Plant Ecol (2015) 216:657–667
DOI 10.1007/s11258-014-0406-z

Principal component analysis with missing values: a comparative survey of methods

Stéphane Dray · Julie Josse

Received: 28 February 2014 / Accepted: 21 August 2014 / Published online: 19 November 2014
© Springer Science+Business Media Dordrecht 2014

Abstract Principal component analysis (PCA) is a standard technique to summarize the main structures of a data table containing the measurements of several quantitative variables for a number of individuals. Here, we study the case where some of the data values are missing and propose a review of methods which accommodate PCA to missing data. In plant ecology, this statistical challenge relates to the current effort to compile global plant functional trait databases producing matrices with a large amount of missing values. We present several techniques to consider or estimate (impute) missing values in PCA and compare them using theoretical considerations. We carried out a simulation study to evaluate the relative merits of the different approaches in various situations (correlation structure, number of variables and individuals, and percentage of missing values) and also applied them on a real data set. Lastly, we discuss the advantages and drawbacks of these approaches, the potential pitfalls and future challenges that need to be addressed in the future.

Keywords Imputation · Ordination · PCA · Traits

Communicated by P. R. Minchin and J. Oksanen.

Electronic supplementary material The online version of this article (doi:10.1007/s11258-014-0406-z) contains supplementary material, which is available to authorized users.

S. Dray (✉)
Université de Lyon, 69000 Lyon, France
e-mail: stephane.dray@univ-lyon1.fr

S. Dray
Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France

J. Josse
Applied Mathematics Department, Agrocampus Ouest, Rennes, France

Introduction

Studies in community ecology aim to understand how and why individuals of different species co-occur in the same location at the same time. Hence, ecologists usually collected and stored data on species distribution as tables containing the abundances of the different species in a number of sampling sites. Additional information (e.g., measures of environmental variables or species traits) can also be recorded to examine the effects of abiotic and biotic features on observed assemblage structures. Since the early work of Goodall (1954) who applied principal component analysis (PCA) to vegetation data, multivariate analyses have been and remain intensively used to summarize the main structures of ecological data sets. Standard multivariate techniques like PCA are based on the eigendecomposition of a cross-product matrix (e.g., covariance matrix) and thus require complete

data sets. Whatever precaution we take, ecological data tables can contain missing values and then need a particular attention during the statistical analysis.

At the time of global change, a better understanding of ecological processes could be provided by studies at larger temporal and/or spatial scales (e.g., Wright et al. 2004). Hence, several projects aim to build worldwide repositories by compiling data from preexisting databases. For instance, the TRY initiative (Kattge et al. 2011) compiles plant traits data on global scale and contains almost three million entries for 69,000 species. However, due to the wide heterogeneity of measurement methods and research objectives, these huge data sets are often characterized by an extraordinarily high number of missing values (Swenson 2014). Hence, in addition to ecological questions, such data sets also present some important methodological and technical challenges for multivariate analysis.

Rubin (1976) distinguished three mechanisms generating missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR means that the probability that an observation is missing is not related to its value or to any other values in the data set. MAR means that the probability that an observation is missing is related to the values for some other observed variables. Finally, MNAR means that the probability that an observation is missing is related to its value. Depending on the proportion and the generating mechanism of missing data, different strategies can be envisaged to apply PCA on an incomplete data set. The most common approach is to delete individuals and/or variables containing missing observations and perform standard PCA. However, this loss of information reduces the ability to detect patterns and can also introduces biases if data are not MCAR (Nakagawa and Freckleton 2008). A second strategy consists in imputing (i.e., estimating) missing values and then applying PCA on the completed data table. The simplest approach is to replace missing values by the mean of the variables but more sophisticated techniques can improve the imputation by considering the correlation structure between the observed variables or external information (e.g., phylogenetic proximities among species in Swenson 2014). Lastly, some procedures adapt the standard PCA algorithm either by skipping or by considering the missing values in the computation of PCA outputs (e.g., Wold and Lyttkens 1969; Kiers 1997).

The aim of this paper was to compare different approaches to perform PCA on an incomplete data set using simulated and real plant traits data sets. Functional trait analyses either focus on the ordination of species (identification of functional types, Diaz and Cabido 1997), on the quantification of traits covariations (Wright et al. 2004), or on the estimation of missing trait values (Shan et al. 2012; Swenson 2014). Hence, contrary to other recent works (e.g., Brown et al. 2012), our study compares methods by considering these three different aspects (imputation of missing values, PCA scores for species and traits) and by evaluating the effect of several parameters (number of traits, number of species, proportion of missing values, correlation structure, and generating mechanism for missing values). We also applied and compared the different approaches on the GLOPNET (a multi-investigator group accumulating and studying global data on plant traits) data set (Wright et al. 2004). We provide the R code and functions to help ecologists to reproduce the analyses and apply methods on other real data sets.

Material and methods

Statistical methods

Let $\mathbf{X} = [x_{ik}]$ contains the measurements of p variables (e.g., traits) for n individuals (e.g., species). We consider that variables are centered and scaled. PCA of $\mathbf{X}$ finds a matrix $\hat{\mathbf{X}}_S$ ($n \times p$), the best approximation of $\mathbf{X}$ of rank S, by minimizing the least squares norm:

$$\|\mathbf{X} - \hat{\mathbf{X}}\|^2.$$

The solution is provided by the singular value decomposition of $\mathbf{X}$:

$$\hat{\mathbf{X}}_S = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}^\top,$$

where the matrices $\mathbf{U}$ ($n \times S$) and $\mathbf{V}$ ($p \times S$) contain the first S left and right singular vectors, and $\boldsymbol{\Lambda}$ ($S \times S$) is the diagonal matrix with the associated eigenvalues (or squared singular values).

The matrix $\mathbf{V}$ contains the loadings and allows to obtain the scores for the individuals ($\mathbf{F} = \mathbf{XV}$). Scores for the variables can also be computed and are equal to $\mathbf{G} = \mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}$. Note that PCA outputs can also be obtained by the eigendecomposition of the correlation matrix $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$.

When the table **X** is not complete, the standard PCA algorithm cannot be applied, and alternative methods should be used. A popular approach consists in deleting individuals and/or variables containing missing values. As this approach leads to an important loss of information, it is not considered in our comparison study. We describe below three other strategies and associated methods to deal with missing values in PCA.

Imputation of missing values prior to standard PCA

A first strategy consists in filling the gaps with plausible values. A completed data set is then obtained, and it can be analyzed by a standard PCA providing loadings and scores for variables and individuals. We consider two methods:

Mean: The *mean imputation* is probably the simplest method. It replaces missing values for each variable by the mean of the observed values. This approach is satisfactory for a small amount of MCAR-generated missing values. However, it distorts the distribution of the data by reducing the variance of the imputed variables and the correlations between variables (Little and Rubin 2002).

JointM: The *joint modeling* approach imputes the missing values with the underlying assumption that data can be described by a multivariate distribution, usually the multivariate normal distribution. The maximum likelihood estimates for the parameters (the vector of means and the covariance matrix) are obtained from the incomplete data set using an expectation–maximization (EM) algorithm (Dempster et al. 1977). More details about this approach can be found in Schafer (1997) and in Little and Rubin (2002). Contrary to the **Mean** method, this approach takes into account the relationships between variables to fill the gaps. The imputation is based on a linear regression that implies two main restrictions: the level of collinearity among variables should be moderate, and the number of individuals should be higher than the number of parameters to estimate.

We used the function `amelia` of the R package Amelia (Honaker et al. 2011) to run this method. This package performs multiple imputation (Rubin 1987) so that several imputed data sets are generated, reflecting the variability of the prediction of the missing values. We considered 5 imputations, and a single estimate of missing values is then obtained by averaging.

PCA algorithms skipping missing values

In the second strategy, the standard PCA algorithm is adapted so that missing values are not considered in the computation.

GowPCoA: This method is based on the equivalence between the representation of individuals (individual scores) produced by PCA and those given by the Principal Coordinates Analysis (PCoA, Gower 1971a) of the Euclidean distance matrix computed between the individuals. In this latter approach, Gower (1971b) suggested to deal with missing values using pairwise deletion so that missing values are removed when computing dissimilarities (Pavoine et al. 2009). More precisely, Euclidean distances ($d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$) can be computed in the presence of missing values as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \delta_{ijk}(x_{ik} - x_{jk})^2 \Big/ \sum_{k=1}^{p} \delta_{ijk}},$$

where δ_{ijk} is equal to 0 if the value of the kth variable is missing for one of the two individuals i and j and 1 otherwise. We considered a constraint to avoid some null distances: each pair of individuals should have data available for at least one common variable, otherwise the distance matrix is not computed. A shortcoming of this method is that there is no guarantee that the distance matrix is Euclidean (i.e., individuals cannot be represented in an Euclidean space) implying that PCoA returns negative eigenvalues. Corrections for negative eigenvalues may be obtained by adding a constant to the distance matrix (Lingoes 1971; Cailliez 1983). This **GowPCoA** strategy provides only scores for the individuals: missing values cannot be imputed, and no outputs concerning the variables are produced.

We used the function `vegdist` of the R package vegan (Oksanen et al. 2013) to compute the distances matrix in combination with the function `lingoes` and `dudi.pco` of the R package ade4 (Dray and Dufour 2007) to perform the PCoA on the transformed distance matrix.

PairCor: The *pairwise correlation* approach computes the correlation matrix using only the observed

values for each pair of variables independently. Hence, each pair of variables should have data available for at least two common individuals. With this method, different correlation coefficients are not necessarily based on the same individuals or the same number of individuals. It can be seen as the equivalent of the **GowPCoA** but for variables. Then, PCA is achieved by the eigendecomposition of the resulting correlation matrix. It can produce negative eigenvalues, and associated dimensions should not be interpreted. This method only provides scores for the variables.

PCA algorithms considering missing values

This last family of methods adapts the PCA algorithm to consider explicitly the missing values. These procedures return scores for both variables and individuals using an incomplete data set.

Nipals: NIPALS (non-linear iterative partial least squares, Wold and Lyttkens 1969) is an algorithm that provides sequentially the PCA scores and loadings of a complete data set. The scores $\mathbf{F}_1$ and the loadings $\mathbf{V}_1$ on the first dimension are estimated by alternating two steps of simple linear regression (it is an alternating least-squares algorithm). The scores and loadings for the second dimension $(\mathbf{F}_2, \mathbf{V}_2)$ are obtained using the same approach with an additional deflation procedure (computation of a residual matrix $\mathbf{X} - \mathbf{F}_1\mathbf{V}_1^\top$) to ensure orthogonality between subsequent dimensions. This method is easily extended to incomplete data sets using weighted regressions with null weights for the missing entries. This simplicity to accommodate missing data may explain its relative success (see Dray et al. 2003 for an application in ecology). This method suffers, however, from several shortcomings in the presence of missing values: means and variances to standardize the data are only computed with observed values, it can encounter problems of convergence, the second eigenvalues can be larger than the first one, etc. We used the function `nipals` of the R package ade4 (Dray and Dufour 2007).

Ipca: The *iterative PCA* method (Kiers 1997) also known as the EM-PCA algorithm (Josse and Husson 2012) is based on a stronger theoretical framework. It provides the scores and loadings minimizing the least squares criterion on the observed entries, $\|\mathbf{W} \circ (\mathbf{X} - \hat{\mathbf{X}})\|^2$, with $w_{ik} = 0$ if x_{ik} is missing and 1 otherwise and $\circ$ denotes the elementwise product. Hence, this approach is optimal according to the PCA criterion. The minimization is achieved through an iterative procedure: missing values are replaced by random values, and then PCA is applied on the completed data set, and missing values are then updated by the fitted values ($\hat{\mathbf{X}}_S = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$) using a predefined number of dimensions S. The procedure is repeated until convergence. This method provides scores for the individuals and the variables, and also an imputation for the missing values. In the case of standardized PCA, estimates of means and variances are also updated at each iteration. This algorithm often leads to overfitting problems that are solved by the regularized iterative PCA proposed by Josse et al. (2009). An important issue also concerns the number of dimensions S that should be defined at the beginning of the (regularized) iterative PCA algorithm but Josse and Husson (2012) suggested methods based on cross-validation to estimate this parameter from an incomplete data set. The method is implemented in the function `imputePCA` of the R package missMDA (Husson and Josse 2010).

Real and simulated data

We used two complementary approaches. Simulated data are used to compare the relative merits of methods in different contexts. Then, we provided a case study by the analysis of real data as would have been collected by a typical user of the methods considered here.

We used the procedure described in Peres-Neto et al. (2005) and Dray (2008) to generate normally distributed data with a specific correlation structure. We varied the number of individuals $n = \{20, 50, 100\}$, the number of variables $p = \{9, 18, 45\}$, and then introduced different proportions of missing values $p_M = \{0.1, 0.2, 0.5\}$ on all the variables. Missing values were randomly assigned to simulate a MCAR mechanism. For $p_M = 0.2$, we also generated MNAR data: the 20 % highest values of the first variable were replaced by missing values. Hence, p_M refers either to the percentage of missing values for the complete data set (MCAR) or for the first variable (MNAR). We considered 3 correlation matrices $(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)$ to generate data. The first two matrices contain three blocks of variables of decreasing size

(respectively, $4p/9$, $3p/9$, and $2p/9$ variables). In each block, the correlation between variables is equal to 0.8 for $\mathbf{M}_1$ and 0.3 for $\mathbf{M}_2$; variables from two different blocks are uncorrelated. In these two scenario, there are three underlying dimensions. In the last correlation matrix $\mathbf{M}_3$, all the variables are highly and equally correlated (0.8), and consequently there is only one dimension. Thus, we obtain 36 combinations of the parameters, and for each of them we generated 500 data sets. To assess the performances of the methods, we reported the number of times that the algorithm does not return outputs (no convergence, not applicable due to the numerical conditions or to the distribution of missing values). When outputs are produced, we computed the RV coefficient (Escoufier 1973) to evaluate the agreement between the scores for the individuals (respectively, the variables) obtained by the different methods and those obtained from the standard PCA of the complete data set. The RV coefficient is an extension of a correlation coefficient for matrices and varies between 0 and 1. A value of 1 indicates a perfect agreement between configurations. These comparisons were based on using the true number of dimensions. We also computed an imputation error defined as the average squared difference between the estimated values and the true ones.

We also analyzed the GLOPNET data set (Wright et al. 2004) that contains 6 traits measured for 2494 plant species: LMA (leaf mass per area), LL (leaf lifespan), Amass (photosynthetic assimilation), Nmass (leaf nitrogen), Pmass (leaf phosphorus), and Rmass (dark respiration rate). The last four variables are expressed per leaf dry mass. GLOPNET is a compilation of several existing data sets and thus contains a large proportion of missing values. All traits were log-normally distributed and log-transformed before analysis. As the real values of missing entries are not known, we computed the RV coefficients between the outputs produced by the different methods to evaluate their agreement on an incomplete real data set.

Results

Simulation study

The full results of the simulation study are provided in Table A1-A12 of electronic supplementary material 1. The **Mean** and **Ipca** approaches return outputs for all simulations. The **PairCor** (6.44 % of non-returned results), **Nipals** (14.62 %), **GowPcoA** (16.19 %), and **JointM** (91.51 %) methods fail to perform PCA in many cases. The lack of convergence for **Nipals** is mainly observed for the correlation matrix $\mathbf{M}_2$ or for a high proportion of missing values (Table A1, A2). **PairCor** did not return results mainly for the highest level of missing values ($p_M = 0.5$) and low number of individuals ($n = 20$), whereas **GowPCoA** could not be performed for data sets with moderate number of variables and many gaps ($p_M = 0.5$, $p = 9, 18$). Estimation by **JointM** can only be performed for $p = 9$ and $n = 100$ due to the limitation of this method concerning the number of individuals and the level of collinearity between variables.

Table 1 Imputation of the missing values: average performance of the methods according to the correlation structure and the mechanism to generate missing values

	Ipca	JointM	Mean
MCAR			
$\mathbf{M}_1$	0.454	–	1.005
	0.365	*0.592*	*0.838*
$\mathbf{M}_2$	1.248	–	1.007
	0.874	*1.379*	*0.845*
$\mathbf{M}_3$	0.230	–	0.990
	0.201	*0.290*	*0.839*
MNAR			
$\mathbf{M}_1$	0.547	–	3.142
	0.618	*0.667*	*3.203*
$\mathbf{M}_2$	2.481	–	3.171
	2.657	*2.771*	*3.270*
$\mathbf{M}_3$	0.466	–	3.138
	0.513	*0.565*	*3.252*

Averages are computed for all combinations of parameters (plain) or only for the combinations for which all methods return results (italics)

Using the full results, some general trends are observed. For the **Mean**, there is no effect of the number of individuals, variables, and proportion of missing values when data are MNAR. For all other methods, as expected, increasing the number of individuals and/or variables and reducing the proportion of missing values provide a better imputation and a better agreement with the PCA outputs of the complete data set (Table A4-A12). Tables 1, 2, and 3 summarize all the results by giving the average of the imputation error and the RV coefficients for the scores

Table 2 Scores for the variables: agreement with the PCA of the complete data set (average RV coefficient) according to the correlation structure and the mechanism to generate missing values

	Ipca	JointM	Mean	Nipals	PairCor
MCAR					
$\mathbf{M_1}$	0.968	–	0.948	0.942	–
	0.988	*0.980*	*0.982*	*0.950*	*0.985*
$\mathbf{M_2}$	0.817	–	0.834	0.785	–
	0.872	*0.866*	*0.887*	*0.835*	*0.891*
$\mathbf{M_3}$	0.632	–	0.179	0.269	–
	0.615	*0.593*	*0.219*	*0.267*	*0.495*
MNAR					
$\mathbf{M_1}$	0.999	–	0.993	0.999	0.996
	1.000	*1.000*	*0.993*	*0.999*	*0.997*
$\mathbf{M_2}$	0.980	–	0.976	0.979	0.975
	0.983	*0.981*	*0.978*	*0.983*	*0.979*
$\mathbf{M_3}$	0.965	–	0.225	0.864	0.576
	0.930	*0.921*	*0.146*	*0.665*	*0.309*

Averages are computed for all combinations of parameters (plain) or only for the combinations for which all methods return results (italics)

Table 3 Scores for the individuals: agreement with the PCA of the complete data set (average RV coefficient) according to the correlation structure and the mechanism to generate missing values

	GowPcoA	Ipca	JointM	Mean	Nipals
MCAR					
$\mathbf{M_1}$	–	0.938	–	0.910	0.907
	0.934	*0.975*	*0.969*	*0.939*	*0.945*
$\mathbf{M_2}$	–	0.782	–	0.801	0.751
	0.868	*0.886*	*0.861*	*0.878*	*0.837*
$\mathbf{M_3}$	–	0.990	–	0.964	0.989
	0.976	*0.994*	*0.992*	*0.976*	*0.994*
MNAR					
$\mathbf{M_1}$	0.990	0.998	–	0.992	0.997
	0.976	*0.996*	*0.996*	*0.981*	*0.992*
$\mathbf{M_2}$	0.971	0.978	–	0.974	0.979
	0.949	*0.963*	*0.959*	*0.956*	*0.964*
$\mathbf{M_3}$	0.997	0.999	–	0.998	0.999
	0.993	*0.999*	*0.999*	*0.994*	*0.998*

Averages are computed for all combinations of parameters (plain) or only for the combinations for which all methods return results (italics)

for the variables and individuals across all the simulations.

The imputation by the **Mean** is not influenced by the correlation structure but strongly deteriorated when missing data are MNAR distributed (Table 1). **Ipca** provides more often and more accurate imputations than **JointM**; both methods perform better when correlations between variables are stronger (Table 1). Concerning the agreement for PCA scores of the variables (Table 2), the **Mean** and **Nipals** methods provide similar results, and the **PairCor** is slightly better, whereas **Ipca** returns the best estimates except when correlations are lower ($\mathbf{M_2}$). **Mean**, **PairCor**, and **Nipals** poorly perform when all variables are highly correlated ($\mathbf{M_3}$). Concerning the scores of individuals (Table 3), differences among methods are less noticeable. All methods are efficient, and the agreement is better when correlations between variables are stronger.

GLOPNET data set

53.38 % of the entries in the GLOPNET data set are missing. Only 72 species have complete information for the 6 traits and the proportion of missing values varied between 4.97 % (LMA) and 89.01 % (Rmass). It was not possible to apply the **GowPCoA** method because it frequently happens that two species have no observation for a common trait. Scores for species on the first two axes are represented in Fig. 1. Configurations obtained by **Nipals**, **JointM**, and **Ipca** are very coherent (RV coefficients between 0.96 and 0.98, Fig. 1b), whereas the graphical representation obtained by the **Mean** imputation highlights a very particular shape indicating that results are not reliable. Considering the variables, we also applied the **PairCor** approach that was used in Wright et al. (2004). This first axis corresponding to the "leaf economic spectrum" separates species with potential for quick returns for investment with high values for Nmass, Amass, Rmass, and Pmass, and low values for LL and LMA (right part) from species with slow returns on the left part (Fig. 2a). Scores for the traits are very consistent between methods, to a lesser extent for the Mean (Fig. 2b). The percentage explained by the leaf economic spectrum compared to the global trait variability was equal to 74.3 % for **PairCor**. It varies between 44.8 % (**Mean**) and 91.2 % for **Ipca** (Fig. 1a). It was not possible to compute this percentage for

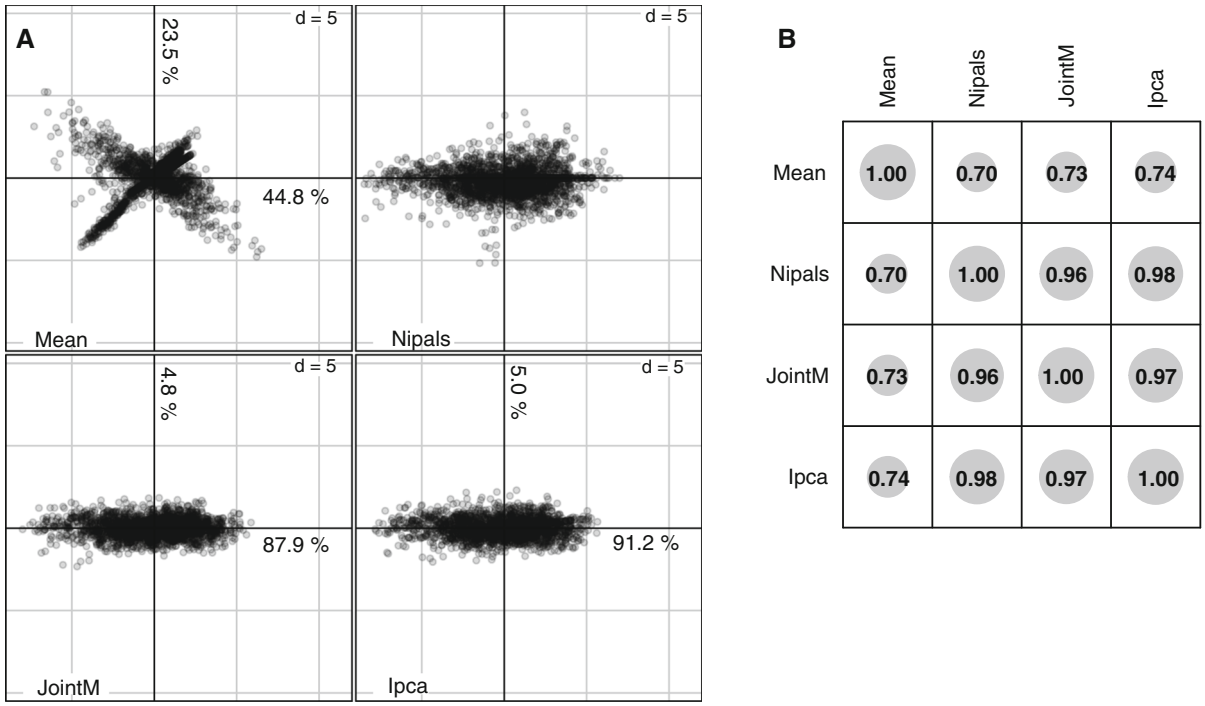

Fig. 1 Analysis of the GLOPNET data set: scores for species. **a** Scores for the two first axes produced by four different procedures. **b** RV coefficients computed between these scores

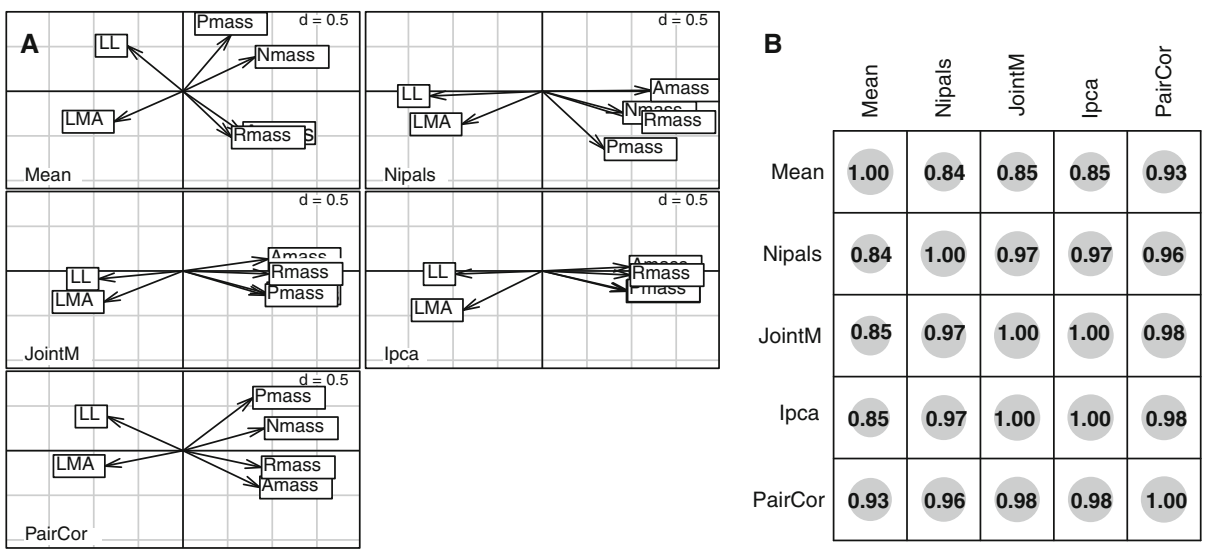

Fig. 2 Analysis of the GLOPNET data set: scores for traits. **a** Scores for the two first axes produced by five different procedures. **b** RV coefficients computed between these scores

Nipals as the algorithm did not converge for the 6 dimensions. Moreover, as **Nipals** does not handle properly the scaling of the variables, its first eigenvalue (6.44) is higher than the expected total variability (6, the number of traits included in the normed PCA) and some associated traits scores are

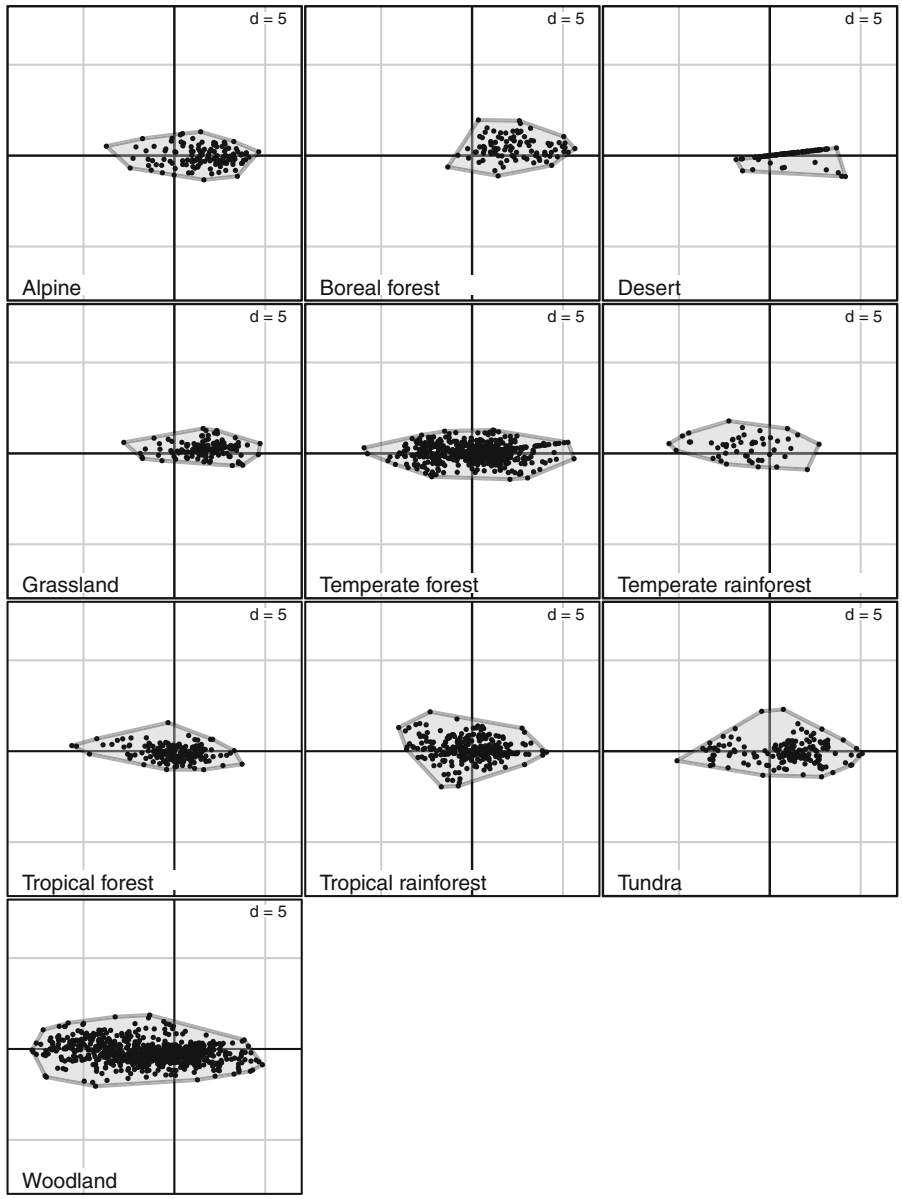

Fig. 3 Analysis of the GLOPNET data set by the Ipca procedure. Representation of species for the two first axes. Species are split by major biomes and surrounded by convex hulls

greater than 1, whereas they are usually interpreted as squared correlations in the case of normed PCA.

Contrary to the **PairCor** approach, other methods are able to produce species scores. This representation can be used to add external information: grouping species by major biomes illustrates the universality of the leaf economic spectrum (several hulls cover a large extent of the first axis as shown in Fig. 3) but also some specificities (e.g., Desert and Boreal forest mainly contain species of the quick-return end).

Discussion

Multivariate analysis of incomplete data sets has received little attention in ecology. This lack of consideration can be explained either by the infrequency of missing values in ecological data sets or by the ignorance of the issues and problems related to this question by ecologists. Both reasons are probably legitimate: as missing values rarely occur in ecological data sets, users have favored simple methods such as

individual and variable deletion or mean imputation to handle incomplete data. The development of collaborative tools to gather existing data sets into worldwide databases provides a great opportunity to study ecological patterns at larger scale but is quite challenging in terms of statistical analysis. These data often characterized by a large proportion of missing values, and require an adequate treatment. Our analysis of the GLOPNET data set illustrates clearly this issue: more than 50 % of the data are missing which makes the deletion of species or traits impossible, whereas the mean imputation (**Mean**) did not provide reliable results. In this context, the use of more sophisticated techniques is required.

The simulation study demonstrated that the mean imputation is not affected by the correlation structure, the number of individuals, and variables or the proportion of missing values. However, when missing values are MNAR, mean imputation estimates are strongly biased, and its use should be avoided (Table 1). In the MNAR scenario, other methods that take into account the relationships between variables perform better. On the other hand, when variables are poorly correlated (correlation matrix $\mathbf{M}_2$), taking into the correlation structure is not an advantage, and **Mean** is a relevant alternative. The analysis of the GLOPNET data set illustrates another problem of the **Mean** approach: when a variable has too many missing values, scores for individuals are located orthogonally to its direction as the replacement by a single value reduces artificially the variation for this variable (Fig. 1a).

The pairwise correlation (**PairCor**) and the Gower PCoA (**GowPCoA**) approaches are *ad hoc* techniques that patch the PCA algorithm to skip missing values. Their main advantage is their simplicity of implementation, and they produce acceptable results compared to more sophisticated methods. Both methods do not impute missing values, and they perform half of a standard PCA by returning scores only for the variables (**PairCor**) or the individuals (**GowPCoA**). These techniques cannot be applied in all situations: **PairCor** (respectively **GowPCoA**) requires that each pair of variables (resp. individuals) has observed measures for two common individuals (resp. variables). Hence, we were not able to run **GowPCoA** on the GLOPNET data set, and we demonstrated that these methods could rarely be ran for an high level of missing values and a low number of variables (**GowPCoA**) or a low number of individuals (**PairCor**). These two approaches do not ensure the positive definiteness of the diagonalized matrix leading to negative eigenvalues. A simple alternative is to achieve positive definiteness by adding or removing a small quantity to the distances (Lingoes 1971; Cailliez 1983) or the correlation matrix (Yuan and Chan 2008; Bentler and Yuan 2011; Yuan et al. 2011). In practice, this corresponds to a modification of the observed values, and it reduces artificially the signal of the structures present in the data. From a more theoretical viewpoint, it should be reminded that PCA eigenvalues are interpreted as variances, and it means that these two approaches would produce negative variances that are undefined. As these techniques did not outperform other methods, we advocated the use of better theoretically grounded approaches such as **Ipca** or **JointM**.

Nipals provides both scores for individuals and variables but fails to converge in many cases when the variables are poorly correlated or when an high proportion of data is missing. Means and variances are only computed with observed values. Moreover, as it is an iterative algorithm, it is not able to provide percentage of variation explained by each dimension if all axes are not computable (Fig. 1a). Lastly, this procedure does not impute the missing values.

Ipca and **JointM** provide the best estimates for missing values when variables are highly correlated. As these methods take into account the correlations among variables during the imputation, they perform well even if data are MNAR, contrary to the **Mean** approach. Both techniques return complete PCA outputs (scores for individuals and variables on all dimensions). **JointM** assumes a multivariate normal distribution for the data and requires that the number of individuals is greater than the number of estimated parameters (i.e., $n > (p^2 + 3p)/2$). Hence, this approach can only be applied on big data sets such as GLOPNET. On the other hand, **Ipca** can also be applied for small data sets even if $n < p$. However, this approach requires the *a priori* choice of the number of dimensions to impute the missing value. In our simulation study, we used the true number of dimensions, and this would probably overestimate the performance of the method and others compared to applications on real data sets where the true value of

this parameter is unknown. To solve this issue, Josse and Husson (2012) suggested the use of cross-validation to estimate this parameter. In the future, further works on the estimation of the number of dimensions in the presence of missing values are needed to improve the performance of imputation methods.

When missing values are imputed, it is natural that users wonder if and how the estimates are reliable. Hence, several authors have tried to estimate a maximum proportion of missing values that can be properly considered in imputation methods (e.g., Strauss et al. 2003). Recent studies (Brown et al. 2012; Clavel et al. 2014) demonstrated that no simple recommendation can be produced since this proportion can be affected by the imputation method, the correlation structure, or the number of individuals and variables. For instance, for a given proportion of missing values, estimates would be more reliable for a big data set with highly correlated variables. An alternative is to evaluate the uncertainty of estimates by measuring the variability of imputed data using a multiple imputation method (Clavel et al. 2014). It could then be possible to provide confidence intervals around the estimated values. Indeed, our study focused on bias, and thus the estimations obtained by multiple imputation methods were averaged prior to PCA. Hence, information about uncertainty around estimates was lost. Two alternatives to get confidence areas around the PCA scores could be envisaged. The first one used the axes defined by the PCA of the average data set and projects the multiple imputed data sets as supplementary information. The second one consists in performing PCA on each imputed data set and finding a common configuration using a multitable approach such as generalized Procrustes rotation (Gower 1975). In the first approach, the stability of the scores for individuals and variables is measured relative to a fixed set of PCA axes. On the other hand, the second method quantifies this variability using common axes defined by the different imputed data sets so that uncertainty of PCA axes can also be studied. However, dealing and combining the results of different imputed data sets in PCA remains still challenging, and further works are welcome to evaluate the advantages and drawback of both strategies (Josse et al. 2011).

We hope that this paper will help ecologists to consider properly missing values in multivariate analysis. Ignoring this issue would undoubtedly introduce some biases in ecological studies. We provide R script as electronic supplementary material 2 so that readers can reproduce our analysis of the GLOPNET data set.

Acknowledgments We would like to thank Peter Minchin and Jari Oksanen for the invitation to participate to this special issue and Gavin Simpson and an anonymous reviewer for comments on an earlier draft of the manuscript. We would like to warmly thank Ian Wright for freely distributing the GLOPNET data set.

References

Bentler P, Yuan K (2011) Positive definiteness via off-diagonal scaling of a symmetric indefinite matrix. Psychometrika 76:119–123

Brown CM, Arbour JH, Jackson DA (2012) Testing of the effect of missing data estimation and distribution in morphometric multivariate data analyses. Syst Biol 61(6):941–954

Cailliez F (1983) The analytical solution of the additive constant problem. Psychometrika 48(2):305–308

Clavel J, Merceron G, Escarguel G (2014) Missing data estimation in morphometrics: how much is too much? Syst Biol 63(2):203–218

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39(1):1–38

Diaz S, Cabido M (1997) Plant functional types and ecosystem function in relation to global change. J Veg Sci 8:463–474

Dray S (2008) On the number of principal components: a test of dimensionality based on measurements of similarity between matrices. Comput Stat Data Anal 52:2228–2237

Dray S, Dufour AB (2007) The ade4 package: implementing the duality diagram for ecologists. J Stat Softw 22(4):1–20

Dray S, Pettorelli N, Chessel D (2003) Multivariate analysis of incomplete mapped data. Trans GIS 7:411–422

Escoufier Y (1973) Le traitement des variables vectorielles. Biometrics 29:751–760

Goodall DW (1954) Objective methods for the classification of vegetation III. An essay on the use of factor analysis. Aust J Bot 2:304–324

Gower J (1971a) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

Gower JC (1971b) Statistical methods of comparing different multivariate analyses of the same data. In: Hodson FR, Kendall DG, Tautu P (eds) Mathematics in the archaeological and historical sciences. Edinburgh University Press, pp 138–149

Gower JC (1975) Generalized procrustes analysis. Psychometrika 40:33–51

Honaker J, King G, Blackwell M (2011) Amelia II: a program for missing data. J Stat Softw 45(7):1–47

Husson F, Josse J (2010) missMDA: handling missing values with/in multivariate data analysis (principal component methods). R package version 1:2

Josse J, Husson F (2012) Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique 153(2):1–21

Josse J, Pagès J, Husson F (2009) Gestion des données manquantes en analyse en composantes principales. Journal de la Société Française de Statistique 150(2):28–51

Josse J, Pagès J, Husson F (2011) Multiple imputation in PCA. Adv Data Anal Classif 5(3):231–246

Kattge J, Díaz S, Lavorel S, Prentice IC, Leadley P, Bönisch G, Garnier E, Westoby M, Reich PB, Wright IJ, Cornelissen JHC, Violle C, Harrison SP, Van Bodegom PM, Reichstein M, Enquist BJ, Soudzilovskaia NA, Ackerly DD, Anand M, Atkin O, Bahn M, Baker TR, Baldocchi D, Bekker R, Blanco CC, Blonder B, Bond WJ, Bradstock R, Bunker DE, Casanoves F, Cavender-Bares J, Chambers JQ, Chapin FS III, Chave J, Coomes D, Cornwell WK, Craine JM, Dobrin BH, Duarte L, Durka W, Elser J, Esser G, Estiarte M, Fagan WF, Fang J, Fernández-Méndez F, Fidelis A, Finegan B, Flores O, Ford H, Frank D, Freschet GT, Fyllas NM, Gallagher RV, Green WA, Gutierrez AG, Hickler T, Higgins SI, Hodgson JG, Jalili A, Jansen S, Joly CA, Kerkhoff AJ, Kirkup D, Kitajima K, Kleyer M, Klotz S, Knops JMH, Kramer K, Kühn I, Kurokawa H, Laughlin D, Lee TD, Leishman M, Lens F, Lenz T, Lewis SL, Lloyd J, Llusià J, Louault F, Ma S, Mahecha MD, Manning P, Massad T, Medlyn BE, Messier J, Moles aT, Müller SC, Nadrowski K, Naeem S, Niinemets U, Nöllert S, Nüske A, Ogaya R, Oleksyn J, Onipchenko VG, Onoda Y, Ordoñez J, Overbeck G, Ozinga WA, Patiño S, Paula S, Pausas JG, Peñuelas J, Phillips OL, Pillar V, Poorter H, Poorter L, Poschlod P, Prinzing A, Proulx R, Rammig A, Reinsch S, Reu B, Sack L, Salgado-Negret B, Sardans J, Shiodera S, Shipley B, Siefert A, Sosinski E, Soussana JF, Swaine E, Swenson N, Thompson K, Thornton P, Waldram M, Weiher E, White M, White S, Wright SJ, Yguel B, Zaehle S, Zanne AE, Wirth C, (2011) TRY—a global database of plant traits. Glob Change Biol 17(9):2905–2935

Kiers HAL (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrika 62(2):251–266

Lingoes J (1971) Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika 36(2): 195–203

Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley series in probability and statistics. Wiley, Hoboken, NJ

Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evolut 23(11): 592–596

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2013) Vegan: community ecology package. R package version 2.0-9

Pavoine S, Vallet J, Dufour AB, Gachet S, Daniel H (2009) On the challenge of treating various types of variables: application for improving the measurement of functional diversity. Oikos 118(3):391–402

Peres-Neto PR, Jackson DA, Somers KM (2005) How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal 49:974–997

Rubin D (1976) Inference and missing data. Biometrika 69(3):581–592

Rubin DB (1987) Multiple imputation for non-response in survey. Wiley, London

Schafer J (1997) Analysis of incomplete multivariate data. Chapman & Hall/CRC, London

Shan H, Kattge J, Reich PB, Banerjee A, Schrodt F, Reichstein M (2012) Gap filling in the plant kingdom—trait prediction using hierarchical probabilistic matrix factorization. Proceedings of the 29th international conference on machine learning (ICML-12). Edinburgh, Scotland, pp 1303–1310

Strauss RE, Atanassov MN, De Oliveira JA (2003) Evaluation of the principal-component and expectation-maximization methods for estimating missing data in morphometric studies. J Vertebr Paleontol 23(2):284–296

Swenson N (2014) Phylogenetic imputation of plant functional trait databases. Ecography 37:105–110

Wold H, Lyttkens E (1969) Nonlinear iterative partial least squares (NIPALS) estimation procedures. Bull Int Stat Inst 43:29–51

Wright IJ, Reich PB, Westoby M, Ackerly DD, Baruch Z, Bongers F, Cavender-Bares J, Chapin T, Cornelissen JHC, Diemer M, Flexas J, Garnier E, Groom PK, Gulias J, Hikosaka K, Lamont BB, Lee T, Lee W, Lusk C, Midgley JJ, Navas ML, Niinemets U, Oleksyn J, Osada N, Poorter H, Poot P, Prior L, Pyankov VI, Roumet C, Thomas SC, Tjoelker MG, Veneklaas EJ, Villar R (2004) The worldwide leaf economics spectrum. Nature 428:821–827

Yuan KH, Chan W (2008) Structural equation modeling with near singular covariance matrices. Comput Stat Data Anal 52(10):4842–4858

Yuan KH, Wu R, Bentler PM (2011) Ridge structural equation modelling with correlation matrices for ordinal and continuous data. Br J Math Stat Psychol 64:107–133