Ecological Informatics 61 (2021) 101235

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Principal component analysis of incomplete data – A simple solution to an old problem

János Podani [a,*], Tibor Kalapos [a], Barbara Barta [b], Dénes Schmera [b]

[a] Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös Loránd University, Pázmány P. s. 1/C, H-1117 Budapest, Hungary
[b] Centre for Ecological Research, Balaton Limnological Institute, Klebelsberg K. u. 3, H-8237 Tihany, Hungary

ARTICLE INFO

Keywords:
Biplot
Correlation
Functional trait
Missing data
Morphometry
Ordination

ABSTRACT

A long-standing problem in biological data analysis is the unintentional absence of values for some observations or variables, preventing the use of standard multivariate exploratory methods, such as principal component analysis (PCA). Solutions include deleting parts of the data by which information is lost, data imputation, which is always arbitrary, and restriction of the analysis to either the variables or observations, thereby losing the advantages of biplot diagrams. We describe a minor modification of eigenanalysis-based PCA in which correlations or covariances are calculated using different numbers of observations for each pair of variables, and the resulting eigenvalues and eigenvectors are used to calculate component scores such that missing values are skipped. This procedure avoids artificial data imputation, exhausts all information from the data and allows the preparation of biplots for the simultaneous display of the ordination of variables and observations. The use of the modified PCA, called InDaPCA (PCA of Incomplete Data) is demonstrated on actual biological examples: leaf functional traits of plants, functional traits of invertebrates, cranial morphometry of crocodiles and fish hybridization data – with biologically meaningful results. Our study suggests that it is not the percentage of missing entries in the data matrix that matters; the success of InDaPCA is mostly affected by the minimum number of observations available for comparing a given pair of variables. In the present study, interpretation of results in the space of the first two components was not hindered, however.

1. Introduction

Principal component analysis (PCA) has long been the most popular tool of multivariate data exploration in biology. Its theoretical foundations, computational details and conditions of application are described in a wide range of references, from standard texts of multivariate statistics (Jolliffe, 1986; Mardia et al., 1979) to more biologically oriented literature sources (Digby and Kempton, 1987; Legendre and Legendre, 1998; Orlóci, 1978). Like many other methods that utilize matrix algebra extensively, the standard PCA algorithm requires access to the entire data matrix to produce an ordination of observations (individuals, objects, sample units, specimens, etc.). Thus, the absence of even a single value may prevent the use of many commercial data analysis packages. It is a problem especially for biologists who are often faced with high percentages of values lacking from their data set for various reasons (see Brown et al., 2012 and Dray and Josse, 2015, for review and discussion with relevance to morphometrics and ecology, respectively).

This issue may be handled in very different ways. The most radical – and common – solution is to delete as many variables and/or objects from the data as necessary to reach completeness. Deletion always involves undesirable loss of information and does not work when too many rows and columns of the matrix have unknown values. The second possibility is imputation: the unknown scores are substituted by the mean of existing values for each variable or estimated through regression analysis. Mathematically sophisticated iterative algorithms starting from randomly imputed, estimated and iteratively refined or simulated values have also been suggested in various forms (Nelson et al., 1996; Grung and Manne, 1998; Oba et al., 2005; Stanimirova et al., 2007; Serneels and Verdonck, 2008; among others). These PCA variants may only be meaningful if the absence of scores is due primarily to random effects (data loss, unavailable information on certain variables for a given observation) – that is, if the values are indeed *missing* from their positions in the data array.

Some entries may also be lacking for logical reasons, however.

Abbreviations: InDaPCA, Principal Component Analysis of Incomplete Data; PairCor, Pairwise Correlation.

* Corresponding author.
E-mail address: podani@ludens.elte.hu (J. Podani).

https://doi.org/10.1016/j.ecoinf.2021.101235
Received 11 November 2020; Received in revised form 18 December 2020; Accepted 19 December 2020
Available online 23 January 2021

Morphometric data matrices used in biological taxonomy, or tables of functional traits in ecology may include variables that do not apply to all observations. For example, the value of "seed size" for ferns in a large plant trait database is not simply missing, but biologically meaningless and therefore cannot be imputed. Consequently, the term "*incomplete* data" refers more appropriately to the general situation when some scores are either unknown or undefined. The third group of methods that require attention here appear particularly useful to compute PCA in such situations. The simplest possibility is to calculate distances between observations using Gower's (1971a) function and perform Principal Coordinates Analysis (Brown et al., 2012; GowPCoA in Dray and Josse, 2015) or to calculate correlations between each pair of variables based only on known scores in PCA (abbreviated as PairCor-PCA in Dray and Josse, 2015). The latter authors commented that in the first case "no outputs concerning the variables are produced" while the second method "only provides scores for variables". This is unfortunate, because one of the most meaningful features of PCA, the simultaneous display of ordinations of variables and observations (the biplot) is lost. However, we show that starting from the PairCor approach to PCA, the calculations can be continued to derive scores of observations and then to draw a biplot diagram as well. The application of this, modified procedure abbreviated as InDaPCA (standing for PCA of INcomplete DAta) is illustrated using four actual data sets representing different research fields of biology. These examples demonstrate that in practical situations the absence of values in the data, due to whatever reason, does not necessarily pose serious problems as earlier thought.

2. Method

The algorithmic details of PCA need not be repeated here, the reader may consult the literature cited in Section 1 or references therein. We place emphasis on the minor, yet important novelty which allows performing a complete PCA study from incomplete data matrices. Before analysis, the data must be checked for extreme situations: variables for which only one value is available are to be excluded. Furthermore, care must be taken to avoid pairs of variables that are known for one observation or none. We assume that after unavoidable deletions (if any) we have a data matrix $\mathbf{X} \equiv \{x_{ij}\}$, with n variables as rows and m observations as columns, and dummy entries inserted for unknown or undefined scores. Array $\mathbf{W}$ of the same size is an indicator matrix with $w_{ij} = 1$ if x_{ij} is known, and $w_{ij} = 0$ otherwise.

- The first step is identical to the pairwise correlation (PairCor) procedure as discussed by Dray and Josse (2015), with the extension that centered PCA starting from covariances (PairCov) may also be performed via the same algorithmic steps. Essentially, either function is computed for each pair of variables based on observations for which the values of both variables are known. The correlation for variables i and h is given by

$$cor_{ih} = \frac{\sum w_{ij}w_{hj}x_{ij}x_{hj} - \frac{\sum w_{ij}w_{hj}x_{ij}\sum w_{ij}w_{hj}x_{hj}}{\sum w_{ij}w_{hj}}}{\sqrt{\left(\sum w_{ij}w_{hj}x_{ij}^2 - \frac{\left(\sum w_{ij}w_{hj}x_{ij}\right)^2}{\sum w_{ij}w_{hj}}\right)\left(\sum w_{ij}w_{hj}x_{hj}^2 - \frac{\left(\sum w_{ij}w_{hj}x_{hj}\right)^2}{\sum w_{ij}w_{hj}}\right)}} \tag{1a}$$

whereas the covariance is calculated as

$$cov_{ih} = \frac{\sum w_{ij}w_{hj}x_{ij}x_{hj} - \frac{\sum w_{ij}w_{hj}x_{ij}\sum w_{ij}w_{hj}x_{hj}}{\sum w_{ij}w_{hj}}}{\sum w_{ij}w_{hj} - 1} \tag{1b}$$

In Eqs. (1a) and (1b), summations are taken from $j = 1$ to m. The use of weights $w_{ij}w_{hj}$ shows that the different coefficients in the correlation (or covariance) matrix $\mathbf{C} \equiv \{c_{ih}\}$ are potentially computed based on different numbers of observations. Even if the sample size is identical for some pairs of variables, their correlation may be based on different subsets of observations.

- Matrix $\mathbf{C}$ is subjected to eigenanalysis to derive a square diagonal matrix Λ with eigenvalues λ_k and matrix $\mathbf{V} \equiv \{v_{ik}\}$ of associated eigenvectors, each normalized to unit length to obtain direction cosines between the components and the original variables in the multidimensional space. The relative magnitude of eigenvalues reflects the percentage of variance accounted for by the components. If $\mathbf{X}$ is incomplete, some of the resulting eigenvalues may be negative, their values actually depending on how much the "estimated" $\mathbf{C}$ deviates statistically from a theoretical $\mathbf{C}$ which could have been calculated if all values were known (in case of missing data). This is measured by the ratio of the sum of negative and positive eigenvalues, multiplied by 100 to obtain a percentage (Podani and Miklós, 2002). For data in which scores are unknown due to logical reasons, $\mathbf{C}$ is considered the only representation of the correlation (covariance) structure of variables. Negative eigenvalues are not detrimental as far as the interpretation of the ordination results is concerned, unless their magnitude is close to the largest positive eigenvalues (see e.g., Digby and Kempton, 1987, Legendre and Legendre, 1998, Podani and Miklós, 2002, for a more detailed discussion of how to handle this problem).

  Correlations between variables and components, $\mathbf{R} \equiv \{r_{ik}\}$ may be derived from the standard deviations of variables, the eigenvalues and the eigenvectors using the following formula

$$r_{ik} = \lambda_k v_{ik} \Big/ s_i\sqrt{\lambda_k} = v_{ik}\sqrt{\lambda_k} \Big/ s_i \tag{2}$$

  In these calculations, we use the standard deviations of variables computed from all available data,

$$s_i = \sqrt{\frac{\sum w_{ij}x_{ij}^2 - \frac{\left(\sum w_{ij}x_{ij}\right)^2}{\sum w_{ij}}}{\sum w_{ij} - 1}} \tag{3}$$

  with summations from $j = 1, \ldots, m$, as above. The correlations may be used as coordinates to display the ordination of variables along components.
- It was implicitly assumed that $\mathbf{C}$ is a meaningful representation of the relationships between variables and, consequently, that $\mathbf{V}$ is an equally useful summary of the relationships between components and original variables, namely, that v_{ik} is a good approximation to the "true" direction cosine between variable i and component k, for all $\lambda_k > 0$. Therefore, it is equally acceptable to calculate the scores of observations from the eigenvectors and the original data using the following formula

$$u_{jk} = \sum_{i=1}^{n} w_{ij}x_{ij}v_{ik} \tag{4}$$

  to yield matrix $\mathbf{U} \equiv \{u_{jk}\}$. The scores in this matrix may be used as coordinates to display the ordination of observations along principal components.
- In PCA, there are several possibilities available to superimpose the ordination of variables over the ordination of observations, thus facilitating joint interpretation of the two ordinations in form of a biplot diagram. The values in $\mathbf{R}$ and $\mathbf{U}$ are scaled differently, so the variable scores are to be multiplied by a scaling factor to make the two ordinations visually comparable. As a further graphical aid, arrows are directed from the origin towards the positions of variables. The simultaneous display of the two ordinations obtained this way is called the Rohlf-biplot (Podani, 2000). In this paper, we shall not be concerned with two other possibilities, the Euclidean-biplot in which the direction cosines v_{ik} are used as variable coordinates, and the

Table 1
Actual data sets, properties and summary of eigenanalysis in standardized (correlation-based) InDaPCA.

Data set	Basic properties					Results	
	Size (n variables × m observations)	Number and % of unknown scores	Number of unknown scores per variable (min – max)	Number of unknown scores per observation (min – max)	Number of objects for which correlations were computed (min – max)	Number and size of negative eigenvalues, and deviation percentages	Highest positive eigenvalues and associated percentages
GLOPNET	12 traits for 2006 species	14,328 (59.5%)	247 to 1865	0 to 11	90 to 1628	3 −0.09 to −0.08 D ≈ 1%	$\lambda_1 = 4.93$ (40.7%) $\lambda_2 = 2.74$ (22.6%) $\lambda_3 = 1.91$ (15.9%)
Functional data of European invertebrates	63 traits for 596 species	388 (1%)	1 to 30	0 to 39	553 to 594	0	$\lambda_1 = 7.23$ (11.5%) $\lambda_2 = 4.64$ (7.4%) $\lambda_3 = 4.22$ (6.7%)
Crocodile cranial morphometry – four data sets with increasing number of missing values	23 characters by 226 individuals	0	0	0	226	0	$\lambda_1 = 20.9$ (90.7%) $\lambda_2 = 1.08$ (4.7%) $\lambda_3 = 0.32$ (1.4%)
		68 (1.3%)	0 to 7	0 to 2	215 to 224	1 −0.0001 D ≈ 0%	$\lambda_1 = 20.9$ (90.7%) $\lambda_2 = 1.08$ (4.7%) $\lambda_3 = 0.33$ (1.4%)
		1040 (20%)	36 to 58	0 to 10	121 to 163	6 −0.06 to −0.002 D ≈ 0.5%	$\lambda_1 = 20.8$ (90.2%) $\lambda_2 = 1.08$ (4.7%) $\lambda_3 = 0.34$ (1.4%)
		2651 (51%)	103 to 138	3 to 18	35 to 74	8 −0.24 to −0.006 D ≈ 2.7%	$\lambda_1 = 20.8$ (87.9%) $\lambda_2 = 1.34$ (5.7%) $\lambda_3 = 0.45$ (1.9%)
Hybrid fish	27 characters and 5 groups	4 (3%)	0 or 1	0 or 1	4 or 5	12 −2.7 to −0.0000008 D ≈ 11%	$\lambda_1 = 19.5$ (64.2%) $\lambda_2 = 7.82$ (25.7%) $\lambda_3 = 1.97$ (6.5%)

Mahalanobis-biplot in which the object scores are standardized to equalize the variances of components.

2.1. Software

The SYN-TAX 2000 package (Podani, 2001) has been modified to perform standardized (correlation-based) and centered (covariance-based) PCA in the manner described above. The unknown scores must be coded by a value of −1, or any other negative score, that is, the original data cannot contain negative values – which rarely appear in biological data anyway. If they do appear (e.g. temperatures below zero on the Celsius scale) then all temperature measurements should be increased by a constant to make all scores positive. This operation does not influence correlations. If the program detects an incomplete data matrix by finding negative scores, then InDaPCA is performed automatically. Supplementary Appendix 1 contains an R script performing PCA with incomplete data, where unknown entries are coded as NA.

3. Material

Four different data sets with very different levels of incompleteness were selected for the present study. Their main features are presented in Table 1.

3.1. Leaf functional traits

The GLOPNET (Global Plant Trait Network) database was constructed as a summary of 12 leaf functional variables including photosynthetic capacity, dark respiration rate, stomatal conductance, N and P concentrations, leaf life span and leaf mass per area of many species, based on the work of several independent authors (Wright et al., 2004). In its original form, many species were included twice or more times with different numbers of missing scores, because the measurements came from different studies. Since the use of this matrix would overemphasize the replicate species in calculating correlations, the data of each species were combined into a single row, resulting in a total of 2008 species, thus decreasing the proportion of missing scores. This proportion is still very high: 59.5% of scores are unknown, so that there are traits for which only 141 scores were available. Some correlations are based only on pairs of trait values for 90 species, whereas others were calculated from 1628 – thus providing a greatly unbalanced computational background for the correlation matrix. The data set used here is not identical to the one used by Dray and Josse (2015): they had six variables and 2494 species (probably not correcting for duplicates) – making the comparison of their results with ours less straightforward.

3.2. Invertebrate functional traits

The European freshwater invertebrate data set (Tachet et al., 2010) was accessed through the www.freshwaterecology.info website (Schmidt-Kloiber and Hering, 2015; Schmidt-Kloiber and Hering, 2019) and was complemented with data from Bonada and Dolédec (2011). The final set consisted of 63 functional traits belonging to 11 trait groups (Supplementary Appendix 2) and 596 taxa. Compared to leaf functional traits, which are based on direct measurements, the traits of freshwater invertebrates were *scored by experts* (see Schmera et al., 2014). This means that experts quantified the relative importance of a particular trait for a species, or in other words, the relative affinity of a species to a particular trait (e.g., is the shredder feeding mode characteristic to a particular species?). Each value in the matrix expresses the affinity of a species to the given trait (e.g., shredder), measured on a linear scale from 0 to 5. The taxonomic resolution of the data was varying, and it included adult and juvenile stages of the same

species. The Mediterranean data set (Bonada and Dolédec, 2011) had only four traits in the trait group of 'Respiration' missing the 'hydrostatic vesicle' trait and it lacked the 'absorber' trait from the 'Feeding habit' trait group. In both cases, 0-s were added to these traits assuming that taxon affinities for these traits were zero (see Schmera et al., 2014). The trait 'Detritus < 1 mm' from the Mediterranean set was matched to the 'Fine detritus (≤ 1 mm)' trait in the Tachet data and the 'Plant detritus ≥ 1mm' was matched 'Dead plant (> 1mm)'. This data matrix is much closer to completeness than the previous one: only 1% of the scores are missing.

3.3. Cranial morphometry

A crocodilian cranial data set was taken from Brown et al. (2012) for 226 skulls representing 21 taxa (species and subspecies) described by 23 osteological variables. From the complete data matrix, we created three incomplete arrays by randomly removing 68 (1.3%), 1040 (20%) and 2561 (51%) values, thus creating a series of data matrices with increasing loss of information. These ordinations are compared with the PCA result of the complete data via Procrustes analysis (Gower, 1971b) using the SYN-TAX 5.0 program package (Podani, 1993). The Procrustes statistic for a given pair of ordinations yields 0 for complete identity and 2 for maximally different configurations, regardless of the number of points. Interpretation of ordination diagrams is facilitated by superimposing convex hulls drawn around points representing the same genus (A–*Alligator*, I–*Caiman*, P–*Paleosuchus*, C–*Crocodylus*, O–*Osteolaemus*, G–*Gavialis*, T–*Tomistoma*). We selected this taxonomic level for demonstration, because for many species only 1–3 individuals were examined and measured.

3.4. Fish morphometry

The fourth data set comes from a study of hybridization between Russian sturgeon (*Acipenser gueldenstaedtii*) and American paddlefish (*Polyodon spathula*) (Káldy et al., 2020). In addition to the parent species, three different groups of hybrids were distinguished in which genome size was large, small, or unknown. Before performing regular PCA for 218 specimens, Káldy et al. (2020) had to remove four variables (scute length and the number of dorsal, lateral and ventral scutes) from the data set because the scute was absent from the American paddlefish and some of the hybrid individuals. Thus, this matrix represents a situation in which the absence of certain values is due to biological reasons. Here, we use the mean values of the 27 characters for the five groups, with four values undefined, as given in Table 3 of Káldy et al. (2020). Thus, the number of missing scores is relatively small (3%) while the correlations are calculated for 4 or 5 observations only.

4. Results

4.1. Leaf functional traits

This example exhibits the highest percentage of missing scores in the present study, but the large number of observations allowed the calculation of pairwise trait correlations based on at least 90 species. As a result, the correlation matrix was a good "estimate" of the "true" correlations that could have been derived from a complete data set (deviation D ≈ 1%). This is clearly shown by InDaPCA which provided only three negative eigenvalues, all negligible in size in comparison with the largest positive eigenvalues (Table 1). Most of the variation (>63%) is accumulated in the subspace of the first two dimensions, thus we restrict interpretation of results to these two with good reason. The species do not form separate groups; they are arranged as a large unstructured cluster around the origin, with a density decreasing towards all directions (Fig. 1.). The juxtaposition of variable positions on the ordination plane provides the Rohlf biplot in which all arrows are long, suggesting balanced contribution of variables to the total variation. The first axis represents a contrast between a group of five variables (Amass, Gs, Nmass, Pmass, Rdmass, on the left) and LL and, to a lesser extent,

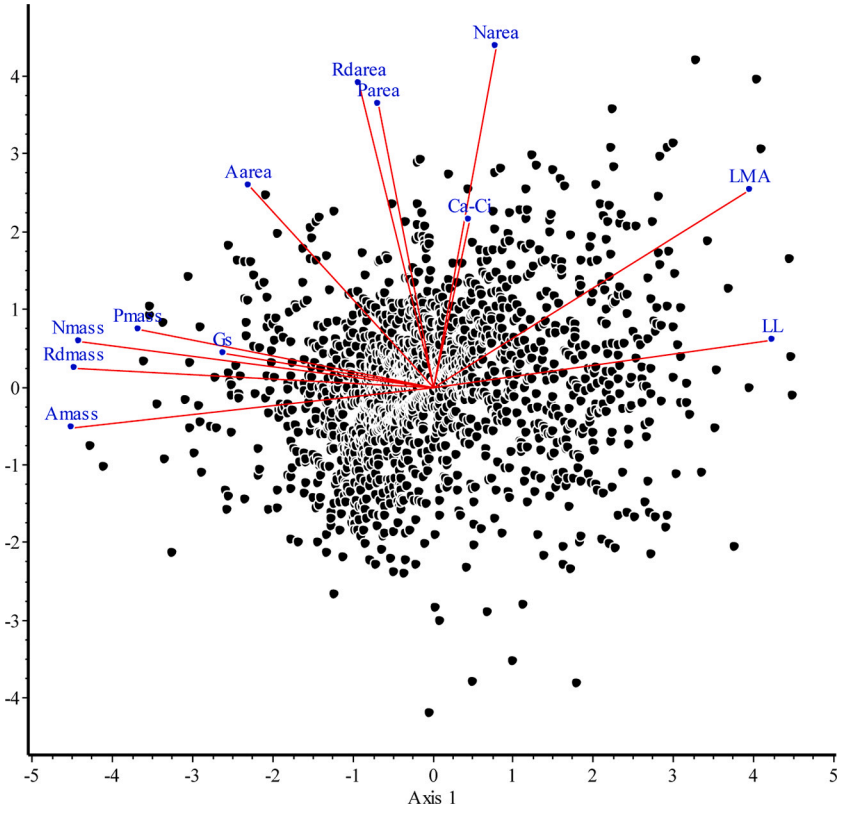

Fig. 1. InDaPCA of 2008 species described by 12 leaf functional traits, based on a data matrix with 59.5% of scores missing. Abbreviations: Aarea: photosynthetic capacity per unit leaf area; Amass: photosynthetic capacity per unit leaf mass; Ca-Ci: CO_2 drawdown in the leaf mesophyll; Gs: stomatal conductance; LL: leaf lifespan; LMA: leaf mass per area; Narea: nitrogen content per unit leaf area; Nmass: nitrogen content per unit leaf mass; Parea: phosphorous content per unit leaf area; Pmass: phosphorous content per unit leaf mass; Rdarea: dark respiration rate per unit leaf area; Rdmass: dark respiration rate per unit leaf mass.

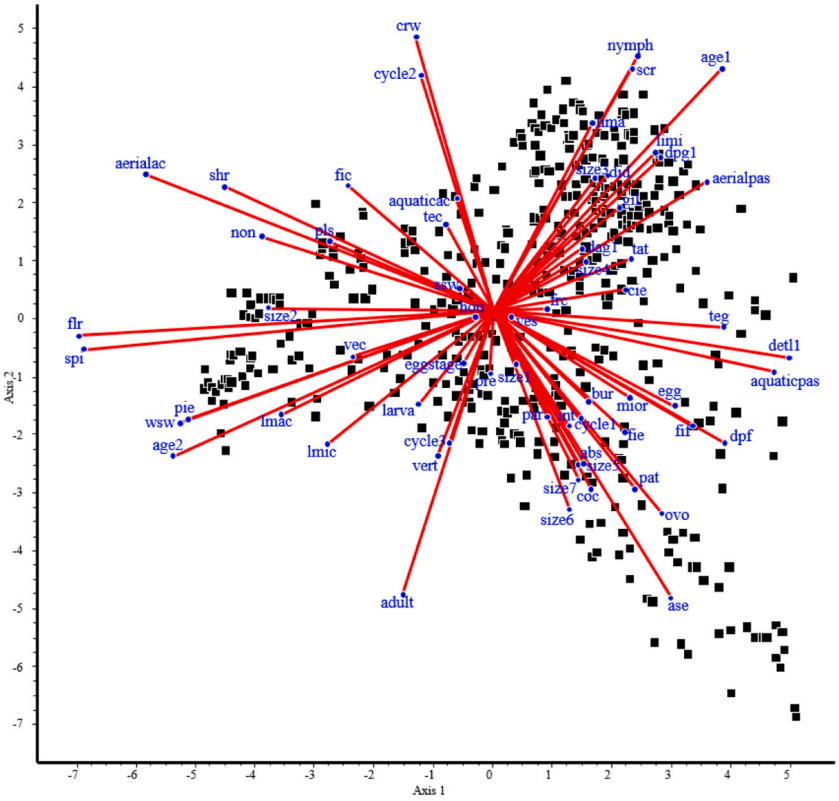

Fig. 2. InDaPCA of functional data of 596 European freshwater invertebrate taxa with 1% of scores missing. Abbreviations are given in the text for traits that have interpretative value; for the others see Supplementary Appendix 2.

LMA (right). This is in line with the general leaf economic spectrum (Wright et al., 2004): long leaf lifespan requires robust leaf structure resulting in high LMA, and these traits are associated with low mass-based N and P concentrations and hence moderate assimilation rates (Amass, Rdmass) since a considerable part of leaf material in long-lived leaves is devoted to compounds serving persistence at the expense of physiological performance (e.g. Lambers and Poorter, 1992; Wright et al., 2004). Although Dray and Josse (2015) used a very different subset of data, our results support their conclusions that LL plus LMA and the mass-based variables are on the opposite ends of axis one (four out of five methods, see their Fig. 2.) The second axis may be identified as an area-related dimension: species with large values of N and P concentrations (Narea, Parea), and photosynthetic capacity (Aarea) and dark respiration rate (Rdarea) per unit leaf area are on the positive side, those with smaller values on the opposite side. Note that Ca-Ci is also correlated with these variables, and – to a lesser extent – with LMA. These were not included in the Dray and Josse study. The correlation of Ca-Ci with area based variables (Narea, Parea, Aarea, Rdarea) along axis 2 reflects that – on a leaf area base – high N and P concentrations and consequently high photosynthetic rate usually result in substantial drawdown of CO_2 concentration (high Ca-Ci value) in the leaf mesophyll relative to that in the ambient air. In a metaanalysis of 60 studies covering 81 species, Niinemets et al. (2009) found positive relationship between LMA and CO_2 drawdown.

4.2. Invertebrate functional traits

The low percentage of missing scores (1%) caused no complications in calculating component scores: all eigenvalues obtained were nonnegative. This demonstrates that the correlation structure of the variables was very closely approximated: correlations were calculated based on at least 553 pairs of scores – apparently a sufficiently large number. Note that the first two axes account for a relatively low percentage of variance, less than 19% (Table 1) which is nevertheless expected for so many variables. For the purposes of the present paper, therefore, it is satisfactory to interpret only these two components. The biplot diagram of InDaPCA (Fig. 2) displays how the different traits are associated and defines specific niche dimensions that represent successful ecological strategies. The diagram also shows that different taxa possess various combinations of dominant strategies, although the taxa do not form separate clusters. The shape of the point cloud is more structured than in the previous example, we may observe three major trends among the taxa, which do not always coincide with components. Taxa with spiracle respiration (aerial) and flier locomotion (flr) appear at the negative end of axis 1, corresponding to one of the major trends. On the other end of the same axis, however, relatively few taxa appear, these are best characterized as detritus feeders (detl1) and by aquatic passive dispersal mode (aquaticpas) and respiration by tegument (teg). Component 2 has no such direct interpretation; instead, the two other trends are disclosed in the (+,+) and (+,–) quadrants of the PCA plane, respectively. The first one is represented by a large group of taxa that may be described by at least ten different traits, with scraper feeding habit (scr), with nymph aquatic stage (nymph), and short life cycle duration (age1) being the most conspicuous. The other group is more elongated and asexual reproduction (ase) and ovovivipary (ovo) are their best explanatory traits. A remarkably large group of taxa is arranged around the origin, in these most of the important traits are lacking.

4.3. Cranial morphometry

The series of InDaPCA biplots (Fig. 3, left column) largely demonstrate that gradual and random loss of entries from the raw data did not cause significant modifications in the correlation structure of variables and in the scatterplot of individuals. There is a very large first eigenvalue (around 90%) dominating all results, thus leaving only ca. 5% of

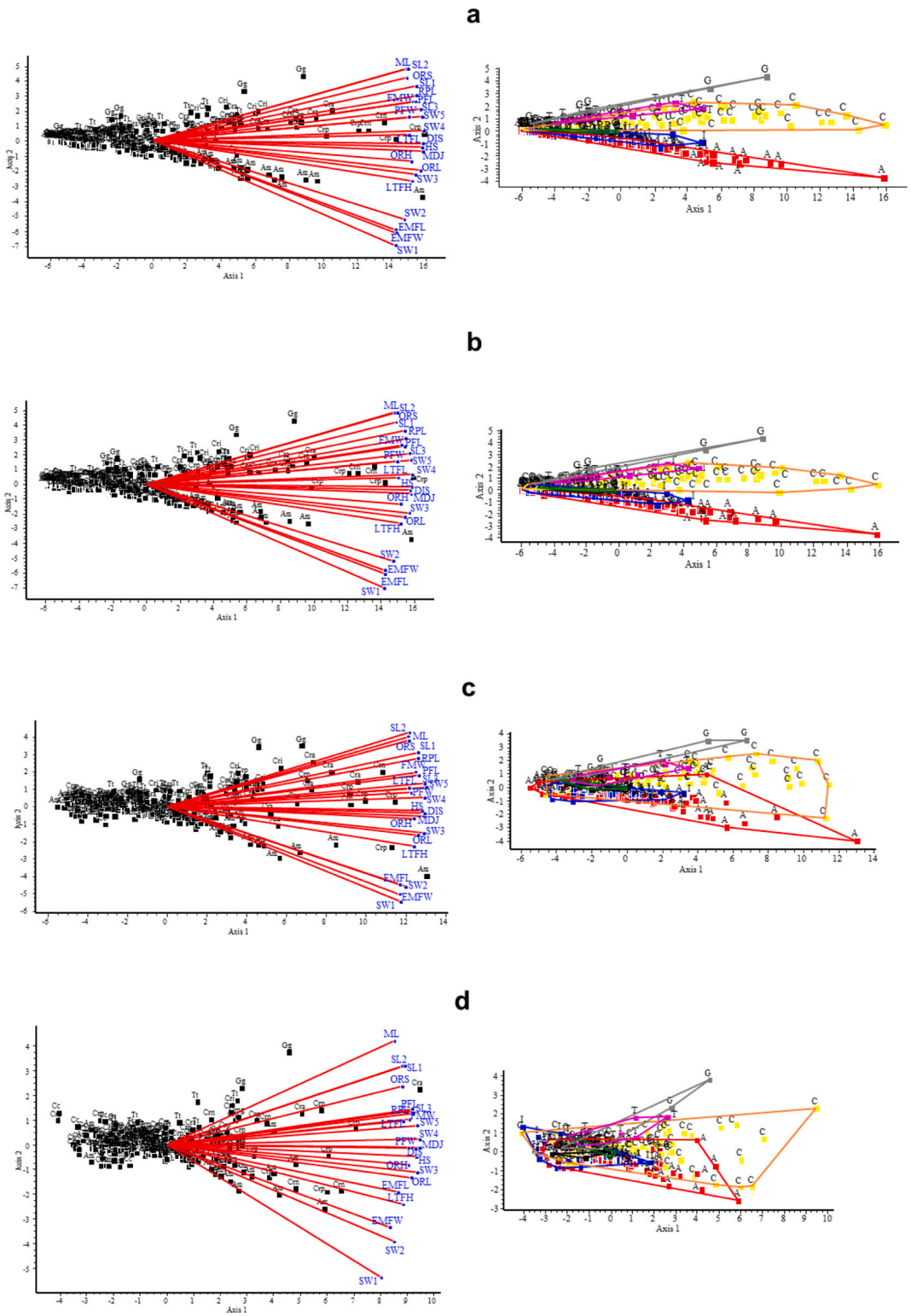

Fig. 3. InDaPCA of 226 skulls, representing 23 species and subspecies of crocodile, based on a data matrix with all scores (**a**), with 1.3% of scores randomly removed (**b**), with 20% of scores randomly removed (**c**), and with 51% of scores randomly removed (**d**). Left column: biplots; right column: ordination of skulls with convex hulls representing genera, all diagrams with axes 1–2. Variables and individuals labeled following Brown et al. (2012), for labels of genera, see text.

variation explained by the second, which is slightly larger than 1 in all cases (Table 1). The third and all subsequent eigenvalues are much lower than 1, i.e., the unit variance of standardized variables themselves, and are therefore disregarded. Along with the increase in the number of missing values, the largest negative eigenvalue also increased but remained relatively low even in the last member of the series (deviation values are in Table 1). Thus, the use of 226 individuals provided roughly the same result as 35–74 individuals in calculating correlations. This interpretation is confirmed by the Procrustes sum of squares statistics calculated for axes 1–2. The sum of squares difference between the original ordination and the ordinations of incomplete data at various levels of missing scores (1.3%, 20% and 51%) is 0.001, 0.024

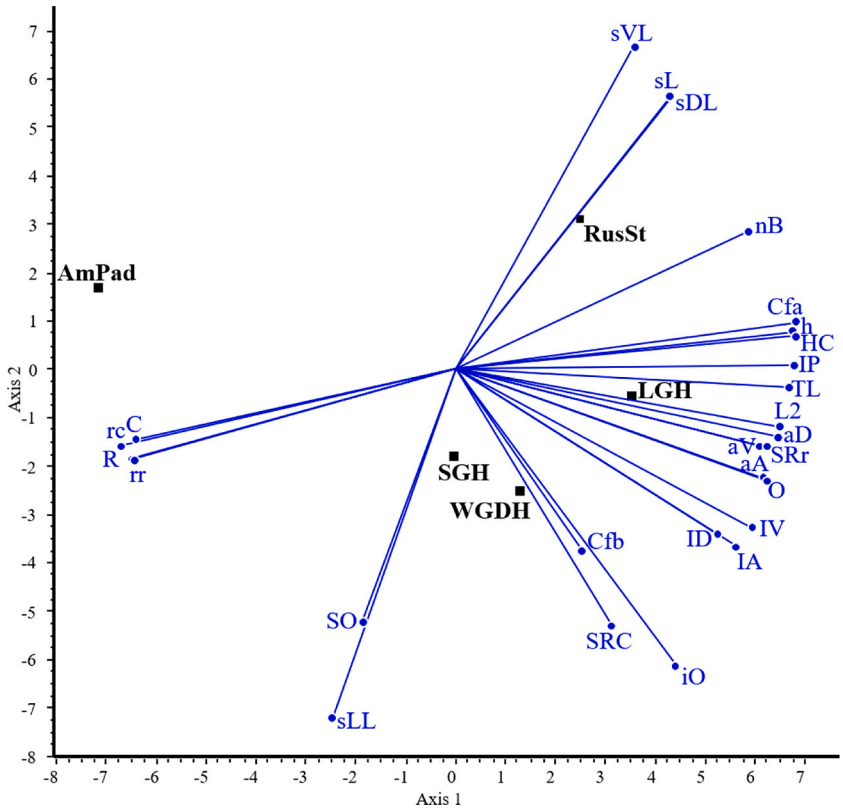

Fig. 4. InDaPCA of five groups of fish (AmPad-American paddlefish, RusSt-Russian sturgeon, LGH-hybrids with large genome size, SGH-hybrids with small genome size, WGDH-hybrids with no genome size data) based on 27 characters. Abbreviations for characters explained in this study: sL–scute length, sDL–number of dorsal scutes, sLL–number of lateral scutes, sVL–number of ventral scutes, SO–mouth width. For all other characters, see Table 3 in Káldy et al. (2020).

and 0.110, respectively, for the specimens. The corresponding Procrustes statistics for the ordination of variables are lower (0.0001, 0.011 and 0.091, respectively) demonstrating that relationships among variables were preserved slightly better than those among individuals.

All variables are positively correlated with axis 1, which may be interpreted as a size component, as in Dodson's (1975) study on alligators. Axis 2 has to do with the shape of the skull, with the gavial (*Gavialis gangeticus* – Gg) and the American alligator (*Alligator mississippiensis* – Am) representing the extremes. The contrast of osteological variables along this weak second axis appears between the group of two skull width characters (SW1, SW2) and the length and width of external mandibular fenestra (EMFL, EMFW) and the group composed of two skull length characters (SL1, SL2), maxilla length (ML) and orbit separation (OS). Changes in the arrangement of individuals along axes 1–2 are efficiently visualized by convex hulls drawn around points representing the same genera (Fig. 3, right column). Based on the full data set, the convex hulls are narrow-linear and long, demonstrating how skull shape of genera differentiates with age. These convex hulls become wider, shorter, and more overlapping when the percentage of deleted scores increases, clearly showing that incompleteness causes increased spatial dispersion of points.

4.4. Fish morphometry

The interplay between the number of observations (5), the number of variables (27) and the number of undefined scores (4) provided an unexpected set of eigenvalues. Normally, for five points one would get only four nonzero eigenvalues at most, because interpoint distances among five points can be faithfully represented in four dimensions or less.[1] The fact that 98 values in the correlation matrix were calculated using 4, rather than 5 objects, is responsible for the occurrence of two additional, slightly positive eigenvalues, for many, mostly negligible negative eigenvalues, and for the relatively high value of deviation (11%). The absolute value of the largest negative eigenvalue (−2.73) exceeds the third largest positive eigenvalue (1.97) suggesting that only the first two ordination axes are interpretable (cf. Legendre and Legendre, 1998, p. 437).

Although we used only group centroids in the ordination, the result of InDaPCA (Fig. 4) of the incomplete data set is largely comparable to the PCA ordination (Fig. 3 in Káldy et al., 2020) of the fish individuals. Along axis 1 separates strongly the American paddlefish from the others; the hybrids are much closer to the Russian sturgeon, especially the one with large genome size. The separation of American paddlefish is explained by four variables in the same way as in Fig. 3 of Káldy et al. (2020). However, inclusion of four more characters, absent from the American species, in the analysis greatly enhanced the biological interpretability of the InDaPCA biplot. Káldy et al. (2020) had to rely on the arithmetic comparison of means, but in this case these comparisons are visualized: scute length and the number of ventral and dorsal scutes are correlated positively with one another, and negatively with the number of lateral scutes and mouth width (SO, not shown by Káldy et al., 2020). The values of the latter two are higher in the hybrids than in the Russian sturgeon. These five variables appear to express contrast along axis 2 – an important feature of the data which was obviously

[1] A regular PCA of the same data set with the four undefined values "estimated" using the mean of known values of the other species supports this statement. Four positive eigenvalues resulted (68.9%, 23.1%, 7.7% and 0.3%).

undetected in the original study.

5. Discussion

Principal Component Analysis can be carried out in two different ways. Although Singular Value Decomposition of the input matrix is considered numerically stable, it is obviously inappropriate in case of incomplete data because matrix algebraic calculations cannot be performed. However, as shown in this paper, the second algorithm, operating through the eigenanalysis of a cross product (covariance or correlation) matrix **C**, may be modified to comply with incomplete data. The idea is to calculate **C** based on as much information as possible by considering for each pair of variables all observations whose scores are known (as in PairCor). This means that eigenanalysis starts from an "estimated" **C** and the success of PCA depends on how close this matrix is to a "theoretical" **C** which could only be obtained if the data matrix were complete. Then, the eigenvalues and eigenvectors can be used to derive ordination scores in the usual way, except that missing values are skipped during the calculations. It means that coordinates of observations and variables in biplots are not calculated on the same basis.

The analysis of four actual data sets by the modified PCA (InDaPCA) from correlation matrices demonstrates that it is not the proportion of unknown scores that matters. Estimating the correlation structure of variables may be less reliable even if the total number of unknown scores is low (only 4 (3%) in the fish example) or reasonably good even if the percentage of missing values is high (59% for the plant functional traits matrix and 20 or 50% for the skull morphometry data). What is more important is the number of observations that are available for calculating pairwise correlations. In the fish example, this number was reduced from five to four for one observation and many pairs of variables, leading to a correlation matrix with "distorted" internal structure and then to a relatively large negative eigenvalue. Five is not a sufficiently large number anyway for calculating correlations.[2] In the other examples, correlations were based on at least 90, 553, 121 and 35 observations, causing negligible or no distortion in the correlation structure. Negative eigenvalues did not appear at all in case of the functional traits of invertebrates, when 553 was the minimum, apparently a very large number. The arrangement of objects is more influenced by unknown scores, as the crocodilian data set illustrates the points become more scattered along with increases in the number of missing values. Nevertheless, Procrustes analysis detected only 6% statistical departure from the reference ordination at most.

Appearance of negative eigenvalues does not necessarily influence the interpretability of results at least in the first two InDaPCA dimensions. For the fish data, the biplot for the first two components was in good agreement with the ordination of individuals, and revealed new information on the relationship of morphometric characters which was not possible in the original study. The series of the craniometric data with increasing percentage of missing scores demonstrated strong dominance of the first, size component which remained stable after data reduction. The correlation structure, and therefore the biplot arrangement of variables did not change considerably even when half of the data scores were lost. The weak second axis also retained its meaning, namely, a dimension associated with skull shape. The InDaPCA of leaf functional traits was also interpretable, the relationships among variables are portrayed in a meaningful manner along axes 1 and 2, and basically reflect the performance vs. persistence trade-off in leaf construction. The analysis of invertebrate data, in addition to revealing the correlation structure of functional traits was useful to indicate tendencies among species for forming different functional groups.

The objective of the present paper was to introduce a straightforward modification of the standard PCA algorithm to accommodate incomplete data sets and to demonstrate its usefulness in examining data from various fields of biology. Most of or conclusions regarding the role of negative eigenvalues and the numbers of observations and variables involved are therefore preliminary; their general relevance requires attention in a different simulation study. The advantage of the new PCA version over the simulation/estimation-based algorithms developed in the past decade is its simplicity, ease of programming and applicability to any situation no matter whether absence of scores from the data is due to chance or biological necessity.

[2] Note that for five observations (df = 3) the correlation must be at least 0.878 to be significant in a two-tailed test at $\alpha = 0.05$, whereas the critical value is 0.195 for df = 100 and 0.062 for df = 500.

Authors' contributions

JP developed the main idea described in the paper, wrote the body of the manuscript. BB, DS and TK prepared parts of the manuscript, interpreted the results and performed literature search. The R script has been written by DS. All authors read and approved the final manuscript.

Funding

This study was supported financially by the Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (NKFIH) K128496 grant.

Declaration of Competing Interest

The authors declare that they have no competing interest.

Acknowledgements

We are grateful to two anonymous referees for their useful comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2021.101235.

References

Bonada, N., Dolédec, S., 2011. Do mediterranean genera not included in Tachet et al. 2002 have mediterranean trait characteristics? Limnetica 30 (1), 129–142.
Brown, C.M., Arbour, J.H., Jackson, D.A., 2012. Testing of the effect of missing data estimation and distribution in morphometric multivariate data analyses. Syst. Biol. 61 (6), 941–954.
Digby, P.G.N., Kempton, R.A., 1987. Multivariate Analysis of Ecological Communities. Chapman and Hall, London, UK.
Dodson, P., 1975. Functional and ecological significance of relative growth in Alligator. J. Zool. 175, 315–355.
Dray, S., Josse, J., 2015. Principal component analysis with missing values: a comparative survey of methods. Plant Ecol. 216, 657–667.
Gower, J.C., 1971a. A general coefficient of similarity and some of its properties. Biometrics 27, 857–871.
Gower, J.C., 1971b. Statistical methods of comparing different multivariate analyses of the same data. In: Hodson, F.R., Kendall, D.G., Tautu, P. (Eds.), Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press, Edinburgh, pp. 138–149.
Grung, B., Manne, R., 1998. Missing values in principal component analysis. Chemom. Intell. Lab. Syst. 42, 125–139.
Jolliffe, I.T., 1986. Principal Component Analysis. Springer, New York.
Káldy, J., Mozsár, A., Fazekas, G., Farkas, M., Fazekas, D.L., Fazekas, G.L., Goda, K., Gyöngy, Z., Kovács, B., Semmens, K., Bercsényi, M., Molnár, M., Patakiné Várkonyi, E., 2020. Hybridization of Russian sturgeon (*Acipenser gueldenstaedtii*, Brandt and Ratzeberg, 1833) and American paddlefish (*Polyodon spathula*, Walbaum 1792) and evaluation of their progeny. Genes 11, 753.
Lambers, H., Poorter, H., 1992. Inherent variation in growth rate between higher plants: a search for physiological causes and ecological consequences. Adv. Ecol. Res. 23 (C), 187–261.
Legendre, L., Legendre, P., 1998. Numerical Ecology, 2nd ed. Elsevier, Amsterdam.
Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic, London.
Nelson, P.R.C., Taylor, P.A., MacGregor, J.F., 1996. *Missing data* methods in PCA and PLS: score calculations with incomplete observations. Chemom. Intell. Lab. Syst. 35, 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X.
Niinemets, Ü., Díaz-Espejo, A., Flexas, J., Galmés, J., Warren, C.R., 2009. Role of mesophyll diffusion conductance in constraining potential photosynthetic

productivity in the field. J. Exp. Bot. 60 (8), 2249–2270. https://doi.org/10.1093/jxb/erp036.
Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Sin, I., 2005. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19, 2088–2096.
Orlóci, L., 1978. Multivariate Analysis in Vegetation Research, 2nd ed. Junk, The Hague.
Podani, J., 1993. SYN-TAX Version 5.0. User's Guide. Scientia, Budapest.
Podani, J., 2000. Introduction to the Exploration of Multivariate Biological Data. Backhuys, Leiden.
Podani, J., 2001. SYN-TAX 2000. Computer Programs for Data Analysis in Ecology and Systematics. Scientia, Budapest. Available at: http://podani.web.elte.hu/SYN2000.html.
Podani, J., Miklós, I., 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology 83, 3331–3343.
Schmera, D., Podani, J., Erős, T., Heino, J., 2014. Combining taxon-by-trait and taxon-by-site matrices for analysing trait patterns of macroinvertebrate communities: a rejoinder to Monaghan & Soares (2014). Freshw. Biol. 59, 1551–1557.
Schmidt-Kloiber, A., Hering, D., 2015. www.freshwaterecology.info - an online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. Ecol. Indic. 53, 271–282. https://doi.org/10.1016/j.ecolind.2015.02.
Schmidt-Kloiber, A., Hering, D. (Eds.), 2019. www.freshwaterecology.info - The Taxa and Autecology Database for Freshwater Organisms, Version 7.0 (accessed on 12.09.2019).
Serneels, S., Verdonck, T., 2008. Principal component analysis for data containing outliers and missing elements. Comput. Stat. Data Anal. 52, 1712–1727.
Stanimirova, I., Daszykowski, M., Walczak, B., 2007. Dealing with missing values and outliers in principal component analysis. Talanta 72, 172–178.
Tachet, H., Bournaud, M., Richoux, P., Usseglio-Polatera, P., 2010. Invertébrés d'eau Douce - Systématique, Biologie, Écologie. CNRS Editions, Paris (Accessed through www.freshwaterecology.info - the taxa and autecology database for freshwater organisms, version 7.0).
Wright, I.J., Reich, P.B., Westoby, M., Ackerly, D.D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J.H.C., Diemer, M., Flexas, J., Garnier, E., Groom, P.K., Gulias, J., Hikosaka, K., Lamont, B.B., Lee, T., Lee, W., Lusk, C., Midgley, J.J., Navas, M.-L., Niinemets, Ü., Oleksyn, J., Osada, N., Poorter, H., Poot, P., Prior, L., Pyankov, V.I., Roumet, C., Thomas, S.C., Tjoelker, M.G., Veneklaas, E.J., Villar, R., 2004. The worldwide leaf economics spectrum. Nature 428, 821–827.