for any extreme outliers. Examining the correlations among the multi-attributes on simple scatter plots between any two variables is the first step in exploring the data. An example of this simple two-dimensional scatter plot showing the correlation between any two attributes are presented Figure 1. The regression line displays significant positive or negative relationship. If the 95% confidence interval lines intersect the y-axis mean (Horizontal line) then the observed correlation is considered significant at 5% level.

Fig. 1 Scatter plot matrix illustrating the degree of linear correlation, among the five attributes



These scatter plots are useful in examining the range variation and the degree of correlations between any two attributes. The scatter plot presented in Fig.1 revealed the strong correlation existed between GLUTEST (X3) and GLUFAST.(X2). The descriptive simple statistics of all multi-attributes generated by the SAS PROC CORR are presented in Table1. The number of observations (N) per variable is useful in checking for missing values for any given attribute and providing information on the size of the $n \times p$ coordinate data. The estimates of central tendency (mean) and the variability (standard deviation) provide information on the nature of multi-attributes that can be used to decide whether to use standardized or un-standardized data in the PCA analysis. The minimum and the maximum values describe the range of variation in each attribute and help to check for any extreme outliers.

Table 1 Descriptive statistics of multi-attributes

| Variable | N | Mean | Std Dev | Simple Statistics | | |
|----------|-----|-----------|-----------|-------------------|-----------|-----------|
| | | | | Sum | Minimum | Maximum |
| X1 | 141 | 0.9588 | 0.14262 | 135.19103 | 0.551 | 1.30318 |
| X2 | 141 | 117.48094 | 50.31322 | 16565 | 75.57512 | 271.77898 |
| X3 | 141 | 533.12341 | 266.1886 | 75170 | 284.09156 | 1294 |
| X4 | 141 | 183.65149 | 111.49252 | 25895 | -43.83149 | 569.70318 |
| X5 | 141 | 179.80537 | 87.96018 | 25353 | -2.03249 | 407.46277 |

Table 2: Pearson Correlations

| | Pearson Correlation Coefficients, N = 141 | | | | |
|-----------------------------|---|----------|----------|---------|---------|
| | X1 | X2 | X3 | X4 | X5 |
| Relative wt | 1 | -0.06216 | -0.04614 | 0.25965 | 0.37687 |
| Fastina Plasma Glucose | -0.06216 | 1 | 0.5869 | 0.0019 | <.0001 |
| test plasma qlucose | 0.464 | <.0001 | 1 | <.0001 | <.0001 |
| Plasma insulin durina test | -0.04614 | 0.949 | -0.27538 | 1 | 0.69322 |
| steady state plasma qlucose | 0.5869 | <.0001 | 0.0009 | 0.0009 | 1 |
| | 0.25965 | -0.32239 | -0.27538 | 0.18196 | 0.37687 |
| | 0.0019 | <.0001 | 0.0009 | 0.18196 | 0.69322 |
| | <.0001 | <.0001 | <.0001 | 0.0308 | 0.18196 |

The degree of linear association among the variables measured by the Pearson correlation coefficient (r) and their statistical significance are presented in Table.2. The value of the r ranged from -0.04 to 0.94. The statistical significance of the r varied from no correlation (p-value: 0.58) to a highly significant correlation (p-value <0.0001). Among the 10 possible pairs of correlations, 8 pairs of correlations were highly significant indicating that this data is suitable for performing PCA analysis

Table 3 Eigenvalues in PCA analysis

Eigenvalues of the Correlation Matrix: Total = 5

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|------------|------------|------------|------------|
| 1 | 2.58361986 | 1.08022348 | 0.5167 | 0.5167 |
| 2 | 1.50339638 | 0.80498625 | 0.3007 | 0.8174 |
| 3 | 0.69841012 | 0.53195131 | 0.1397 | 0.9571 |
| 4 | 0.16645881 | 0.11834399 | 0.0333 | 0.9904 |
| 5 | 0.04811483 | | 0.0096 | 1 |

Principal Component Analysis (PCA)

In the PCA analysis [5], the dimensions of standardized multi-attributes define the number of eigenvalues. An eigenvalue greater than 1 indicates that PC accounts for more of the variance than one of the original variables in standardized data. This can be confirmed by visually examining the improved scree plot (Figure 2) of eigenvalues and the parallel analysis plot

Fig 2 : Intersection of scree plot illustrating the relationship between number of PC and the rate of decline of eigenvalue and the parallel analysis plot

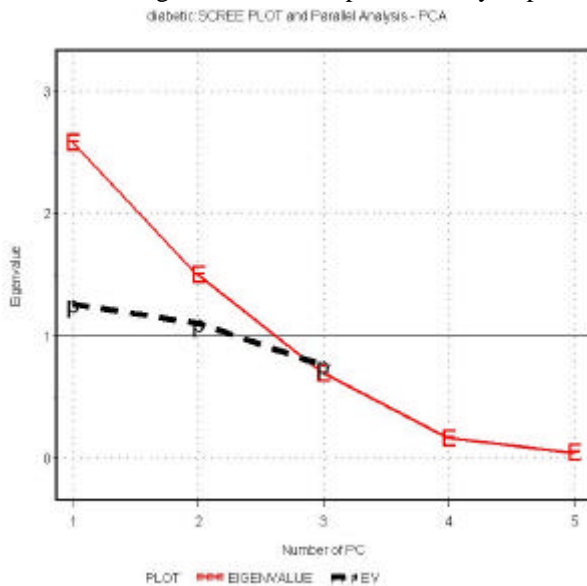


Table 4 Table of eigenvectors

| Eigenvectors | | 1 | 2 | 3 |
|--------------|-----------------------------|---------|---------|---------|
| X1 | Relative wt | 0.05206 | 0.6498 | -0.7026 |
| X2 | Fasting Plasma Glucose | 0.59668 | -0.1355 | 0.05848 |
| X3 | test plasma glucose | 0.60343 | -0.0949 | 0.09925 |
| X4 | Plasma insulin during test | -0.1603 | 0.62072 | 0.6942 |
| X5 | steady state plasma glucose | 0.50145 | 0.40633 | 0.10577 |

Table 5 PC loadings PCA analysis

| Factor Pattern | | FACTOR1 | FACTOR2 |
|----------------|-----------------------------|----------|----------|
| X1 | Relative wt | 0.08368 | 0.79674 |
| X2 | Fasting Plasma Glucose | 0.95909 | -0.16609 |
| X3 | test plasma glucose | 0.96993 | -0.11641 |
| X4 | Plasma insulin during test | -0.25758 | 0.76109 |
| X5 | steady state plasma glucose | 0.80602 | 0.49822 |

of eigenvalues. This added scree plot shows the rate of change in the magnitude of the eigenvalues for an increasing number of PC. The rate of decline levels off at a given point in the scree plot that indicates the optimum number of PC to extract. Also, the intersection point between the scree plot and the parallel analysis plot reveals that the first two eigenvalues that account for 81.7% of the total variation and could be retained as the significant PC (Table 3).

The new variables PC1 and PC2 are the linear combinations of the five standardized variables and the magnitude of the eigenvalues accounts for the variation in the new PC scores. The eigenvectors presented in Table 4, provide the weights for transforming the five standardized variables into PC. For example, the PC1 is derived by performing the following linear transformation using these eigenvectors.
 $PC1 = 0,052*x1 + 0.596*x2 + 0.603*x3 - 0.160*x4 + 0.501*x5$ The sum of the squared of eigenvectors for a given PC is equals to one.

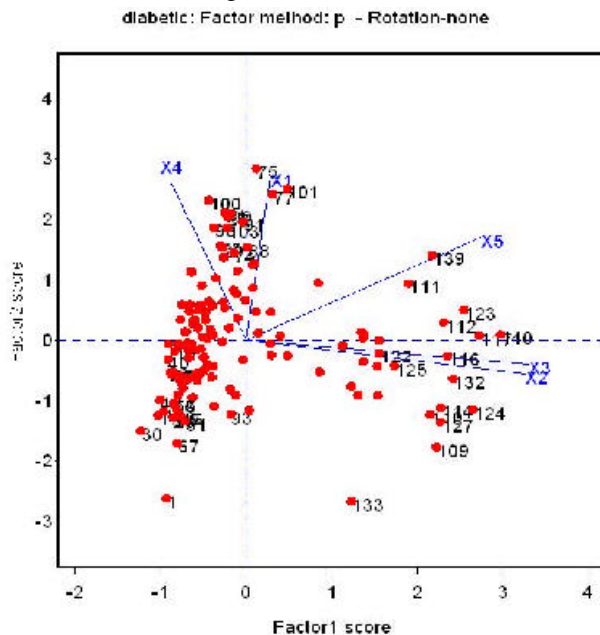
PC loadings presented in Table5 are the correlation coefficient between the first two PC scores and the original variables. They measure the importance of each variable in accounting for the variability in the PC. That is, the larger the loadings in absolute term, the more influential the variables are in forming the new PC and vice versa. A high correlation between PC1 and 'X2, X3, & X5', indicate that these variables are associated with the direction of the maximum amount of variation in this dataset. The first PC loading patterns suggest that fasting plasma glucose, test plasma glucose and steady state plasma glucose are strongly influencing high degree of diabetics..

A Bi-plot display [6] of both PC (PC1 and PC2) scores and PC loadings (Figure4) is very effective in studying the relationships within observations, between variables, and the inter-relationship between observations and the variables.

The X-Y axis of the bi-plot of PCA analysis represents the standardized PC1 and PC2 scores respectively. In order to display the relationships among the variables, the PC loading values for each PC are overlaid on the same plot after being multiplied by the corresponding maximum value of PC. For example, PC1 loading values are multiplied by the maximum value of the PC1 score and the PC2 loadings are multiplied by the maximum value of the PC2 scores. This transformation places both the variables and the observations on the same scale in the bi-plot display since the range of PC loadings are usually shorter than the PC scores. Observations having larger (> 75% percentile) or smaller (< 25% percentile) PC scores are only identified by their ID numbers on the bi-plot to avoid crowding of too many id valus. Patients with similar characteristics are displayed together in the bi-plot observational space since they have similar PC1 and PC2 scores..

The correlations among the multivariate attributes used in the PCA analysis are revealed by the angles between any two PC loading vectors. For each variable, a PC load vector is created by connecting the x-y origin (0,0) and the multiplied value of PC1 and PC2 loadings in the bi-plot. The angles between any two variable vectors will be: 1) narrower (< 45°) if the correlations between these two attributes are positive and larger (E.G: X2 and X3); 2) wider (around 90°) if the correlation is not significant (E.G: X5 and X4), 3) closer to 180° (>135°) if the correlations between these two attributes are negative and stronger (E.G:X4 and X2).

Figure 3 Bi-plot display of inter-relationship between the first and second PC scores and PC loadings two PC scores and PC loadings.



Summary:

A user-friendly SAS MACRO “FACTOR” is now available for downloading and performing PCA with data exploration [7,8] in medical research. Medical researchers can effectively and quickly apply data mining techniques in PCA. This MACRO can help them spend more time in exploring data, interpretation of graphs and output rather than debugging their program errors.

References

- [1] Alter, O., P. Brown, and D. Botstein (2000) Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 97 (18), 10101-10106.
- [2] Sharma, S., *Applied Multivariate Techniques*. New York: Wiley 1996
- [3]. Reaven, G. M. and Miller, R. G. (1979) “An attempt to define the nature of Chemical Diabetes using multi-dimensional analysis. *Diabetologica* 16: 17-24.
- [4]. SAS Institute Inc. (1999) SAS/STAT Users Guide, Version 8, Cary NC .SAS Institute Inc
- [5] SAS Institute Inc Comparison of the PRINCOMP and FACTOR Procedures In SAS ONLINE Documentation at <http://v8doc.sas.com/sashtml/stat/chap6/sec12.htm> (Last accessed 05-21-02)
- [6] Gabriel, K. R Bi-plot display of multivariate matrices for inspection of data and diagnosis. In V. Barnett (Ed.). *Interpreting multivariate data*. London: John Wiley & Sons. 1981.
- [7] Fernandez G (2002) Data mining using SAS applications CRC Press NY Chapter 4
- [5] Fernandez G (2002) Data mining using SAS applications - CD-ROM CRC Press NY

Appendix:

Instructions for downloading and running the Factor SAS macro:

As an alternative to the point-and-click menu interface modules, a user-friendly SAS macro application to perform a complete PCA analysis developed by the author is now available [7]. This macro approach integrates the statistical and graphical analysis tools available in SAS systems and provides complete data analysis tasks quickly without writing SAS program statements by running the SAS macros in the background. The main feature of this approach is that the users can perform graphical PCA analysis quickly using the SAS macro-call file available for downloading from the book's website. Using this MACRO APPROACH, the analysts can effectively and quickly perform complete data analysis and spend more time in exploring data, interpretation of graphs and output rather than debugging their program errors etc.

Requirements:

This macro application was developed in Win/NT SAS version 6.12 and was tested in both version 6:12, 8.2, and SAS (Learning edition). The requirements for using these SAS MACROs are :

- 1) A valid license to run the SAS software on your PC.
- 2) SAS modules such as SAS/BASE, SAS/STAT, and SAS/GRAPH should be installed in your computer to get the complete results.
- 3). Down-load the macro call file 'Factor.sas' from the Data mining using SAS application [7] book's website. Instructions are given in the book for downloading the macro-call file.
- 3) Active internet connection to access the SAS macro from the book's website. If active internet connection is not available, obtain the actual SAS macro code from the Data mining using SAS application - CD-ROM [8].

The steps for performing PCA analysis by the user-friendly SAS MACRO "FACTOR":

Step 1: Create a SAS data set

This data should contain the following variables:
An ID and continuous or numeric multi attribute variable.

Step 2: Click the RUN icon to open the Step 2: Open the FACTOR macro call window (Figure 4) in SAS by running the macro-call file 'Factor.sas'.

Step 3: Input the required values by following the instructions provided in the SAS MACRO-CALL window and using the help file given in the data mining using SAS application book [7].

Step 4: Submit the SAS MACRO:

After inputting all required fields, move your cursor to the last MACRO field, and hit the Enter key to run the SAS MACRO. The MACRO-CALL window file, automatically accesses the appropriate SAS MACROS from the Internet server, College of Agriculture, University of Nevada and provide the users the required exploratory graphs, and complete PCA analysis results.

Figure 4: Screen copy of 'FACTOR' macro call window showing the macro-call parameters required for performing PCA.

