Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 2228 – 2237

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On the number of principal components: A test of dimensionality based on measurements of similarity between matrices

Stéphane Dray*

Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon; Université Lyon 1; CNRS; UMR 5558,
43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France

Received 5 October 2006; received in revised form 23 July 2007; accepted 23 July 2007
Available online 6 August 2007

Abstract

An important problem in principal component analysis (PCA) is the estimation of the correct number of components to retain. PCA is most often used to reduce a set of observed variables to a new set of variables of lower dimensionality. The choice of this dimensionality is a crucial step for the interpretation of results or subsequent analyses, because it could lead to a loss of information (underestimation) or the introduction of random noise (overestimation). New techniques are proposed to evaluate the dimensionality in PCA. They are based on similarity measurements, singular value decomposition and permutation procedures. A simulation study is conducted to evaluate the relative merits of the proposed approaches. Results showed that one method based on the RV coefficient is very accurate and seems to be more efficient than other existing approaches.

Keywords: Co-inertia criterion; Permutation procedure; RV coefficient; Singular value decomposition; Simulation study; Stopping rules

1. Introduction

In many fields such as ecology, chemometry or economy, multivariate analyses are widely used to describe and summarize large data sets (many variables and/or individuals) by removing any redundancy in the data. When only quantitative variables are considered, principal component analysis (PCA, Hotelling, 1933) is a standard approach. PCA searches for linear combinations of original variables to construct a set of new axes (principal components). The principal components are orthogonal by construction and account for decreasing amounts of variance in the data. Hence, PCA allows the reduction of the dimensionality (number of variables) of the data set but retains most of the original variability of the data. Classical output of PCA consists of graphical summaries that are then interpreted for the first few axes in order to reveal the underlying structure of a large data set. PCA is often used as a first step of data reduction in order to replace original variables by the first few principal components in subsequent analyses.

One crucial step of PCA concerns the choice of the number of axes to be retained for interpretation and subsequent analyses. This decision is often made according to practical considerations (e.g., two axes retained because only two dimensions can be represented on a sheet of paper) and not statistical ones. The consequences of this choice

* Tel.: +33 472 43 27 57; fax: +33 472 43 13 88.
E-mail address: dray@biomserv.univ-lyon1.fr
URL: http://biomserv.univ-lyon1.fr/∼dray/.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.07.015

are important: if the number of axes is not correctly estimated, one can introduce noise (overestimation) or loss of information (underestimation) in the analysis. A number of approaches to estimate the dimensionality of a data table (i.e., number of axes) have been proposed and evaluated in the literature (e.g., Jackson, 1993; Peres-Neto et al., 2005; Ferré, 1995). Jolliffe (2002, pp. 112–132) reviews the most frequently used approaches and distinguishes three types of rules. The first type corresponds to ad hoc rules which are intuitively plausible and work quite well in practice. The other two types have a formal basis. Some of these approaches make distributional assumptions (the second type) that are often unrealistic and, in practice, frequently overestimate the dimensionality. The third type corresponds to methods that do not require distributional assumptions. These methods use computationally intensive procedures such as permutation, cross-validation, bootstrap, or jackknife. In this paper, I will focus on the third type of methods. Recently, Peres-Neto et al. (2005) conducted an extensive simulation study and compare the merits of 20 tests of dimensionality. They have shown that the results of the different approaches are very sensitive to the level of correlations among the variables and to the number of observations and variables.

In this paper, I propose a new approach to estimate the dimensionality of a data table based on the link between PCA and the approximation of a matrix by another of lower rank (Eckart and Young, 1936) using singular value decomposition (SVD, Good, 1969). The similarity between a matrix and its approximation is measured, and its significance is evaluated by a permutation procedure (i.e., computationally intensive procedure). Finally, I conduct a simulation study to evaluate this new test of dimensionality.

2. Principal component analysis

2.1. Diagonalization of a covariance matrix

Let $\mathbf{X}$ be a table with the measurements of p centered variables (columns) for n individuals (rows). I consider the covariance matrix $\mathbf{C}=(1/n)\mathbf{X}^{\mathrm{t}}\mathbf{X}$. Note that if the variables in $\mathbf{X}$ have been standardized to mean 0 and variance 1, then $\mathbf{C}$ is a correlation matrix.

PCA is based on the diagonalization of $\mathbf{C}$ ($\mathbf{CA}=\mathbf{A\Lambda}$), where $\mathbf{\Lambda}$ is a diagonal matrix ($r \times r$) with the r non-null eigenvalues of $\mathbf{C}$ ($\mathbf{\Lambda}=diag(\lambda_1,\lambda_2,\ldots,\lambda_r)$) sorted in decreasing order ($\lambda_1>\lambda_2>\cdots>\lambda_r>0$) and $\mathbf{A}$ ($p\times r$) contains the r associated orthonormal eigenvectors ($\mathbf{A}^{\mathrm{t}}\mathbf{A}=\mathbf{I}_r$).

2.2. Singular value decomposition

Now, I consider the SVD of $\mathbf{X}^*=(1/\sqrt{n})\mathbf{X}=\mathbf{UDV}^{\mathrm{t}}$, where $\mathbf{D}$ is a diagonal matrix ($r\times r$) with the r non-null singular values ($\mathbf{D}=diag(d_1,d_2,\ldots,d_r)$) sorted in decreasing order ($d_1>d_2>\cdots>d_r>0$). The column vectors in $\mathbf{U}=[\mathbf{u}_1|\cdots|\mathbf{u}_r]$ ($n\times r$) and $\mathbf{V}=[\mathbf{v}_1|\cdots|\mathbf{v}_r]$ ($p\times r$) are orthonormal and verify $\mathbf{U}^{\mathrm{t}}\mathbf{U}=\mathbf{V}^{\mathrm{t}}\mathbf{V}=\mathbf{I}_r$. It can be easily demonstrated that $\mathbf{\Lambda}=\mathbf{D}^2$ and $\mathbf{A}=\mathbf{V}$. Hence, computation procedures involved in PCA can also be achieved through the SVD of $\mathbf{X}$. Moreover, it is well known that the SVD is intimately linked to the approximation of $\mathbf{X}$ by a matrix $\hat{\mathbf{X}}_m$ of rank m. The best approximation of $\mathbf{X}$ in the sense of least squares (i.e., minimization of $\|\mathbf{X}-\hat{\mathbf{X}}_m\|^2$) is given by (e.g., Good, 1969):

$$\hat{\mathbf{X}}_m=\sum_{i=1}^{m} d_i\mathbf{u}_i{\mathbf{v}_i}^{\mathrm{t}}=\sum_{i=1}^{m}\mathbf{X}_i,$$

where $\mathbf{X}_i=d_i\mathbf{u}_i{\mathbf{v}_i}^{\mathrm{t}}$. We denote the residuals:

$$\mathbf{R}_i=\mathbf{X}-\sum_{j=1}^{i-1} d_j\mathbf{u}_j{\mathbf{v}_j}^{\mathrm{t}}=\sum_{j=1}^{r} d_j\mathbf{u}_j{\mathbf{v}_j}^{\mathrm{t}}-\sum_{j=1}^{i-1} d_j\mathbf{u}_j{\mathbf{v}_j}^{\mathrm{t}}=\sum_{j=i}^{r} d_j\mathbf{u}_j{\mathbf{v}_j}^{\mathrm{t}}=\sum_{j=i}^{r}\mathbf{X}_j.$$

We can then rewrite $\mathbf{X}$:

$$\mathbf{X}=\mathbf{R}_1=\mathbf{X}_1+\mathbf{R}_2=\mathbf{X}_1+\mathbf{X}_2+\mathbf{R}_3=\sum_{j=1}^{i}\mathbf{X}_j+\mathbf{R}_{i+1}.$$

Using these results, the test of dimensionality in PCA can be restated as a problem of approximation of $\mathbf{X}$: Does the addition of an element $\mathbf{X}_i$ introduce relevant information or random noise? In other words, is $\hat{\mathbf{X}}_i$ a significantly better approximation than $\hat{\mathbf{X}}_{i-1}$? Using this point of view, I propose a new methodology to answer this question based on the measurement of the similarity between two matrices.

3. Measuring the similarity between two matrices

Let $\mathbf{Y}_1$ $(n \times p)$ and $\mathbf{Y}_2$ $(n \times q)$ be two matrices corresponding to two sets of observation made on the same n individuals. Let $\zeta(\mathbf{Y}_1)$ and $\zeta(\mathbf{Y}_2)$ be two associated configurations in $\Re^p$ and $\Re^q$, respectively. I assume that all variables have been centered to mean 0.

The RV coefficient (Escoufier, 1973; Robert and Escoufier, 1976) is a measurement of the closeness between the two configurations $\zeta(\mathbf{Y}_1)$ and $\zeta(\mathbf{Y}_2)$ and is defined by:

$$\mathrm{RV}(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{\mathrm{tr}(\mathbf{Y}_1{\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_2{\mathbf{Y}_2}^{\mathrm{t}})}{\sqrt{\mathrm{tr}({\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_1{\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_1)\mathrm{tr}({\mathbf{Y}_2}^{\mathrm{t}}\mathbf{Y}_2{\mathbf{Y}_2}^{\mathrm{t}}\mathbf{Y}_2)}}.$$

The numerator of the RV coefficient corresponds to the co-inertia criterion (Dray et al., 2003a), which is a measurement of the link between the two tables $\mathbf{Y}_1$ and $\mathbf{Y}_2$:

$$\mathrm{COI}(\mathbf{Y}_1, \mathbf{Y}_2) = \mathrm{tr}(\mathbf{Y}_1{\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_2{\mathbf{Y}_2}^{\mathrm{t}}) = \mathrm{tr}({\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_2{\mathbf{Y}_2}^{\mathrm{t}}\mathbf{Y}_1).$$

Another measurement of similarity is given by Gower (1971) and Lingoes and Schönemann (1974):

$$\mathrm{RLS}(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{\mathrm{tr}(({\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_2{\mathbf{Y}_2}^{\mathrm{t}}\mathbf{Y}_1)^{1/2})}{\sqrt{\mathrm{tr}({\mathbf{Y}_1}^{\mathrm{t}}\mathbf{Y}_1)\mathrm{tr}({\mathbf{Y}_2}^{\mathrm{t}}\mathbf{Y}_2)}}.$$

Peres-Neto and Jackson (2001) showed that a testing procedure based on RLS is as powerful or more powerful than the usual Mantel test (Mantel, 1967). Some elements of comparison between these three measurements can be found in Dray et al. (2003b). Note that RLS and RV coefficients are "scaled" so that they vary between 0 and 1.

4. Test of dimensionality in PCA

4.1. The RVDIM, COIDIM and RLSDIM statistics

In order to know if an element $\mathbf{X}_i$ adds relevant information to the decomposition $\hat{\mathbf{X}}_{i-1}$ of rank $i-1$, I propose to use the RV, COI and RLS coefficients. The test proposed, for the ith dimension, is based on the similarity between $\mathbf{X}_i$ and $\mathbf{R}_i$. The configuration $\zeta(\mathbf{X}_i)$ corresponds to the representation of the individuals in the unidimensional space formed by the ith principal axis, while $\zeta(\mathbf{R}_i)$ is the configuration of individuals in the $(r-i+1)$-dimensional space formed by the last $(r-i+1)$ principal axes. If the ith dimension adds relevant information, the two configurations are close and their similarity can be measured by the RV, COI or RLS coefficient. If the RV coefficient is used, the corresponding RVDIM statistic is defined by

$$\mathrm{RVDIM}(i) = \mathrm{RV}(\mathbf{X}_i, \mathbf{R}_i) = \frac{\mathrm{tr}({\mathbf{X}_i}^{\mathrm{t}}\mathbf{R}_i{\mathbf{R}_i}^{\mathrm{t}}\mathbf{X}_i)}{\sqrt{\mathrm{tr}({\mathbf{X}_i}^{\mathrm{t}}\mathbf{X}_i{\mathbf{X}_i}^{\mathrm{t}}\mathbf{X}_i)\mathrm{tr}({\mathbf{R}_i}^{\mathrm{t}}\mathbf{R}_i{\mathbf{R}_i}^{\mathrm{t}}\mathbf{R}_i)}}.$$

After some matrix manipulations, RVDIM can be rewritten:

$$\mathrm{RVDIM}(i) = \mathrm{RV}(\mathbf{X}_i, \mathbf{R}_i) = \frac{{d_i}^4}{\sqrt{{d_i}^4 \sum_{j=i}^{r} {d_j}^4}} = \frac{{\lambda_i}^2}{\lambda_i\sqrt{\sum_{j=i}^{r} {\lambda_j}^2}} = \frac{\lambda_i}{\sqrt{\sum_{j=i}^{r} {\lambda_j}^2}}.$$

For the COI and RLS coefficients, the corresponding COIDIM and RLSDIM statistics are

$$\mathrm{COIDIM}(i) = \mathrm{COI}(\mathbf{X}_i, \mathbf{R}_i) = \mathrm{tr}(\mathbf{X}_i{}^{\mathrm{t}}\mathbf{R}_i\mathbf{R}_i{}^{\mathrm{t}}\mathbf{X}_i) = \lambda_i{}^2$$

and

$$\mathrm{RLSDIM}(i) = \mathrm{RLS}(\mathbf{X}_i, \mathbf{R}_i) = \frac{\mathrm{tr}((\mathbf{X}_i{}^{\mathrm{t}}\mathbf{R}_i\mathbf{R}_i{}^{\mathrm{t}}\mathbf{X}_i)^{1/2})}{\sqrt{\mathrm{tr}(\mathbf{X}_i{}^{\mathrm{t}}\mathbf{X}_i)\mathrm{tr}(\mathbf{R}_i{}^{\mathrm{t}}\mathbf{R}_i)}} = \frac{\sqrt{\lambda_i}}{\sqrt{\sum_{j=i}^{r} \lambda_j}}.$$

4.2. Testing procedure

I propose two randomization procedures to test the significance of the RVDIM, COIDIM and RLSDIM coefficients. I present the procedures only for the RVDIM statistic but it is strictly equivalent for COIDIM and RLSDIM. The first procedure is based on the permutation of values within each column of the original table **X**. The complete procedure consists of:

(1) Perform the SVD of **X**.
(2) Compute the observed values of RVDIM(i) for each axis i.
(3) Repeat a large number of times (e.g., 999 times):
  (3.1) Randomize the values within each column of **X**.
  (3.2) Perform the SVD of the permuted matrix.
  (3.3) Compute RVDIM(i) for each axis i of the permuted matrix.
(4) Estimate the p-value p_i for the ith axis (e.g., (number of random values equal to or larger than the observed $+1$)/1000). Note that the observed value is included as one of the possible values of the randomization procedure.
(5) Choose a significance level α_i for the ith axis. The procedure is defined by keeping the axes $1, \ldots, i$ where $p_i < \alpha_i$ and $p_{i+1} > \alpha_{i+1}$. If $p_1 > \alpha_1$, then no axes are retained.

The decomposition of **X** on the ith axis ($\mathbf{X}_i$) corresponds to the first order decomposition of $\mathbf{R}_i$. Hence, it is equivalent to test the ith dimension of **X** or the first dimension of $\mathbf{R}_i$. The second randomization procedure is based on this equivalence. The test of the ith dimension is achieved by permuting within each column of $\mathbf{R}_i$, while the previous randomization is based on the permutations of the values of the original table. The complete procedure consists of:

(1) Perform the SVD of **X**.
(2) Compute the observed values of RVDIM(i) for each axis i.
(3) For each dimension i ($1 \leqslant i \leqslant r$), repeat a large number of times (e.g., 999 times):
  (3.1) Randomize the values within each column of $\mathbf{R}_i$.
  (3.2) Perform the SVD of the permuted matrix.
  (3.3) Compute RVDIM(*1*) for the first axis of the permuted matrix.
(4) Estimate the p-value p_i for the ith axis (e.g., (number of random values equal to or larger than the observed $+1$)/1000). Note that the observed value is included as one of the possible values of the randomization procedure.
(5) Choose a significance level α_i for the ith axis. The procedure is defined by keeping the axes $1, \ldots, i$ where $p_i < \alpha_i$ and $p_{i+1} > \alpha_{i+1}$. If $p_1 > \alpha_1$, then no axes are retained.

For both randomization procedures, one can choose a significance level of $\alpha = \alpha_i = 0.05$. In order to control the increase in Type I error due to multiple testing, a correction can be used. For instance, one can adjust the significance levels α_i using the sequential Bonferroni procedure (Holm, 1979). The adjusted significance levels are then $\alpha_i = \alpha / i$.

It should be noticed that the permutation procedure (permute values within each column) breaks the links between the variables but does not modify the structure of each variable. Hence, the procedure is only designed to test for the dimensionality that is due to the correlations between variables but not the part that could be related to their variances. The proposed test can then be used for PCA on correlation matrices, but it is not suitable for covariance matrices.

5. Simulation study

I have conducted a simulation study to evaluate the performance of the testing procedures based on RVDIM, COIDIM, and RLSDIM. I also included the classical procedure based on the eigenvalues (LBD), which has been also evaluated in Peres-Neto et al. (2005). The protocol is the same as the one used by Peres-Neto et al. (2005). I briefly present this protocol. For more details, the reader should consult Peres-Neto et al. (2005). I considered either 9 or 18 variables. Various scenarios have been considered using 18 structures of correlation matrices where between-groups (0.5, 0.3, 0.2, 0.1, or 0) and within-groups (0.8, 0.5 or 0.3) correlations of variables were fixed (Fig. 1). Sample sizes have been

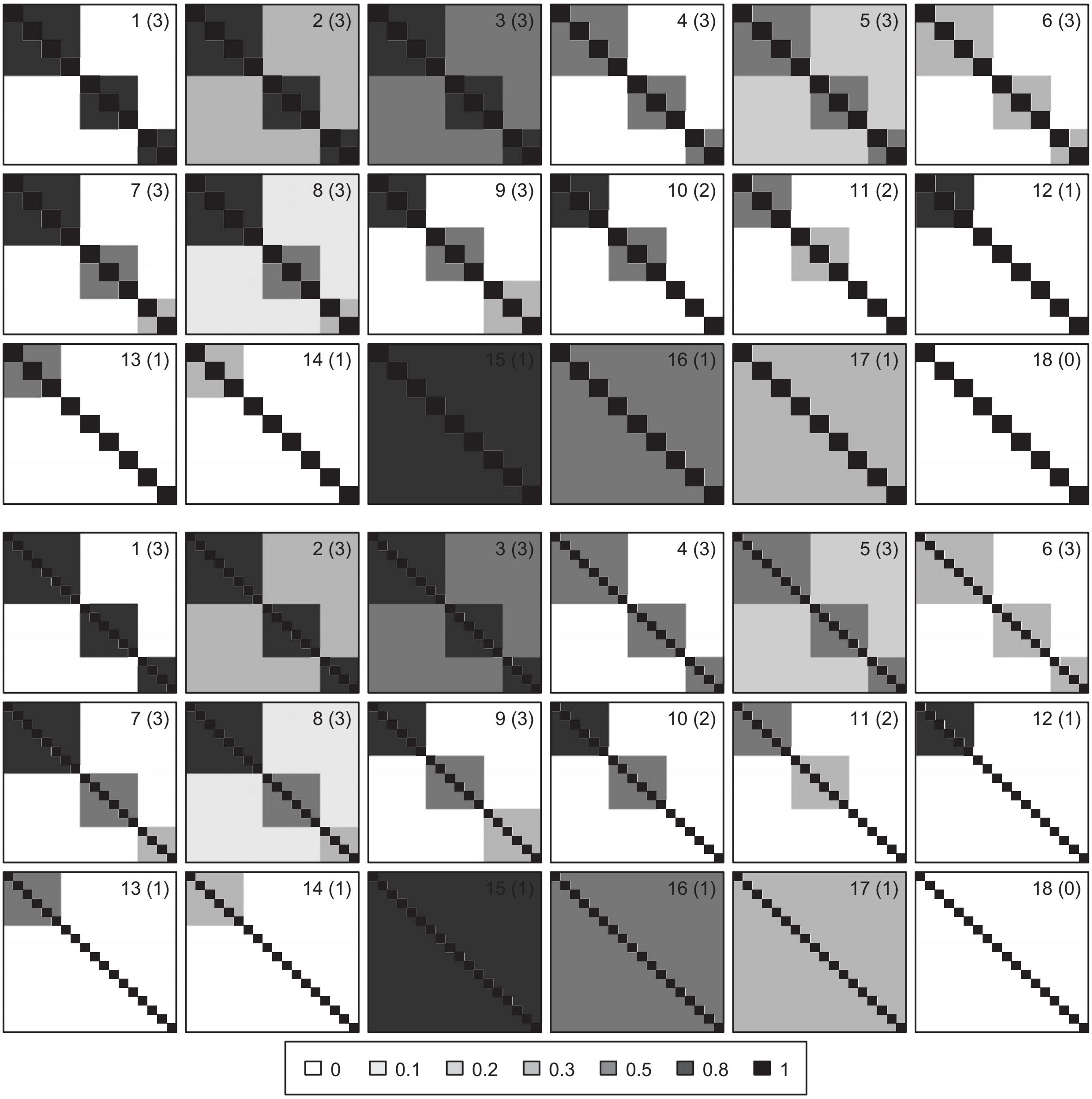

Fig. 1. The 18 correlation matrices considered in the simulation study (9 variables above, 18 variables below). Values of the correlations are represented by the grayscale. Numbers in parentheses correspond to the known dimensionality.

Table 1
Results for the first (-1) and the second (-2) permutation procedures without or with the sequential Bonferroni adjustment (B-)

Procedures	9 variables						18 variables					
	Normal		Exponential		Exponential^3		Normal		Exponential		Exponential^3	
	$n = 30$	$n = 50$	$n = 30$	$n = 50$	$n = 30$	$n = 50$	$n = 60$	$n = 100$	$n = 60$	$n = 100$	$n = 60$	$n = 100$
COIDIM-1[a]	0.757	0.534	0.766	0.539	**0.829**	**0.630**	0.179	0.079	0.181	0.086	0.288	0.214
RVDIM-1	0.940	1.254	1.079	1.470	1.981	2.540	2.511	3.613	2.914	4.243	6.484	8.193
RLSDIM-1	1.183	1.572	1.403	1.867	2.498	3.051	3.472	4.294	4.038	5.112	8.246	9.475
B-COIDIM-1[b]	0.795	0.566	0.803	0.571	0.871	0.666	0.191	0.083	0.192	0.089	0.286	0.203
B-RVDIM-1	0.683	0.869	0.801	1.011	1.457	1.776	1.589	2.767	1.872	3.118	4.555	6.148
B-RLSDIM-1	0.838	1.172	0.996	1.354	1.830	2.233	2.592	3.485	2.938	3.982	6.201	7.462
COIDIM-2[c]	2.182	2.840	2.219	2.865	1.854	2.185	2.403	4.544	2.633	4.912	1.570	2.372
RVDIM-2	0.719	0.628	0.776	0.703	1.050	0.909	0.157	0.213	0.172	0.246	0.328	0.330
B-COIDIM-2[d]	1.443	2.202	1.515	2.264	1.312	1.424	0.597	1.871	0.701	2.179	0.470	0.579
B-RVDIM-2	**0.681**	**0.442**	**0.723**	**0.488**	1.041	0.819	**0.076**	**0.068**	**0.085**	**0.082**	**0.271**	**0.201**

Values of the average absolute difference between the known dimensionality and the result of the tests are presented according to number of variables, sample size and distribution type. Values in bold indicate the best procedure for each scenario.

[a] Same results for LBD-1.

[b] Same results for B-LBD-1.

[c] Same results for RLSDIM-2 and LBD-2.

[d] Same results for B-RLSDIM-2 and B-LBD-2.

set to 30 and 50 observations (9 variables) and 60 and 100 observations (18 variables). For each scenario, I generated 1000 samples following the normal, exponential and exponential^3 distributions. The two testing procedures have been used with 3999 permutations and $\alpha = 0.05$ while Peres-Neto et al. (2005) used only 999 permutations.

The efficiency of each method was evaluated by computing the absolute difference between the known dimensionality and the estimated number of axes. Averages of these absolute differences over the 1000 samples for the 18 correlation matrices are presented in Table 1. These values produce a measure of the overall quality of the different methods but give no indication of the eventual tendency of an approach to underestimate or overestimate the number of principal components. The procedure based on the RVDIM statistic using the second testing procedure with the Bonferroni adjustment (B-RVDIM-2) is the most efficient one. It produces the best results for 10 out of 12 scenarios, and its average absolute difference is always less than one component except for one case. In the case of correlation matrices with 18 variables, the results become very accurate for B-RVDIM-2. It is noticeable that for some cases (e.g., RVDIM-1, RLSDIM-1, COIDIM-2, LBD-2), results become worse when the number of observations and/or the number of variables increases. This result could appear quite surprising, however, it reflects the fact that the second permutation procedure is more suitable when the statistic is “scaled” (RLS, RV) while the first permutation procedure is not adapted for these cases. Lastly, it could seem surprising that for some statistics (e.g., COIDIM) results are worse with the Bonferroni-corrected procedure (e.g., COIDIM-1 versus B-COIDIM-1). The Bonferroni adjustment aims to reduce Type I error by using a smaller significance level for each axis (except for the first one). This choice implies that the number of significant axes as well as the estimated dimensionality will be smaller. Results, not presented in this paper, show that the classical procedure (COIDIM-1) tends to underestimate the true dimensionality. Then, the sequential Bonferroni correction (B-COIDIM-1) would increase this underestimation, and this explains why the procedure produces worse results.

Detailed results for the B-RVDIM-2 procedure are presented in Tables 2 and 3 for normally distributed samples. For each correlation matrix, we calculated the percentage of deviations between the result of the testing procedure B-RVDIM-2 and the known dimensionality.

The method is very precise for uniform matrices (matrices 15–17) and uncorrelated data (18). For samples with 50 observations (9 variables), the method is quite efficient for non-uniform matrices when between-groups correlations are greater than 0.3 (percentage of correct assessments varies between 50.3% and 97.9% for matrices 1–5,12–13). When between-groups correlations are equal to 0.3, the dimensionality is often underestimated. The results are very accurate for samples of 100 observations (18 variables). In this case, the percentage of correct assessments varies between 84.8%

Table 2
Percentage of deviations between the result of the testing procedure B-RVDIM-2 and the known dimensionality

Matrix	9 variables ($n = 30$)							9 variables ($n = 50$)						
	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$
1		3.7	0.4	80.3	13.5	1.7	0.4		0.2		77.7	18.7	2.4	1.0
2		7.3	0.7	79.6	10.7	1.4	0.3		0.2		80.9	15.5	2.5	0.9
3		7.0	6.6	73.5	11.3	1.4	0.2		0.3	0.6	81.0	15.2	2.2	0.7
4	9.9	26.0	33.7	28.8	1.4	0.2			5.9	16.7	68.0	8.0	0.7	0.7
5	1.6	42.0	38.4	16.7	1.1	0.2			14.4	29.2	50.3	5.5	0.6	
6	45.9	42.8	10.3	1.0				18.9	45.4	29.4	6.0	0.3		
7		25.1	59.1	12.4	2.7	0.4	0.3		2.6	66.1	17.4	7.1	1.9	4.9
8		21.3	64.0	12.0	2.0	0.6	0.1		3.0	68.4	17.4	5.3	1.7	4.2
9	1.7	33.2	39.8	23.4	1.4	0.4	0.1		6.5	30.1	54.2	5.2	1.2	2.8
10		0.5	34.6	63.1	1.8					8.1	91.0	0.8	0.1	
11		41.5	49.5	8.9	0.1				7.4	55.7	36.0	0.6	0.3	
12			0.2	96.3	3.4	0.1					97.9	2.1		
13			38.9	59.6	1.5					8.0	89.1	2.9		
14			80.8	18.7	0.5					61.2	38.3	0.5		
15				94.1	5.8	0.1					92.6	7.2	0.2	
16				95.6	4.2	0.2					95.6	4.2	0.2	
17			1.4	94.4	3.9	0.3					94.2	5.7	0.1	
18				95.0	5.0						93.7	6.2	0.1	

Results are presented for each correlation matrix based on normally distributed samples for 9 variables (30 and 50 observations). Differences of zero indicate perfect assessment.

Table 3
Percentage of deviations between the result of the testing procedure B-RVDIM-2 and the known dimensionality

Matrix	18 variables ($n = 60$)							18 variables ($n = 100$)						
	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$	$\leqslant 3$	-2	-1	0	1	2	$\geqslant 3$
1				92.6	7.3	0.1					91.4	8.2	0.4	
2				92.6	7.0	0.4					91.3	8.3	0.4	
3				92.9	6.1	1.0					91.6	8.2	0.2	
4				93.8	6.1	0.1					92.3	7.4	0.3	
5			1.2	93.8	4.7	0.3					93.8	5.8	0.3	0.1
6		1.3	18.7	77.1	2.9					1.0	96.5	2.4	0.1	
7			5.4	87.3	6.2	1.1				0.1	84.8	10.1	3.6	1.4
8			12.6	81.2	4.7	1.3	0.2			0.7	85.9	9.9	1.8	1.7
9			0.2	95.4	4.1	0.3					93.6	5.9	0.4	0.1
10				98.7	1.3						98.6	1.4		
11			2.9	94.6	2.5						98.3	1.7		
12				96.1	3.9						97.0	3.0		
13				95.8	4.1	0.1					97.9	2.1		
14			4.2	93.3	2.5						98.4	1.6		
15				96.5	3.4	0.1					95.8	4.2		
16				97.4	2.6						95.2	4.7	0.1	
17				95.9	4.1						95.7	4.2	0.1	
18				94.0	6.0						95.8	4.2		

Results are presented for each correlation matrix based on normally distributed samples for 18 variables (60 and 100 observations). Differences of zero indicate perfect assessment.

and 98.6% for the 18 correlation matrices and the majority of errors consists of an overestimation of the dimensionality. Results in Table 1 suggest that the accuracy of the B-RVDIM-2 procedure is more sensitive to the number of variables than to the number of individuals. For 18 variables, the results did not vary much with the number of observations (Table 3). For 9 variables, results are slightly better for a higher number of observations (Table 2).

Table 4
Pollution data: values of the four statistics and p-values obtained with 9999 permutations for the first (p_1) and the second (p_2) permutation procedures

Axis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
LBD	4.611	3.096	2.506	1.334	1.227	0.787	0.704	0.435	0.263	0.200	0.178	0.129	0.112	0.082	0.050	0.020
p_1	0.0001	0.0001	***0.0001***	0.9037	0.8365	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
p_2	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	***0.0001***
COI	21.266	9.584	6.278	1.779	1.506	0.620	0.495	0.189	0.069	0.040	0.032	0.017	0.013	0.007	0.002	0.000
p_1	0.0001	0.0001	***0.0001***	0.9037	0.8365	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
p_2	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	***0.0001***
RV	0.712	0.682	0.754	0.611	0.710	0.646	0.757	0.716	0.621	0.602	0.670	0.656	0.753	0.836	0.927	1.000
p_1	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0042	**0.0001**	0.0169	0.0009	0.0001	*0.0003*	0.8145
p_2	0.0001	0.0001	***0.0001***	0.1736	0.0941	0.0005	0.0001	0.0001	0.0040	0.2268	0.0002	0.0009	0.0001	0.0001	0.0001	0.0001
RLS	0.541	0.528	0.559	0.492	0.541	0.516	0.569	0.544	0.504	0.509	0.558	0.573	0.651	0.734	0.844	1.000
p_1	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001	0.0020	0.0002	0.0001	***0.0003***	1.0000
p_2	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	***0.0001***

Bold values indicate the dimension retained with the sequential Bonferroni adjustment, while italic values correspond to the dimension retained without adjustment.

Here, the results have been evaluated in terms of the estimation of the number of components. Tables 2 and 3 can also be used to evaluate the testing procedure in a more traditional way. Overestimation corresponds to cases where the null hypothesis was true and was rejected (i.e., Type I error). On the contrary, underestimation corresponds to cases where the null hypothesis was false and has not been rejected (i.e., Type II error). In this context, results in Tables 2 and 3 show that the testing procedure has correct Type I error in most cases, but it is too liberal for matrices with 3 dimensions and especially for samples with 50 observations (even if the sequential Bonferroni correction is used). For these cases, estimations could be improved by reducing the significance level (α). As said before, Type II error was quite high when correlations are equal to 0.3, but it decreases when the number of variables and individuals increases. If one wants to reduce underestimation, a higher value for the significance level can be chosen. However, it must be noticed that increasing α would also increase the proportion of false positives (i.e., overestimation). The effect of sample size can also be evaluated. Increasing the number of observations produces a higher power. Hence, for a given correlation structure, underestimation is always higher for a smaller number of observations.

Relative merits of the B-RVDIM-2 procedure can be compared to those of other approaches described in Peres-Neto et al. (2005) since I use the same protocol for the simulation study. However, it must be noticed that I use 3999 permutations for the testing procedures while Peres-Neto et al. (2005) used only 999 permutations. Results for the LBD-1 procedure (which has been used by Peres-Neto et al., 2005) are very consistent between the two papers, except for the exponential^3 distribution. This unexplained difference is quite important and I decided to exclude the exponential^3 distribution for comparison purposes. The global accuracy of the method can be assessed by averaging among all scenarios (except the exponential^3 distribution) of the values presented in Table 1. This average value is equal to 0.390 for LBD-1 (0.396 in Peres-Neto et al., 2005). Among all the methods evaluated in the two papers, the B-RVDIM-2 procedure produces the best results with an average value of 0.331.

6. Example

I illustrate the use of the different procedures on a real data set. I consider the data of McDonald and Schwing (1973), assembled to study air pollution, which are available at http://lib.stat.cmu.edu/datasets/pollution and have been used by Besse (1992) and Besse and de Falguerolles (1993) in the context of evaluation of dimensionality in PCA. Sixteen variables were measured for 60 areas in the USA and three of them (HC, NO_x, SO_2) were log transformed. I used PCA on a correlation matrix (i.e., variables were centered and scaled to unit variance) and applied the different testing procedures with 9999 permutations. The screeplot (Table 4) and procedures proposed by Besse and de Falguerolles (1993) suggests three dimensions. This is confirmed by six procedures (RVDIM-2, COIDIM-1, LBD-1, B-RVDIM-2, B-COIDIM-1, B-LBD-1). For the 10 other procedures, the estimated dimensionality varies between 11 and 16.

7. Discussion and conclusions

Results obtained with the B-RVDIM-2 procedure are very promising. Simulation results showed that B-RVDIM-2 tends to underestimate the dimensionality in the case of low correlation structures (scenarios with 9 variables). When the number of variables associated with each component increases (scenarios with 18 variables), the method is very accurate (percentage of correct assessments varies between 84.8% and 98.6%) with a small tendency to overestimation which corresponds to Type I error. This leads to the inclusion and the interpretation of random noise. Fava and Velicer (1992) showed that consequences of overestimation are more problematic in cases of low correlation structures and small sample sizes. In this context, results obtained by the B-RVDIM-2 procedure are quite satisfactory, because it tends to underestimate the number of components with low correlations and small sample size. All of these results have been obtained by a simulation study, which is the *only way to test and to compare the methods* (Ferré, 1995, p. 670). Joliffe (2002, p. 130) stated that *simulation of multivariate data sets can always be criticized as unrepresentative, because they can never explore more than a tiny fraction of the vast range of possible correlations and covariance structures*. Like other simulation studies, this work suffers from this problem, and it must be stated that the hypothetical superiority of the B-RVDIM-2 procedure cannot be extended into a more general context than the one presented in this paper.

The approach proposed is a computationally intensive procedure. For the first testing procedure, n_p SVDs must be performed while the second one requires $n_p \cdot n_a$ SVDs where n_p is the number of permutations and n_a the number of tested dimensions. Moreover, the use of the Bonferroni correction implies a test where $\alpha_i = \alpha / i$, and a great number of permutations could be required to have enough precision to properly test a principal component for large i. I have implemented an R function which calls C code and LAPACK Fortran routines. The test of the 20 axes of a 500 by 20 table using the B-RVDIM-2 procedure with 999 permutations requires 36.34 s on a desktop computer with an Intel Pentium-4 3.00 GHz processor (R version 2.4.0 on a Debian distribution). In order to reduce the computation time of the second procedure, one can define a reduced number of tested dimensions. Looking at the screeplot and/or using the testing procedure with less permutations could help to choose this number. When the number of individuals (respectively, variables) is much larger than the number of variables (respectively, individuals), SVD can be very time consuming. In order to speed up this step, one can use eigendecomposition in the smaller dimension and matrices multiplications to compute eigenvalues and matrix approximations instead of SVD.

For the two procedures, the estimated dimensionality is determined by the number of subsequent significant tests. In some cases, this approach can be problematic. For instance, if there is a set of nearly equal eigenvalues that are well-separated from all other eigenvalues, the associated subspace is well-defined and stable. However, individual components are unstable and the procedure, which tests components one by one, could miss important information. In this case, it would be more suitable to test the subspace induced by these components rather than individual components. The procedure would also be inefficient if a random component has been permuted with a relevant one. In this case, the test of the random component would be non-significant and the relevant one could not be selected even if its associated test is highly significant. This is due to the procedure of estimation of the dimensionality which does not authorize the selection of the subspace associated with the first k components if one of the first $k-1$ components is not significant. In this context, it is useful to study the stability of the subspace associated with the selected components. Several approaches have been proposed for measuring the stability of components. For instance, Daudin et al. (1988) used the bootstrap method and a stability measure that is equivalent to the RV coefficient. Besse (1992) and Besse and de Falguerolles (1993) proposed another measure of stability and gave an asymptotic jackknife approximation. Sinha and Buchanan (1995) presented another approach and recommended that their measure be used in conjunction with some ad hoc rules (e.g., elbow in the screeplot) to determine the interpretability of principal components. I agree that the stability of components is an important criterion which must be considered for dimensionality estimation. However, most approaches devoted to evaluating the stability of principal components do not contain a statistical procedure to evaluate the dimensionality and require a pragmatic step for this task (e.g., visual comparison of boxplots in Besse and de Falguerolles, 1993). The use of the B-RVDIM-2 procedure in conjunction with a stability measurement could provide an efficient way to estimate the dimensionality. Further work is required to evaluate whether the RVDIM statistic could be used for this task.

The objective of this paper was to provide a new point of view for testing the dimensionality in PCA. The simultaneous use of similarity measurements, SVD and randomization procedures offers new perspectives on this problem. The coefficients proposed are quite simple to compute, as they are functions of eigenvalues. Moreover, its use is facilitated by the implementation of the complete procedure as a function of the R package ade4 (Chessel et al., 2004). The use of similarity measurements could be used to evaluate the dimensionality in other multivariate methods. Numerous techniques such as correspondence analysis (Benzécri, 1969), multiple correspondence analysis (Tenenhaus and Young, 1985) or coupling techniques such as co-inertia analysis (Dolédec and Chessel, 1994; Dray et al., 2003a), redundancy analysis (van den Wollenberg, 1977) or canonical correlation analysis (Hotelling, 1936) are based on the SVD of a product of matrices. Further works are needed to develop randomization procedures adapted to these other multivariate methods.

Acknowledgments

I would like to thank Pedro Peres-Neto for discussions and Daniel Chessel for comments on earlier drafts of the manuscript. Access to the computers of CCIN2P3 for the simulation study is gratefully acknowledged, with special thanks to Simon Penel. I acknowledge the constructive comments of the two anonymous reviewers and the thoughtful reading and corrections by the associate editor, which improved the manuscript.

References

Benzécri, J., 1969. Statistical analysis as a tool to make patterns emerge from data. In: Watanabe, S. (Ed.), Methodologies of Pattern Recognition. Academic Press, New York, pp. 35–60.

Besse, P., 1992. Pca stability and choice of dimensionality. Statist. Probab. Lett. 13, 405–410.

Besse, P., de Falguerolles, A., 1993. Application of resampling methods to the choice of dimension in principal component analysis. In: Hardle, W., Simar, L. (Eds.), Computer Intensive Methods in Statistics. Physica-Verlag, Heidelberg, pp. 167–176.

Chessel, D., Dufour, A.-B., Thioulouse, J., 2004. The ade4 package—I: one-table methods. R News 4, 5–10.

Daudin, J., Duby, C., Trecourt, P., 1988. Stability of principal component analysis studied by the bootstrap method. Statistics 19, 241–258.

Dolédec, S., Chessel, D., 1994. Co-inertia analysis: an alternative method for studying species–environment relationships. Freshwater Biol. 31, 277–294.

Dray, S., Chessel, D., Thioulouse, J., 2003a. Co-inertia analysis and the linking of ecological data tables. Ecology 84, 3078–3089.

Dray, S., Chessel, D., Thioulouse, J., 2003b. Procrustean co-inertia analysis for the linking of multivariate data sets. Ecoscience 10 (1), 110–119.

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1 (3), 211–218.

Escoufier, Y., 1973. Le traitement des variables vectorielles. Biometrics 29, 750–760.

Fava, J., Velicer, W., 1992. The effects of overextraction on factor and component analysis. Multivariate Behavioral Res. 27, 387–415.

Ferré, L., 1995. Selection of components in principal component analysis: a comparison of methods. Comput. Statist. Data Anal. 19, 669–682.

Good, I., 1969. Some applications of the singular decomposition of a matrix. Technometrics 11, 823–831.

Gower, J., 1971. Statistical methods of comparing different multivariate analyses of the same data. In: Hodson, F., Kendall, D., Tautu, P. (Eds.), Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press, Edinburgh, pp. 138–149.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6, 65–70.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educational Psych. 24, 417–441.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Jackson, D., 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74 (8), 2204–2214.

Jolliffe, I., 2002. Principal Component Analysis. second ed. Springer, Berlin.

Lingoes, J., Schönemann, P., 1974. Alternative measures of fit for the Schönemann–Carrol matrix fitting algorithm. Psychometrika 39, 423–427.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27 (2), 209–220.

McDonald, G., Schwing, R., 1973. Instabilities of regression estimates relating air pollution to mortality. Technometrics 15, 463–481.

Peres-Neto, P., Jackson, D., 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129, 169–178.

Peres-Neto, P., Jackson, D., Somers, K., 2005. How many principal components? stopping rules for determining the number of non-trivial axes revisited. Comput. Statist. Data Anal. 49, 974–997.

Robert, P., Escoufier, Y., 1976. A unifying tool for linear multivariate statistical methods: the RV coefficient. Appl. Statist. 25, 257–265.

Sinha, A., Buchanan, B., 1995. Assessing the stability of principal components using regression. Psychometrika 60, 355–369.

Tenenhaus, M., Young, F., 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50 (1), 91–119.

van den Wollenberg, A., 1977. Redundancy analysis, an alternative for canonical analysis. Psychometrika 42 (2), 207–219.