Article

Permutation tests to estimate significances on Principal Components Analysis

Vasco M. N. C. S. Vieira

CCMAR-Centre of Marine Sciences, University of Algarve, Campus of Gambelas, 8005-139 Faro, Portugal

E-mail: vvieira@ualg.pt

Received 24 January 2012; *Accepted 1 March 2012*; *Published online 1 June 2012*

IAEES

Abstract

Principal Component Analysis is the most widely used multivariate technique to summarize information in a data collection with many variables. However, for it to be valid and useful the meaningful information must be retained and the noisy information must be sorted out. To achieve it an index from the original data set is estimated, after which three classes of methodologies may be used: (i) the analytical solution to the distribution of the index under the assumption the data has a multivariate normal distribution, (ii) the numerical solution to the distribution of the index by means of permutation tests without any assumption about the data distribution and (iii) the bootstrap numerical solution to the percentiles of the index and the comparison to its assumed value for the null hypothesis without any assumption about the data distribution. New indices are proposed to be used with permutation tests and compared with previous ones from application to several data sets. Their advantages and draw-backs are discussed together with the adequacy of permutation tests and inadequacy of both bootstrap techniques and methods that rely on the assumption of multivariate normal distributions.

Keywords multivariate; permutation tests; principal components analysis; randomization; significance; stopping rules.

1 Introduction

Often the ecologist is faced with the problem of having data with many variables on a certain subject. It is his wish that information can be summarized into a small number of components. Optimality is achieved when the researcher gets the smaller possible number of components with the least loss of relevant information (Gauch Jr, 1982; Jackson, 1991; Manly, 1986; Zhang, 2011c). The Principal Components Analysis (PCA) is the benchmark in these linear dimension reduction techniques. However, when performing it the ecologist faces several problems. Among these, three are most common:

(i) Is it worth undergoing a PCA? It depends whether the variables are correlated enough so that the *p* variables can be reduced to *k* <*p* principal components or otherwise the *p* variables are badly correlated and an analysis on the principal components brings no advantage from a separate analysis on each of the variables (Jackson 1991).

(ii) Which principal components to use? The researcher whishes to know which eigenvectors are expressing meaningful correlations among variables and which are must probably just explaining error (Gauch Jr, 1982; Zwick and Velicer, 1986; Jackson, 1991; Jolliffe, 2002; Peres-Neto et al., 2005).

(iii) What is each principal component's interpretability? The researcher wishes to understand what is the pattern expressed by each principal component and in doing so he needs to know which variables are of real interest to that particular principal component (Jackson, 1991, Manly, 1986, Jolliffe, 2002, Peres-Neto et al., 2003).

It is the objective of this work to discuss simple solutions to the questions above in a language accessible to the common researcher. The answers are given by permutation tests, a class of randomization tests which application to PCA has already proven to give reliable results (ter Braak, 1988; Dijksterhuis and Heiser, 1995; Peres-Neto et al., 2005). These were tested on many data sets among which a set of twelve abiotic variables and a set of eleven seagrass variables both from Ria Formosa lagoonary system (Cabaço et al., 2008), a set of nine variables characterizing employment in Europe in the 1970's (Manly, 1986), Bumpus' set of five morphometric measurements on female sparrows (Manly, 1986) and an artificially generated set of ten uncorrelated variables with twenty sampling units. The full methodology was implemented in Matlab® environment with a software package developed in event oriented code. It is available together with its tutorial in the supplementary material randPCA.rar.

2 Methods

Answering the questions above always involved a test to tell how likely was the researcher to be wrong when the answer was of a certain kind. These tests always had three features: (i) a statistic taken from the original data set, (ii) a null hypothesis that a similar result could be obtained from a PCA upon a data set of uncorrelated variables, in which case the result was the product of error variation and (iii) the distribution of the statistic when taken from *n* data sets of uncorrelated variables. It was against this distribution that the alternative hypothesis was tested. So, the permutation tests consisted of generating *n* (usually in the order of thousands) new data sets from the original data set, but where all *p* variables were uncorrelated. This was achieved by randomly permuting the *m* measurements on each variable, independently from the others, which did break any correlation between them (Zhang, 2011a, b, c; Zhang and Zheng, 2011). Afterwards, the statistic estimated for the original data set was compared with the ones estimated for the *n* randomly redistributed data sets sorting all and checking for its rank (ter Braak, 1988; Dijksterhuis and Heiser, 1995; Peres-Neto et al., 2005). Suppose a certain statistic is monotonically ascendant in relation to correlation between variables. Then, if the value estimated for the original data set is within the upper *x*% of all the values, it means that there is an *x*% probability of being wrong if the researcher rejects the null hypothesis.

2.1 When to start a PCA?

When variables were uncorrelated each principal component (pc) tended to explain as much variation as any of the variables and a departure from this expected behavior could only arise from random error. In this advent it was useless to apply a PCA to the specified data set as the extracted pc did not synthesize information. On the other hand, when groups of variables showed factual correlations the bigger pc tended to describe it. Hence, the variation explained by these pc was bigger than the variation of a single variable, which made it worthwhile to perform a PCA (Jackson, 1991, Manly, 1986). Seven statistics were used to estimate whether it was useful to undergo a PCA to a specified set of variables. These statistics were:

i) The *generalized variance* (Jackson, 1991) is the determinant of the covariance matrix and it is proportional to the area, volume or hyper-volume generated by a data set. The more uncorrelated the variables the bigger the volume generated while the more correlated the variables the more the volume tends to contract to $k<p$ dimensions. Hence, the *generalized variance* is monotonically descendent with increasing covariation among variables. Its maximum value (in the advent of absolutely none covariation) is the sum of the main diagonal of the covariance matrix. Its minimum value (in the advent of total covariation) is zero.

ii) The *scatter coefficient* (Frisch, 1929) is similar to the *generalized variance* statistic in all but the fact that it is applied to the correlation matrix instead of its covariance matrix. The diagonal line of the correlation matrix is all ones and hence the maximum value of the scatter coefficient can not be greater than *p*. The minimum is zero.

iii) The ψ *index* is a statistic here proposed which relies on the magnitude of the *eigenvalues* taken from the correlation matrix for the data set. If variables are uncorrelated each of the extracted pc tends to explain as much variance as any single variable and thus their associated eigenvalues tend to 1. When these are plotted in a ranked order, they follow a curve line which approximates the *y*=1 line. On the other hand, the more the correlation between variables the bigger some eigenvalues while the smaller the remainder and the steeper the curve line when these are plotted in a ranked order (Fig.1a). The area between the ranked eigenvalues line and the *y*=1 line is a good estimate of the degree of correlation between variables and, as the number of pc is a discrete quantity, one may simply estimate that area as a Riemann's integral: $\sum(\lambda_i-1)$. However, the area relative to the eigenvalues smaller than 1 (in the right hand side of Fig.1a) is negative and thus is subtracted to the area relative to the eigenvalues bigger than 1 (in the left hand side of Fig.1a). This may render the statistic less sensitive. The ψ *index* solves this question by simply squaring each parcel: $\psi = \sum(\lambda_i-1)^2$ (Fig.1b). This statistic is monotonically ascendant with increasing correlation. Its maximum value is *p*(*p*-1) and its minimum value is zero.

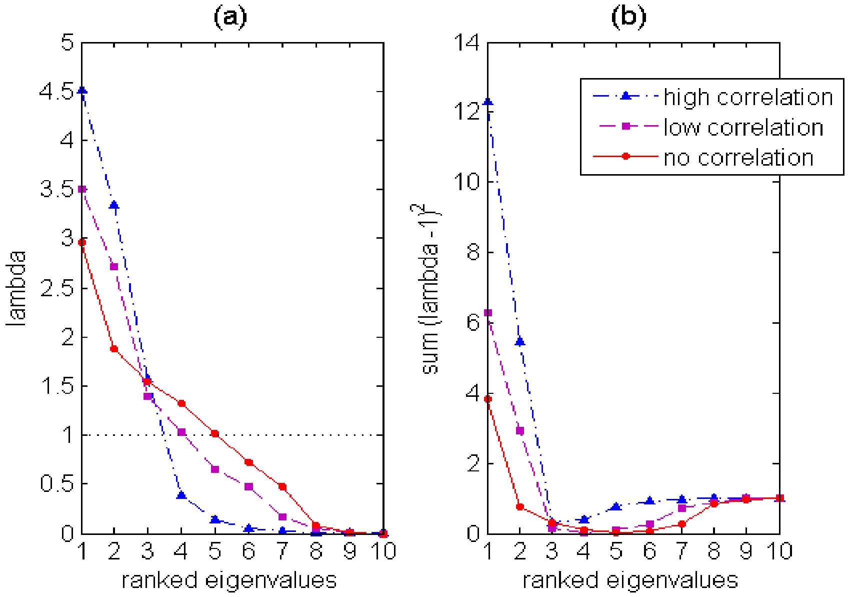

Fig. 1 (a) eigenvalues and (b) ISE for three data sets of ten variables each, manipulated to have different degrees of correlation.

iv) The φ *statistic* (Gleason and Staelin, 1975) is much similar in principles and behavior to the ψ *index*. This statistic is also monotonically ascendant with increasing correlation. Its maximum value is 1 and its minimum value is zero. Applied to the correlation matrix its formulae is given by

$$\varphi = \sqrt{\frac{\sum \lambda_i^2 - p}{p(p-1)}} \qquad (1)$$

v) The *index of a matrix* (Jackson, 1991) compares the biggest and smallest eigenvalues but neglects the distribution of the variation allocated to all the other pc. This statistic is monotonically ascendant with increasing correlation. Its maximum value is +∞ and its minimum value is zero. It is given by the equation

$$\sqrt{\frac{\lambda_1}{\lambda_p}} \qquad (2)$$

vi) The *information statistic* (Kullback, 1959) to be applied upon the correlation matrix is monotonically ascendant with increasing correlation. Its maximum value is +∞ and its minimum value is zero. It is given by the equation

$$-\frac{1}{2}\sum \ln(\lambda_i) \qquad (3)$$

vii) The *divergence statistic* (Kullback, 1959) to be applied upon the correlation matrix is monotonically ascendant with increasing correlation. Its maximum value is +∞ and its minimum value is zero. It is given by the equation

$$\sum \frac{1-\lambda_i}{2\lambda_i} \qquad (4)$$

2.2 Which principal components to use?

Even if at least some of the variables were significantly correlated, this did not mean that all of the pc were of interest. In fact only the first few should explain a true correlation and if re-sampling was done it should be expected to obtain approximately these same pc. As for the remainder pc, these were most probably just explaining residual error and if re-sampling was done one should not expect to obtain similar ones (Gauch Jr, 1982; Jackson, 1991). Four criteria were used to estimate for which pc there was a low probability of being wrong if one said they account for concrete correlations and not just residual error.

(i) The *rank of roots* (ter Braak, 1988) is very simple as it states that the kth biggest pc is most likely explaining meaningful correlations or covariances if it tends to be greater than the pc of the same rank taken from the *n* randomly generated data sets, which are known to only explain error.

(ii) The *equality of roots* (Jackson, 1991) relies on that when at least two roots of the characteristic equation (the eigenvalues) are nearly equal the orientation of the correspondent eigenvectors is not defined with much precision within the plane generated by these same eigenvectors. Hence, attempts to interpret these may be unwise. The most common occurrence will be for the larger first few roots accounting for most of the variability to be fairly well separated while the remainder would all be small, of the same order of magnitude and assumed to represent residual error. With the present randomization method the researcher first calculates the difference between all consecutive roots for the original data set. It is then expected for the eigenvectors accounting for meaningful correlations to have their differences significantly larger than the differences between the same pair of consecutive roots for the *n* randomly permuted data sets.

(iii) The *ratio of roots* (Jackson, 1993) is similar to the test for the *equality of roots* in every thing but the dissimilarities between roots being estimated as a ratio (λ_a/λ_b) and not as a difference (λ_a-λ_b). This is supposed to make the statistic more sensitive and precise for the smaller roots.

(iv) The *pseudo F ratio of roots* (ter Braak, 1990) estimates the ratio between the kth root and the sum of all the roots smaller than the kth one. This should be more reliable because when λ_k is compared with the sum of all smaller roots, results may not be misled by odd events in λ_{k+1}.

2.3 Which variables to each principal component?

Once the pc were chosen they needed to be interpreted as a measure of something which was either a weighted average or a weighted contrast between variables that could have rather different contributions to the amount of variation explained by the pc (Jackson, 1991, Manly, 1986). Therefore, it was desirable to quantify how likely was each variable to be associated with each pc. Yet, the magnitude of the correspondent loadings alone was a poor estimator. For this purpose two monotonically ascendant statistics were used:

i) The *correlations of the pc with the variables* (Jackson, 1991) is a statistic that estimates the correlation of each pc with each variable.

$$r_{ij} = \frac{u_{ij}\sqrt{\lambda_i}}{s_j} \qquad (5)$$

where r_{ij} is the correlation coefficient of the ith pc with the jth variable, u_{ij} is the loading of the jth variable in the ith pc, λ_i is the eigenvalue of the ith pc and s_j is the standard deviation of the jth variable. If one is working with the correlation matrix all s_j equal 1. Axis reflexion, that is the permutation of signs between loadings (Jackson, 1995, Mehlman et al., 1995, Peres-Neto et al., 2003 and 2005), may occur. So it was also tested the use of the absolute values of this statistic. However, both forms could fail when applied with the permutation tests because they over-accounted the loadings relative to the eigenvalues. This could get problematic as each pc explaining random error tended to be associated to a particular variable and thus exhibited a high loading but a low eigenvalue.

ii) To solve the problem above it was developed the *index of the loadings* (*IL*), which enhanced both the loadings and the eigenvalues by squaring them. This way, only high loadings from pcs with high eigenvalues had good chances to be considered significant whereas the remainder events only had mild chances of doing so.

$$IL_{ij} = \frac{{u_{ij}}^2 . \lambda_i^2}{s_j} \qquad (6)$$

Two approaches could be taken in the process of choosing the right variables to each pc according to significance: (a) to choose only the variables for which there was a good probability of being right if one said they were associated to the specific pc, or (b) to reject only the variables for which there was a good probability of being wrong if one said they were associated to the specific pc. Also, it was found that a 0.05 significance level could be too restrictive. It is up to the researcher to choose the approach supported by his objectives and better knowledge of the processes under study.

2.4 Non orthogonal simplification

Once the significance of the loadings was estimated each original pc could be changed to a simplified vector (sv) by discarding the variables with non-significant loadings. By doing so they could also lose their orthogonality. It was then a matter of judging what was more important: to assure orthonormality or to simplify the structure of the vectors. If the choice was to simplify the pc, three criteria needed to be fulfilled for the sv to be considered adequate approximations: (i) each sv should have a high correlation with the pc it was approximating, (ii) each sv should have a low correlation with the other pc and (iii) each sv should have a low correlation with the other sv.

2.5 Vector rotation (Factor Analysis)

Regarding a Factor Analysis the methodologies above answered the questions about whether it was worthwhile trying to reduce the dimensionality, how many factors should be use in the model and what was the structure of the provisional factors. Moreover, the identification of the significant loadings was a good estimator of the interpretability of the rotated factors.

2.6 The z scores

The variables found not significantly contributing to a certain pc were not accounted for its z scores. Theoretically there were two distinct ways to do it which nonetheless yielded the same practical results: (i) The z-scores were estimated upon the sv, for which the non significant loadings were set to zero. (ii) Alternatively, the z scores were estimated upon the orthonormal pc, for which case the sample residuals for each badly correlated variable were taken out. Hence only the variables' means were used. As the means were also standardized to be zero, either ways resulted in parcels of the z-scores which equaled zero when they corresponded to a non-significant association between a variable and a pc. Throughout this work it was focused on the most common approach of a PCA over the correlation matrix, for which the z scores have zero mean. It is equivalent to a PCA over the covariance matrix of variables standardized to zero mean and unit variance. It is also possible to perform a PCA and estimate the z scores under different options. These are well explained by Jackson (1991). The following methods are valid but require slight adaptations for those cases.

2.7 Variance partition

The eigenvalues are estimates of the variances explained by the pc. Therefore, if the variances of their associated z scores are estimated these will equal the eigenvalues (Manly, 1986; Jackson, 1991). However, the PCA numerical method estimates the structure of the pc (the eigenvectors) given the constrain that:

$$1 = u_1^2 + u_2^2 + ... + u_j^2 \qquad (7)$$

And so:

$$\text{var}(pc_i) = \lambda_i . \left(u_{i,1}^2 + u_{i,2}^2 + ... + u_{i,j}^2\right) \qquad (8)$$

which leads to the variance of variable x_j explained by pc_i being given by $\lambda_i.u_{i,j}^2$. Consequently, the total amount of variation of variable x_j explained by the PCA model is given by:

$$\text{var}\left(x_j\right) = u_j^2 \lambda_u + v_j^2 \lambda_v + ... + w_j^2 \lambda_w \qquad (9)$$

Notice the $u_{1,j}$, $u_{2,j}$, … , $u_{i,j}$ were replaced by u_j, v_j, … , w_j in order to simplify the presentation. Each package of variation was classified according to the significance of its correspondent pc (**u**, **v**, … , **w**) and loading (the *j* index). A non-significant package meant that amount of variation corresponded to error and was attributed to that pc by a random process. Equation (9) could equally be achieved starting from the z scores. When the kth sample was defined by the column vector of *x* (**X**, where the x_{jk} were in standardized units) the sample z scores (**Z**) for the pc (**u**, **v**, … , **w**) could be estimated with the aid of the square matrix of the loadings (**U'**) by **Z**=**U'X**:

$$\begin{bmatrix} z_{uk} \\ z_{vk} \\ ... \\ z_{wk} \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & ... & u_j \\ v_1 & v_2 & ... & v_j \\ ... & ... & ... & ... \\ w_1 & w_2 & ... & w_j \end{bmatrix} \times \begin{bmatrix} x_{1k} \\ x_{2k} \\ ... \\ x_{jk} \end{bmatrix} \qquad (10)$$

Because **U** is orthonormal and hence $\mathbf{U}^{-1}$=**U'**, the column vector of *x* could be recalculated back from **UZ**=**UU'X**=$\mathbf{UU}^{-1}\mathbf{X}$=**IX**=**X**, where **I** is the identity matrix:

$$\begin{bmatrix} u_1 & v_1 & ... & w_1 \\ u_2 & v_2 & ... & w_2 \\ ... & ... & ... & ... \\ u_j & v_j & ... & w_j \end{bmatrix} \times \begin{bmatrix} z_{uk} \\ z_{vk} \\ ... \\ z_{wk} \end{bmatrix} = \begin{bmatrix} x_{1k} \\ x_{2k} \\ ... \\ x_{jk} \end{bmatrix} \qquad (11)$$

This could equally be written as a system of linear equations. Furthermore, it was straight forward to develop this system into a system of non-linear equations where each reported to the variance contained in each variable:

$$\begin{cases} \frac{1}{n}\sum_k (x_{1k})^2 = \frac{1}{n}\sum_k (u_1.z_{uk} + v_1.z_{vk} + ... + w_1.z_{wk})^2 \\ \frac{1}{n}\sum_k (x_{2k})^2 = \frac{1}{n}\sum_k (u_2.z_{uk} + v_2.z_{vk} + ... + w_2.z_{wk})^2 \\ \qquad\qquad ... \\ \frac{1}{n}\sum_k (x_{jk})^2 = \frac{1}{n}\sum_k (u_j.z_{uk} + v_j.z_{vk} + ... + w_j.z_{wk})^2 \end{cases} \tag{12}$$

Each of the equations of the system could be further developed to:

$$\frac{1}{n}\sum_k (x_{jk})^2 = u_j^2 \sum_k \frac{(z_{uk})^2}{n} + v_j^2 \sum_k \frac{(z_{vk})^2}{n} + ... + w_j^2 \sum_k \frac{(z_{wk})^2}{n} + 2u_j v_j \sum_k (z_{uk} z_{vk}) + ... \\ + 2u_j w_j \sum_k (z_{uk} z_{wk}) + ... + 2v_j w_j \sum_k (z_{vk} z_{wk}) + ... \tag{13}$$

Knowing that $\frac{1}{n}\sum_k (x_{jk})^2$ was the variation of variable x_j, that $\sum_k \frac{(z_k)^2}{n} = \lambda$ was the eigenvalue of the respective pc and that $\sum_k (z_{uk} z_{vk})$, $\sum_k (z_{uk} z_{wk})$, … and $\sum_k (z_{vk} z_{wk})$ all equal zero as a consequence of the orthogonality of the pc; each of the equations of the system could be rewritten in the form already presented in equation (9). An alternative method enabled to decompose the variance contained in each pc into parcels relating to variances and covariances of the variables (equation 14).

$$\lambda = \frac{1}{n}\sum_{k}^{n} z_k^2$$

$$= \frac{1}{n}\sum_{k}^{n}\left(\begin{bmatrix} u_1 & u_2 & \dots & u_j \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_j \end{bmatrix}_k\right)^2$$

$$= \frac{1}{n}\sum_{k}\Big(u_1^2 x_{1,k}^2 + u_2^2 x_{2,k}^2 + \dots + u_j^2 x_{j,k}^2 + 2u_1u_2x_{1,k}x_{2,k} + \dots$$

$$\dots + 2u_1u_jx_{1,k}x_{j,k} + \dots + 2u_2u_jx_{2,k}x_{j,k} + \dots\Big)$$

$$= u_1^2\sum_{k}^{n}\frac{x_{1,k}^2}{n} + u_2^2\sum_{k}^{n}\frac{x_{2,k}^2}{n} + \dots + u_j^2\sum_{k}^{n}\frac{x_{j,k}^2}{n} + 2u_1u_2\sum_{k}^{n}\frac{x_{1,k}x_{2,k}}{n} +$$

$$\dots + 2u_1u_j\sum_{k}^{n}\frac{x_{1k}x_{jk}}{n} + \dots + 2u_2u_j\sum_{k}^{n}\frac{x_{2,k}x_{j,k}}{n} + \dots \qquad (14)$$

$$= u_1^2\,\mathrm{var}(x_1) + u_2^2\,\mathrm{var}(x_2) + \dots + u_j^2\,\mathrm{var}(x_j) + 2u_1u_2\,\mathrm{cov}(x_1, x_2) + \dots$$

$$\dots + 2u_1u_j\,\mathrm{cov}(x_1, x_j) + \dots + 2u_2u_j\,\mathrm{cov}(x_2, x_j) + ..$$

3 Results

3.1 When to start a PCA?

The results from the application of all the seven statistics agreed that it was worthwhile to under go a PCA to the four factual data sets while it was not worthwhile to do it for the random data set (Table 1). When comparing the values and critical thresholds (for p=0.05), attention to the fact that the first two statistics are monotonically descendent while the other five are monotonically ascendant.

3.2 Which principal components to use?

The results have shown only the test for the *rank of the roots* was consistent and in accordance with what was previously known about the data sets (Table 2 to Table 6). The test for the *equality of the roots* could be misleading when the biggest pc was close to the second biggest but was reliable as a stopping rule for identifying the last non-trivial pc. Both the tests for the *ratio of the roots* and the *pseudo F ratio of the roots* often gave false results about the probability of Type I error in trivial components. The test for the *ratio of the roots* also showed it could easily give Type II error in the non-trivial components. The results from the permutation tests were compared with other methodologies. The *broken stick* model was a reliable stopping rule for the present data sets and for other data sets of few variables. It always agreed with the tests for the rank of the roots and the *equality of the roots* (Fig. 2). However, it was not trust-worthy when applied to very large data sets (*not shown*). Hence, for both data sets about Ria Formosa should only be used the first three pc; in the case of employment in Europe in the 1970’s, only the first two pc; in the case of Bumpus’ sparrows, only the first pc and in the case of the 10 uncorrelated variables, none. It was also demonstrated that accepting every component with a root bigger than one was a bad criterion. In three of the five data sets were obtained pc with roots bigger than 1 which still only related to error. Also bad criterion would be to retain all the pc until a pre-established cumulative proportion of variation. In the particular case of the randomly generated data set 100% of its variation is related to error, all the pc are trivial and still four had roots bigger than 1.

Table 1 “Whether to start a PCA” for five different data sets and according to seven different statistics and randomization tests with 10000 simulations

		Gen. variance	Scatter coeff.	Ψ index	Φ statistic	Index of a matrix	Inf. statistic	Div. statistic
12 abiotic variables from Ria Formosa	value	$2.7*10^{14}$	$4.4*10^{-8}$	20.9	0.4	26.1	8.47	133
	critical (0.05)	$3.6*10^{17}$	$6*10^{-5}$	8.8	0.26	14.1	4.16	45
	sig.	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0222**	**<0.0001**	**0.0209**
11 seagrass variables from Ria Formosa	value	$7.3*10^{14}$	$5.8*10^{-7}$	17	0.39	27.6	7.18	155
	critical (0.05)	$3.5*10^{18}$	0.0028	7.5	0.26	6.19	2.58	11.52
	Sig	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**0.001**	**<0.0001**	**<0.0011**
Emp. in Europe	value	152	$2.3*10^{-6}$	10.4	0.38	276	6.47	10964
	critical (0.05)	$5.5*10^{6}$	0.087	3.5	0.22	3	0.99	3
	sig	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
Bumpus’ female sparrows	value	2.5	0.04	8.63	0.66	4.69	1.65	4.57
	critical (0.05)	45.3	0.67	0.63	0.18	1.68	0.16	0.33
	sig	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
Random generated	value	$5.9*10^{9}$	$4.3*10^{-19}$	8.18	0.3	$1.5*10^{8}$	20.98	$3.8*10^{15}$
	critical (0.05)	0	0	9.34	0.32	$4.2*10^{8}$	21.45	$4.7*10^{15}$
	sig	**0.78**	**0.78**	**0.43**	**0.43**	**0.96**	**0.78**	**0.89**

3.3 Which variables to each principal component?

The *index for the loadings* gave more reliable results when applied to the five data sets here presented (Tables 7 to Table 10). The *correlations of the pc with the variables* (Jackson 1991) showed a tendency to overlook high loadings (type I error) in non-trivial pc while accepting high loadings (type II error) from trivial pc. However, posterior application of both statistics to a much larger data set (with many more variables and sampling units) showed the *index for the loadings* to accept variables that were not so well correlated (Guerra, 2010). On the contrary, the *correlations of the pc with the variables* by being more strict yielded more striking patterns.

Table 2 Choosing the principal components for the “Bumpus' sparrows” data sets. Type I error probability: 0.05 (in black bold) and 0.1 (in grey bold). Both the test for the equality of roots and the ratio of roots were done using λ_i and λ_{i+1}.

eigenvalue	percentage of variation	broken stick	comulative percentage of variation	rank of roots	equality of roots	ratio of roots	pseudo F ratio of roots
3.62	72.32	45.7	72.3	**0.001**	**0.001**	**0.001**	**0.001**
0.53	10.63	25.7	82.9	1	0.600	**0.077**	**0.011**
0.39	7.73	15.7	90.7	1	0.806	0.267	**0.059**
0.30	6.03	9	96.7	1	0.640	**0.027**	**0.027**
0.16	3.29	4	100	1			

Table 3 Choosing the principal components for the “employment in Europe in the 1970’s” data set. Type I error probability: 0.05 (in black bold) and 0.1 (in grey bold). Both the test for the equality of roots and the ratio of roots were done using λ_i and λ_{i+1}.

eigenvalue	percentage of variation	broken stick	comulative percentage of variation	rank of roots	equality of roots	ratio of roots	pseudo F ratio of roots
3.49	38.75	31.43	38.75	**1E-04**	**6.E-04**	**0.017**	**1E-04**
2.13	23.67	20.32	62.41	**6.E-04**	**4.E-04**	**4.E-04**	**1E-04**
1.10	12.21	14.77	74.63	0.9951	0.843	0.793	**0.004**
0.99	11.05	11.06	85.68	0.9108	**0.025**	**0.002**	**1E-04**
0.54	6.04	8.28	91.71	1	0.545	0.160	**0.002**
0.38	4.26	6.06	95.97	1	0.493	**0.051**	**6.E-04**
0.23	2.51	4.21	98.48	1	0.766	0.162	**0.002**
0.14	1.52	2.62	1E+02	1	0.506	**1E-04**	**1E-04**
5E-05	5E-04	1.23	100	1			

Table 4 Choosing the principal components for the “12 abiotic variables from Ria Formosa” data set. Type I error probability: 0.05 (in black bold) and 0.1 (in grey bold). Both the test for the equality of roots and the ratio of roots were done using λ_i and λ_{i+1}.

eigenvalue	percentage of variation	broken stick	comulative percentage of variation	rank of roots	equality of roots	ratio of roots	pseudo F ratio of roots
4.21	35.07	25.86	35.07	**3.E-04**	**0.051**	0.211	**3.E-04**
2.99	24.94	17.53	60.01	**2.E-04**	**0.021**	**0.079**	**1E-04**
2.01	16.77	13.36	76.78	**0.056**	**0.009**	**0.008**	**1E-04**
1.17	9.71	10.58	86.49	0.979	0.270	0.111	**1E-04**
0.80	6.63	8.50	93.11	0.998	0.157	**0.009**	**1E-04**
0.41	3.44	6.83	96.56	1	0.418	**0.003**	**2.E-04**
0.18	1.54	5.44	98.10	1	0.912	0.310	**0.065**
0.12	1.02	4.25	99.12	1	0.827	**0.018**	**0.040**
0.05	0.38	3.21	99.50	1	0.986	0.591	0.801
0.03	0.25	2.29	99.75	1.000	0.996	0.884	0.825
0.02	0.19	1.45	99.95	0.995	0.933	0.359	0.359
0.01	0.05	0.69	100	0.960			

Table 5 Choosing the principal components for the “11 seagrass variables from Ria Formosa” data set. Type I error probability: 0.05 (in black bold) and 0.1 (in grey bold). Both the test for the equality of roots and the ratio of roots were done using λ_i and λ_{i+1}.

eigenvalue	percentage of variation	broken stick	comulative percentage of variation	rank of roots	equality of roots	ratio of roots	pseudo F ratio of roots
3.46	31.43	27.45	31.43	**0.008**	0.867	0.933	**0.008**
3.22	29.25	18.36	60.69	**10^{-4}**	**0.005**	**0.061**	**10^{-4}**
2.10	19.05	13.82	79.74	**0.002**	**0.0002**	**0.000**	**10^{-4}**
0.92	8.34	10.79	88.08	1	0.340	**0.073**	**0.0004**
0.59	5.35	8.51	93.43	1	0.239	**0.002**	**0.001**
0.26	2.35	6.70	95.78	1	0.960	0.689	0.438
0.21	1.91	5.18	97.69	1	0.904	0.472	0.172
0.15	1.33	3.88	99.02	1	0.878	0.391	**0.066**
0.09	0.81	2.75	99.83	0.999	0.707	**0.004**	**0.003**
0.01	0.13	1.74	99.96	1	0.987	0.351	0.351
0.00	0.04	0.83	100	0.998			

Table 6 Choosing the principal components for the “ten random variables” data set. Type I error probability: 0.05 (in black bold) and 0.1 (in grey bold). Both the test for the equality of roots and the ratio of roots were done using λ_i and λ_{i+1}.

eigenvalue	percentage of variation	broken stick	comulative percentage of variation	rank of roots	equality of roots	ratio of roots	pseudo F ratio of roots
2.22	22.17	29.29	22.17	0.652	0.951	0.956	0.652
2.10	20.99	19.29	43.16	**0.056**	0.237	0.314	0.101
1.63	16.28	14.29	59.44	0.136	**0.053**	**0.051**	0.108
1.08	10.79	10.96	70.24	0.893	0.279	0.178	0.674
0.77	7.75	8.46	77.98	0.977	0.792	0.716	0.969
0.66	6.63	6.46	84.61	0.847	0.895	0.882	0.941
0.59	5.93	4.79	90.54	0.411	0.280	0.283	0.704
0.38	3.83	3.36	94.37	0.622	0.775	0.789	0.968
0.30	3.03	2.11	97.40	0.296	0.905	0.951	0.951
0.26	2.60	1	100	**0.028**			

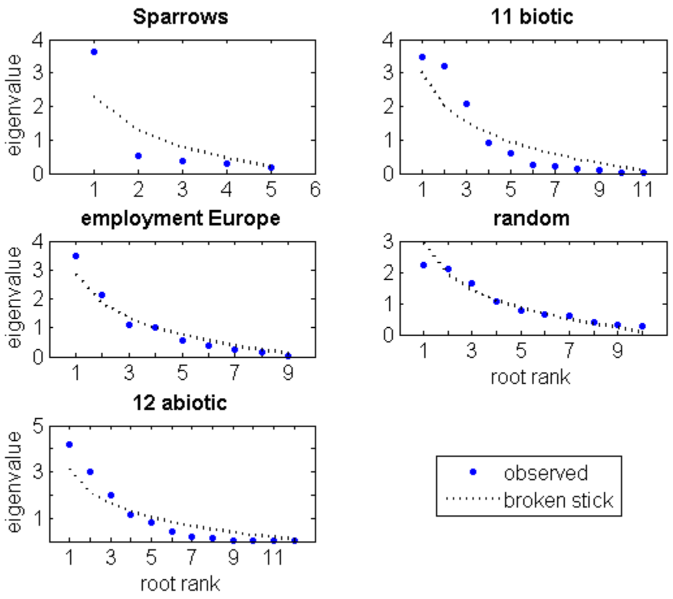

Fig. 2 SCREE plot for the five different data sets and their broken stick model predictions

Table 7 Interpreting the principal components for the data set of 5 morphometric variables from Bumpus' female sparrows: significant loadings in black bold (0.05 significance level) and in grey bold (0.1 significance level).

Variables total lenght	alar extent	length of beak and head	length of humerus	length of keel of sternum
significance based on the correlation of the pcs with the variables				
0.001	**0.0002**	**0.001**	**0**	**0.07**
0.53	0.35	0.34	0.41	0.93
0.85	0.66	0.28	0.29	0.46
0.39	0.71	0.26	0.64	0.55
0.42	0.65	0.59	0.33	0.53
significance based on the index of the loadings				
0	**0**	**0**	**0**	**0**
0.97	0.82	0.80	0.88	0.47
0.57	0.77	0.70	0.73	0.87
0.80	0.74	0.71	0.81	0.97
0.90	0.85	0.91	0.82	0.95
choosen loadings based on the correlation of the pcs with the variables				
0.45	**0.46**	**0.45**	**0.47**	**0.40**
choosen loadings based on the index of the loadings				
0.45	**0.46**	**0.45**	**0.47**	**0.40**

The biggest pc extracted from the data set of five morphometric variables on Bumpus' female sparrows (Table 7) was a weighted average of those variables. This pc was a measure of size as already stated by Manly (1986).

The two pc taken from the data set of the nine variables about employment in Europe in the 1970's (Table 8) matched their interpretation by Manly (1986). The first pc was a contrast between occupation on the primary sector (agriculture) and occupation on the secondary (manufacturing and construction) and tertiary (services, social and personal services, and transports and communications) sectors. The second pc was a contrast between occupation on the secondary sector (mining) and tertiary sector (finance).

The three pc taken from the data set of the twelve abiotic variables from Ria Formosa (Table 9) were describing three common processes in lagoonary systems. The first pc was a measure of the stream inputs to the system, namely fresh water and nutrient loadings. The second pc was a contrast between phosphate in the sediments, ammonia in the sediments and temperature in the low tide on one hand and the redox potential in the sediments in the other. This second pc could be said to be a measure of the anoxia induced by temperature. According to the Boudin diffusive model the oxygen concentration in the sediment is largely dependent on oxygen diffusion from the water column immediately above. So, when a major increase in temperature in the water column decreases its oxygen solubility, it also decreases its diffusion to the sediment because although diffusivity increases with temperature, the oxygen gradient is much smaller and thus there is less oxygen being "pumped" to the sediments. With increasing anoxia conditions in the sediment, ammonia tends to be the dominant inorganic nitrogen form and the phosphate adsorbed to the sediment particles is released to the interstitial water. The third pc was a contrast between temperature and organic matter content. It was a measure of the effect of temperature on remineralization as most of the remineralization taking place was done by bacteriological activity which was temperature dependent.

Table 8 Interpreting the principal components for the data set of 9 variables about the employment in Europe in the 1970's: significant loadings in black bold (0.05 significance level) and in grey bold (0.1 significance level). agr: agriculture; min: mining; man: manufacturing; ps: power supplies; con: construction; ser: service industries; fin: finance; sps: social and personal services; tc: transport and communications.

Variables								
agr	min	man	ps	con	ser	fin	sps	tc
significance based on the correlation of the pcs with the variables								
1.00	0.4964	**0.09**	0.22	0.12	**0.06**	0.42	**0.048**	0.12
0.45	**0**	0.13	0.23	0.44	0.85	0.94	0.73	0.47
0.46	0.68	0.63	0.93	0.36	0.61	0.94	0.22	0.28
0.53	0.56	0.18	0.85	**0.03**	0.45	0.44	0.87	0.74
0.35	0.64	0.80	0.26	0.15	0.73	0.27	0.69	0.16
0.42	0.57	0.72	0.24	0.40	**0.10**	0.85	0.70	0.56
…	…	…	…	…	…	…	…	…
significance based on the index of the loadings								
0	1.00	**0**	0.23	**0.04**	**0.001**	0.80	**0.0007**	**0.004**
0.87	**0**	0.19	0.39	0.88	0.20	**0.045**	0.47	0.51
0.92	0.67	0.75	0.18	0.75	0.80	0.17	0.49	0.41
0.94	0.88	0.38	0.32	**0.07**	0.90	0.89	0.29	0.43
…	…	…	…	…	…	…	…	…
choosen loadings based on the correlation of the pcs with the variables								
-0.52	0.00	**0.35**	0.26	0.33	**0.38**	0.07	**0.39**	0.37
0.05	**0.62**	0.36	0.26	0.05	-0.35	-0.45	-0.22	0.20
choosen loadings based on the index of the loadings								
-0.52	0.00	**0.35**	0.26	**0.33**	**0.38**	0.07	**0.39**	**0.37**
0.05	**0.62**	0.36	0.26	0.05	-0.35	**-0.45**	-0.22	0.20

The three pc taken from the data set of the eleven seagrass variables from Ria Formosa (Table 10) were describing three processes in seagrass meadows. The first pc was a measure of shape. It evaluated the contrast between seagrasses with relatively short roots and big leafs and seagrasses with relatively long roots and small leafs. Later, this contrast was determined not to be between individuals but rather within each individual on a seasonal basis. The second pc was a measure of plant overall size (leafs and roots) which was found to vary over a spatial gradient. The third pc was a measure of population size (biomass and density) which varied seasonally and spatially.

3.4 Vector simplification

The estimation of the correlations involving the pc and their sv is illustrated with an example from the data set with twelve abiotic variables from Ria Formosa. The significant loadings were chosen according to the *index of the loadings* with a 0.1 significance level. To estimate the correlation of $\mathbf{sv_1}$ with $\mathbf{pc_1}$ a matrix $\mathbf{T}$ was constructed:

Table 9 Interpreting the principal components for the data set of 12 abiotic variables from Ria Formosa: significant loadings in black bold (0.05 significance level) and in grey bold (0.1 significance level). OM: organic matter; EH: redox potential; sed: measured on the sediment; lt: measured on the water column during the low tide.

NH_4	NO_3	PO_4	NH_4	NO_3	PO_4		EH		Temp	Sal	
sed	sed	sed	lt	lt	lt	OM sed	sed	pH sed	sed	lt	Temp lt
significance based on the correlation of the pcs with the variables											
0.59	0.92	0.50	1.00	0.99	0.99	0.62	0.52	0.39	0.20	**10^{-4}**	0.48
0.05	0.57	**10^{-3}**	0.46	0.35	0.41	0.22	1.00	0.87	0.18	0.45	0.17
0.90	0.43	0.36	0.30	0.28	0.28	0.99	0.27	0.52	**0.04**	0.66	**0.09**
0.81	**0.08**	0.72	0.70	0.28	0.78	0.53	0.67	0.98	0.59	0.27	0.14
0.23	**0.07**	0.20	0.73	0.17	0.83	0.49	0.20	0.12	0.48	0.38	0.72
…	…	…	…	…	…	…	…	…	…	…	…
significance based on the index of the loadings											
0.74	**0.04**	0.98	**0**	**10^{-4}**	**10^{-4}**	0.69	0.95	0.74	0.24	**10^{-4}**	0.35
0.04	0.84	**0**	0.88	0.65	0.79	0.36	**10^{-4}**	0.17	0.26	0.89	**0.09**
0.15	0.83	0.71	0.57	0.53	0.54	**0.02**	0.50	0.95	**0.06**	0.64	**0.07**
0.42	0.22	0.59	0.63	0.59	0.47	0.95	0.70	**0.07**	0.82	0.58	0.96
…	…	…	…	…	…	…	…	…	…	…	…
choosen loadings based on the correlation of the pcs with the variables											
-0.09	-0.33	-0.01	-0.45	-0.42	-0.42	-0.11	-0.02	0.09	0.25	**0.45**	0.21
0.39	-0.05	**0.50**	0.04	0.12	0.07	0.23	-0.50	-0.30	0.26	0.04	0.35
-0.38	0.06	0.10	0.16	0.18	0.17	-0.54	0.19	-0.02	**0.46**	-0.13	**0.45**
choosen loadings based on the index of the loadings											
-0.09	**-0.33**	-0.01	**-0.45**	**-0.42**	**-0.42**	-0.11	-0.02	0.09	0.25	**0.45**	0.21
0.39	-0.05	**0.50**	0.04	0.12	0.07	0.23	**-0.50**	-0.30	0.26	0.04	**0.35**
-0.38	0.06	0.10	0.16	0.18	0.17	**-0.54**	0.19	-0.02	**0.46**	-0.13	**0.45**

$$\mathbf{T}=\begin{bmatrix}\mathbf{pc_1}\\ \mathbf{sv_1}\end{bmatrix}=\begin{bmatrix}-0.09 & -0.33 & -0-01 & -0.45 & -0.42 & -0.42 & -0.11 & -0.02 & 0.09 & 0.25 & 0.45 & 0.21\\ 0 & -0.33 & 0 & -0.45 & -0.42 & -0.42 & 0 & 0 & 0 & 0 & 0.45 & 0\end{bmatrix}$$

With **S** being the covariance matrix of the original variables, the product **TST'** yielded the covariance matrix of $\mathbf{pc_1}$ with $\mathbf{sv_1}$:

$$\mathbf{TST'}=\begin{bmatrix}9969 & 9850\\ 9850 & 9751\end{bmatrix}$$

which tells us that their correlation was:

$$\mathrm{r}=9850/\sqrt{9969\times 9751}=0.999$$

The correlations between the sv and the pc were performed with the sv edited according to the *index of the loadings* (Table 11.a) or according to the *correlations of the pc with the variables* (Table 11.b). This particular example is interesting because though the vectors were simplified they were still orthogonal, even if they do not look so when projected in a plane which was not their own (Fig.3). This happens whenever the sv do not share variables but even if they do they may still keep orthogonality. The easiest way to verify it is to calculate the inner product of the sv. If it equals zero they are orthogonal. In this example as well as with all other tested data sets, editing the sv based on the *index of the loadings* was the best option as these sv were better correlated with their respective pc and lesser correlated with the other pc and sv. It was a direct consequence of the *index of the loadings* accepting more loadings as significant: the closer a sv was to its original pc the better correlated they were and the worst correlated with all the others. Both the orthogonal varimax rotation and the non-orthogonal promax rotation, gave results identical to the vectors simplified according to the *index of the loadings* and did not add any relevant information (Table 12).

Table 10 Interpreting the principal components for the data set of 11 seagrass variables from Ria Formosa: significant loadings in black bold (0.05 significance level) and in grey bold (0.1 significance level). Biom: biomass, Dens: density, Inter: internode and diam: diameter.

Above Biom	Below Biom	Algae biom	Dens.	Leaf length	Leaf width	Leaf nº	Sheat length	Inter. length	Inter. Diam.	Root lenght
significance based on the correlation of the pcs with the variables										
0.996	0.989	0.367	0.636	0.105	0.827	0.209	0.190	0.257	**0.100**	**0.033**
0.267	0.624	0.284	0.529	0.216	**0.001**	0.812	0.103	0.168	0.248	0.480
0.015	**0.042**	0.439	**0.029**	0.771	0.506	0.422	0.890	0.890	0.928	0.891
…	…	…	…	…	…	…	…	…	…	…
significance based on the index of the loadings										
0.996	0.987	0.266	0.574	**0.027**	0.802	0.103	**0.087**	0.149	**0.027**	**0.003**
0.127	0.532	0.148	0.421	**0.081**	**0.000**	0.763	**0.012**	**0.046**	0.105	0.358
0.002	**0.014**	0.379	**0.007**	0.742	0.451	0.360	0.875	0.878	0.919	0.877
…	…	…	…	…	…	…	…	…	…	…
choosen loadings based on the correlation of the pcs with the variables										
0.00	-0.01	0.29	0.18	-0.40	0.08	0.35	-0.36	0.33	**0.40**	**0.44**
-0.30	0.15	-0.29	0.18	-0.32	**-0.51**	-0.07	-0.38	-0.35	-0.31	0.21
0.55	**0.51**	-0.23	**0.52**	-0.09	0.20	-0.24	0.04	0.04	-0.03	0.04
choosen loadings based on the index of the loadings										
0.00	-0.01	0.29	0.18	**-0.40**	0.08	0.35	**-0.36**	0.33	**0.40**	**0.44**
-0.30	0.15	-0.29	0.18	**-0.32**	**-0.51**	-0.07	**-0.38**	**-0.35**	-0.31	0.21
0.55	**0.51**	-0.23	**0.52**	-0.09	0.20	-0.24	0.04	0.04	-0.03	0.04

3.5 The z scores

The graphical representation of the z scores elucidates about the error that was filtered out with this methodology. For the data set of twelve abiotic variables from Ria Formosa the evolution of the processes in space and time was equally evident as only the less influential variables were kept out of each pc (Fig.4). Nevertheless, differences between sampling units turned more accurate. These processes showed the annual cycle more conspicuous closest to the waste water treatment plant. For the data set of the nine variables about employment in Europe in the 1970’s the main patterns were equally evident (Fig.5). Four groups of countries could be identified: (1) eastern block countries, (2) western block, peripheral, underdeveloped countries, (3) western block, central, developed countries, and (4) Yugoslavia and Turkey outliers. Nevertheless, the relative positions of the countries became more accurate. For the data set of five morphometric measures on Bumpus’ female sparrows all variables were found to be significantly contributing to the chosen pc. Hence, both ways of estimating the z scores led to the exact same results as also did any graphical interpretation or statistical inference from these.

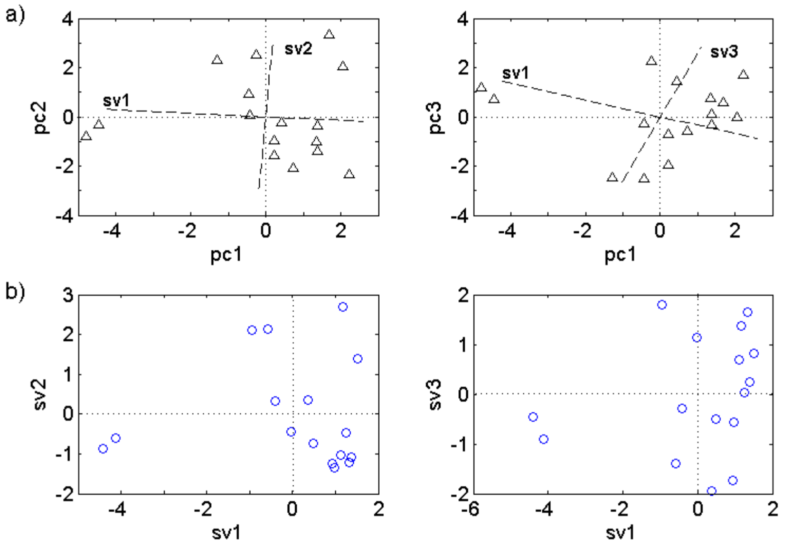

Fig. 3 Projection in principal component generated planes of the 16 observations from Ria Formosa and the simplified vectors.

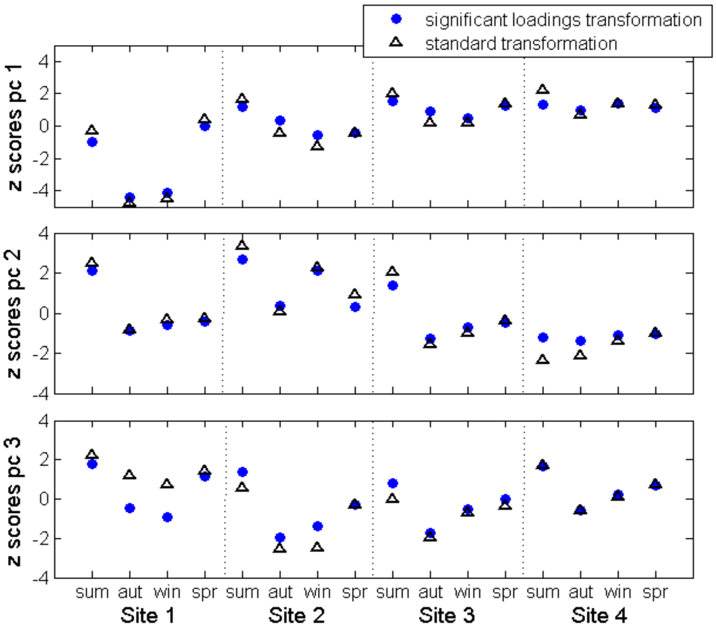

Fig. 4 z scores for the three main principal components taken from the data set of 12 abiotic variables from Ria Formosa. Distance from a waste water treatment plant: Site 1 (closest) to Site 4 (furthest). sum: summer, aut: autumn, win: winter and spr: spring.

Table 11 correlations between the simplified vectors (sv) and the principal components (pc) for the data set of 12 abiotic variables from Ria Formosa when the simplified vectors are edited according to the significance of the loadings estimated by (a) the 'index of the loadings' and (b) the 'correlations of the loadings with the pcs', with a 0.05 significance.

(a)	sv1	sv2	sv3	(b)	sv1	sv2	sv3
pc1	**0.93**	0.12	0.37	pc1	**0.56**	0.05	0.48
pc2	0.06	**0.87**	0.15	pc2	-0.06	**0.58**	0.34
pc3	-0.27	-0.10	**0.84**	pc3	-0.21	-0.26	**0.63**
sv1		0.10	0.05	sv1		-0.13	-0.13
sv2			0.19	sv2			-0.20

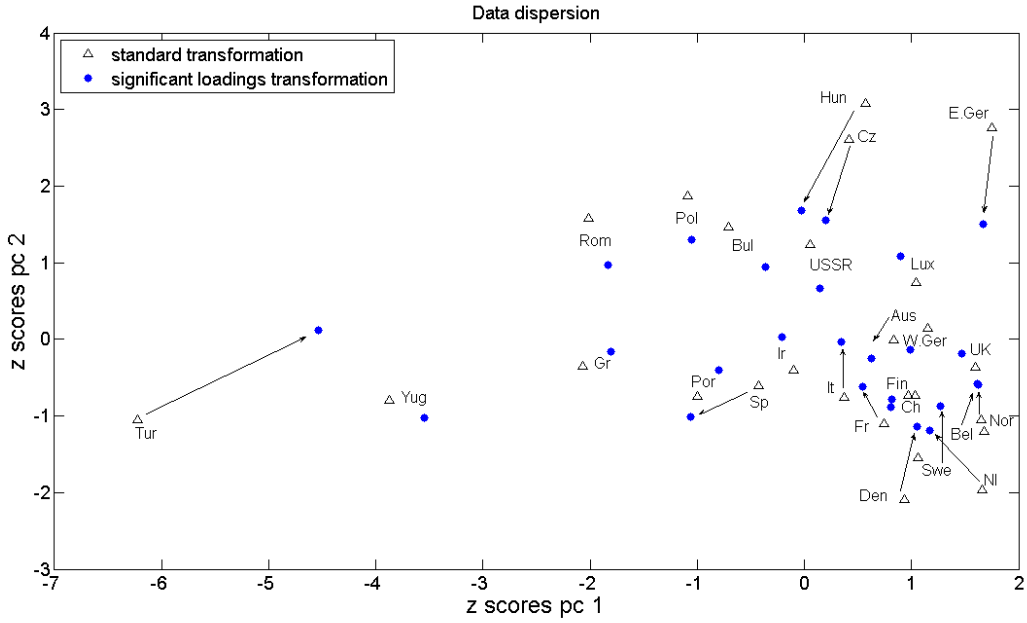

Fig. 5 Relative position of each country according to the two principal components about employment in Europe in the 1970's.

Table 12 Factor Analysis on the data set of 12 abiotic variables from Ria Formosa. Non-orthogonal promax rotation.

	factor loadings			communality			specificity
	f1	f2	f3	f1	f2	f3	e
NH4 sed	-0.03	-0.17	**0.96**	0.00	0.03	**0.92**	0.05
NO3 sed	0.44	-0.16	-0.06	0.19	0.03	0.00	**0.78**
PO4 sed	0.10	**0.50**	**0.61**	0.01	**0.25**	**0.37**	**0.36**
NH4 lt	**1.00**	0.00	-0.02	**1.00**	0.00	0.00	0.00
NO3 lt	**0.74**	0.05	0.08	**0.54**	0.00	0.01	**0.45**
PO4 lt	**0.99**	0.06	-0.01	**0.99**	0.00	0.00	0.01
OM sed	-0.07	-0.41	**0.72**	0.00	0.17	**0.52**	**0.31**
EH sed	0.04	-0.22	**-0.68**	0.00	0.05	**0.47**	**0.49**
pH sed	-0.14	-0.19	-0.17	0.02	0.03	0.03	**0.92**
Temp sed	-0.04	**0.96**	-0.15	0.00	**0.92**	0.02	0.06
Sal lt	**-0.92**	0.13	0.05	**0.85**	0.02	0.00	0.13
Temp lt	-0.01	**1.00**	-0.04	0.00	**1.00**	0.00	0.00

3.6 Variance partition

The variance partition is shown for the data set of twelve abiotic variables from Ria Formosa (Table 13). As an example, for the temperature in the low tide 0.4 of its variance was allocated to the third pc and 0.36 to the second pc. Together they made up to the 0.76 found significantly correlated with the PCA model. The remainder was error composed by the 0.19 randomly correlated to the first pc and the 0.05 randomly correlated to the smaller pc which only described error. Adding up all meaningful packages of variance the first pc accounted for 3.64 standardized variables instead of the 4.21 estimated by the eigenvalue, the second pc accounted for 2.29 instead of the 2.99 and the third pc accounted for 1.41 instead of the 2.01. It was also evident the variables and associations significantly contributing to the same pc could still do it in rather different measures.

Table 13 Partition of the variation among variables and principal components for the data set of 12 abiotic variables from Ria Formosa. PCA upon the correlation matrix. “Packages” of variation categorized according to the **0.1** significance level

	NH4	NO3	PO4	NH4	NO3	PO4	OM
pc:	sed	sed	sed	lt	lt	lt	sed
1st	0.03	**0.47**	0.00	**0.85**	**0.73**	**0.73**	0.05
2nd	**0.46**	0.01	**0.74**	0.00	0.04	0.01	0.15
3rd	0.29	0.01	0.02	0.05	0.06	0.06	**0.58**
4th	0.10	0.24	0.05	0.04	0.05	0.08	0.00
5th	0.05	0.22	0.07	0.04	0.09	0.09	0.00
6th	0.01	0.02	0.08	0.01	0.00	0.01	0.17
7th	0.01	0.02	0.03	0.01	0.00	0.00	0.03
8th	0.05	0.00	0.01	0.00	0.00	0.00	0.01
9th	0.00	0.01	0.00	0.00	0.02	0.00	0.00
10th	0.00	0.00	0.01	0.00	0.00	0.00	0.00
11th	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12th	0.00	0.00	0.00	0.00	0.00	0.00	0.00
total:	1	1	1	1	1	1	1
sig:	**0.46**	**0.47**	**0.74**	**0.85**	**0.73**	**0.73**	**0.58**
	EH	pH	Temp	Sal			
	sed	sed	sed	lt	Temp lt	total:	sig:
1st	0.00	0.03	0.26	**0.86**	0.19	4.21	**3.64**
2nd	**0.74**	0.26	0.21	0.00	**0.36**	2.99	**2.29**
3rd	0.07	0.00	**0.43**	0.04	**0.40**	2.01	**1.41**
4th	0.02	0.51	0.01	0.05	0.00	1.17	0
5th	0.07	0.14	0.00	0.01	0.00	0.80	0
6th	0.00	0.01	0.07	0.01	0.02	0.41	0
7th	0.06	0.02	0.00	0.01	0.00	0.18	0
8th	0.02	0.02	0.01	0.00	0.00	0.12	0
9th	0.00	0.00	0.00	0.01	0.00	0.05	0
10th	0.00	0.00	0.00	0.00	0.01	0.03	0
11th	0.00	0.00	0.01	0.01	0.00	0.02	0
12th	0.00	0.00	0.00	0.00	0.00	0.01	0
total:	1	1	1	1	1	12	0
sig:	**0.74**	0	**0.43**	**0.86**	**0.76**	0	**7.35**

4 Discussion

The statistics taken from a PCA used to be tested against their estimated normal or normal-derived distributions. It required manipulation of the statistic and the assumption of multivariate normal distribution of the data (Jackson, 1991; Dijksterhuis and Heiser, 1995; Peres-Neto et al., 2005). Alternatively, *n* data sets with the same size of the original data set could be generated where variables were randomly and normally distributed. The statistics taken from the original data set were compared with their distributions from the *n* randomly generated data sets but constrain of normality was still present (Zwick and Velicer, 1986; Grossman et al., 1991; Peres-Neto et al., 2005). Multivariate normal distribution of the data seldom is true and the researcher is forced to trust the robustness of the method to departure from the assumption of normality (Jackson, 1991; Peres-Neto et al., 2005). The application of randomization tests unties the data from the restriction of multivariate normal distribution (Dijksterhuis and Heiser, 1995; Peres-Neto et al., 2005). However, it then emerges the question about the proper class of randomization tests. Bootstrap methodologies have been used to propose significances for metrics taken from PCA by comparing their confidence intervals with thresholds as the roots greater than unity, the *broken stick* model, loadings that overlap zero, among others (Stauffer et al., 1985; Jackson, 1993; Peres-Neto et al., 2003; Peres-Neto et al., 2005). However, bootstrap techniques simply randomly resample with reposition each sampling unit (Manly, 1991; Jackson, 1993; Yu et al., 1998; Peres-Neto et al., 2005; Lebart, 2006; Ferrarini, 2011; Zhang, 2011a, b. c; Zhang and Zheng, 2011) which does not break the correlations between variables latent within the sampling units (Dijksterhuis and Heiser, 1995). Therefore, there is never the statement of a null hypothesis against which to test the alternative hypothesis but only a twisted and computationally complicated way to try to go around a problem without truly addressing it. Yet, bootstrap was not developed to do this but only to estimate the error and confidence intervals around a metric taken from a sample (Efron and Tibshirani, 1986; Manly, 1991; Lebart, 2006). On the contrary, permutation tests do break any correlation in the original data set by randomly and independently permuting each variable so that the alternative hypothesis of meaningful correlations can be truly tested against the null hypothesis of no correlations (Dijksterhuis and Heiser, 1995; Peres-Neto et al., 2005). Furthermore, bootstrap methods have to account for axis reflection and reordering whereas the permutation tests gave simple solutions to these problems. Axis reflection, that is the permutation of signs between loadings (Jackson, 1995; Mehlman et al., 1995; Peres-Neto et al., 2003, 2005), was solved by the *index for the loadings* as both eigenvalues and loadings are squared. Axis reordering, i.e: permutations on the rank of the pc (Jackson, 1995; Peres-Neto et al., 2003, 2005), is specific to bootstrap. It does not occur with the permutation tests as the new generated axis relate only to error. Mixed techniques have been tested that compare the average or confidence interval of a statistic taken from *n* randomly permuted data set with the average of the statistic taken from *m* bootstrapped data sets (see Peres-Neto et al., 2005). Comparison with the bootstrapped average seems odd as its expected value is that of the statistic taken from the original data set if bootstrap resampling is honest and any departure from this may only come from error. Maybe a comparison with a bootstrapped statistic other than the average would seem more natural.

Permutation tests proved to be an efficient tool whether it was to decide if it was worthwhile to submit the data set to a multivariate analysis, to decide about how many dimensions the data set should be reduced to or to interpret the new extracted dimensions. Besides the data sets presented in the examples, others were tested which comprehended both real data and manipulated data to simulate different degrees and patterns of correlations. The results were always coherent with what was expected. Generally, when submitted to a data set, most statistics tended to yield the same results. In particular, when a data set had been optimized (in the sense that correlated variables were kept while uncorrelated ones were sorted out) these statistics tended to converge. On the contrary, when a data set contained much noisy information, like many uncorrelated variables or variables which tended to be linear combinations of others, the results given by different statistics tended to diverge. This was a “rule of thumb” to find the optimized data set. As an example, both the data sets about employment in Europe and the twelve abiotic variables from Ria Formosa had originally more variables than the presented and were reduced with the aid of this criterion.

About whether to start a PCA, all statistics including the *ψ index* were coherent among each other despite their different equations. However, when submitted to other data sets where the number of variables exceeded the number of samples, only the *ψ index* and the *φ statistics* remained valid and coherent.

About the number of dimensions to retain, Peres-Neto et al. (2005) had already proven the permutation tests to give better results than bootstrap techniques, comparisons to predefined models as the *broken stick*, or tests under the assumption of multivariate normal distributions. The present results showed the permutation tests agree with the SCREE plot and the *broken stick* model when applied to data sets with a relatively small number of variables, whereas they perform better with data sets with a large number of variables. The results confirmed the proportion of variance explained is not an acceptable rule to use (see Peres-Neto et al., 2005) as there is nothing to ascertain that all data sets always contain the same predetermined fraction of meaningful variation. Neither it is wise to retain a pc just because it justifies more variation than one standardized variable. The present results confirmed that in data sets of many variables may be obtained many pc with eigenvalues higher than 1 which still relate only to error (Jackson, 1991; Dijksterhuis and Heiser, 1995; Peres-Neto et al., 2005). From the four statistics applied to the permutation tests only the *rank of the roots* proved to be a reliable stopping rule. The assumptions that sustain the test for the *equality of the roots*, the *ratio of the roots* and the *pseudo F ratio of the roots* were proven to be false: the results showed two consecutive pc can justify variation of approximate magnitude and still be non-trivial. Therefore, these statistics promoted type II error when applied to larger non-trivial pc whereas they promoted type I error when applied to smaller, trivial pc. Another methodology to choose non-trivial pc is to retain all which have at least two loadings found significant (Zwick and Velicer, 1986; Jackson, 1993; Peres-Neto et al., 2005). This was set aside as with data sets with many variables and relatively few sampling units it is very likely to happen by chance alone.

The interpretation of the new dimensions was given by the association between the original variables and the extracted eigenvectors. When applied to relatively small data sets the *correlations of the pc with the variables* promoted Type II error in the non-trivial pc and Type I error in the trivial pc resulting from the weakness of this metric against strong correlations within error. In this case the *index for the loadings* performed better. On the other hand, for data sets with many variables the *correlations of the pc with the variables* is probably a better option as it showed more power of synthesis whereas the *index for the loadings* had a tendency to promote Type I error in the non-trivial pc. Overall, there is no definite rule or best index to use. It is always advisable to use both indexes and compare their results with each other, with the correlation matrix and with the previous knowledge about the data set and research subject.

Estimating the z scores from the sv allowed sorting out error and thus enhanced any graphical representation or statistical inference from these scores. It also permitted recalculating the original variables from the z-scores differentiating the predicted residuals of each original variable in each sample as a function of both the processes represented by each pc and by error. It further enabled to divide the total variation into packages of variation distributed among variables and pc. Comparing with the terminology from Factor Analysis, the significant packages could be thought as the communality of the respective variable whiles the non-significant packages as the specificity of the variable.

Acknowledgements

The Author is grateful to Bryan Manly and Margarida Castro for their valuable comments regarding this work.

References

Cabaço S, Machás R, Vieira V, et al. 2008. Impacts of urban wastewater discharges in seagrass meadows (Zostera noltii). Estuarine, Coastal and Shelf Science, 78: 1-13

Dijksterhuis GB, Heiser WJ. 1995. The role of permutation tests in exploratory multivariate data analysis. Food Quality and Preference, 6: 263-270

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1(1): 54-75

Ferrarini A. 2011. A fitter use of Monte Carlo simulations in regression models. Computational Ecology and Software, 1(4): 240-243

Frisch R. 1929. Correlation and scatter in statistical variables. Nordic Statistical Journal, 8: 36-102

Gauch Jr HG. 1982. Noise reduction by eigenvector ordinations. Ecology, 63(6): 1643-1649

Gleason TC, Staelin R. 1975. A proposal for handling missing data. Psychometrika, 40: 229-252

Grossman GD, Nickerson DM, Freeman MC. 1991. Principal component analyses of assemblage structure data: Utility of tests based on eigenvalues. Ecology, 72(1): 341-347

Guerra LMA. 2010. Análise multivariada da Sucessão dunar na Península do Ancão e Ilha da Barreta. PhD Thesis. Universidade do Algarve. Faculdade de Ciências do Mar e Ambiente, Faro, Portugal

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

Jackson DA. 1995. Bootstrapping principal components analysis: reply to Mehlman et al. Ecology, 76(2): 644-645

Jackson JE. 1991. A User's Guide to Principal Components. John Wiley & Sons, New York, USA

Jolliffe IT. 2002. Principal Component Analysis. Springer-Verlag, New York, USA

Kullback S. 1959. Information Theory and Statistics. John Wiley & Sons, New York, USA

Lebart L. 2007. Which Bootstrap for Principal Axes Methods? In: Selected Contributions in Data Analysis and Classification (Brito P, Cucumel G, Bertrand P, de Carvalho F, eds). 581–588, Springer, Berlin, Heidelberg, Germany

Manly BJF. 1986. Multivariate Methods. Chapman & Hall, London, UK

Manly BJF. 1991. Randomization and Monte Carlo Methods in Biology. Chapman & Hall, London, UK

Mehlman DW, Shepherd UL, Kelt DA. 1995. Bootstrapping principal components analysis: A comment. Ecology, 76(2): 640-643

Peres-Neto PR, Jackson DA, Somers KM. 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology, 84: 2347-2363

Peres-Neto PR, Jackson DA, Somers KM. 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Computational Statistics and Data Analysis, 49: 974-997

Stauffer DF, Garton EO, Steinhorst RK. 1985. A comparison of principal components from real and random data. Ecology, 66(6): 1693-1698

ter Braak CFJ. 1988. CANOCO – a Fortran program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1). Agricultural Mattematic Group, Report LWA-88-02, Wagningen, Netherlands

ter Braak, CFJ. 1990. Update notes: CANOCO (version 3.1). Agricultural Mattematic Group, Report LWA-88-02, Wagningen, Netherlands

Yu CC, Quinn JT, Dufournaud CM, et al. 1998. Effective dimensionality of environmental indicators: a principal component analysis with bootstrap confidence intervals. Journal of Environmental Management, 53: 101-119

Zhang WJ. 2011a. A Java algorithm for non-parametric statistic comparison of network structure. Network Biology, 1(2): 130-133

Zhang WJ. 2011b. A Java program to test homogeneity of samples and examine sampling completeness. Network Biology, 1(2): 127-129

Zhang WJ. 2011c. Simulation of arthropod abundance from plant composition. Computational Ecology and Software, 1(1): 37-48

Zhang WJ, Zheng H. 2012. A program for statistic test of community evenness. Computational Ecology and Software, 2(1): 80-82

Zwick RW, Velicer WF. 1986. Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99: 432-442