Some simple rules for interpreting outputs of principal components and correspondence analysis

Gilbert Saporta

IX International Symposium on Applied Stochastic Models and Data Analysis,
June 14-17, 1999, Lisbon, Portugal

SOME SIMPLE RULES FOR INTERPRETING OUTPUTS OF PRINCIPAL COMPONENTS AND CORRESPONDENCE ANALYSIS

Gilbert Saporta
CEDRIC, Conservatoire National des Arts et Métiers
292 rue Saint Martin, 75141 Paris Cedex03, France
saporta@cnam.fr

A large literature has been devoted to the assessment of the right number of eigenvalues in PCA and CA (two-way and multiple). Most of the publications are based on distributional assuptions for the sample, or on bootstrap techniques. After having recalled some of the most important results, we present simple thresholds based on a « control chart » approach for eigenvalues as well as for contributions, distances…

1. Principal components

We deal here with standardised data, and *n* equally weighted observations of *p* variables.

1.1 Choosing eigenvalues :

It is commonly accepted that significant eigenvalues should be greater than one and well separated. Kaiser's rule consists in discarding eigenvalues less than one. Confidence intervals based on Anderson's asymptotic distributions are frequently used in this context, despite it is well known that these results do not hold for a correlation matrix but only for a covariance matrix : if l_i is the true ith value and $\hat{l}_i$ is its estimation with a sample of size n :

$$\hat{l}_i \exp(-1.96\sqrt{\frac{2}{n-1}}) < l < \hat{l}_i \exp(1.96\sqrt{\frac{2}{n-1}})$$

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 the square correlation coefficient between two independent normal variables is $\frac{1}{n-1}$ we find that in this situation

$$E(\sum \hat{l}_i^2) = p + \frac{p(p-1)}{n-1}$$

and the variance 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 one if :

$$\hat{l}_i > 1 + 2\sqrt{\frac{p-1}{n-1}}$$

1.2 Contributions

Principal components *C* may often be considered as normally distributed if p and n are large, with zero mean and variance equal to l_k. The contribution of an observation i to the variance being defined by $\frac{1}{n}\frac{c_i^2}{l_k}$, $\frac{c_i^2}{l_k}$ is distributed as c_1^2. Hence a contribution might be considered as significantly large with α=0.05 if it is greater than 3.84/n.

1.3 Distance to the centroid

For normally distributed observations, the square distance to *0* is a weighted sum of p independent c_1^2 variables : $\sum_{i=1}^{p} l_i c_{i.1}^2$. Its expectation is $\sum_{i=1}^{p} l_i = p$ and its variance $2\sum_{i=1}^{p} l_i^2$. Observations with a square distance greater than :

$$p + 2\sqrt{2\sum_{i=1}^{p} l_i^2}\ .$$

may be considered as outliers.

In the case of independence, we may replace $2\sum_{i=1}^{p} l_i^2$ by its expectation and the upper bound becomes :

$$p + 2\sqrt{2p(1+\frac{p-1}{n-1})} \text{ or } p + 2.8\sqrt{p} \text{ For large } n$$

1.3 Quality of representation

A common but questionable index of proximity between an observation and a principal axis is *cos²(q)*. For the first axis we have $\tan^2(q) = \frac{l_1 c_1^2}{\sum_{i=1}^{p} l_i c_{i.1}^2}$. A crude approximation of

the expected value of $cos^2(q)$ gives $\frac{l_1}{p}$. (This is an exact value for the mean of the square correlations between variables and principal components). No simple formula seems available for the variance ; however we may use empirical 2σ-bounds. For an axis $cos^2(q)$ seems approximately distributed like beta.

A much better way to know if an observation is well represented in a subspace is to examine its square distance to this subspace which may be compared to a linear combination of c_1^2 variables. For the first principal plane we may consider that points with a square distance greater than

$$\sum_{i=3}^{p} l_i + 2\sqrt{2\sum_{i=3}^{p} l_i^2}$$

are not correctly projected.

2.Correspondence analysis of contingency tables

CA being a weighted PCA where weights depend on the data, results using chi-square distributions do not generally hold for contributions and quality of representation.This is also due to the usually small number of rows and columns in contingency tables. However it is still possible to derive empirical 2σ-bounds.

2.1 Distribution of eigenvalues

For a contingency table with m_1 rows and m_2 columns, the assumption that N is a realization of a multinomial distribution $M(n ; p_{ij})$ is realistic. In this framework the observed eigenvalues $\hat{l}_i$ are estimates of the eigenvalues l_i of nP, where P is the table of unknown joint probabilities. Lebart and O'Neill have proved the following results : if $l_i = 0$ then $\hat{l}_i$ has the same distribution as the corresponding eigenvalue of a Wishart matrix $W_{(m_1-1)(m_2-1)}(r, I)$ where r=min(m_1-1 ; m_2-1).

If $l_i \neq 0$, then $\sqrt{\hat{l}_i}$ is asymptotically normally distributed, but with parameters which depend on the unknown p_{ij}. Since it is difficult to test this hypothesis, some authors have proposed a bootstrap approach which unfortunately is not valid : since the empirical eigenvalues, on which the replication is based, are generally not null, we cannot observe the distribution based on Wishart matrices.

2.2 Malinvaud's test

Based upon the reconstitution formula, which is a weighted singular value decomposition of **N**: $n_{ij} = \left(n_{i.} n_{.j} / n\right)\left(1 + \sum_{i,j} a_{ik} b_{jk} / \sqrt{\hat{l}_k}\right)$, we may use a chi-square test comparing the observed n_{ij} from a sample of size n to the expected frequencies under the hypothesis H_k of only k non zero l_i .Weighted least squares estimates ot these expectations are precisely the $\tilde{n}_{ij}$ of the reconstitution formula with its first k terms. We then compute the classical chi-square goodness of fit statistic:

$$Q_k = \sum_{i,j} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}}$$

If k= 0 (independence) Q_0 is compared to a chi-square with (p-1)(q-1) degrees of freedom .

Under H_k, Q_k is asymptotically distributed like a chi-square with (p-k-1)(q-k-1) degrees of freedom.

E.Malinvaud 1987 proposed to use $n_{i.} n_{.j}/n$ instead of $\tilde{n}ij$ for the denominator which leads to a simple relation with the some of the discarded eigenvalues :

$$Q'_k = \sum_{i,j} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\frac{n_{i.} n_{.j}}{n}} = n(\mathbf{I} - l_1 - l_2 - \ldots - l_k) = n(l_{k+1} + l_{k+2} + \ldots + l_r)$$

Many experiences have proved that this procedure is efficient (Saporta, Tambrea, 1995).

3. Multiple correspondence analysis

3.1 Eigenvalues

Let $\mathbf{X} = \left(\mathbf{X}_1 | \mathbf{X}_2 | \ldots | \mathbf{X}_p\right)$ be a disjunctive table of p variables and q be the number of non trivial eigenvalues $q = \sum_{i=1}^{p} m_i - p$

Despite that MCA is an extension of CA, results of part 2 are not valid and one cannot use Malinvaud's test : elements of X being 0 or 1 and not frequencies, Q_k and Q'_k do not follow a chi-square distribution.

However it is possible to get informations about the dispersion of the q eigenvalues in particular cases (Ben Ammou, Saporta 1998).

It is well known that :

$$\sum_{i=1}^{q} l_i = \frac{1}{p}\sum_{i=1}^{p} m_i - 1 \text{ and } \sum_{i=1}^{q} l_i^2 = \frac{1}{p^2}\sum_{i=1}^{p}(m_i - 1) + \frac{1}{p^2}\sum\sum j_{ij}^2$$

When variables are pairwise independent $n j_{ij}^2$ is distributed as $c^2_{(m_i-1)(m_j-1)}$ and the expected value is $(m_i\text{-}1)(m_j\text{-}1)$, hence :

$$E(\sum_{i=1}^{q} l_i^2) = E(\frac{q}{p^2} + \frac{1}{p^2}\sum_{i\neq j}\sum\frac{c_{ij}^2}{n}) = \frac{q}{p^2} + \frac{1}{p^2}\frac{1}{n}\sum_{i\neq j}\sum(m_i - 1)(m_j - 1)$$

Let $S_l^2 = \frac{1}{q}\sum_{i=1}^{q}(l_i - \frac{1}{p})^2 = \frac{1}{q}\sum_{i=1}^{q} l_i^2 - \frac{1}{p^2}$. Then $E(S_l^2) = \frac{1}{q}E(\sum_{i=1}^{q} l_i^2) - \frac{1}{p^2}$

and we get :

$$E(S_l^2) = \frac{1}{p^2}\frac{1}{n}\frac{1}{q}\sum_{i\neq j}\sum(m_i - 1)(m_j - 1)$$

With $E(S_l^2) = \sigma^2$ we may assume that the $\frac{1}{p} \pm 2\sigma$ 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%.

3.2 Other statistics

Since MCA is similar to PCA, we may apply results of part 1 for distances, and contributions.

4 Concluding remarks

Of course, most of the preceding results are crude approximations and one has to be careful when using it. They work well for moderately large samples , but not for too large sample sizes : it is well known that in this case, even small and useless departures from the mean are statistically significant. Further developments are needed, but we think that a reasonable use of 2σ-bounds should be proposed in softwares.

References

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

E.Malinvaud, (1987) Data analysis in applied socio-economic statistics with special consideration of correspondence analysis, *Marketing Science Conference*, Jouy en Josas, France, 1987

L.Lebart (1976). The significance of Eigenvalues issued from Correspondence Analysis *COMPSTAT*, Physica Verlag, Vienna, p 38-45 .

L.Lebart, A.Morineau, M.Piron (1995). *Statistique exploratoire multidimensionnelle.* Dunod, Paris

M.E.O'Neill. (1978). Asymptotic distributions of the canonical correlations from contingency tables. *Australian Journal of. Statistics*. 20(1) p 75-82.

M.E.O'Neill (1978). Distributional expansion for canonical correlations from contingency tables . *Journal of the Royal. Statistical Society. B.* 40, n°3 p 303-312.

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