Simple component analysis based on RV coefficient

Analisi delle Componenti "semplici" basate sul coefficiente RV

Michele Gallo
Department of Social Science, University of Naples "L'Orientale", Italy
E-mail: mgallo@unior.it

Pietro Amenta
Department of Analysis of Economic and Social Systems, University of Sannio, Italy
E-mail: amenta@unisannio.it

Luigi D'Ambra
Department of Mathematics and Statistics, University of Naples "Federico II", Italy
E-mail: dambra@unina.it

Riassunto: l'analisi delle componenti principali (ACP) è, tra le tecniche di riduzione delle dimensioni, quella maggiormente utilizzata in quanto presenta delle proprietà ottimali rispetto alle altre presenti in letteratura. In molti casi reali, tuttavia, l'ACP genera dei risultati difficilmente interpretabili. Al fine di migliorare l'interpretabilità dei risultati, in letteratura sono state proposte diverse tecniche che fanno uso di criteri sub-ottimali tra i quali l'analisi delle componenti semplici (SCA - Rousson and Gasser, 2003). Obiettivo del presente lavoro è quello di confrontare le diverse tecniche proposte in letteratura e di proporre una variante di SCA che fa uso del coefficiente di correlazione vettoriale RV.

Keywords: Principal components, Dimensionality reduction methods, Interpretability of components, Simplicity, RV coefficient.

1. Introduction

When the number of observed variables is very large, it may be advantageous to find linear combinations of the explanatory variables having the property to account for most of the variance in the observed variables.
Among linear dimensional reduction techniques, Principal Component Analysis (PCA) presents many optimal proprieties. Unfortunately, in many applicative case, as in case of customer satisfaction data, all variables are strongly correlated and consequently the first principal component may be correspond to overall size. In these case, PCA doesn't produce full interpretable results. To resolve these kinds of problems, many authors proposed some procedure that will be suboptimal compared to the principal components obtained by PCA but make more interpretable principal components.
In literature three different strategies are proposed. The first is based on replacement of some elements of the correlation matrix with other than produce more simple and generally more interpretable results (Hausman, 1982; Vines, 2000; Rousson and Gasser,

2003). The second is based on centroid methods (Choulakian, D'Ambra and Simonetti, 2004). The last is based on a rotation methods of the loading matrix (Jolliffe, 1989; Jolliffe and Uddin, 2000 and 2003).
Following Rousson and Gasser (2003), in this paper we propose to modify the algorithm used for the Simple Component Analysis (SCA) based on RV coefficients (SCA-RV) in order to improve interpretability of results. Moreover, to compare PCA, SCA, SPCA (Zous et al., 2004) ScoTLASS (Jolliffe and Uddin, 2003), Factorial Analysis based on varimax criterion (Hausman, 1982) with SCA-RV, an application on the level of the patient satisfaction in the service of a Neapolitan Children's Hospital will be given.

2. Simple and Interpretable versus optimality

Simplicity does not assure more interpretable components but allows to analyze the case variables measure different aspects of a same theme and all the elements of the correlation matrix are strictly positive. A correlation matrix with this structure gives more interpretability problems when principal components are extracted. Unfortunately, there is a trade-off between the simplicity and optimality. When the simplicity is searched by rotated principal components (or block-components), the optimality is worsen because the block-components are correlated and less variability is extracted from the original variables, when the loss of extracted variability is small and the correlation between the components are low, it is advantageous use SCA-RV.
Analogously to SCA, SCA-RV gives more simple loading matrix, only three kinds of values (positive, negative and zero) and the sum of loadings for each component is always zero. In this way, the block-components is just an averages and difference-components just a simple contrast of variables. Differently to factorial analysis, SCA-RV gives the possibility to choose the number of block-components. So the correlation between them is cut off.

3. SCA-RV

Let Y be a matrix with p standardize random variables $(Y_1,\ldots,Y_p)$, such that $C = Y'Y$ is the correlation matrix with rank q $(q \le p)$. Moreover, Let p_j (with $j = 1,\ldots,q$) be the column of a pxq projection matrix P. PCA points out a solution P with the following major properties: 1) the columns of P are orthogonal, 2) the projected data YP are uncorrelated, 3) the vector p_1 is chosen to maximize the variance of Yp_1 and p_j is chosen to maximize the variance of Yp_j with $p_j' p_{j'} = 0$ (for $j \neq j'$ and $j = 1,\ldots,q$ and $j' = 2,\ldots,q$). These proprieties are desirable and in this sense PCA is a reduction technique with optimal proprieties.
When all variables measure different aspects of a same theme all the elements of C are strictly positive. This structure of the correlation matrix gives more problems of interpretability of PCA results. In order to get more simple and generally more interpretable components sometimes a suboptimal solution is preferred to the optimal solution of PCA. Rousson and Gasser (2003) proposed a procedure that maximizes

$p_1'Cp_1 + \sum_{j=2}^{q} p_j'C_{(j-1)}p_j$, divide by the sum of variances of original variables. This criterion assures equivalent results to PCA only in case of uncorrelated components, while it is a penalized version of PCA criterion for correlated components. Seeking a system of q simple components with b blocks maximizing the cited criterion of optimality, the two stages SCA algorithm provides an approximation to the optimal system of simple components. With fixed values of b and q, in the first stage of the algorithm b simple block components (components whose non zero loadings have all the same sign) are defined while in the latter $(q-b)$ simple difference components (components which have some strictly positive and some strictly negative loadings) are described.

First stage of SCA is to classify p variables into b disjoint blocks. The approximate block-structure in the correlation matrix leads to a maximal within block correlations and in the meantime to a minimal between blocks correlations. Authors solved this problem with an agglomerative hierarchical procedure based on a dissimilarity measure between clusters called "median linkage" alternative to the possible single or complete linkages. Coming from the matrix of loadings corresponding to the b simple block components of the first stage, the second phase of the algorithm is based on a suitable difference component shrinkage procedure of the sequential first components of the residual variables obtained by regressing step by step the original variables on the first $(j-1)$ simple components.

Our proposal modifies the criterion of solution at the first stage. Instead to use an agglomerative hierarchical procedure, which can lead to a non unique solution with a choice of a possible different link criterion, we propose to use the RV vectorial correlation coefficient proposed by Robert and Escoufier (1976). Several proposal in literature considered the use of the RV coefficients for variable selection; for example in order to select subsets of variables in the context of PCA (Bonifas et al, 1984; Mori et al. 1999). Robert and Escoufier (1976) have derived a measure of similarity of the two configurations, taking into account the possibly distinct metrics to be used on them to measure the distances between points. The measure is computed as

$$RV(V,Z) = tr\left(\overline{V}\overline{V}'\overline{Z}\overline{Z}'\right) \Big/ \left[tr\left(\overline{V}\overline{V}'\right)^2 tr\left(\overline{Z}\overline{Z}'\right)^2\right]^{1/2}$$

where $\overline{V}$ and $\overline{Z}$ are centered matrices of $\overline{V}$ and $\overline{Z}$, respectively. This measure respects all the four conditions for a vectorial correlation coefficient proposed by Renyi (1959): a vectorial correlation coefficient *r* is an application

$$r : \Psi_{(nxs)} \text{ x } \Psi_{(nxs)} \to [0,1]$$
$$(V,Z) \to r(V,Z)$$

where $\overline{V}$ and $\overline{Z} \in \Psi_{(nxs)}$ are not simultaneously null and Ψ_n is the set of the squared real matrices of order n. This application verifies the following properties: 1) $\forall (a,b) \in \Re^2$, $r(aV,Z) = r(V,bZ) = r(V,Z)$; 2) $r(V,Z) = r(Z,V)$; 3) if $V = bZ$ then $r(V,Z) = 1$; 4) $r(V,Z) = 0 \Leftrightarrow V'DZ = 0$ with D weights diagonal matrix.

A lot of vectorial correlation coefficients proposed in literature do not respect all the cited properties (Amenta, 1993).

It can be shown that the RV coefficient is equivalent to squared Pearson's correlation coefficient between the association matrices, if these are rearranged as vectors and it is

invariant to a change of scale. Moreover the RV coefficient can be linked also to a concept of proximity between matrices (Robert and Escoufier, 1976): $dist(A,B)=0 \Leftrightarrow RV(A,B)=1$.

In this sense, the first stage of SCA can be synthesize in three most important steps:

1. Start with p blocks $B_1,...,B_p$ where each block contains one of the original variables;
2. Select two blocks B_I and B_J for which a measure of RV is biggest and aggregate them into a new block $B_{(I,J)}$;
3. If b blocks remain then stop the loop, otherwise go back to step 2.

Similarly to SCA, the agglomeration process could be continued until the RV correlation between some block components is larger than a prefixed value (e.g. 0,3 or 0,4). All the proprieties of RV coefficient and the full algorithm will be given on the extended version of the paper. Moreover, to compare PCA, SCA, SPCA (Zous et al., 2004), ScoTLASS (Jolliffe and Uddin, 2003), Factorial Analysis based on variamax criterion (Hausman, 1982) and SCA-RV an application on the level of the patient satisfaction in the service of a Neapolititan Children's Hospital will be given.

References

Amenta, P., (1993) *Il coefficiente di correlazione lineare tra matrici di dati nel contesto multivariato*. XVII Convegno A.M.A.S.E.S., Ischia.

Bonifas, I., Escoufier, Y., Gonzalez, P.L. et Sabatier, R. (1984), Choix de variables en analyse en composantes principales. *Revue de Statistique Appliquée*, 23, 5-15.

Choulakian, H. A., D'Ambra, L. and Simonetti, B. (2004) *The extended centroid method*. Submitted.

Hausman, R. E. (1982) *Constrained Multivariate Analysis*. Optimisation in Statistics, eds. S. H. Zanckis and J. S. Rustagi, Amsterdam: North Holland, 137-151.

Jolliffe, I. T. (1989) Rotation of Ill-Defined Principal Components. *Applied Statistics*, 38, 139 - 147.

Jolliffe, I. T. and Uddin, M. (2000) The Simplified Component Technique: An Alternative to Rotated Principal Components. *Journal of Computation and Graphical Statistics*, 9, 689-710.

Jolliffe, I. T., Uddin, M. (2003) A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12, 531 - 547.

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y. (1999) *Variable Selection in "Principal Component Analysis Based on a Subset of Variables"*. Bulletin of the International Statistical Institute (*52nd Session Contributed Papers Book2*), 333-334.

Renyi, A. (1959) *On measures of dependence*. Acta Mathematica of the Academy of Science of Hungary, 10.

Robert, P. and Escoufier, Y. (1976) A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied Statistics*, 25, 257-65.

Rousson, V. and Gassen, T. (2003) *Some case studies of simple component analysis*. Manuscript on http://www.unizh.ch/biostat/Manuscripts.

Vines, S. K. (2000) Simple Principal Components. *Applied Statistics*, 49, 441 - 451.

Zous, H., Hastie, T. and Tibshiraniz, R. (2004) *Sparse Principal Component Analysis*. Manuscript on http://www-stat.stanford.edu/~hastie/pub.htm.