The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 637 - 644

Partial Quantification in Principal Component Analysis*

Hye-Sun Suh[1] and Myung-Hoe Huh[2]

Abstract

Sometimes, the first principal component may come logically from the established knowledge and premises. For example, for the high school students' test scores of Korean, English, Mathematics, Social Study, and Science, it is natural to define the first principal component as the average of all subject scores. In such cases, we need to respect both the background knowledge and the data exploration.

The aim of this study is to find the remaining components in principal component analysis of multivariate data when the first principal component is defined *a priori* by the researcher. Moreover, we study related matrix decomposition and their application to the graphical display.

1. Introduction

Principal component analysis(PCA) was first developed by Karl Pearson and Harold Hotelling (Jolliffe, 1986), as a way to reduce the dimensionality of the data set, achieved by principal components which are linear combinations of variables. In a similar vein, Mardia, Kent and Bibby (1979), Anderson (1984) and Jolliffe (1986) explain principal component analysis as a way to find linear combinations which have the maximum variance, or the best explanatory power.

In contrast, Lebart (1984) of French school of data analysis re-derives PCA as a descriptive analysis tool of multivariate data in the Euclidean space. Thus, PCA is regarded as a special case of correspondence analysis.

In this paper, we propose a special kind of PCA that is suitable in situations that the first principal component is given *a priori*. Such cases arise when the presupposition requires that the first principal component should be of certain form. Then, we need to harmonize both the background knowledge and the data exploration.

---

1) Postdoctoral Researcher, Institute of Statistics, Korea University. Seoul 136-701, Korea.
2) Professor, Dept. of Statistics, Korea University. Anam-dong 5-1, Seoul 136-701, Korea.
* This research was supported by the Institute of Statistics, Korea University.

In psychometrics and sociometrics, "quantification" is the terminology referring to the assignment of numeric values to qualitative objects. Thus, PCA can be viewed as a quantification method of n rows and p columns of the data matrix. By that reason, we name the statistical procedure of this study "partial quantification" in PCA.

The partial quantification procedure is developed in Section 2, followed by a data matrix decomposition in Section 3 and a graphical representation of the multivariate data in Section 4. A numerical illustration is given in Section 5.

In the writing hereafter, we use the following notations:

· X: $n \times p$ data matrix, $X = \begin{pmatrix} \vec{x}_1^{\,t} \\ \vdots \\ \vec{x}_n^{\,t} \end{pmatrix} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$.

· Thus $\vec{x}_1, \cdots, \vec{x}_n \in R^p$; $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p \in R^n$.

Here, we assume that X is centered and scaled, if necessary.

2. Partial Quantification

Suppose that the first principal component is defined *a priori* by

$$s_i = \vec{x}_i^{\,t} \vec{v}_1, \quad i = 1, \cdots, n$$

or the row elements of $X \vec{v}_1$, where $\vec{v}_1$ is the size p unit vector. Then, for the next steps to extract remaining information contained in multivariate observations, decompose $\vec{x}_i$ into two parts:

$$\vec{x}_i = s_i \vec{v}_1 + (\vec{x}_i - s_i \vec{v}_1), \quad i = 1, \cdots, n.$$

Hence we may restart usual PCA on

$$\vec{x}_i - s_i \vec{v}_1, \quad i = 1, \cdots, n.$$

Therefore, we obtained the following partial quantification algorithm in PCA.

Step 1: When unit vector $\vec{v}_1$ is given as the first principal component coefficients, replace the data matrix X by

$$X_{[1]} = X(I_p - P_{\vec{v}_1}), \tag{2.1}$$

where $P_{\vec{v}} = \vec{v}(\vec{v}^{\,t}\vec{v})^{-1}\vec{v}^{\,t}$, the projection matrix on $\vec{v}$.

Step 2: Find the size p unit vector $\vec{v}_2$ so that

$$\min \ \| \ X_{[1]} - X_{[1]} P_{\vec{v}_2} \ \|^2 \ !$$

or

$$\max \ \| \ X_{[1]} P_{\vec{v}_2} \ \|^2 \ !$$

Here, $\vec{v}_2$ is the second principal component coefficient vector, obtained by solving an eigensystem

$$X_{[1]}{}^t X_{[1]} \ \vec{v}_2 = \eta \ \vec{v}_2 .$$

Next, replace $X_{[1]}$ by

$$X_{[2]} = X_{[1]} (I_p - P_{\vec{v}_2}) \ . \tag{2.2}$$

Step 3: Repeat Step 2 until we obtain $\vec{v}_3, \cdots, \vec{v}_p$, remaining principal component coefficient vectors.

Figure 1 and Figure 2 illustrate the geometirc logic of the above algorithm. Let us denote $\vec{x}_{i.1}{}^t$ $(i=1, \cdots, n)$ as the i-th row of $X_{[1]}$. Then, in Figure 1, we see that

$$\vec{x}_{i.1} \ \perp \ \vec{v}_1 , \quad i = 1, \cdots, n,$$

where $\vec{v}_1$ is the given unit vector. In Figure 2, $\vec{x}_{i.1}$ $(i=1, \cdots, n)$ are projected on another unit vector $\vec{v}_2$: we easily see that the norms of projected vectors become largest when it is parallel to $\vec{v}_2{}^*$ that is orthogonal to $\vec{v}_1$. In the following, when there is no confusion, we use the notation $\vec{v}_2$ for $\vec{v}_2{}^*$. Thus,

$$\vec{v}_1{}^t \ \vec{v}_2 = 0,$$

and the second principal component scores are given by

$$\vec{x}_{i.1}{}^t \vec{v}_2 = (\vec{x}_i - k_1 \vec{v}_1)^t \vec{v}_2 = \vec{x}_i{}^t \vec{v}_2 \ , \ i=1, \cdots, n.$$

More generally, $\vec{v}_j$ $(j=1, \cdots, p)$ are orthogonal to each other, and the principal component scores are given by

$$\vec{x}_{i.1 \cdots j-1}{}^t \vec{v}_j = (\vec{x}_i - \sum_{l=1}^{j-1} k_l \vec{v}_l)^t \vec{v}_j = \vec{x}_i{}^t \vec{v}_j \ , \ j=2, \cdots, p \ ; \ i=1, \cdots, n.$$

Practically, $\vec{v}_j$ $(j=2, \cdots, p)$ can be obtained by solving the eigensystem

$$X_{[1]}{}^t X_{[1]} \ \vec{v}_j = \eta_j \ \vec{v}_j , \quad \eta_2 \geq \cdots \geq \eta_p .$$

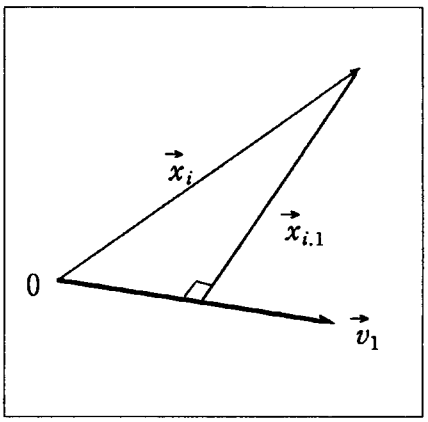


Figure 1
Step 1 of Partial Quantification.

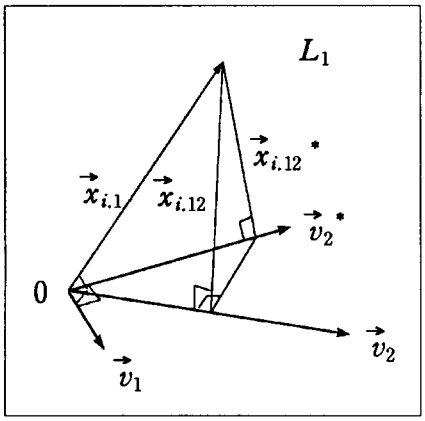


Figure 2
Step 2 of Partial Quantification.

3. A Decomposition of Data Matrix

We will develop a decomposition of the matrix X, according to the procedure of Section 2. From (2.1),

$$X = X \vec{v}_1 \vec{v}_1^{\,t} + X_{[1]}$$

and from (2.2),

$$X_{[1]} = X_{[1]} \vec{v}_2 \vec{v}_2^{\,t} + X_{[2]}.$$

Since $\vec{v}_1^{\,t} \vec{v}_2 = 0$, it turns out that

$$X = X \vec{v}_1 \vec{v}_1^{\,t} + X \vec{v}_2 \vec{v}_2^{\,t} + X_{[2]}.$$

Then, when $v_3, \cdots, v_p$ are included to approximate X further,

$$X = X \vec{v}_1 \vec{v}_1^{\,t} + X \vec{v}_2 \vec{v}_2^{\,t} + \cdots + X \vec{v}_{p-1} \vec{v}_{p-1}^{\,t} + X \vec{v}_p \vec{v}_p^{\,t} + X_{[p]}.$$

Since R^p is p-dimensional, the remainder $X_{[p]}$ is $\mathbf{0}$. Therefore

$$X = \sum_{j=1}^{p} X \vec{v}_j \vec{v}_j^{\,t}.$$

Now, define λ_j and unit vector $\boldsymbol{u}_j$ by

$$\lambda_j = \vec{v}_j^{\,t} X^t X \vec{v}_j \ , \quad \boldsymbol{u}_j = X \vec{v}_j / \sqrt{\lambda_j} \ , \quad j = 1, \cdots, p \ .$$

Then, we can write

$$X = \sum_{j=1}^{p} \sqrt{\lambda_j} \ \boldsymbol{u}_j \vec{v}_j^{\,t} = U D_{\sqrt{\lambda}} V^t \ , \tag{2.3}$$

where

$$U=(\ \boldsymbol{u}_1,\cdots,\ \boldsymbol{u}_p\ ),\ V=(\ \vec{v}_1,\cdots,\ \vec{v}_p\ ),\ D_{\sqrt{\lambda}}=diag(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_p}).$$

Since $\vec{v}_j{}^t\,\vec{v}_l=0\ (j\neq l)$, $V^tV=I_p$. Moreover, for $j,\ l=2,\cdots,p,\ j\neq l$,

$$\boldsymbol{u}_j{}^t\,\boldsymbol{u}_l\ \propto\ \vec{v}_j{}^tX^tX\,\vec{v}_l\ =\ \vec{v}_j{}^tX_{[1]}^tX_{[1]}\,\vec{v}_l=\ \eta_j\,\vec{v}_j{}^t\,\vec{v}_l=0\ .$$

But, in general,

$$\boldsymbol{u}_1{}^t\,\boldsymbol{u}_j\neq 0,\quad j=2,\cdots,p.$$

As a result

$$V^tV=\ VV^t=I_p\quad\text{and}\quad U^tU=\begin{pmatrix}1 & \boldsymbol{m}^t\\ \boldsymbol{m} & I_{p-1}\end{pmatrix},$$

where $\boldsymbol{m}=(m_2,\cdots,m_p)^t$, $m_j=\boldsymbol{u}_1{}^t\,\boldsymbol{u}_j\ (j=2,\cdots,p)$. Therefore, we call (2.3) as a quasi-singular value decomposition (SVD) of the matrix X and λ_j as quasi-eigenvalues.

4. Biplot Representation

The biplot, proposed by Gabriel (1971), is a statistical graph showing the relative positions of rows (observations) with the directions of columns (variables) of the data. Its validity is based on a matrix decomposition of the data matrix

$$X=GH^t \tag{2.4}$$

which can be obtained in several ways via usual singular value decomposition (SVD) of X, i.e.

$$X=MD_{\sqrt{\xi}}N^t, \tag{2.5}$$

where

$$M^tM=N^tN=NN^t=I_p,\quad D_{\sqrt{\xi}}=diag(\xi_1,\cdots,\xi_p),\ \xi_1\geq\cdots\geq\xi_p.$$

After applying the partial quantification, we obtain quasi-SVD (2.3) rather than SVD (2.5). However, by taking $G=UD_{\sqrt{\lambda}}$ and $H=V$, we have (2.4) or

$$x_{ij}\ =\ \vec{g}_i{}^t\,\vec{h}_j\ ,\quad i=1,\cdots,n\ ,\ j=1,\cdots,p,$$

where $\vec{g}_i{}^t$ is the i-th row vector of matrix G, and $\vec{h}_j$ is the j-th row vector of matrix H. Hence

$$x_{ij}\ \approx\ \vec{g}_{i(r)}{}^t\,\vec{h}_{j(r)}\ ,\quad i=1,\cdots,n\ ,\ j=1,\cdots,p, \tag{2.6}$$

where $\vec{g}_{i(r)}$ and $\vec{h}_{j(r)}$ are, respectively, the size r subvectors of $\vec{g}_i$ and $\vec{h}_j$ that keep the first r elements. Therefore, n rows (observations) of X are represented in the r-dimensional

subspace by $\vec{g}_{i(r)}$, $i=1,\cdots,n$, and, p columns (variables) of X are represented in the r-dimensional subspace by $\vec{h}_{j(r)}$, $j=1,\cdots,p$.

Equivalently, (2.6) can be expressed as

$$X \approx G_{(r)} H_{(r)}^{t},$$

where $G_{(r)}$ and $H_{(r)}$ are respectively the submatrices of G and H with first r columns retained.

Thus, the goodness-of-approximation index of the r-dimensional biplot for the representation of rows may be defined by

$$GOA_{(r)} = \frac{\| G_{(r)} \|^2}{\| G \|^2} = \frac{\| U_{(r)} D_{\sqrt{\lambda_{(r)}}} \|^2}{\| U D_{\sqrt{\lambda}} \|^2} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

since

$$\| U D_{\sqrt{\lambda}} \|^2 = tr(D_{\sqrt{\lambda}} U^t U D_{\sqrt{\lambda}})$$

$$= tr\left(D_{\sqrt{\lambda}} \begin{pmatrix} 1 & \boldsymbol{m}^t \\ \boldsymbol{m} & I_{p-1} \end{pmatrix} D_{\sqrt{\lambda}} \right) = \sum_{j=1}^{p} \lambda_j .$$

5. A Numerical Example

To illustrate the proposed partial quantification method, consider the ability test data in du Toit, et al. (1986). It consists of data obtained during a long-term research project in South Africa, called Project Talent Survey in which 21 variables were measured from approximately 2800 white pupils. The first 18 variables constituting the test scores were obtained from the pupils in the various ability tests, at three distinct educational levels. The list of variables are:

X_1, X_7, X_{13}: Number Series, X_4, X_{10}, X_{16}: Classification (Word Pairs),

X_2, X_8, X_{14}: Figure Analogies, X_5, X_{11}, X_{17}: Verbal Reasoning,

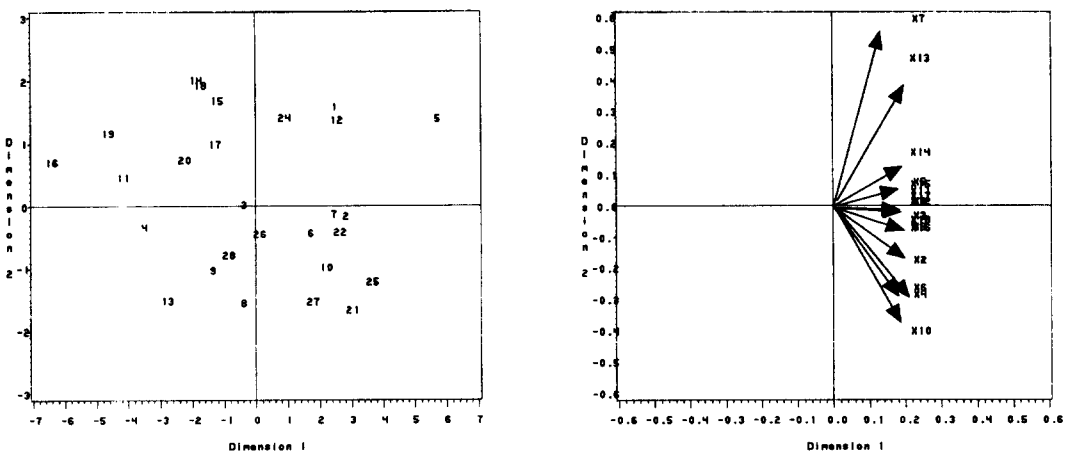
X_3, X_9, X_{15}: Pattern Completion, X_6, X_{12}, X_{18}: Word Analogies.

This numerical example deals with the partial quantification of the subset data from 28 pupils, for the case that the first principle component is defined *a priori* as the simple average for all 18 variables.

As reported in Table 1, the goodness-of-approximation in two dimensional subspace is 56% for the row representation, of which nearly 20% is contributed by the second axis. In Figure 3, observe that the number series (X_7, X_{13}) and the figure analogy (X_{14}) are in one side of the second axis, and word pairs (X_{10}, X_4) and verbal reasoning (X_6) are in the other side. So, the second principal component represents the differential orientation toward the quantitative aptitude (in positive direction) versus the verbal aptitude (in negative direction).

Table 1. Principal Components and Quasi-Eigenvalues by Partial Quantification

Variables	Principal Component Coefficient Vectors	
	First Axis	Second Axis
X_1	0.235	-0.172
X_2	0.235	-0.174
X_3	0.235	-0.030
X_4	0.235	-0.284
X_5	0.235	0.077
X_6	0.235	-0.259
X_7	0.235	0.597
X_8	0.235	-0.065
X_9	0.235	0.016
X_{10}	0.235	-0.398
X_{11}	0.235	0.048
X_{12}	0.235	0.014
X_{13}	0.235	0.470
X_{14}	0.235	0.170
X_{15}	0.235	0.069
X_{16}	0.235	-0.069
X_{17}	0.235	0.031
X_{18}	0.235	-0.041
Quasi-Eigenvalues	223.0	48.6
Cumulative %	46%	10%



(1) Row Plot (2) Column Plot

Figure 3. Partial Quantification Plots

6. Concluding Remarks

Principal component analysis, in fully exploratory use, is adopted to find out a number of linear combinations that deliver maximal information contained in the data set. But, in the fields of social science, the first principal component may come logically from the established knowledge and natural premises. For example, the simple average of given variables may be designated as the first principal component.

The variance explained by the first principal axis of conventional principal component analysis is always larger than that of the partial quantification PCA. Thus, there is less danger of over-interpretation by partial quantification in PCA.

Although the partial quantification method of Section 2 is formulated from the geometry in the row space R^p, we can construct similarly another partial quantification method in the column space R^n. Details can be found in the first author's doctoral thesis (Suh, 1997).

This study derives remaining principal components in the data, given the first principal component. It is obvious that the algorithm of Section 2, if modified slightly, will work out smoothly for the situation that the first two principal components are specified *a priori*.

References

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* Second Edition. Wiley, New York.

[2] du Toit, S. H. C., Steyn, A. G. W. and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis.* Springer-Verlag, New York.

[3] Gabriel, K. R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, Vol. 58, 453-467.

[4] Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

[5] Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices.* Wiley, New York.

[6] Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

[7] Suh, H. S. (1997). *Three Aspects of Statistical Methods for Social Research.* Unpublished Doctoral Thesis, Dept. of Statistics, Korea University. (written in Korean)