Variable Selection and Principal Component Analysis

Noriah Al-Kandari
University of Kuwait, Department of Statistics and OR
P.O. Box.: 5969 Safat-Code No. 13060
Kuwait
e-mail: noriah@kuc01.kuniv.edu.kw

1. Introduction

In most of applied disciplines, many variables are sometimes measured on each individual, which result a huge data set consisting of large number of variables, say *p* [Sharma (1996)]. Using this collected data set in any statistical analysis may cause several troubles.

The dimensionality of the data set can often be reduced, without disturbing the main features of the whole data set by Principal Component Analysis (PCA) technique [Rencher (1996)]. Dimensionality reduction is affected if *k* (<< *p*) of the Principal Components (PCs) convey virtually all the information inherent in the *p* variables [Jolliffe (1986) and Krzanowski (1988)]. However, the constructed PCs may not be easy to interpret in terms of all the original *p* variables. Therefore, it is useful to reduce the number of variables as much as possible whilst capturing most of the variation of the complete data set, *X*. The present paper aims to determine for a given structured correlation data sets which Variable Selection Criteria (VSC) yield the best subset of *q* (*q* < *p*) variables capturing most of the variation in the complete data set and aid the simultaneous interpretation of the *first two correlation* PCs.

2. Variable Selection Criteria

Jolliffe (1972, 1973) and McCabe (1984) suggest the variable selection criteria (VSC). Al-Kandari (1998) provided a few more VSC, using the concept of "Principal Variables" given by McCabe (1984). By applying these VSC, several subsets of size *q* are retained and a final selection between these subsets is desirable. For this purpose two measures of efficiency are considered, namely Measure (1) and Measure (2). Recently, Cadima and Jolliffe (1995) discussed approximation methods to interpret PCs using subset of *q* variables, namely truncated and regressed approximations.

3. Data Simulations and Analysis

In the present study, several 6-dimensional data sets are generated such that the variables are allocated in two-equal sized clusters with high correlations within clusters and low correlations among clusters. Also, two structures of correlation eigenvectors are considered, namely *Block-Structure* (BS) and Uniform-Structure (US). These data sets are generated using C program language [Al-Kandari (1998)].

PCA was performed on each generated data set and on the average, the first two correlation PCs y_1 and y_2 explained about 89% of the total variation of the complete data sets. The VSC were to retain subsets of size 2, 3 and 4. The resultant subsets are used to interpret the first two PCs using the two methods of approximation. Next, the efficiencies of the resultant subsets of size *q*, which yielded good correlations between the first two PCs and their two approximations, were determined.

Based on the obtained efficiencies, the VSC were investigated in order to determine which criteria yielded the best subset of size *q* under each measure.

Moving to the interpretations of the first two PCs y_1 and y_2 using the truncated and regressed approximations, several interpretation problems were noted.

6. Conclusion

The statistical analysis of a huge sheer volume of a data set can be simplified by retaining only few variables. For the considered correlation structure in the present paper, only few VSC played main roles in retaining subsets of size *q*. The selection among these criteria depends on the desired size of the retained subset and the measure of efficiency of interest. In turns, the decision here depends on the user's target from the whole study. It was noted that there is a relationship between the interpretation of correlation PCs and the structure of the associated eigenvectors. Less number of interpretation problems can be obtained if the eigenvectors tend to have block structure.

REFERENCES

Al-Kandari, N. M. (1998). Variable selection and interpretation of principal component analysis. Ph.D. Thesis, University of Aberdeen. Aberdeen.

Cadima, J. F. C. L. and Jolliffe, I. T. (1995). Loadings and Correlations in the interpretation of Principal Components. Journal of Applied Statistics, 22(2), 203-214.

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis I: Artificial Data. Applied Statistics, 21, 160-173.

Jolliffe, I. T. (1973). Discarding Variables in a Principal Component Analysis II: Real Data. Applied Statistics, 21, 160-173.

Krzanowski, W. J. (1988). Princip;es of Multivariate Analysis: a user's perspective. Clarendon Press, Oxford.

McCabe, G. P. (1984). Principal Variables. Technometrics, 26(2), 137-144.

Sharma, S. (1996). Applied Multivariate Techniques. John Wiely & Sons, Inc. New York.

Rencher, A. C. (1995). Methods of Multivariate Analysis. John Wiley & Sons. New York.