A Study on the Test of the Number of Principal Components Using Measurements of Similarity Between Matrices

Lee, Seong Keon
Sunshin Women's University, Dept. of Statistics
Dongseon-Dong 3 ga
Seoul (136-742), Korea
E-mail: sklee@sungshin.ac.kr

Kim, So Young
Sunshin Women's University, Dept. of Statistics
Dongseon-Dong 3 ga
Seoul (136-742), Korea
E-mail: stat@sungshin.ac.kr

Kang, Hyun Cheol
Hoseo University, Dept. of Information Statistics
Asan(336-795)l, Korea
E-mail: hychkang@hoseo.ac.kr

Kim, Eun Seok
GDS Korea
Seoul, Korea
E-mail: eskim@gds.co.kr

Introduction

In many fields, multivariate analyses are widely used to describe and summarize large data sets (many variables and/or individuals) by removing any redundancy in the data. Principal Component Analysis is to reduce the dimensionality of a data set that has a number of correlated variables and analyze complex construction between correlated variables. This method is achieved by linearizing and transforming to a new set of variables, which is principal component, which are uncorrelated and sorted in descending order. The first few retain most of the variation present in all of the original variables.

One crucial step of PCA concerns the choice of the number of axes to be retained for interpretation and subsequent analyses. This decision is often made according to practical considerations (e.g., two axes retained because only two dimensions can be represented on a sheet of paper) and not statistical ones. The consequences of this choice are important: if the number of axes is not correctly estimated, one can introduce noise (overestimation) or loss of information (underestimation) in the analysis. A number of approaches to estimate the dimensionality of a data table (i.e., number of axes) have been proposed and evaluated in the literature (e.g., Jackson, 1993; Peres-Neto et al., 2005). Jolliffe (2002, pp. 112–132) reviews the most frequently used approaches that how many principal components should be retained and distinguishes three types of rules.

Classical methods that are choosing of the number of principal component are using a ‘scree plot ‘or eigenvectors.(e.g., if cumulative rates of 3 eigenvectors are more than 80%, then the number of principal component is 3.) One of types corresponds to methods that do not require distributional assumptions. S.Dray(2008) focused on the type of method and proposed a new approach to estimate the dimensionality of

a data set based of the link between PCA and the approximation of a matrix by another of lower rank (Eckart andYoung, 1936) using singular value decomposition (SVD, Good, 1969). In this paper, we will compare methods suggested by S.dray(2008) and proposed by us.

Modified RV

If sample size is decreasing, then RV coefficient would become high. So RV is sensitive for sample size. Also increasing the number of variable makes RV coefficient to be high. Because of these problems, especially in high dimensionality, A. K. Smilde et al (2009) suggested modified RV coefficient. Let X ($I \times J_1$) and Y ($I \times J_2$) be two matrices corresponding to two sets of observation made on the same I individuals. The RV coefficient can be approximated as

$$RV(X,Y) \approx \frac{(J_1 J_2)}{(J_1^2 + 2J_1 + (I-1)J_1)^{1/2}(J_2^2 + 2J_2 + (I-1)J_2)^{1/2}}.$$

From the above formula, it can be seen that the value of RV coefficient for random data matrices depends on I: for small I, the RV coefficient is close to 1, whereas, as I increases the denominator increases and the value approaches zero. In the other word, the accuracy of these approximations depends on I.

If these diagonal elements are ignored or, equivalently, set to zero, then the problem will disappear. So we use $\widetilde{XX}' = [\,XX' - \mathrm{daig}(XX')]$ instead of using XX' and modified-RV is

$$modified\ RV(X,Y)' = \frac{tr(\widetilde{XX}'\widetilde{YY}')}{\sqrt{tr(\widetilde{XX}'\widetilde{XX}')tr(\widetilde{YY}'\widetilde{YY}')}} \quad .$$

Results

The number of principal component	Matrix 1		Matrix 2	
	RVDIM	Modified RV	RVDIM	Modified RV
1	0.0002500	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0.0642660	0.10325	0.78125	0.72975
5	0.4456114	0.3685	0.24075	0.18625
6	0.9892473	0.81625	0.0515	0.60475
7	0.883971	0.2235	0.746	0.60675
8	0.7964491	0.65375	0.157	0.15525
9	0.9892473	0.94825	0.1555	0.943

<p-values of RVDIM and Modified RVDIM in Matrix 1, 2>

Original RVDIM and modified RVDIM show similar results. Even using modified RVDIM is more accurate outcome. The details will be shown in the presentation.

Conclusion

We want to compare p-values that one of these is calculated to use standardized RV-coefficient suggested by Korth and Tucker (1976) and another is calculated by permutation procedure (e.g. (number of random values equal to or larger than the observed +1)/ (repeat time)). However standardized RVDIM is so

large that we can't compare two p-values. And using of covariance matrices is elicited different results with original RVDIM. As a result, choosing of the number of principal component computed by covariance matrices doesn't apply. Values of modified RVDIM that is improving disadvantages of RVDIM are similar to original RV-coefficient's values and the results of modified are also nearly same.

REFERENCES

Jolliffe, I., 2002. Principal Component Analysis. *second ed. Springer*, *Berlin.*

Peres-Neto, P., Jackson, D., Somers, K., 2005. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput. Statist. Data Anal. 49, 974–997*.

Hervé Abdi, 2007. RV Coefficient and Congruence Coefficient. Encyclopedia of Measurement and Statistics.

J. Josse , F. Husson and J. Pagμes,2007. Testing the significance of the RV coefficient, *IASC 07, Aveiro, Portugal.*

Stephane Dray, 2008. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. C*omputational Statistics & Data Analysis, 52, 2228 – 2237.*

A. K. Smilde, H. A. L. Kiers, S. Bijlsma, C. M. Rubingh and M. J. van Erk, 2009. Matrix correlations for high-dimensional data: the modified RV-coefficient. *bioinformatics ,25, 401–405.*