A control chart approach to select eigenvalues in Principal Component and Correspondence Analysis

Gilbert Saporta
CNAM, Chaire de statistique Appliquée & CEDRIC
292 rue Saint Martin
75141 Paris cedex 03, France
saporta@cnam.fr

1. Introduction

A vast literature (Cattell, Horn, Velicer) has been devoted to the assessment of the proper number of eigenvalues that have to be retained in Principal Components Analysis. Most of the publications are based on either (non-realistic) distributional assumptions for the underlying populations or on empirical criteria. Techniques that are based on bootstrap or cross-validation have been proposed (Diana, Krzanowski ,Wold) but requires a lot of computation. For Multiple Correspondence Analysis, the problem is similar, but there are few publications. In this paper a simple technique based on a control chart approach is proposed for selecting the number of principal components to retain for the analysis.

2. A new rule for PCA

In PCA with *p* standardised variables the most common rule is the Kaiser's criterion, which selects components that correspond to eigenvalues larger than 1. This rule if often supplemented by the consideration of the confidence interval based on Anderson's asymptotic result which states that with .95 confidence level the true eigenvalue l_i is such that $\hat{l}_i \exp(-2\sqrt{\frac{2}{n-1}}) < l < \hat{l}_i \exp(2\sqrt{\frac{2}{n-1}})$. Hence one should have $\hat{l}_i > \exp(2\sqrt{\frac{2}{n-1}}) \approx 1+2\sqrt{\frac{2}{n-1}}$ for large *n*

Forgetting about the fact that the $\hat{l}_i$ are an ordered sample of non independent variables, we may notice that they have a mean equal to 1 and that $\sum \hat{l}_i^2 = p + 2\sum_{i>j} r_{ij}^2$. Since the expectation of R^2 between two independent variables is $(n-1)^{-1}$ (exact for normal distributions and approximately true in other cases), we have $E(\sum \hat{l}_i^2) = p + \frac{p(p-1)}{n-1}$ and the variance of the set of the p $\hat{l}_i$ has thus an expectation equal to $\frac{p-1}{n-1}$. Like in control charts, we may assume that an eigenvalue is significantly greater than 1 if $\hat{l}_i > 1+2\sqrt{\frac{p-1}{n-1}}$. This method is thus a modification of Kaiser's rule.

3. The case of Multiple Correspondence Analysis

Let $\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2 | \ldots | \mathbf{X}_p)$ be the disjunctive table (indicator matrix) of p categorical variables . Then the number of non trivial eigenvalues is $q = \sum_{i=1}^{p} m_i - p$ and it is well known that : $\sum_{i=1}^{q} l_i = \frac{1}{p}\sum_{i=1}^{p} m_i - 1$ and $\sum_{i=1}^{q} l_i^2 = \frac{1}{p^2}\sum_{i=1}^{p}(m_i - 1) + \frac{1}{p^2}\sum\sum_{i \neq j} j_{ij}^2$

Let $S_l^2 = \frac{1}{q}\sum_{i=1}^{q}(l_i - \frac{1}{p})^2 = \frac{1}{q}\sum_{i=1}^{q} l_i^2 - \frac{1}{p^2}$ and denote $\sigma^2 = E(S_l^2) = \frac{1}{q}E(\sum_{i=1}^{q} l_i^2) - \frac{1}{p^2}$. When variables are pairwise independent nj_{ij}^2 is distributed as $c^2_{(m_i-1)(m_j-1)}$ which has an expectation equal to $(m_i\text{-}1)(m_j\text{-}1)$. Hence $E(\sum_{i=1}^{q} l_i^2) = E(\frac{q}{p^2} + \frac{1}{p^2}\sum\sum_{i\neq j}\frac{c_{ij}^2}{n}) = \frac{q}{p^2} + \frac{1}{p^2}\frac{1}{n}\sum\sum_{i\neq j}(m_i-1)(m_j-1)$ and

$s^2 = \frac{1}{qp^2}\frac{1}{n}\sum\sum_{i\neq j}(m_i-1)(m_j-1)$ (cf. Saporta & Ben Ammou).

Like in PCA we may assume that the $\frac{1}{p} \pm 2s$ interval should contain about 95% of the eigenvalues. Since the kurtosis of the set of eigenvalues is lower than for a normal distribution, the actual proportion is larger than 95%. The modification of Kaiser's rule consists here in retaining the eigenvalues greater than $\frac{1}{p} + 2s$.

4. Discussion

The proposed technique is distribution free since it uses only properties of the mean and of the dispersion of eigenvalues. In a recent paper (Karlis and al.), we have shown with extensive simulations that the method works better than other existing methodologies for PCA and it is conservative in the sense that it may reject eigenvalues that are larger than one but very close to one (such components are usually of little interest).

REFERENCES

Cattell, R.B. (1966) The Scree Test for the Number of Factors. *Multivariate Behavioral Research,* **1**, 245-276.

Diana G., Tommasi C. (2002) Cross validation methods in principal component analysis: a comparison, *Statistical Methods and Applications*, **11**, n°1,71-82

Horn, J.L. (1965) A Rationale and Test for the Number of Factors in Factor Analysis, *Psychometrika*, **30**, 179-185

Karlis, D., Saporta, G. and Spinakis A. (2003) A Simple Rule for the Selection of Principal Components, *Communications in Statistics, Theory and Applications*, **32**, 3, 643-666

Krzanowski, W. J. (1987) Cross-Validation for Principal Components Analysis, *Biometrics*, **43**, 575-584

Saporta G., Tambrea N. (1993) About the selection of the number of components in correspondence analysis, in J.Janssen & C.H.Skiadas, eds. *Applied Stochastic Models and Data Analysis*, World Scientific, p. 846-856

Saporta G., Ben Ammou S. (1998) Sur la normalité asymptotique des valeurs propres en ACM sous l'hypothèse d'indépendance des variables. *Revue de Statistique Appliquée* , **XLVI**, n°3, 21-35

Wold, S. (1978) Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Analysis, *Technometrics*, **20**, 397-405.

RESUMÉ

Le calcul de la dispersion des valeurs propres permet de modifier la règle classique de Kaiser de retenir les valeurs propres plus grandes que leur valeur moyenne (qui vaut 1 en ACP ou 1/p en ACM). En adoptant comme dans les cartes de contrôle une limite supérieure égale à la moyenne plus deux fois l'espérance de l'écart-type on a une règle simple et efficace, indépendante de la distribution des observations.