Principal Components Analysis onto Reference subspace in presence of multicollinearity

Antonello D'Ambra
Seconda Università degli Studi di Napoli
antonello.dambra@unina2.it

Pasquale Sarnacchiaro
Seconda Università degli Studi di Napoli
sarnacch@unina.it

Abstract: *The Principal Component Analysis onto References Subspaces is a multivariate method to analyze two sets of quantitative variables when between the two sets exists a directional relationship. When the explicative variables are affected by multicollinearity this technique is not recommended.*
In literature exist many methods to resolve this problem (Ridge Regression, Principal Component Regression, Partial Least Square, Latent Root Regression), this work shows an alternative method based on simple linear regression.

Keywords: PCA, PCAR, CPR, Linear Regression

Introduction

Both statisticians and researchers of the many disciplines that employ Regression Analysis, Principal Component Analysis should be aware of the adverse effect of multicollinearity and of the pitfalls that may exist in the detection of linear dependencies.
The Principal Component Analysis onto References subspaces (PCAR) is a multivariate method to analyze two sets of quantitative variables in which one is composed by explicative variables and the other by dependent variables.
An often overlooked, but nevertheless significant problem in analyzing data, is that of multicollinearity, or non-orthogonality, among the independent variables (X).
The condition of multicollinearity is not, of course, limited to the PCAR but is, in fact, a pervasive and potential problem in all research in the family studies field.
This paper is concerned with multicollinearity within the context of PCAR. The paper is organized along the following lines. In Paragraph one it is illustrated PCAR and it is introduced the problem of multicollinearity among explicative variables. In paragraph two, a multivariate approach based on linear regression, to solve the multicollinearity problem, is proposed. Starting from this proposal, in the last paragraph a weighted version of PCAR is presented.

1. Principal Component Analysis onto reference Subspaces

Hotelling's canonical correlation analysis (1933) and its various generalizations (Carol, 1968 and Kettenring, 1971) seem not suitable for the analysis of systems involving several sets of variables describing them, when the hypothesis of symmetrical relationships on which the corresponding theory rests cannot be assumed (D'Ambra L., Lauro N.C., 1984).
Let two subsystems composed by the same n observations described by q dependent variables, stored in a $n \times q$ matrix denoted Y, and p predictors collected in the $n \times p$ matrix X.

$$X = \begin{bmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_{11} & \cdots & \cdots & y_{1q} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ y_{n1} & \cdots & \cdots & y_{nq} \end{bmatrix}$$

Variables are assumed to have mean zero and to be divided by the normalizing scale factor $\sqrt{n}$.
Object of the analysis is to single out later subsystem structure with respect to the former, in terms of the principal component associated with it.
Let R be a vector space sized $p+q$ and R_x the R vector subspaces spanned by matrix column, the image of subsystem described by Y on this subspace is obtained through the orthogonal projector-operator $P_x = X\left(X'X\right)^{-1}X'$.
The Y projector on R having been effected

$$\hat{Y} = P_x Y \tag{1}$$

The Principal Components Analysis onto the Reference subspace (PCAR) can be made through the following steps:

- Search a subspaces of R_x (principal axes) through the extraction of the eigenvalues and eigenvectors $(\alpha = 1...h)$ of the expression:

$$Y'P_xYu_\alpha = \lambda_\alpha u_\alpha \tag{2}$$

$Y'P_xY$ being symmetrical and positive semidefinite, it derives that $\lambda_\alpha \geq 0$ and $u'_\alpha u_{\alpha'} = 0$ with $\alpha \neq \alpha'$;

An often overlooked, but nevertheless significant problem in analyzing data is that of multicollinearity, or non-orthogonality, among the independent variables (X). The condition of multicollinearity is not, of course, limited to the PCAR but is, in fact, a pervasive and potential problem in all research in the family studies field. In particular, in multiple linear regression is used to analyze a data set, as the magnitude of the relationships among the independent variables increases, less and less reliance can be placed on the results generated by an ordinary least squares solution (OLS). In particular, as multicollinearity increases, (a) the standard errors inflate, resulting in unstable parameter estimates; (b) the regression estimates tend to have large sampling variability; (c) the regression estimates may become so unstable that an incorrect sign can result; (d) important predictors may be eliminated from the model because of statistical nonsignificance. In short, there can be little confidence in the reliability of one's findings when working with highly multicollinear data. Most researchers confronted with multicollinearity have had only two possible "cures" between which to choose. They could either arbitrarily drop one or more of the "offending" variables from the model under consideration or factor analyze the highly correlated variables. Although either choice has the advantage of improving the accuracy of the statistical analysis, such improvement is always at the cost of creating potentially serious methodological and theoretical problems. The technique of factoring or clustering variables, while very useful as a data reduction technique, may produce highly misleading results when the goal is building an explanatory model.

2. Weighted Principal Component Analysis onto reference subspace

In order to overcome multicollinearity problem of explicative variables in PCAR, we proposed an alternative approach. Starting from the X and Y centred matrices we perform $p \times q$ simple linear regression between variables $y_k, k \in [1,\ldots,q]$ on $x_j, j \in [1,\ldots,p]$, than we ranged the regression coefficient in the matrix B

$$B = \begin{bmatrix} \beta_{11} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & \beta_{jk} & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \beta_{p,q} \end{bmatrix}$$

This approach can be seen as a generalization of the univariate Partial Least Square proposed by Garthwaite (1994). Thus, we construct a matrix $\tilde{X}$ obtained by the X matrix repeated q times. Multiplying the matrices $\tilde{X} \times B$ we obtain regression theorycal value $\tilde{y}_{kj}$:

$$\tilde{Y} = \begin{bmatrix} x_1\left(x_1'x_1\right)^{-1}x_1y_1 & \cdots & x_p\left(x_p'x_p\right)^{-1}x_py_1 \\ \vdots & \ddots & \vdots \\ x_1\left(x_1'x_1\right)^{-1}x_1y_q & \cdots & x_p\left(x_p'x_p\right)^{-1}x_py_q \end{bmatrix}$$

Using the orthogonal projector-operator we can write the preceding matrix as follow:

$$\tilde{Y} = P_j y_k \tag{3}$$

Where the matrix $P_j = x_j\left(x_j'x_j\right)^{-1}x_j'$ is the orthogonal projector-operator obtained by simple linear regression. The last step of our approach consists making a PCA of matrix $\tilde{Y}$. The criteria to optimize can be written in different ways:

$$\min\left[\sum_{k=1}^{q}\sum_{j=1}^{p}\left(P_j y_k - \phi_\alpha\phi_\alpha' P_j y_k\right)\right] = \min\left[\left(\tilde{X}B - \phi_\alpha\phi_\alpha'\tilde{X}B\right)\right] = \min\left[\left(\tilde{Y} - \phi_\alpha\phi_\alpha'\tilde{Y}\right)\right] \tag{4}$$

Where ϕ_α are the coordinates of the n observations on the α-th axis with $\alpha = 1...h$.
Thus our approach allows to observe the problem by different point of views: the minimization of y_k image onto the p reference subspaces constructed with the explicative variables, the principal component analysis of the matrix $\tilde{X}$ using as metric B and the principal component analysis of the matrix $\tilde{Y}$. Following the second remark we can define our proposal as a Weighted Principal Component Analysis (WPCA) of the matrix $\tilde{X}$ weighted by the regression coefficients y_k on x_j.
The WPCA can be made through the search of principal axis through the extraction of the eigenvalues and eigenvectors $(\alpha = 1,\ldots,h)$ of the expression:

$$B'\tilde{X}'\tilde{X}Bu_\alpha = \lambda_\alpha u_\alpha \tag{5}$$

Moreover, $B'\tilde{X}'\tilde{X}B$ being symmetrical and positive semidefinite, it derives that $\lambda_\alpha \geq 0$ and $u'_\alpha u_{\alpha'} = 0$ with $\alpha \neq \alpha'$;
The transition formulas are:

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} B\tilde{X}'v_\alpha \qquad v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \tilde{X}Bu_\alpha \tag{6}$$

and the principal components in subspace $R^{p\times q}$ are:

$$\phi_\alpha = \tilde{X}Bu_\alpha \tag{7}$$

While, the principal components in subspaces R^n are

$$\psi_\alpha = \left[X \mid Y\right]' v_\alpha \tag{8}$$

Bibliography

Belinfante, A. and Coxe, K. L. (1986). Principal components regression selection rules and application. Proc. Bus. & Econ. Sec., Amer. Stat. Assoc., 429-431.

Carrol, J.D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. Proc Am. Psychol. Ass.,

D'Ambra L., Lauro N., (1982), Analisi in componenti principali in rapporto ad un sottospazio di riferimento. Rivista di Statistica Applicata, n 1.

D'Ambra L., Sabatier R., Amenta P., (2001), Three Way Factorial Analysis: Synthesis and New Approaches. Rivista di Statistica Applicata, Vol. 13, n. 2.

Garthwaite P.H., (1994), An interpretation of partial least squares. JASA, 89

Hotelling, H., (1933), Analyses of complex of statistical variables in to principal components. Journal of Educational Psychology, vol. 24.

Kattering ,H. (1971). Canonica Analysis of several sets of variables. Biometrka n 58.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiaah (Ed.). Multivariate Analysis. (pp.391-420) New York: Academic Press.

