Principal Component Analysis

Ahmed Saia'rah
Department of Math and computer science
Palestine Polytechnic University
Hebron, Palestine

Rae'd Sharabati
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
Hebron , Palestine

Mohammed Jiebreen
Department of Math and computer science
Palestine Polytechnic University
Hebron, Palestine

Supervisor: Monjed Samuh
Department of Math and computer science
Palestine Polytechnic University
Hebron, Palestine
monjedsamuh@ppu.edu

Abstract— **The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation presented in all of the original variables.**

Keywords-component; correlation matrix; principal components; reduction of dimensionality; variance covariance matrix.

I. Introduction

Principal Components Analysis (PCA) is a traditional multivariate statistical method commonly used to reduce the number of predictive variables. PCA looks for a few linear combinations of the variables that can be used to summarize the data without losing too much information in the process. This method of dimension reduction is also known as "parsimonious summarization".

Principal components depend solely on the covariance matrix $\sum$ (or the correlation matrix ρ) of $X_1, X_2, \dots, X_p$.

II. *Population* Principal *Components*

Let the random vector $X^T = [X_1, X_2, \dots, X_p]$ have the covariance matrix $\sum$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the linear combinations :

$$Y_1 = a_1^T X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Y_2 = a_2^T X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = a_p^T X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Where

$$Var(Y_i) = a_i^T \sum a_i, \qquad i = 1,2,\dots,p$$

$$Cov(Y_i, Y_k) = a_i^T \sum a_k, \qquad i,k = 1,2,\dots,p.$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, \dots, Y_p$ whose variances are as large as possible.

First principal component is the linear combination $a_1^T X$ that maximizes $Var(a_1^T X)$ subject to $a_1^T a_1 = 1$.

Second principal component is the linear combination $a_2^T X$ that maximizes $Var(a_2^T X)$ subject to $a_2^T a_2 = 1$ and $Cov(a_1^T X, a_2^T X) = 0$.

At the i^th step, the i^th principal component is the linear combination $a_i^T X$ that maximizes $Var(a_i^T X)$ subject to $a_i^T a_i = 1$ and $Cov\,(a_i^T X, a_k^T X) = 0 \quad \text{for } all\ k < i.$

Theorem 2.1.1:

Let $\sum$ be the covariance matrix associated with the random vector $X^T = [X_1, X_2, \dots, X_p]$. Let $\sum$ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots,$

(λ_p, e_p) where $\lambda_1 \ge \lambda_2 \ge \ \dots \ \ge \lambda_p \ge 0$. Then the ith principal component is given by:

$$Y_i = e_i^T X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p \qquad i = 1,2,\dots,p$$

where

$$Var\ (Y_i) = e_i^T \Sigma e_i = \lambda_i \qquad i = 1,2,\dots,\mathrm{p}$$

$$Cov(Y_i, Y_k) = e_i^T \Sigma e_k = 0 \qquad i \neq k$$

III. Methods for Discarding Components

When carrying out a principal component analysis , the researcher must decide how many components to use to represent the data; the other components will be discarded. Fava and Velicer (1992) studied the effects of using too many components. Note that before discarding components, we may wish to examine the "smallest" one for the information it carries.

A. *Percent of Variance*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. *Average Eigenvalue*

We could retain those components whose eigenvalues are greater than the average eigenvalue, $\bar{\lambda} = \sum_{j=1}^{p} \lambda_j \big/ p$ which is also the average variance of the variables, since $\sum_j \lambda_j = tr(s)$ For a correlation matrix, $\bar{\lambda} = 1$.

The average eigenvalue method often works well in practice .When this method errors, it is likely to be on the side of retaining too many components. Cattell and Jaspers (1967), Browne (1968), and Linn (1968) have studied the performance of this criterion in situations where the true dimensionality is known. They found the method to be fairly accurate when the number of variables is ≤ 30 and the variables are rather highly correlated. For larger numbers of variables that are not as highly correlated, the technique tends to overestimate the number of components.

C. *Scree Graph*

We could plot the eigenvalues in an attempt to find a visual break between the "large" eigenvalues and the "small" eigenvalues. This plot *is called a scree graph. The term scree, suggested by Cattell (1966), refers to the geological term for the debris at the bottom of a rocky cliff.*

An ideal scree graph is shown in Figure 2.4, in which it is easy to distinguish the large eigenvalues from the small ones. The first two eigenvalues *form a steep curve; the remaining eigenvalues exhibit a linear trend with small slope. In such a case, it is clear that we should delete the components corresponding to the small eigenvalues on the straight line. In practice,this ideal pattern may not appear, and this approach may not be conclusive.*

The accuracy of the scree method in choosing the correct number of components has been investigated in several studies. Cattell and Jaspers (1967) found it to give the correct number in 6 out of 8 cases. Linn *(1968) found it to be correct in 7 of 10 cases, and Tucker et al. (1969) found it to be accurate in 12 of 18 cases. Hakstian et al. (1982), comparing the average eigenvalue method and the scree method, found both to be accurate when n > 250 and the variables are at least moderately intercorrelated. When the correlations were smaller so that more components are needed, both methods were less accurate, the average eigenvalue method performing slightly better than the scree method.*

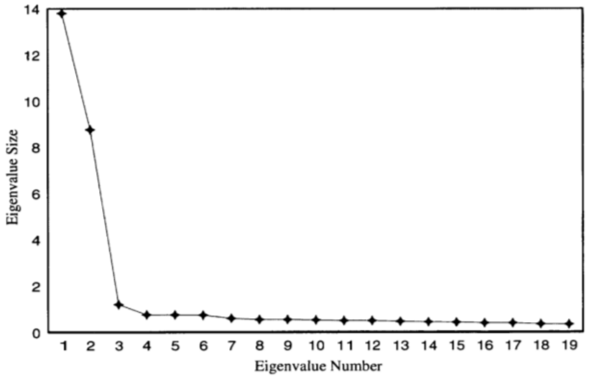

Figure 1

IV. Principal Component Regression

Recall that a multiple linear regression model with k predictor variables and a response variable y, can be written as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where y is the value of response (dependent) variable, $X_1, X_2, \dots, X_k$ are k predictor (independent) variables, $\beta_1, \beta_2, \dots, \beta_k$ are parameters (regression coefficients) and ε is the random error term, with mean $E(\varepsilon)$ =0 and variance $\sigma^2(\varepsilon) = \sigma^2$.

One of the problems in a multiple linear regression is multicollinearity that occurs when one or more of the independent variables are highly correlated with one or more of the other independent variables.

Multicollinearity effects on the regression analysis .Such that the regression coefficients will be unstable from sample to sample because the standard errors of the regression coefficients are very large. which means that the coefficients

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors) can't be estimated with great accuracy as well as the interpretation of the results. And the multicollinearity leads to high variance of coefficients and these reduce the accuracy of estimation.

One way of solving the problem of multicollinearity is principal component regression, in which y is regressed on the principal components of the $x's$. The standard errors of the regression coefficients on principal component regression become small.

The estimator for β is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

and consider the covariance matrix of $\hat{\beta}$ is

$$cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

and the total variance of $\hat{\beta}_j$ is equal $\sigma^2 \sum_{j=1}^{q} \frac{1}{\lambda_j}$, $j = 1, 2, \ldots, q$.

where $\frac{1}{\lambda_j}$ is the ith eigenvalue of $(X^T X)^{-1}$. If one or more of the $\lambda_j's$ is small, the total variance of the $\hat{\beta}_j's$ will be large.

A small eigenvalue of $X^T X$ induces multicolinearity among the $x's$.

Example

To illustrate principal component regression we use a data set given by Longer (1967) that has high multicolinearity and has often been used to test regression software for numerical accuracy. This data set has been used to illustrate principal component regression by Hill, Fompy, and Johnson (1977).

The variables are: y is the number of federal government employees, x_1 is the GNP price deflator, x_2 is the gross national product, x_3 is the unemployed, x_4 is the size of armed forces, x_5 is the population 14 years and over, x_6 is the year. The data were collected for 16 consecutive years.

$$R_{xx} = \begin{bmatrix} 1 & 0.9916 & 0.6206 & 0.4647 & 0.9792 & 0.9911 \\ 0.916 & 1 & 0.6043 & 0.4464 & 0.9911 & 0.9953 \\ 0.6206 & 0.6043 & 1 & -0.1774 & 0.6866 & 0.6683 \\ 0.4647 & 0.4464 & -0.1774 & 1 & 0.3644 & 0.4172 \\ 0.9792 & 0.9911 & 0.866 & 0.3644 & 1 & 0.9940 \\ 0.9911 & 0.9953 & 0.6683 & 0.4172 & 0.9940 & 1 \end{bmatrix}$$

The presence of multicolinearity is indicated by several high correlations.

In Table 3.2, we compare the principal component regression coefficients and the least square regression coefficients.

Table 3.2 Comparison of least squares and principal component regression

Standardized principal component			Least Squares	
Variable	Coefficient	standard error	Coefficient	standard error
x_1	36.61	9.04	29.02	0.01
x_2	1063.98	119.23	25.95	0.03
x_3	118.18	2.24	-32.45	0.20
x_4	141.23	0.24	109.14	0.14
x_5	-179.94	26.61	16.87	0.02
x_6	-936.77	50.6	23.38	0.01

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.