Computers and Geosciences 123 (2019) 161–171

Contents lists available at ScienceDirect

Computers and Geosciences

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

Enhanced Principal Tensor Analysis as a tool for 3-way geological data reconstructions

S. Kotov[*], H. Pälike

MARUM, Bremen Uni., Leobener Str. 8, 28359, Bremen, Germany

ARTICLE INFO

Keywords:
Tensor decomposition
Singular spectrum analysis
Sediments
Climate

ABSTRACT

Principal tensor analysis (PTA) is considered as a potentially useful tool in geosciences, particularly for reconstructions of multi-way (time (depth), space, and proxies) datasets. We introduce an extension of PTA: PTA enhanced with Singular Spectrum Analysis (SSA). Using this enhancement, we were able to isolate clear patterns in noisy multi-way data. As an illustrative example, the method has been applied to 4-way data tensor (time, space, proxies, and delay-time) constructed from marine sediment proxies. Possible restrictions and main reasons for the method's limitations are discussed. The algorithm has been implemented as an R function which is provided in the supplementary data and appendix.

1. Introduction

Statistical methods for dealing with multidimensional data such as Principal Component Analysis (PCA) and factor analysis are now standard tools in geosciences. Attempts to generalise the approach from classical two-way data matrix analysis to arbitrary number of dimensions have been made since the middle of previous century (Hitchcock (1928); Cattell (1944); Tucker (1964)). The most common term of this approach is known now as Principal Tensor Analysis (PTA). Principal tensor analysis is becoming a standard technique in many scientific disciplines, such as psychometrics, linguistics, chemometrics, and others (e.g., Carroll and Chang (1970); Harshman (1970); Smilde et al. (2004)). A detailed historical review and introduction to the methods can be found in e.g. Cichocki et al. (2015), Kroonenberg (2008), Kolda and Bader (2009). At the same time, geological sciences are usually restricted to classical two-way multi-dimensional methods. We believe that the main reason of this restriction is that raw geological data are always space distributed. It is usually a hard and time-consuming task to construct correct and accurate general (normally time) scales to build a reliable n-way data hypercube for multi-way data analysis. Some details and explanations will be provided in the “Data” section of this article. Additionally, standard statistical packages such as Statistica and SPSS Statistics do not offer needed functionality, which generally restricts the popularity of the method, not only in geosciences.

In the following method section, we derive the simplest form of tensor decomposition as a starting point, followed by PTA enhancement by Singular Spectrum Analysis (Kotov and Pälike (2017)). The data section describes measured physical properties from a set of deep ocean sediment cores, which were utilised in a decomposition example. Finally, we discuss the results from both methods in the results section and show that the enhanced PTA yields more appropriate results.

2. Method

2.1. 3-Way tensor decomposition

As a starting point, we used a simple form of Tensor Decomposition known from different sources as Polyadic decomposition, PARAFAC, CANDECOMP, or CP decomposition (Cichocki et al. (2015); Kolda and Bader (2009)). The general idea of the method is to represent an observed tensor χ as a sum of R orthogonal rank one tensors (an N-way tensor is rank one if it can be written as the outer product of N vectors):

$$\chi \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{1}$$

the symbol ‘∘’ represents the vector outer product.

Fig. 1 shows a schematic tensor decomposition of a three-way array.

We assume that the vectors **a**, **b**, **c** are normalised to length one with the weights in the vector of singular values λ:

$$\chi \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{2}$$

It is interesting to note that for a two-way data matrix, the task is

[*] Corresponding author.
E-mail addresses: skotov@marum.de (S. Kotov), hpaelike@marum.de (H. Pälike).

https://doi.org/10.1016/j.cageo.2018.11.001
Received 19 January 2018; Received in revised form 12 October 2018; Accepted 1 November 2018
Available online 09 November 2018

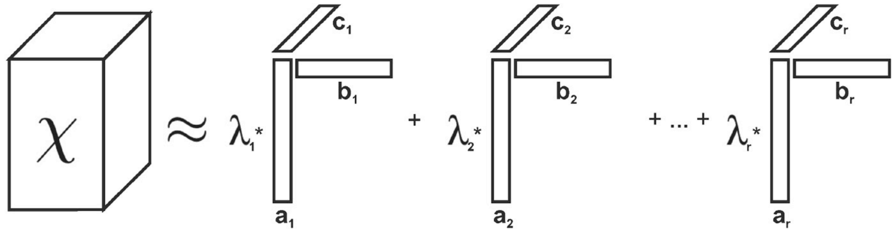

Fig. 1. Schematic tensor decomposition of a three-way array: **r** is the minimal number of needed tensors to represent the data tensor (modified from Kolda and Bader (2009)).

reduced to the classical Principal Component Analysis, where $\mathbf{a}_r \circ \mathbf{b}_r$ is the outer product of the scores and loadings of the r-th Principal Component correspondingly.

Computational nuances are beyond the scope of this article. In our research, we used the R package PTAk (Principal Tensor Analysis on k-modes) after Leibovici (2010). The PTAk-modes method is a way to generalise Singular Value Decomposition from matrices to arrays of arbitrary dimension. The numerical method is based on the algorithm of "reciprocal averaging" (Leibovici (2010); Hill (1973)).

It will be demonstrated in section 4 that it can be difficult to interpret raw results of PTA applied to real geological data. The two main reasons are the complexity of geological systems and noise components that are difficult to separate from true signal. We introduce here an algorithm to improve the outcome by merging PTA with Singular Spectrum Analysis (SSA) elements. General principles of Singular Spectrum Analysis will be briefly described in the next section.

2.2. SSA enhanced PTA

Singular Spectrum Analysis (SSA) is a modern method of time series analysis aiming at decomposing one-dimensional signals into orthogonal time components using empirical orthogonal functions (Ghil et al. (2002); Allen and Smith (1996); Danilov and Zhigljavsky (1997); Golyandina and Osipov (2007)). "SSA is designed to extract information from short and noisy time series and thus provide insight into the unknown or only partially known dynamics of the underlying system that generated the series", Ghil et al. (2002).

The main task of SSA is calculating the principal directions in the reconstructed phase space. Let us highlight the main steps of the classical SSA while they will be incorporated into our algorithm (Ghil et al. (2002)):

a. Construction of the so-called trajectory matrix (a sequence of M-dimensional vectors $\{X(t)\}$) from one-dimensional sequence of observations $X(t)$ using delay-time procedure:

$$\mathbf{X}(t) = (X(t), X(t+1), ..., X(t+M-1)) \tag{3}$$

the vectors $X(t)$ are indexed by $t = 1, ..., N'$, where $N' = N - M + 1$, N – length of the time series $X(t)$, M – length of delay time window.

b. Estimation of the matrix of eigenvectors $\mathbf{E}_x$ and eigenvalues $\mathbf{\Lambda}_x$ of the covariance matrix $\mathbf{C}_x = 1/N(\mathbf{X}^T\mathbf{X})$ (PCA approach):

$$\mathbf{E}_X^t \mathbf{C}_X \mathbf{E}_X = \mathbf{\Lambda}_X, \tag{4}$$

or singular values decomposition of the trajectory matrix $\mathbf{X}(t)$ (SVD approach):

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{5}$$

where $\mathbf{U}$ is an $N' \times M$ matrix of left singular vectors, $\mathbf{S}$ is an $M \times M$ diagonal matrix with singular values on the diagonal, and $\mathbf{V}^T$ is also an $M \times M$ matrix of right singular vectors (e.g., Wall et al. (2003)). We use the second approach because it allows us to merge SSA and PTA very easily. Following Ghil et al. (2002), we will call the rows of $\mathbf{V}$ empirical orthogonal functions (EOFs).

c. Projecting the time series onto EOF (PCA approach) to get principal components (PCs) is not needed in the SVD approach. Instead, we use columns of the matrix $\mathbf{U}$ as raw PCs (A_k).

d. Reconstruction of a time series based on a set of selected PCs (A_k) and EOFs (ρ_k):

$$R_{\mathscr{K}}(t) = \frac{1}{M_t} \sum_{k \in \mathscr{K}} \sum_{j=L_t}^{U_t} A_k(t-j+1)\rho_k(j) \tag{6}$$

where

$$(M_t, L_t, U_t) = \begin{cases} (1/t, 1, t), & 1 \le t \le M-1, \\ (1/M, 1, M), & M \le t \le N', \\ (1/(N-t+1), t-N+M, M), & N'+1 \le t \le N. \end{cases} \tag{7}$$

– normalisation factor for the windowed bounds of the time series, see notations for equation (3).

Reconstruction is also needed to capture the phase of the time series, which is uncertain in raw PCs (Ghil et al. (2002)). Having in mind that tensor decomposition is simply a generalisation of Singular Value Decomposition from matrix to an array of arbitrary dimension, we now have merged ideas of PTA and SSA:

1. Construction of 4-way tensor from our 3-way tensor with the same procedure of time delay as in SSA, but using 3-way tensor instead of the time series and 1-st dimension (vertical) for delaying,[1] see Fig. 1. In tensor notation, the main task can be reformulated as

$$\chi \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \circ \mathbf{d}_r \tag{8}$$

where $\mathbf{d}_r$ stands for delay coordinates or EOFs.

2. Second, we apply PTAk method to constructed 4-dimensional data array to estimate R singular values and raw vectors **a**, **b**, **c** and **d** followed by reconstruction of obtained components **a'** (tensor scores) using SSA technique to get reconstructed components RCs, equation (6) and (7). The algorithm is implemented as an R function. An example case for the function is given in Appendices A and B and in the supplemental files.

3. Data

The International Ocean Discovery Program (IODP) is an international marine research collaboration that explores Earth's history and dynamics using data recorded in seafloor sediments and rocks (https://www.iodp.org). IODP databases contain gigabytes of information, particularly on climate history of the Earth, written in different physical and chemical properties (proxies) of ocean sediments, contained (what

[1] Theoretically, it could be possible to extend our tensor to 5 and even more dimensions, e.g. delaying the space dimension (unordered now), thus incorporating some elements of kriging. Unfortunately, we are restricted in computational resources, see section 4. A trade off between the size of the data tensor and calculation time can be very important.

is important for us) in space (sediment core position) and depth distributed data sets.[2] Unfortunately, it is not easy to use data directly from measurements as it could be in other disciplines, like documentation of encephalograms or other direct measurement. Being originally time archives, geological data are non-linearly transformed into space depending on the local geological history with different sedimentation rates, tectonic processes, events etc. To decipher geological history, or in mathematical terms, to build an inverse mapping function from space to time, is a principal task in geosciences.

Let us briefly highlight some details which are specific to IODP cores, IODP-MI (2011), but can be very illustrative to understand geological scale problems in general, especially in chronostratigraphy. Most of the measurements, especially made onboard, are collected in the so-called meters below sea floor scale (mbsf-scale). Already here, one can recognise a source of uncertainty – seafloor, which is not possible to define precisely due to a soft layer of unconsolidated sediments on the top. Every borehole consist of several cores with unavoidable disturbances and sediment lost between cores. Later, every core is cut in sub-sections for sampling and measurement. Effects such as sediment contraction or expansion due to degassing are very common for deep sea cores (one more scale). Then, a composite depth scale is constructed based on available data from several adjacent boreholes from one site (to avoid possible gaps in data). In particular, the most difficult task is to put data from different sites onto a common scale, normally time. This task requires accurate mapping (or time scale) from space to time taking in account possible hiatuses, changing sedimentation rates and other artefacts. In the case of IODP paleo-climate reconstruction problems on the scale of thousands of years, this mapping normally involves the construction of a rough model based on some stratigraphical markers followed by more precise tuning of data to some target (for example, to some astronomically driven climatic solution according to the Milankovitch theory[3]).

Raw proxy-data for the site 1385 are shown in Fig. 2 as functions of meters below sea floor (mbsf), where δ^{18} stands for the delta 18 of oxygen in benthic foraminiferas, MS – magnetic susceptibility, GRA – density, NGR – natural gamma ray, RSC – reflectance, and SRM – remanent magnetisation. An example data tensor was constructed using measurements of these six proxies from six sites (925, 927, 929, 1143, 1146, 1385) as illustrated in Fig. 3.

Raw data were mapped from the depth below sea floor (mbsf-scale) to meters composite depth (mcd-scale), de-trended (de-trending is a standard procedure in time series analysis aimed to refine high frequency components in the observed signal and to avoid low frequency poorly resolved pseudo components due to linear trends in data), cleaned from outliers, and normalised. Transformation from depth to time (age) was made using piecewise linear mapping based on time scales from Lisiecki and Raymo (2005) and Hodell et al. (2015). To construct the final data tensor, data were interpolated with an even step of 2 kilo-years from 8 to 1000 kilo-years before present. The choice of the data step is dictated by the size, sampling resolution of raw data, and power of the computer. The time step should not be too large to avoid loss of high frequencies, but not too small to keep the reasonable computing time. See the computing time illustration in Fig. 9. The resulting tensor is represented as a 3-way (Age, Site, Proxy) R array with dimensions 497 × 6 × 6.

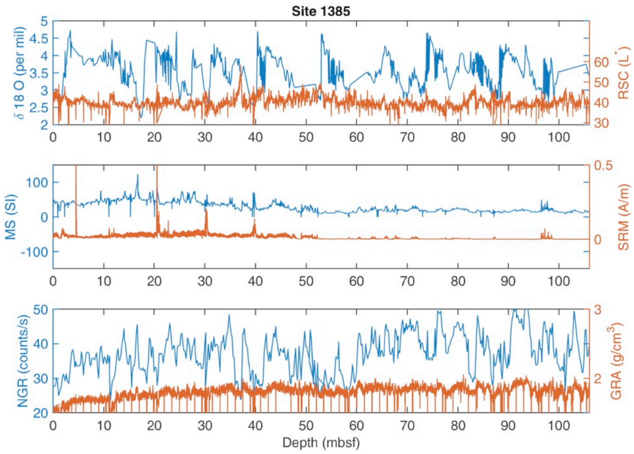

Fig. 2. Raw proxy-data versus depth, site 1385.

We now use the methods described in Section 2 on the real-world data to test the effectiveness of the SSA enhanced PTA method.

4. Results

4.1. 3-Way PTAk

Fig. 4 shows results of tensor decomposition (PTA3) on three modes applied to the original data set. They can be interpreted in a very similar way as in classical Principal Component Analysis, except that here we have two sets of loadings (sites are in red colour, proxies in green) and vectors of loadings are normalised to the length one. Fig. 5 shows the distribution of scores in time. We discuss interpretations of principal tensors later. Let us just note that we have split the original data tensor on 5 orthogonal sub-tensors explaining about 45% of the global variability – some of them demonstrate quasi-periodical behaviour (PT 1), some contain a non-linear trend (PT 2). The rest (PT 3–5) looks like noise and is difficult to interpret. One possible solution for dealing with large amounts of noise could be additional data processing, such as filtering, de-noising, etc. Nowadays, this technique is standard in time series processing, which is well described in scientific literature, and beyond the scope of this paper. We introduce another method – Enhanced Principal Tensor Analysis based on the idea of merging of PTA with Singular Spectrum Analysis, see section 2.

4.2. 4-Way PTAk (enhanced tensor decomposition)

Fig. 6 shows results of tensor decomposition applied to 4-way data tensor, constructed from the original 3-way data tensor by the function PTA_SSA (). Additionally to sites and proxies (red and green colours), there are Empirical Orthogonal Functions (EOFs) printed in black. It is possible to see how different EOFs catch different frequencies, see Figs. 6 and 7: the higher the number of Principal Tensor (PTs are ordered in decreasing singular values, or explained variability of data) the higher frequency it covers. Such a power frequency distribution is very typical for many paleo-climatic signals, especially for those driven by astronomical parameters of the solar system, Pälike (2005). The exception is PT 2 – the non-linear trend in time. Reconstructed components and corresponding periodograms are shown in Fig. 8.

To simplify the interpretations of tensors based on the values of loadings from Figs. 6 and 7, we use the "digest" representation: structures of tensors are represented as fractions with leading (with absolute values >0.3 here) positive loadings in the numerator and negative in the denominator, see equation (9)–(13). The percentages of explained variability is shown in round brackets before the fractions. Different

[2] The reason why we use proxies from paleo-climatic sciences is that these can serve as an appropriate and illustrative example: 1) it is straightforward to create a multi-way array (tensor) from different proxies and sites distributed in depth and time, 2) the data contains multiple periodical components, which is essential for SSA, and 3) the records contain natural noise that is typical for many geological situations.

[3] Milankovitch theory describes the collective effects of changes in the Earth's movements on its climate over thousands of years. According to it, the main driving climate parameters are periodic changes in the eccentricity of the orbit (≈ 100, 400 kyr periodicity), precession (≈ 19, 22, 24 kyr) and tilt of the Earth's rotation axis (≈ 41 kyr), see e.g. Pälike (2005).

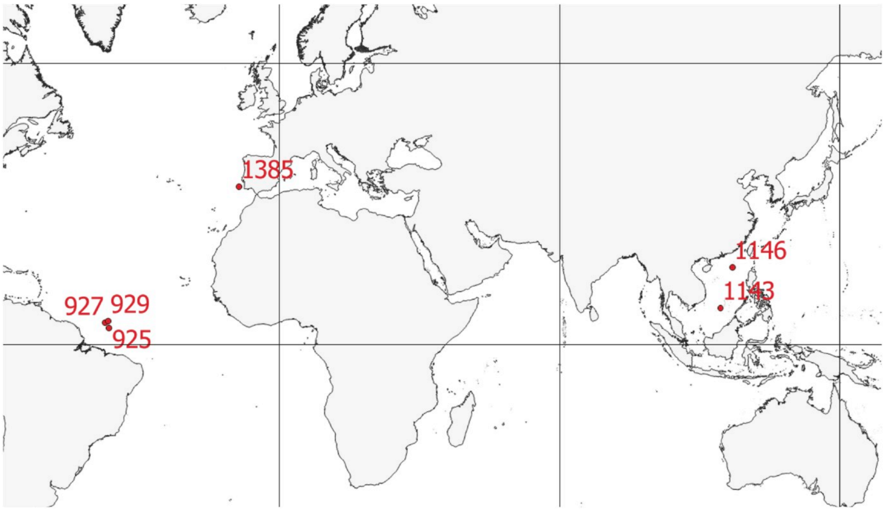

Fig. 3. Positions of sites indicated by their IODP numbers.

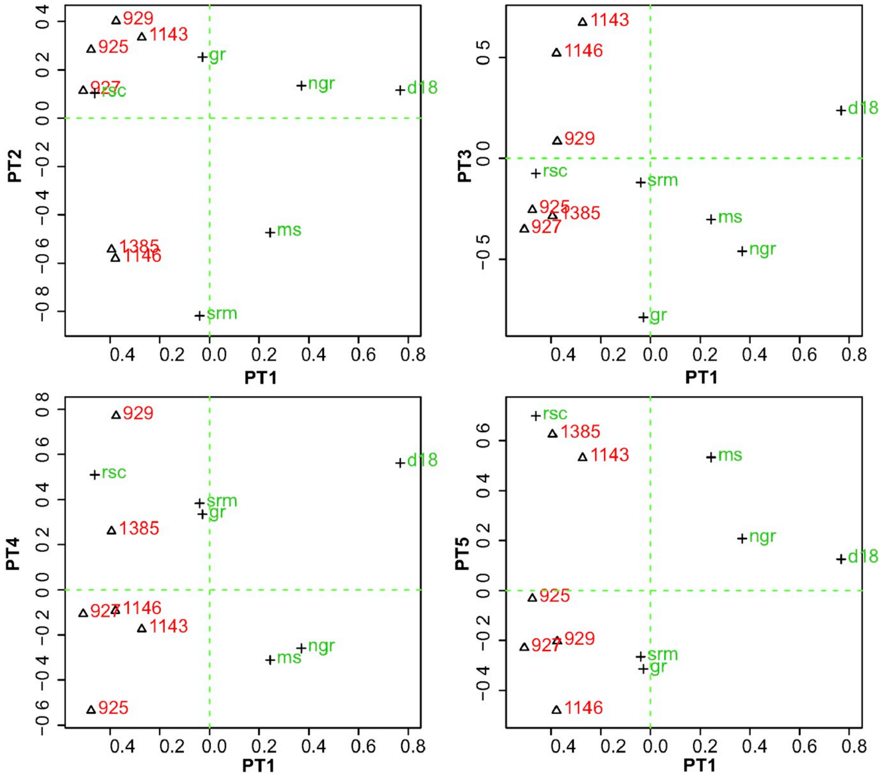

Fig. 4. Loadings of 3-way PTA: PT 1 vs PT 2, 3, 4, 5; sites are in red, proxies in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

kinds of loadings (sites and proxies) are grouped in brackets.[4]

Now, let us have a closer look at results and interpretations. First, results are much smoother and the noise components were discarded when incorporating SSA. The percentages of explained variability are rather low - 13% in total against 45% in case of 3-way PTA, but it is because, as we believe, of the artificial increase of the total data variability by multiple data replication in delaying process. Methods of possible normalisation of the results need additional investigations but are outside of the scope of this article. Second, tensor loadings and frequencies are easier to interpret. While the interpretation of tensor loadings is not the principal task of this article, we give here some possible geological explanations as an example.

As illustrated by the periodogram, see Fig. 8, and equation (9), $PT1$ represents climate variations with 100 kyr periodicity (Eccentricity), that affect mainly temperature plus ice volume expressed by delta of oxygen (- d18), and lightness of sediments (RSC) in all sites.

[4] For example, first fraction tells us that first Principal Tensor is responsible for more than 7% of variability of our 4-way data tensor, it is manifested synchronous in all investigated sites (927, …, 1143; all loadings are positive, minus in the denominator stands for the absence of negative loadings for sites), positively correlated with the reflectance of sediments, and negatively correlated with delta 18 of oxygen, and natural gamma ray activity.

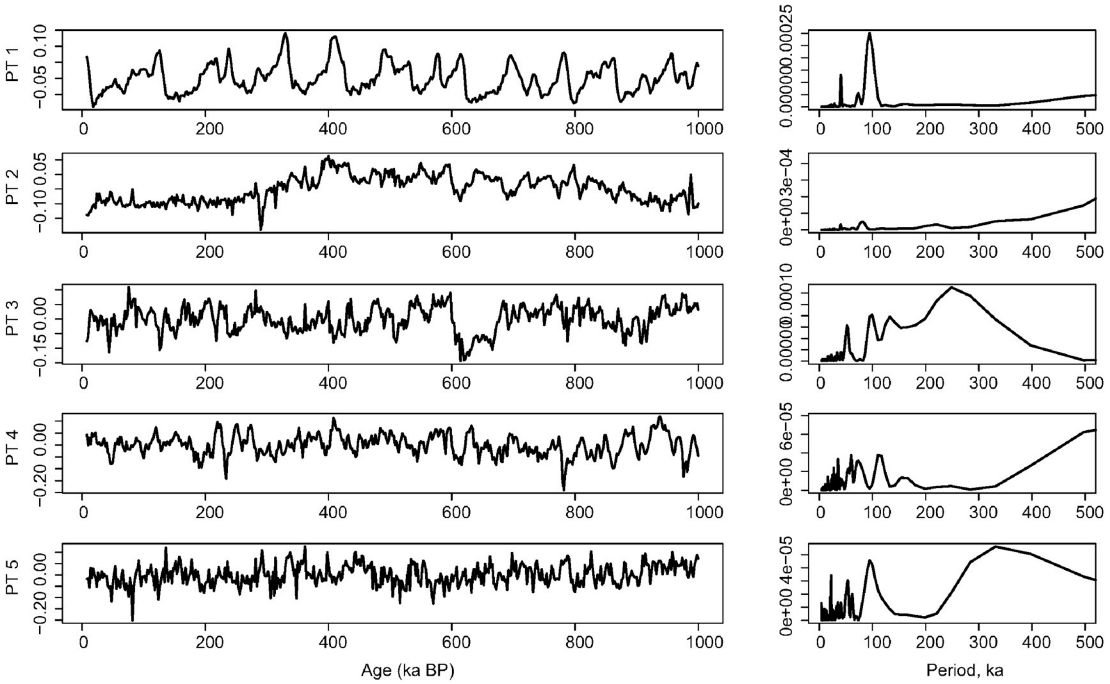

Fig. 5. Scores of 3-way PTA vs age (left) and their associated periodograms (right).

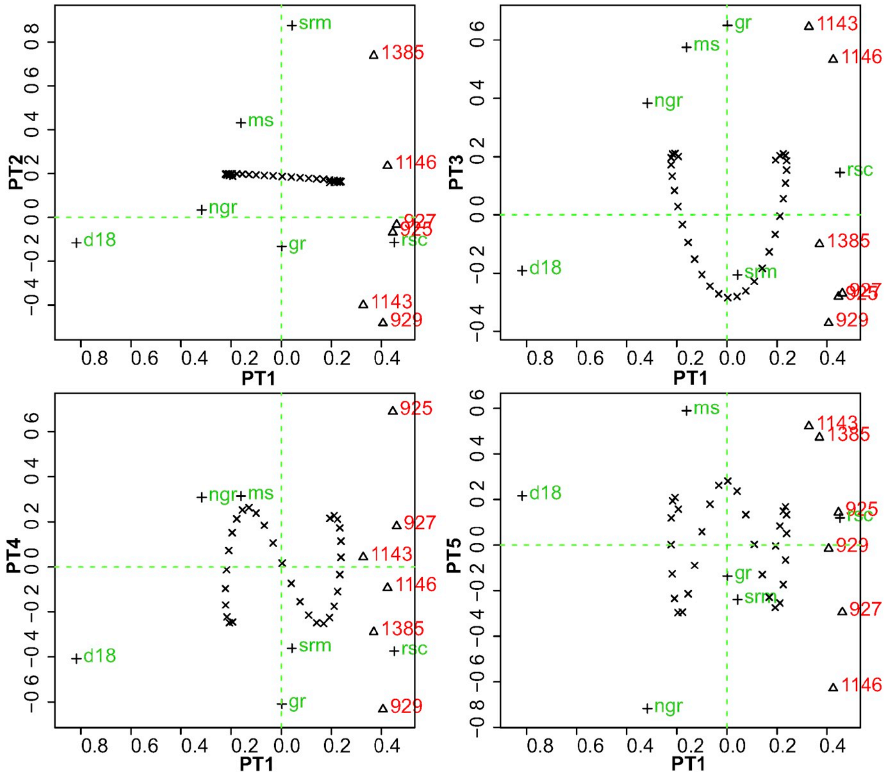

Fig. 6. Loadings of 4-way PTA: PT 1 vs PT 2, 3, 4, 5; sites are in red, proxies in green, EOFs in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

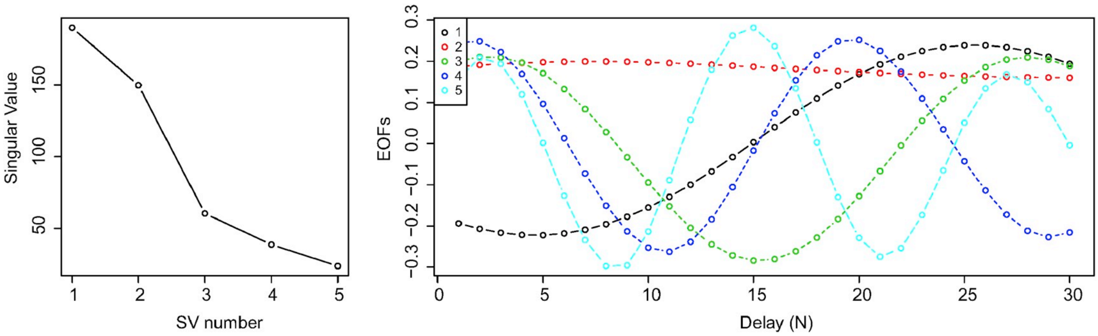

Fig. 7. Singular values (left) and Empirical Orthogonal Functions (right) for PT 1–5.

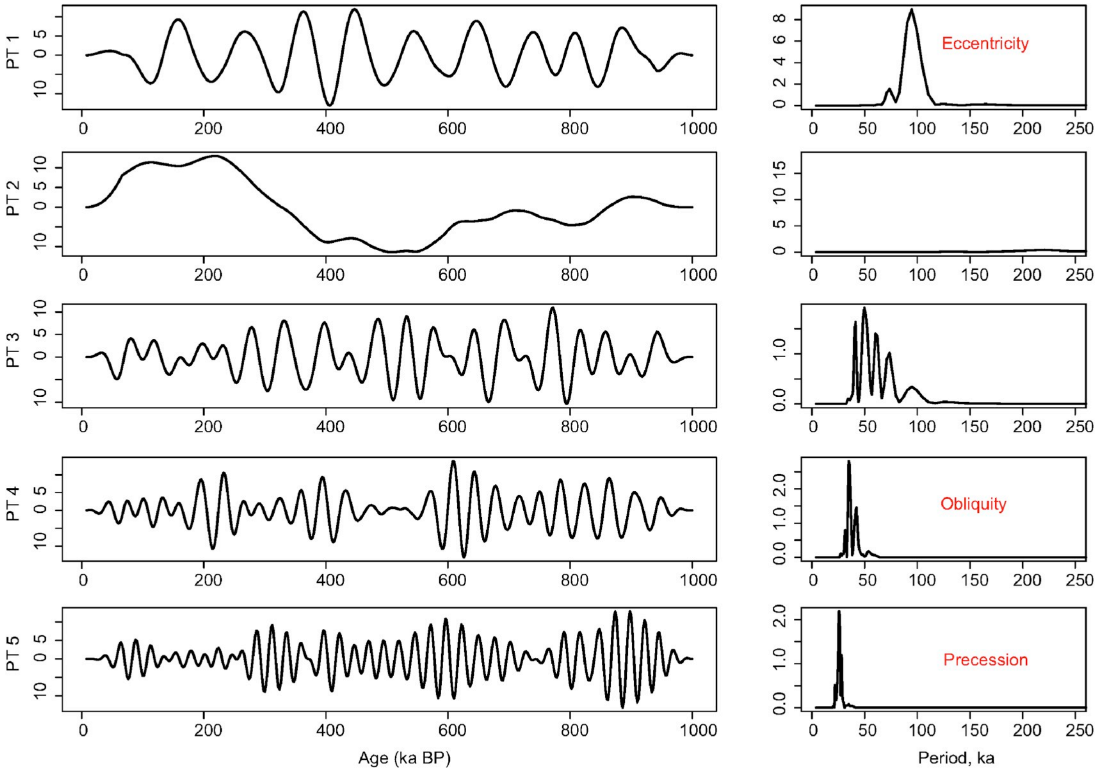

Fig. 8. Scores of 4-way PTA vs age (left) and their associated periodograms (right).

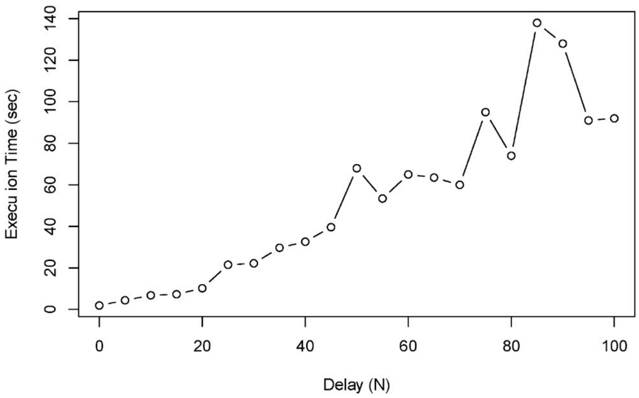

Fig. 9. Execution time versus the length of delay window.

$$PT1(7.2\%)\frac{(927_{0.46}925_{0.44}1146_{0.42}929_{0.4}1385_{0.37}1143_{0.33})(RSC_{0.45})}{(-)(d18_{0.81}NGR_{0.31})} \tag{9}$$

$PT2$ (eq. (10)) possibly represents long non-linear trend of different terrestrial input with magnetic minerals to sediments (SRM, MS) from Europe to Eastern Atlantic (sites 1385, 929) with a switch in the middle of observed interval, which can be related to the middle Pleistocene transition (MPT) – a long-term average ice volume gradual increase, see e. g. Clark et al. (2006).

$$PT2(4.5\%)\frac{(1385_{0.73})(SRM_{0.89}MS_{0.43})}{(929_{0.48})(-)} \tag{10}$$

$PT3$ (eq. (11)) represents quasi-periodical specifics of sedimentation (GR, MS, NGR) in China Sea (sites 1143, 1146).

$$PT3(0.72\%)\frac{(1143_{0.64}1146_{0.53})(-)}{(-)(GR_{0.65}MS_{0.57}NGR_{0.38})} \tag{11}$$

$PT4$ (eq. (12)) represents climatic variations with 41 kyr periodicity (Obliquity) in Atlantic (temperature + ice volume (d18) + terrestrial supply (GR, RSC, sites 925, 929).

$$PT4(0.3\%)\frac{(929_{0.63})(GR_{0.61}d18_{0.41}RSC_{0.37}SRM_{0.36})}{(925_{0.69})(MS_{0.31}NGR_{0.31})} \tag{12}$$

$PT5$ (eq. (13)) represents mainly China Sea water streams oscillations with 23 kyr periodicity (Precession), different terrestrial supply (NGR, MS), sites 1143,1146.

$$PT5(0.11\%)\frac{(1143_{0.52}1385_{0.47})(NGR_{0.72})}{(1146_{0.62})(MS_{0.59})} \tag{13}$$

Thus, just in one computer run we have extracted and refined several informative components from the noisy climatic data, including those with main Milankovitch frequencies, separated non-linear trend, located sites and recognised appropriated proxies best suited for paleo-climatic reconstructions. This was clearly not possible using the ordinary PT decomposition on 3-way data as we demonstrated in section 4.1.

As mentioned above, one source of the uncertainty in the method lays in original data, which includes natural noise, and in data pre-processing steps – such as de-trending and, especially, re-mapping from the original scale to a general scale (time in our case) or tuning. Extended method of PTA can resolve some of these uncertainties. But there is also another source related to the numerical algorithms used for the tensor decompositions (Principal Tensor Analysis on k-modes in this paper). Fig. 9 shows the execution time versus the length of delay window. The algorithm was tested on a computer with the 3.4 GHz Intel Core i7, 24 GB DDR3, data array of doubles 497 × 6 × 6. It is possible to see that we need a trade off between the size of the data tensor, delay window and execution time. The dependence is rather quasi-linear, but the execution time can be unpredictably long due to problems of the convergence of the algorithm, see Leibovici (2010) for details. Using another algorithms and packages can produce slightly different results.

5. Conclusions

Principal tensor analysis (PTA) is a modern tool for multi-way data reconstructions. It can also be useful in geosciences, particularly for extraction of information from multi-way (time (depth), space, and proxies) datasets. As opposed to other disciplines, like psychometrics, medicine and others with direct multiple data measurements in the same scale, it is not easy to use the similar direct approach with geological data. To construct an initial data tensor, all the measurements have to be re-mapped from the original (normally, depth or other distance) to some general scale (most often, time). We illustrate this using an example set from the IODP data base constructed using some climatic physical proxies from several deep ocean sediment cores.

Different solutions to run PTA are available, e.g. as R or Python packages. We illustrated here using an R package PTAk that direct application of the method on real geological data produces results which are difficult to be interpreted, mainly due to high levels of noise present in the data (sub-section 4.1, 3-way PTAk). We introduce an advanced method of PTA: PTA enhanced with Singular Spectrum Analysis (SSA). Using this enhancement, we were able to isolate clear patterns in noisy multi-way data. As an illustrative example, the method has been applied to 4-way data tensor (time, space, proxies, and delay-time) constructed from marine sediment proxies (sub-section 4.2, 4-way PTAk). In one computer run, we extracted and refined several informative components from the noisy climatic data, including those with main Milankovitch frequencies, separated non-linear trend, located sites and recognised appropriated proxies best suited for paleo-climatic reconstructions.

Theoretically, it could be possible to extend the data tensor to 5 and even more dimensions, e.g. delaying the space dimension, thus incorporating some elements of kriging. Practical problems can be expected because of the restrictions in computational resources. A trade off between the size of the data tensor and calculation time can be very important.

The algorithm has been implemented as an R function and available for use in the supplementary data and appendix. The applicability of the method is not limited to geosciences. It can be successfully used to reconstruct arbitrary multi-way datasets, including those that contain high level noise and non-linear trends.

Acknowledgements

We thank three anonymous reviewers for their careful reading of our manuscript and many useful comments and suggestions.

This work was conducted under support of ERC Consolidator Grant "EarthSequencing" (Grant Agreement no. 617462).

Appendix A

```
#********************************************************
# Function: PTA_SSA() - calculates Tensor Decomposition of 3-way data
    tensor
# (Principal Tensor Analysis merged with Singular Spectrum Analysis):
# 3-way data tensor is transformed to 4-way data tensor
# using delay time technique followed by Tensor Decomposition (PTAk
    package)
# and SSA reconstruction (Leibovici, 2010; Ghil et al., 2002; Kotov &
    Paelike (2017)).
#*******************************************************
# INPUT: myt - 3-way data tensor, first dimension is time/space
    ("score") scale
#        nbPT - number of extracted tensors
#        N - size of delay time window
#        summ - print summary (TRUE/FALSE)
#        plt - plot loadings (TRUE/FALSE)
#        dn2 - vector of names for 2-nd dimension
#        dn3 - vector of names for 3-rd dimension
# OUTPUT: list of decomposition matrices and vector of singular values
#        $rc - reconstructed orthogonal components (scores)
#        $l2 - matrix of loadings for 2-nd dimension
#        $l3 - matrix of loadings for 3-rd dimension
#        $eof - matrix of loadings for 4-th dimension (EOFs)
#        $sv - singular values
# ********************************************************
# Copyright 2017-2018 by Sergey Kotov (MARUM, Bremen Uni.)
#*********************************************************

# install.packages(’PTAk’) # Uncomment and install if not installed
library(PTAk)

PTA_SSA <- function(myt, nbPT, N, summ = TRUE, plt = TRUE, dn2 = NULL,
    dn3 = NULL) {

  N1 = dim(myt)[1] # time length of the 3-w data tensor (1-st dim)
  N2 = dim(myt)[2] # number of sites (2-nd dim)
  N3 = dim(myt)[3] # number of proxies (3-rd dim)

# Creation of delay time 4d tensor
  myt4 <- NULL

  for (i in 1:N) {
    myt4 <- c(myt4, myt[(N-i+1):(N1-i+1),,])
  }

   myt4 <- array(myt4,c(N1-N+1 , N2, N3, N))
   dnames4 <- list(NULL, dn2, dn3, NULL)
   dimnames(myt4) <- dnames4

# Run PTAk designed for PARAFAC/CANDECOMP model,
# see Leibovici (2010)
  pta4 <- PTAk(myt4, nbPT = c(0,nbPT), nbPT2 = 0, minpct = 0.0)

# Reconstructed Components RC
  RC <- matrix(0.0, N1, nbPT)
  for (j in 1:nbPT) {
    for (i in 1:(N-1)) {Mt=1/i; Lt=1; Ut=i; MyS <- 0
    for (k in Lt:Ut) {MyS <- MyS+pta4[[1]]$v[j,i-k+1]*pta4[[4]]$v[j,k]}
      RC[i,j] <- MyS/Mt
    }
    for (i in N:(N1-N+1)) {Mt=1/N; Lt=1; Ut=N; MyS <- 0
    for (k in Lt:Ut) {MyS <- MyS+pta4[[1]]$v[j,i-k+1]*pta4[[4]]$v[j,k]}
      RC[i,j] <- MyS/Mt
    }
    for (i in (N1-N+2):N1) {Mt=1/(N1-i+1); Lt=i-N1+N; Ut=N; MyS <- 0
    for (k in Lt:Ut) {MyS <- MyS+pta4[[1]]$v[j,i-k+1]*pta4[[4]]$v[j,k]}
      RC[i,j] <- MyS/Mt
    }
  }

  if (summ) summary.PTAk(pta4,testvar = 0)
  if (plt) {
    plot(pta4[[4]]$d,type = ’b’, xlab = ’SV number’, ylab = ’Singular
        Value’)
    for (i in 2:nbPT) plot(pta4, mod=c(2,3,4), nb1 = 1, nb2 = i, xpd=NA,
        lengthlabels = 4, cex=1.0)
  }

  l2 <- pta4[[2]]$v # matrix of loadings for 2-nd dimension
  l3 <- pta4[[3]]$v # matrix of loadings for 3-rd dimension
  eof <- pta4[[4]]$v # matrix of loadings for 4-th dimension (EOFs)
  sv <- pta4[[4]]$d # singular values

  result <- list(rc=RC, l2=l2, l3=l3, eof=eof, sv=sv)
  return(result)
}
```

Appendix B

```
#********************************************************
# Example on how to use the PTA_SSA() function.
# Function: PTA_SSA() - calculates Tensor Decomposition of 3-way data
    tensor
# (Principal Tensor Analysis merged with Singular Spectrum Analysis):
# 3-way data tensor is transformed to 4-way data tensor
# using delay time technique followed by Tensor Decomposition (PTAk
    package)
# and SSA reconstruction (Leibovici, 2010; Ghil et al., 2002).
# Example of 3-way tensor is constructed on data from IODP deep ocean
    sediments
# (6 sites, 6 proxies, see Kotov & Paelike (2017))
# ********************************************************
# Copyright 2017-2018 by Sergey Kotov (MARUM, Bremen Uni.)
#*********************************************************
rm(list=ls ())
graphics.off()

runtime = Sys.time() # tic

# Creation of 3-way data tensor ----------------------------------------
  # Read data from text files
  d18 = data.matrix(read.table("tensor_d18.txt", sep="\t", header=FALSE))
  gr = data.matrix(read.table("tensor_GRA.txt", sep="\t", header=FALSE))
  ms = data.matrix(read.table("tensor_MS.txt", sep="\t", header=FALSE))
  ngr = data.matrix(read.table("tensor_NGR.txt", sep="\t", header=FALSE))
  rsc = data.matrix(read.table("tensor_RSC.txt", sep="\t", header=FALSE))
  srm = data.matrix(read.table("tensor_SRM.txt", sep="\t", header=FALSE))
  # Merging matrixes into 3-way data tensor
  myt = array(c(d18,gr,ms,ngr,rsc,srm), c(nrow(d18), ncol(d18), 6))
  # Set dimension names
  DN2 = c( 925, 927, 929, 1143, 1146, 1385) # names of 2-nd dim (sites)
  DN3 = c("d18", "gr", "ms","ngr", "rsc", "srm") # names of 3-rd dim
      (proxies)

# Preparation of parameters ------------------------------------------
  nbPT = 5 # Number of Principal Tensors
  N =  30 # delay time window size
  summ = TRUE # print summary
  plt = TRUE # plot singular values and loadings

# PTA_SSA run --------
  source("PTA_SSA.R")
  pta = PTA_SSA(myt,nbPT, N, summ, plt, DN2, DN3)

# Print run time ----------------------------------------------------
  runtime = Sys.time() - runtime
  print(runtime)  # toc
```

```
# Some extra plots ----------------------------------------------------
  # Singular values (automatically plotted when plt = TRUE)
  plot(pta$sv,type = 'b', xlab = 'SV number', ylab = 'Singular Value')

  # PT1 vs PT2 loadings (automatically plotted when plt = TRUE)
  x2 = pta$l2[1,]
  y2 = pta$l2[2,]
  x3 = pta$l3[1,]
  y3 = pta$l3[2,]
  plot(x2,y2, col = "green",xlim = c(min(min(x2), min(x3))-0.1,
      max(max(x2), max(x3)+0.1)),
      ylim = c(min(min(y2), min(y3))-0.1, max(max(y2), max(y3))+0.1),
          xlab = 'PT1 loadings', ylab = 'PT2 loadings')
  text(x2, y2-0.05, DN2)
  points(x3,y3, col = "red")
  text(x3, y3-0.05, DN3)

  # EOFs plots
  eofs =
      cbind(pta$eof[1,],pta$eof[2,],pta$eof[3,],pta$eof[4,],pta$eof[5,])
  matplot(eofs, type = c("b"),pch=1,col = 1:5, xlab = 'Delay (N)', ylab
      = 'EOFs')
  legend("topleft", legend = 1:5, col=1:5, pch=1)

  # RC plots
  par(mfrow=c(nbPT,1)) # Reconstructed Scores
  time = seq(8,1000,2)
  for (i in 1:nbPT) plot(time,pta$rc[,i], type = 'l',lwd=2, xlab = 'Age
      (ka BP)', ylab = paste('PT',i))
```

References

Allen, M.R., Smith, L.A., 1996. Monte Carlo SSA: detecting irregular oscillations in the presence of colored noise. J. Clim. 9 (12), 3373–3404 http://journals.ametsoc.org/doi/abs/10.1175/1520-0442{%}281996{%}29009{%}3C3373{%}3AMCSDIO{%}3E2.0.CO{%}3B2.

Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika 35 (3), 283–319.

Cattell, R.B., 1944. "Parallell proportional profiles" and other principles for determining the choice of factors by rotation. Psychometrika 9 (4), 267–283.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., Phan, H.A., 2015. Tensor decompositions for signal processing applications: from two-way to multiway component analysis. IEEE Signal Process. Mag. 32 (2), 145–163.

Clark, P.U., Archer, D., Pollard, D., Blum, J.D., Rial, J.A., Brovkin, V., Mix, A.C., Pisias, N.G., Roy, M., 2006. The middle Pleistocene transition: characteristics, mechanisms, and implications for long-term changes in atmospheric pCO2. Quat. Sci. Rev. 25 (23–24), 3150–3184. https://www.sciencedirect.com/science/article/pii/S0277379106002332.

Danilov, D., Zhigljavsky, A.E., 1997. Principal Components of Time Series: the "Caterpillar" Method. SPbU Press, St.-Petersburg (in Russian)., St.-Petersburg.

Ghil, M., Allen, M.R., Dettinger, M.D., Ide, K., Kondrashov, D., Mann, M.E., Robertson, A.W., Saunders, A., Tian, Y., Varadi, F., Yiou, P., 2002. Advanced spectral methods for climatic time series. Rev. Geophys. 40 (1), 3 1–3.41.

Golyandina, N., Osipov, E., 2007. The Caterpillar-SSA method for analysis of time series with missing values. J. Stat. Plann. Inference 137 (8), 2642–2653. http://www.sciencedirect.com/science/article/pii/S037837580700016X.

Harshman, R.A., 1970. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis. UCLA Work. Pap. Phonetics 16, 1–84.

Hill, M.O., 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61 (1), 237–249.

Hitchcock, F.L., 1928. Multiple invariants and generalized rank of a p-way matrix or tensor. J. Math. Phys. 7 (1), 39–79.

Hodell, D., Lourens, L., Crowhurst, S., Konijnendijk, T., Tjallingii, R., Jiménez-Espejo, F., Skinner, L., Tzedakis, P.C., Abrantes, F., Acton, G.D., Zarikian, C.A., Bahr, A., Balestra, B., Barranco, E.L., Carrara, G., Ducassou, E., Flood, R.D., José-Abel, Flores, Furota, S., Grimalt, J., Grunert, P., Hernández-Molina, J., Kim, J.K., Krissek, L.A., Kuroda, J., Li, B., Lofi, J., Margari, V., Martrat, B., Miller, M.D., Nanayama, F., Nishida, N., Richter, C., Rodrigues, T., Rodríguez-Tovar, F.J., Roque, A.C.F., Goñi, M.F., Sierro, F.J., Singh, A.D., Sloss, C.R., Stow, D.A., Takashimizu, Y., Tzanova, A., Voelker, A., Xuan, C., Williams, T., 2015. A reference time scale for site U1385 (shackleton site) on the SW iberian margin. Global Planet. Change 133, 49–64.

IODP-MI, 2011. IODP depth scales terminology. https://www.iodp.org/policies-and-guidelines/142-iodp-depth-scales-terminology-april-2011/file.

Kolda, T.G., Bader, B.W., 2009. Tensor decompositions and applications. SIAM Rev. 51 (3), 455–500. http://epubs.siam.org/doi/abs/10.1137/07070111X.

Kotov, S., Pälike, H., 2017. Principal tensor analysis as a tool for paleoclimatic reconstructions. In: 18th International Association for Mathematical Geosciences Conference 2017. IAMG, Perth, pp. 67.

Kroonenberg, P.M., 2008. Applied Multiway Data Analysis. Wiley-Interscience.

Leibovici, D.G., 2010. Spatio-temporal multiway decompositions using principal tensor analysis on k-modes: the R package PTAk. J. Stat. Software 34 (10), 1–34. http://www.jstatsoft.org/v34/i10/.

Lisiecki, L.E., Raymo, M.E., 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic d18O records. Paleoceanography 20 (1), 1–17.

Pälike, H., 2005. Orbital Variation (Including Milankovitch Cycles).

Smilde, A.K., Bro, R., Geladi, P., 2004. Multi-way Analysis with Applications in the Chemical Sciences. Wiley.

Tucker, L.R., 1964. The extension of factor analysis to three-dimensional matrices. Contributions to Mathematical Psychology 109–127.

Wall, M.E., Rechtsteiner, A., Rocha, L.M., 2003. Singular value decomposition and principal component analysis. In: A Practical Approach to Microarray Data Analysis. Kluwer, Norwell, MA, pp. 91–109 https://www.cs.cmu.edu/{%7E}tom/10701{_}sp11/slides/pca{_}wall.pdf.