Journal of Statistical Software

May 2010, Volume 34, Issue 10. *http://www.jstatsoft.org/*

Spatio-Temporal Multiway Decompositions Using Principal Tensor Analysis on k-Modes: The R Package PTAk

Didier G. Leibovici
University of Nottingham

Abstract

The purpose of this paper is to describe the R package **PTAk** and how the spatio-temporal context can be taken into account in the analyses. Essentially `PTAk()` is a multiway multidimensional method to decompose a multi-entries data-array, seen mathematically as a tensor of any order. This PTAk-modes method proposes a way of generalizing SVD (singular value decomposition), as well as some other well known methods included in the R package, such as PARAFAC or CANDECOMP and the PCAn-modes or Tucker-n model. The example datasets cover different domains with various spatio-temporal characteristics and issues: (i) medical imaging in neuropsychology with a functional MRI (magnetic resonance imaging) study, (ii) pharmaceutical research with a pharmacodynamic study with EEG (electro-encephaloegraphic) data for a central nervous system (CNS) drug, and (iii) geographical information system (GIS) with a climatic dataset that characterizes arid and semi-arid variations. All the methods implemented in the R package **PTAk** also support non-identity metrics, as well as penalizations during the optimization process. As a result of these flexibilities, together with pre-processing facilities, **PTAk** constitutes a framework for devising extensions of multidimensional methods such as correspondence analysis, discriminant analysis, and multidimensional scaling, also enabling spatio-temporal constraints.

Keywords: multiway analysis, multi-entries data, spatio-temporal data, variance decomposition, multiway interaction, tensor decomposition, **PTAk**, R programming.

1. Introduction

Multiway data are common in different scientific and non-scientific domains where modelling interactions is crucial for better understanding of the studied phenomena. By "multiway" it is understood that observations are described by a series of characteristics dependent within the

design. The different characteristics are also called the entries, domains or modes of the data; their expressions can be called items, modalities, traits, variables and may be issued after a selected or a random sample within the domain represented. Multiway data occur when "repeated" measurements are made because of the design and/or because of the nature of the measurements. A typical example of multiway data are spatio-temporal data where some variables (1st mode) are measured (or evaluated) at a set of spatial locations (2nd mode) at different dates (3rd mode) (the choice of mode number being arbitrary). Besides testing an hypothesis or a model, on a multiway dataset, one may also be interested to "look at the data itself" and have a descriptive approach at least to formulate future hypotheses. Extracting and describing interactions of the data-modes is of prime interest, for example to derive the dynamics of a multivariate spatio-temporal dataset.

Another classical situation where multiway data are to be analyzed, is within a multidimensional scaling (MDS) approach, where a matrix of similarities (or dissimilarities) between a set of variables, objects or items is available for each subject or sub-samples of a given sample. Then the interest is not only on mapping the proximities of the variables, but also on the pattern of the subjects or associated sub-samples. To do this, the INDSCAL method, for example, uses the multiway decomposition CANDECOMP (Carroll and Chang 1970), see also Borg and Groenen (2005) and De Leeuw and Mair (2009) for recent descriptions of MDS: three-way and other multidimensional scaling methods.

Dealing with multiway data using multidimensional methods may be restrictive. As when analysing two-way tables, the multi-way table has to be collapsed or unfolded in a table with two modes, thereby looking at interactions of order 2 in a multiple fashion instead of looking at multiple interactions. This is the case, for example, for multiple correspondence analysis (MCA), see for example Le Roux and Rouanet (2004), as compared to simple correspondence analysis (CA or FCA). The former is not a *stricto sensu* extension of the latter when dealing with more than 2 categorical variables, but rather a "flatter" extension of it, where only 2-way marginals lack of independence are considered. The R (R Development Core Team 2009) package **PTAk**, available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=PTAk`, aims at decomposing interactions of order $k > 2$ (Leibovici and Sabatier 1998; Leibovici 2001, 2009). For example, the method `FCAk()` within the package decomposes the lack of independence measured by a χ^2 for the k variables in the k-way table. This particular PTAk-modes model will be described in Section 7, for general purposes but also for analyzing spatial patterns of categories from local occurrences of their associations.

Beforehand, the algebraic background extending matrix calculus will be shortly described in Section 2 along with the multiway methods implemented in the R package **PTAk**. In Section 3 the optimization procedure of the main method of the R package will be detailed: `PTAk()`. A brief comparison with some other well known multiway methods will also be made in this section. Sections 4 and 5 will give an overview of using the R package, whilst Sections 6, 7, and 8 will describe some generic approaches to derive decomposition models useful in a spatio-temporal context. The framework used within the PTAk-modes model and so within the R package **PTAk** extends some duality principles (Cailliez and Pagès 1976; Escoufier 1987; Dray and Dufour 2007), therefore extending the approaches of multidimensional analysis focusing on spatial-temporal data, such as the methods decomposing local and global variances as in **ade4** (Chessel *et al.* 2007). We may use indifferently the notations PTAk-modes and PTAk to describe the same model, decomposition or statistical method.

2. Basics of multiway data analysis

Spatio-temporal measurements are naturally linked to multiway data. For example, Tunisian climatic data, analyzed in further sections, deals with 10 climatic indicators measured on a spatial domain, a pixel grid of size 2599, and summarized by their 12 monthly average over 50 years.

In R this can be stored in a multiple-entries table, an "`array`" object, here of dimension $2599 \times 12 \times 10$, where the first entry refers to *space*, the second to *month*, and the third to *indicator*. Multiway data can occur in other contexts and appear usually when repeating the same measurements on some statistical units, spatially, at different times, and/or different conditions. For the CNS drug data one can obtain an array with 5 entries: *subject*, *drug-dose*, *time*, *electrode*, and *EEG spectral band*; the interest for this pharmaco-dynamic study is about identifying differences in doses with a spatial zone of activation for a specific EEG band pattern. Multiway data, stored in an "`array`" object, can be collapsed to a "`matrix`" object, allowing the use of multivariate methods, inferential or descriptive such as multidimensional analysis, or even into a single "`vector`" to use univariate models such as ANOVA taking in account the complex design as covariates in the inferential procedure.

2.1. Models for multiway interactions

Multiway data analysis acknowledges the multiple interactions of the data. ANOVA (analysis of variance) which deals with decomposition and interaction is not a multidimensional method; pursuing this kind of approach FANOVA methods (F for factorial) added nonetheless a factorial decomposition for two modes (Gollob 1968).

In mathematical algebra, an array can be seen as a multilinear form or tensor (Lang 1984). The properties of tensor algebra enable to derive multiple-entries table calculus, therefore extending matrix calculus (Franc 1992; Leibovici 1993; Dauxois *et al.* 1994; Leibovici and Sabatier 1998). A multiway multidimensional method or multiway method for short, deals directly with the multilinear aspects of multiple-entries array data by proposing a tensorial decomposition, i.e., a multilinear decomposition of the form:

$$A(x, y) = \sum_{u}^{r} P_u(x, y) + \epsilon \tag{1}$$

for a matrix A (a tensor of order 2) acting as a bilinear form on vectors x, y of appropriate dimensions, which extends as:

$$B(x, y, z) = \sum_{v}^{r} T_v(x, y, z) + \epsilon \tag{2}$$

for the three-way array B (a tensor of order 3) acting as a trilinear form. The structures of P_u and T_v express the model by being some "simple element" of the tensor space, usually rank-one tensors, see further. The number of components r is part of the model in terms of approximation level and ϵ is the "residual".

According to different model optimizations and constraints, one gets different forms of decomposition and properties for the elementary tensors (the P_u or the T_v). For example in PCA optimization, the components will have the form $P_u = \sigma_u \psi_u \, {}^t\phi_u$ where ψ_u is a principal

components normed to 1 and ϕ_u is the corresponding factor loadings or factor component also normed to 1 (notation: ${}^t v$ is the "row" vector, transpose of a "column" vector v). The σ_u value correspond to the square root of the variance associated with this principal component, and we have:

$$P_u(x,y) = \sigma_u({}^t x\psi_u)({}^t\phi_u y) = {}^t x(\sigma_u \psi_u {}^t\phi_u)y \tag{3}$$

which is equal to σ_u if $x = \psi_u$ and $y = \phi_u$. Equation 3 enables the study of the properties of $\hat{A} = \sum_u^r \sigma_u \psi_u {}^t\phi_u$ as an approximation of A. For trilinear or higher order forms, the notation used in Equation 3 becomes:

$$\begin{aligned} T_u(x,y,z,t) &= \sigma_u({}^t x\psi_u)({}^t\phi_u y)({}^t\varphi_u w)({}^t\xi_u z) \\ &= \sigma_u(\psi_u \otimes \phi_u \otimes \varphi_u \otimes \xi_u)..(x \otimes y \otimes z \otimes t) \\ &= \sigma_u[(\psi_u \otimes \phi_u \otimes \varphi_u \otimes \xi_u)..x]..(y \otimes z \otimes t) \end{aligned} \tag{4}$$

where $\otimes$ and .. are respectively called tensor product and contraction. The tensor product is also known as the outer product and the contraction generalizes the operation performed when transforming a vector by a matrix.

The models PARAFAC/CANDECOMP (refer to Carroll and Chang 1970; Harshman 1970), PARAFAC-orthogonal and PTAk-modes (Leibovici 1993; Leibovici and El Maâche 1997) follow this generic presentation as well as PCAn-modes (or Tucker-n model) (Kroonenberg and De Leeuw 1980; Kroonenberg 1983; Kaptein *et al.* 1986) but the latter is usually presented in a condensed way using tensor product of matrices, see further, Equation 10. The estimation procedure is usually an alternating least squares (ALS) optimization, i.e., after initialization, optimizing one set of components at a time, the other being fixed by the previous optimization.

Setting no particular constraints between vector components within each mode or entry of the table, PARAFAC/CANDECOMP, where the number of components for each mode has to be equal, performs the optimization by alternating multivariate regression techniques. Generic PCAn-modes will not impose equality of number of components for each mode but stating orthogonality within each mode, performs optimization by alternating eigen-decomposition of a particular symmetric matrix (Leibovici 1993). PARAFAC-orthogonal can be seen as a PARAFAC/CANDECOMP where orthogonality between the components of each mode is imposed, or as a PCAn-modes where the *core* tensor C (see Equation 10) expressing cross-links between components is imposed to be hyper-diagonal (only $C_{iii} \neq 0$). PARAFAC-orthogonal can be obtained using a PTAk-modes, by keeping only "main" principal tensors (see further). PTAk-modes proceeds also using ALS (alternating least squares) technique but step by step instead of optimizing the full set of components at each optimization. The algorithm involved in the PTAk-modes model is explained in more detail in Section 3, in Equations 8, 9 and 11; the expression of the CANDECOMP/PARAFAC model and the PCAn-modes model are also explicit in Equation 9 and 10.

2.2. Manipulation of tensors in R

Within R, the tensor product can be performed using the outer product (`%o%`) or using the Kronecker product (`%x%`). As the tensor is an algebraic operation, it is up to the computational step to choose one or the other:

```
R> c(1, 2, 3) %x% c(4, 5)
```

```
[1]  4  5  8 10 12 15
```

The result with the outer product is an array, here a matrix emphasized the bilinear property. The vectorization of the array is a permuted version of the Kronecker product:

```
R> c(1, 2, 3) %o% c(4, 5)

     [, 1] [, 2]
[1, ]    4    5
[2, ]    8   10
[3, ]   12   15

R> dim(c(1, 2, 3) %o% c(4, 5) %o% c(3, 1))

[1] 3 2 2

R>  all(as.vector(t(c(1, 2, 3) %o% c(4, 5))) == c(1, 2, 3) %x% c(4, 5))

[1] TRUE
```

Notice the above matrix is of rank one. The tensor product of any number of vectors gives what is called a rank-one tensor, as in fact any bilinear function resulting from collapsing the array into a matrix will be always of rank one.

When storing a dataset into an "`array`" object it is also essential to know that the left index runs faster: try `array(1:24, c(2, 3, 4))`. Performing a contraction of tensor of dimension $(30, 10, 4, 2)$ by a vector of dimension 4 can be done by collapsing the tensor into a matrix of dimension $(600, 4)$, then performing the multiplication of the vector by the matrix, then by reforming the array of dimension $(30, 10, 2)$. This kind of operation is facilitated by the packages **tensor** (Rougier 2002) and **tensorA** (van den Boogaart 2007).

The outer product concatenates dimensions and multiplies the left matrix by each element of the right matrix; the Kronecker product multiplies dimensions and multiplies the right matrix by each element of the left matrix:

```
R> A <- matrix(1:8, 4, 2)
R> B <- matrix(c(1, 2, 0, 1), 2, 2)
R> class(B %o% A)

[1] "array"

R> dim(B %o% A)

[1] 2 2 4 2

R> class(A %x% A)

[1] "matrix"
```

```
R> dim(A %x% A)

[1] 16  4
```

Note the essential tool for data analysis is the `array()` function to store the dataset and its related methods such as `aperm()` to permute the dimensions of the array. The operators briefly described above will be used within the methods to decompose the multi-entry table.

3. Extension of PCA as proposed by PTAk-modes

The PTAk model approach is similar to a step by step PCA, but for tensors. In order to describe the generalization proposed with the PTAk-modes model, let us first rewrite the PCA method within a tensorial framework.

3.1. PCA of tensor of order 2

For a given matrix X of dimension $n \times p$, the first principal component is a linear combination (given by a p-dimensional vector φ_1) of the p columns ensuring maximum sum of squares of the coordinates of the n-dimensional vector obtained. The square root of this sum of squares is called the first singular value σ_1. One has: ${}^t(X\varphi_1)(X\varphi_1) = \sigma_1^2$ and $X\varphi_1/\sigma_1$ is the principal component normed to 1. This maximization problem can be written either in matrix or tensor form:

$$\begin{aligned} \sigma_1 &= \max_{\substack{\|\psi\|_n=1 \\ \|\varphi\|_p=1}} ({}^t\psi X\varphi) = \max_{\substack{\|\psi\|_n=1 \\ \|\varphi\|_p=1}} X..(\psi \otimes \varphi) \\ &= {}^t\psi_1 X\varphi_1 = X..(\psi_1 \otimes \varphi_1) \end{aligned} \tag{5}$$

In Equation 5, X represents either the data matrix or the data tensor of the same data table. Another easy way of understanding computationally the algebraic operators ".." and "$\otimes$" is to see them as the following operations: $\psi_1 \otimes \varphi_1$ is a np vector of the n blocks of the p vectors $\psi_{1i}\varphi_1$, $i = 1, ...n$ (this is the computational description using the Kronecker product); ".." called a contraction, generalizes the multiplication of a matrix by a vector and in the case of equal dimensions (as above), it corresponds to the natural inner product (X is then also seen an np vector). ψ_1 is termed first principal component, φ_1 first principal axis, $(\psi_1 \otimes \varphi_1)$ is called first principal tensor.

Notice here the description of a tensor of order 2, a bilinear map, as associated to a matrix is usually associated to one linear map. The duality diagram (Cailliez and Pagès 1976; Escoufier 1987; Dray and Dufour 2007) comes to complete the association with another linear map on the dual spaces involved to define the other linear map: expressed by the transposed matrix.

The contraction, "..", is implemented within the function `CONTRACTION()` and it uses the R package **tensor** (Rougier 2002).

3.2. PTAk-modes of a tensor of order $k > 2$

Now if X is a tensor of higher order, say 3 here we can look for the first principal tensor

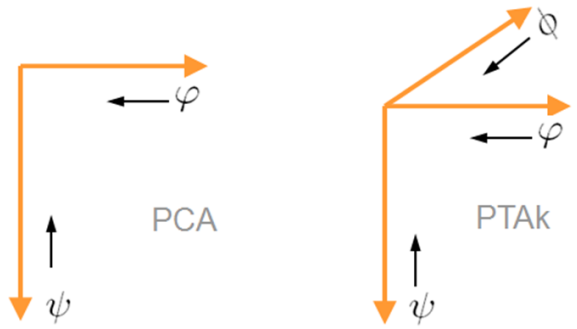

Figure 1: Illustrative comparison between PCA and PTAk (here with $k = 3$) when computing singular values by Complete Contractions given in the Equations 5 and 6: the basis of the *RPVSCC* algorithm.

associated with the singular value with the optimization form:

$$\begin{aligned} \sigma_1 &= \max_{\substack{\|\psi\|_s=1 \\ \|\varphi\|_v=1 \\ \|\phi\|_t=1}} X..(\psi \otimes \varphi \otimes \phi) \\ &= X..(\psi_1 \otimes \varphi_1 \otimes \phi_1) \end{aligned} \tag{6}$$

This is a direct extension of Equation 5, as expressed by the practical schemas in Figure 1, with contractions made either on a matrix table or on a tensor of order 3. The further extension to $k > 3$ is straightforward. `CONTRACTION.list()` is convenient relatively to Equations 5 and 6 as it performs the contraction without computing the tensor product of the vectors in the first place as algebraically:

$$X..(\psi \otimes \varphi \otimes \phi) = (X..\psi)..(\varphi \otimes \phi) = (X..\varphi)..(\psi \otimes \phi) = (X..\phi)..(\psi \otimes \varphi) = ((X..\psi)..\varphi))..\phi \tag{7}$$

The function `SINGVA()` computes the best rank-one approximation of the given tensor X together with its singular value, given by Equation 6 (and a similar equation for higher orders). The therein algorithm, called RPVSCC in Leibovici (1993), is inspired from the algorithm of "reciprocal averaging" (Hill 1973) also known as the "transition formulae" in modern correspondence analysis and in the signal processing community as the "power method".

Adding an orthogonality constraint to Equation 6 allows us to carry on the algorithm to find the second principal tensors and so on. The optimization becomes is equivalent but working on $P_{(\psi_1^\perp \otimes \varphi_1^\perp \otimes \phi_1^\perp)}X$: the orthogonal projection of X onto the orthogonal tensorial of the principal tensor, *i.e* $(\psi_1^\perp \otimes \varphi_1^\perp \otimes \phi_1^\perp)$; this projector can also be written as $(Id - P_{\psi_1}) \otimes Id - P_{\varphi_1}) \otimes (Id - P_{\phi_1})$. Following this algorithm schema, given in Equation 8, the PTAk-modes decomposition obtained offers a way of synthesizing the data according to uncorrelated sets of components. Within this schema implemented for the functions `PTA3()` and `PTAk()` one can distinguish main principal tensors from associated principal tensors. The latter are associated with main principal tensors as they show one or more component of this main principal tensor in their sets of components. The associated principal tensors are obtained by a PTA$(k-1)$-modes decomposition once the k-modes data array has been "contracted" by the given component.

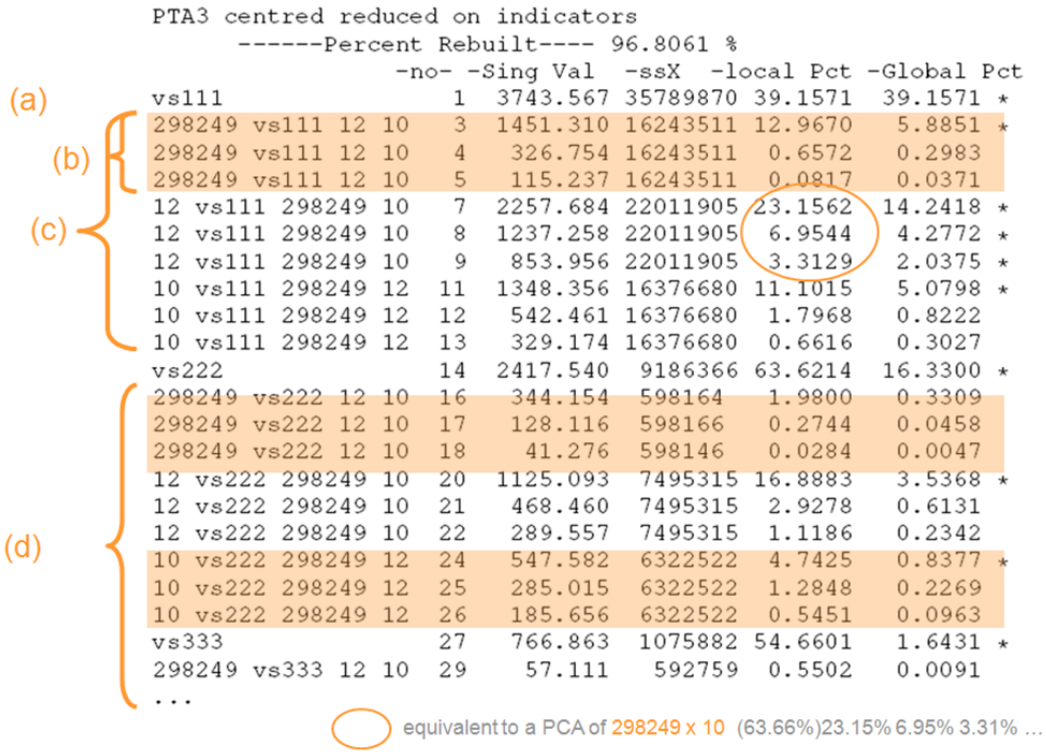

Figure 2: Output summary from the function `summary()` on a "`PTAk`" object, here the climatic data described in Section 4.3: (a) is the first principal tensor, (c) represents all the associated principal tensors to first one such like (b) are the spatial-mode associated principal tensors, (d) corresponds to a PTAk-modes decomposition of the initial data tensor projected onto the orthogonal tensorial of the first principal tensor.

This makes the algorithm a recursive algorithm with the following procedure, where here $k = 3$:

$$\begin{aligned} \text{PTA3}(X) &= \sigma_1(\psi_1 \otimes \varphi_1 \otimes \phi_1) \\ &+ \psi_1 \otimes_1 \text{PTA2}(P(\varphi_1^{\perp} \otimes \phi_1^{\perp})X..\psi_1) \\ &+ \varphi_1 \otimes_2 \text{PTA2}(P(\psi_1^{\perp} \otimes \phi_1^{\perp})X..\varphi_1) \\ &+ \phi_1 \otimes_3 \text{PTA2}(P(\psi_1^{\perp} \otimes \varphi_1^{\perp})X..\phi_1) \\ &+ \text{PTA3}(P(\psi_1^{\perp} \otimes \varphi_1^{\perp} \otimes \phi_1^{\perp})X) \end{aligned} \tag{8}$$

The notation $\otimes_i$ means that the vector on the left hand will take the ith place, among the k places here, in each full tensorial product, e.g., $\varphi_1 \otimes_2 (\alpha \otimes \beta) = \alpha \otimes \varphi_1 \otimes \beta$. More details on the properties of the method and on each function of the R package is given in the references Leibovici and Sabatier (1998); Leibovici (2009).

Equation 8 and Figure 2 illustrate the multi-hierarchical decomposition obtained with the PTAk-modes model. In Figure 2, in almost the same way as for PCA, one gets a hierarchy of principal tensors corresponding to a hierarchy of sum of squares, i.e., by the square of the singular values (σ) under the column `-Sing Val` associated with each principal tensor. It is a multilevel hierarchy in agreement with Equation 8. Percents of variability associated with

each principal tensor can be used to retain the main variability within the data tensor X. These percentages are in the `-Global Pct` column of Figure 2 whereas `-local Pct` are relative to the sum of squares given in column `-ssX` linked to the current tensorial optimization as defined in Equation 8. Plots of the vector components of a particular principal tensor allows the description of the extracted variability for each principal tensor.

`PROJOT()` is the function within `PTAk()` performing the orthogonal tensor projection of Equation 8 but can also be used for any structure or design associated with each mode to perform a linear constrained analysis in the same way as for PCAIV (principal component analysis on instrumental variables), see Leibovici (2000) for a full description of using `PTAIVk()` and in the **PTAk** manual for `PROJOT()` where a quick implementation is given as an example.

3.3. A brief comparison of multiway models

Before expressing in detail the R usage of the main methods within **PTAk** a practical comparison of the multiway models already described is of use. The models behind the methods `PTAk()`, `CANDPARA()` (PARAFAC/CANDECOMP) and `PCAn()` (Tucker-n model) are equivalent when looking for best rank-one approximation. This can be demonstrated from the expression of the models associated with these methods and can be understood from Equations 9 and 10. Using an example this would be:

```
R> library("PTAk")
R> PTAk(X, nbPT = 1, nbPT2 = 0) == CANDPARA(X, dim = 1)
R> PTAk(X, nbPT = 1, nbPT2 = 0) == PCAn(X, dim = rep(1, length(dim(X))))
R> CANDPARA(X, dim = 1) == PCAn(X, dim = rep(1, length(dim(X))))
```

This cannot be strictly verified using the package **PTAk** as `CANDPARA()` and `PCAn()` in their implementation only accept rank approximation greater than 1. Working around using a tensor "nearly" of rank-one is:

```
R> X <- c(1, 2, 3) %o% c(2, 4, 6) %o% c(3, 7) + rnorm(18, sd = 0.0001)
R> sol1 <- PTAk(X, nbPT = 2, nbPT2 = 0)
R> sol2 <- CANDPARA(X, dim = 2);
R> sol3 <- PCAn(X, dim = c(2, 2, 2))
R> sol1[[1]]$v[1, ]

[1] 0.2672617 0.5345234 0.8017830

R> sol2[[1]]$v[1, ]

[1] -0.2672617 -0.5345234 -0.8017830

R> sol3[[1]]$v[1, ]

[1] -0.2672617 -0.5345234 -0.8017830

R> sol1[[3]]$d
```

```
[1] 2.132416e+02 2.086484e-04
```

showing the first mode component for the first principal tensor given by `sol1[[1]]$v[1,]` as equal to the other approximations, (it is the same with mode 2 `sol[[2]]` and mode 3 `sol[[3]]`). The tensor can be said to be "nearly" of rank-one as the ratio the two singular values, `sol1[[3]]$d` is 10^6.

This gives a numerical proof of equivalence between PTAk, PARAFAC/CANDECOMP and Tucker-n when looking for the best rank-one approximation. Then the methods differ as also differs the rank definition attached to each model. PTAk will *try* to look for best approximation according to the orthogonal rank, i.e., the rank-one tensors (of the decomposition) are orthogonal; Tucker-n or PCAn-modes will look for best approximation according to the space-ranks, i.e., ranks of every bilinear form deducted from the original tensor (folding the multi-array into a matrix), that is the number of components in each space; PARAFAC/CANDECOMP will look for best approximation according to the rank, i.e., the rank-one tensors are not necessarily orthogonal.

It is said here "PTAk will *try*" as it has been shown recently on an example that the orthogonal-rank was not necessarily providing a nested decomposition as PTAk-modes implies (Kolda 2003). One can also notice that PTAk model extends the PARAFAC-orthogonal if one only retains in the decomposition the main principal tensors (not the associated ones), i.e., by setting `nbPT2 = 0` in the `PTAk()` call or by ignoring them.

The function `REBUILD()` will return the approximated or filtered dataset according to the method used, either `PTAk()`, `CANDPARA()`, or `PCAn()`; the parameters of the method are the list of tensors and/or a global threshold for percentage of variability explained by each elementary tensors. For `PCAn()` the function calls `REBUILDPCAn()` which does not use these parameters.

```
R> Xapp <- REBUILD(sol1, nTens = c(1, 2), testvar = 1e-12)

-- Variance Percent rebuilt X  at  100 %
-- MSE  4.378514e-09
-- with  2  Principal Tensors out of  2  given
-- compression     0  %
```

For `PTAk()` and `CANDPARA()`, the approximation is done according to the equation model, here written for a tensor of order 4:

$$X = \sum_{i \in \varsigma} \sigma_i \psi_i \otimes \varphi_i \otimes \phi_i \otimes \xi_i + \epsilon \tag{9}$$

where ς is a set of the selected elementary tensors. The `PCAn()` rebuilt approximation is a direct generalization of model from Kroonenberg and De Leeuw (1980):

$$X = (\Psi \otimes \Upsilon \otimes \Phi \otimes \Xi)..C + \epsilon \tag{10}$$

where the components here are matrices of components with as many columns in each mode-space as asked for during the optimization analysis (the space-ranks), and C being the *core* tensor with dimensions corresponding to the space-ranks.

4. Running a general PTAk

`SINGVA()`, `PROJOT()` described above are part of the main functions for 2-modes analysis, such as `SVDGen()`, and k-modes analysis with `PTAk()`, `CANDPARA()` and `PCAn()`. They can also be used to devise new analysis. So once you have loaded or scanned the dataset from other sources or format, put it in a multi-array, an "`array`" object in R you can run the `PTAk()` decomposition or the other multiway methods.

4.1. Structure of the PTAk package

The package can be summarized as four series of methods: (i) preprocessing the methods `Multcent()`, `IterMV()`, `Detren()` for data preprocessing but also some metrics preparation such as `CauRuimet()`, (ii) the multiway analysis methods `PTAk()`, `FCAk()`, `CANDPARA()`, and `PCAn()` which output objects of class "`PTAk`" and appropriate subclasses given by the name of the analysis along with `S3` methods associated with them (iii) `plot()`, `summary()`, and `REBUILD()`, the other methods (iv) are either internal or used within main methods but they can be used for developing further methods.

The principal argument of the preprocessing methods is an "`array`" object which has been prepared beforehand for data analysis: the array will be the multiway table of the measurements arranged by their modes, i.e., the "dimensions" deserving interest, e.g., time(s), variable(s), subject(s), space(s), countrie(s), or condition(s). Whichever name used to describe an entry of the table, it has a particular semantic according to the study. For example *time* for the ecolimatic study example corresponds to 12 months, and for the pharmaco-dynamic study it is the hour and minutes at measurements. Some examples of preprocessing are given in Section 4.3, see the help files of the package for a detailed description of the other arguments.

The principal arguments for the analysis methods are first of all, either an "`array`" object of the multiway dataset or a "`list`" object with `$dat` containing an "`array`" object and `$met` containing the metrics associated with each entry of the array, then the "amount" of approximation chosen. A metric is a semi-definite positive symmetric matrix allowing to perform non-canonical scalar product, i.e., covariance, "*sum of squares of products*", within the corresponding vectorial space (see Section 6 for further explanation). The arguments related to this "amount" of approximation chosen are: `dim` an integer for `CANDPARA()` and a "`vector`" object of integers for `PCAn()` fixing respectively as described previously the number of elementary tensors to fit, and the size of the core tensor (therefore the number components in each space); for `PTAk()` one chooses the number of principal tensors at each "level" of analysis by `nbPT`, the last level (2-modes analysis) is fixed by `nbPT2`. Note that for `PTA3-modes()` `nbPT` has to be just an integer but for $k > 3$ it can be a vector (of length $(k-2)$) specifying this choice for each level above 2-modes analysis. The current version of the package doesn't give much support for other plots or interpretations for `CANDPARA()` and `PCAn()`. For example the `summary()` method on a `PCAn` object doesn't properly describe the core tensor and no jointplot method (see Kroonenberg 1983) has been implemented yet in the package **PTAk**. Further practical use of the functions are described in the help files of the package Leibovici (2009), but some practical examples are given in the next section.

4.2. Practical example

This illustrative session uses the dataset related to an ecoclimatic delineation problem (Lei-

bovici *et al.* 2007), where dynamics over a typical year of 10 climatic indicators were analysed in the circum-saharan zone, using their monthly average estimates. The problematic explained in Leibovici *et al.* (2007) is to delineate homogeneous zones in relation to the ecolimatic profile (rainfall, temperature, evapotranspiration, etc.). So finding main spatial patterns via the spatial components associated with a climatic profile and a seasonal pattern, was the aim of this analysis. Here the studied zone has been limited to Tunisia; the shapefile contains a regular grid with the multivariate values. The dataset `Zone_climTUN` was obtained using the call `read.shape("E:/R_GIS/R_GilHF/TUN/tunisie_climat.shp")` based on the `read.shape()` function from the **MAPS** package. For replication, the data are also available in `PTAk`:

```
R> library("PTAk")
R> library("maptools")
R> data("Zone_climTUN")
```

The next command produces a plot of a `MAP` object not shown here:

```
R> plot(Zone_climTUN, ol = NA, auxvar = Zone_climTUN$att.data$PREC_OCTO,
+    nclass = 20)
```

The data are transformed into an `array` object:

```
R> Zone_clim <- Zone_climTUN$att.data[, c(2:13, 15:26, 28:39, 42:53,
+    57:80, 83:95, 55:56)]
R> Zot <- Zone_clim[, 85:87]
R> temp <-colnames(Zot)
R> Zot <- as.matrix(Zot) %x% t(as.matrix(rep(1, 12)))
R> colnames(Zot) <- c(paste(rep(temp [1], 12), 1:12),
+    paste(rep(temp [2], 12), 1:12), paste(rep(temp [3], 12), 1:12))
R> Zone_clim <- cbind(Zone_clim[, 1:84], Zot)
R> dim(Zone_clim)

[1] 2599 120

R> Zone3w <- array(as.vector(as.matrix(Zone_clim)), c(2599, 12, 10))
R> dim(Zone3w)

[1] 2599   12   10

R> dimnames(Zone3w) <- list(rownames(Zone3w), 1:12, c("P", "Tave", "ETo",
+    "PETo", "Tmax", "Tmin", "Q3", "Alt", "dM2T", "dMETo"))
R> Zone3w.PTA3-modes <- PTA3-modes(Zone3w, nbPT = 3, nbPT2 = 3,
+    minpct = 0.1, addedcomment="centrée réduite sur var")

 ---Final iteration---  7
 --Singular Value--  59898.86  -- Local Percent --  97.62936 %
 ---Final iteration---  26
 --Singular Value--  2860.392  -- Local Percent --  68.66842 %
```

```
 ---Final iteration---  39
 --Singular Value--  401.1593  -- Local Percent --  38.09571 %
  ++ Last 3-modes vs < 0.1 % stopping this level and under ++
 -----Execution Time----- 7.43

R> summary(Zone3w.PTA3-modes, testvar = 0.01)

 ++++ PTA3-modes ++++
               data =   Zone3w   2599 12 10
    PTA3-modes centrée réduite sur var
                ------Percent Rebuilt---- 99.97716 %
                ------Percent Rebuilt from Selected ---- 99.95512 %
                 -no- --Sing Val--    --ssX-- --local Pct-- --Global Pct--
vs111                1      59898.9 3674994157      97.62936      97.629361
2599 vs111 12 10     3       3243.0 3598688392       0.29226       0.286187
12 vs111 2599 10     6       7354.4 3652184965       1.48097       1.471774
12 vs111 2599 10     7       3142.0 3652184965       0.27031       0.268629
vs222               11       2860.4   11915003      68.66842       0.222636
12 vs222 2599 10    16       1677.1   11037709      25.48250       0.076536
 ++++                 ++++
 Shown are selected over  15  PT  with var> 0.01 % total
```

The first principal tensor captures most of the variability, 97.6%, nearly as much as the decomposition up to 3 main principal tensors and 3 for each associated, i.e., at each second level analysis (a PCA). Notice that the listing should be 30 lines long, as for each main principal tensor, 9 associated principal tensors are requested (`nbPT2 = 3`), but redundant tensors are removed automatically and out of the 21 potential principal tensors a selection has been performed here: `Global Pct` $> 0.01\%$. The listing `summary()` mentions "`...over 15 PT`" as in the call function, the parameter `minpct = 0.1` forces the algorithm to stop a $k \geq 3$-level (no sub-level analysis), if this percentage of variability is not met: it was the case here for `vs333`. The full description of the ouput `summary()` is explained in the Section 4.4 where the listing ouput provides a more complete form.

This first PTAk analysis is not very useful as the variations and range of values can be very different from one climatic variable to another. So the main variations captured by the principal tensors will be towards this differentiation without necessarily expressing the interactions between the variables and them with the spatio-temporal domain which may only be detected in some principal tensors (main or associated) with comparatively small singular values. As for PCA, centering and scaling the variables, preprocessing transformation may be crucial as part of the modelling and analysis process.

4.3. Array data and preprocessing

In the ecoclimatic data example the variables of interest are in mode 3, the climatic indicators, as the other tow modes are their support, the spatial-locations and the months. How does one center and scale the 10 indicators? It depends on the variability of the data one put the focus on, so this has to be considered as a part of the model. Here we are focusing on the

spatio-temporal dynamics of the indicators, so we are looking for spatial patterns and temporal patterns of the correlations of the variables. For a given spatial-location or spatial trend one would like to detect the mean temporal patterns of evolution or seasonality, therefore it is not desirable to center and scale the indicators for each month over the spatial-locations. It is also desirable to detect spatial mean patterns for a given month or temporal trend. Therefore, centering and scaling the indicators along the whole spatio-temporal observations seems appropriate. This would be the transformation to do, to perform a PCA on the indicators with spatial observations repeated over the 12 months.

Another interesting preprocessing would have been to perform a bi-centering along spatial-location and month modes for each indicator in order to emphasize only on interactions but not on marginal effects.

Performing centering and scaling can be done with the function `Multcent()` which proposes a centering and/or a scaling along the `by` mode(s) combination, before and/or after, `xxxBA` some possible "multi-centering" along each `bi` combined with `by`. For example a bi-centering on a three-way table corresponding to an ANOVA way of removing each of the two first factor effects and the mean effect for each level of the third factor can be done with:

```
R> Zone3w.bi <- Multcent(dat = Zone3w, bi = c(1, 2), by = 3, centre = mean,
+    centrebyBA = c(FALSE, FALSE), scalebyBA = c(FALSE, FALSE))
```

More advanced centering and scaling can be used iteratively with `IterMV()` as each transformation may destroy the other one, but one could possibly reach convergence in this pre-modeling step. For example removing a smooth trend of the months and scaling spatially the results would be:

```
R> Zone3w.DS <- IterMV(n = 10, dat = Zone3w, Mm = c(1, 3), Vm = c(2, 3),
+    usetren = TRUE, tren = function(x) smooth.spline(as.vector(x),
+    df = 5)$y, rsd = TRUE)
```

Simple centering and scaling as mentioned before has been performed for the ecoclimatic data for Tunisia. This allows us to extract spatio-temporal trends and spatio-temporal interactions with the indicator mode.

```
R> Zone3w <- Multcent(dat = Zone3w, bi = NULL, by = 3, centre = mean,
+    centrebyBA = c(TRUE, FALSE), scalebyBA = c(TRUE, FALSE))
R> Zone3w.PTA3-modes <- PTA3-modes(Zone3w, nbPT = 3, nbPT2 = 3)

---Final iteration---  37
--Singular Value--  362.1039  -- Local Percent --  42.04291 %
---Final iteration---  25
--Singular Value--  276.2334  -- Local Percent --  62.08935 %
---Final iteration---  56
--Singular Value--  28.11064  -- Local Percent --  28.82325 %
-----Execution Time----- 9.57

R> summary(Zone3w.PTA3-modes, testvar = 0.01)
```

```
  ++++ PTA3-modes ++++
               data =   Zone3w   2599 12 10
                ------Percent Rebuilt---- 97.82399 %
                ------Percent Rebuilt from Selected ---- 97.82059 %
                 -no- --Sing Val--   --ssX-- --local Pct-- --Global Pct--
vs111               1     362.1039 311870.00      42.04291      42.042911
2599 vs111 12 10    3      59.7161 136876.70       2.60528       1.143430
2599 vs111 12 10    4      35.1733 136876.70       0.90385       0.396692
12 vs111 2599 10    6     155.7611 156121.23      15.54019       7.779373
12 vs111 2599 10    7      17.9319 156121.23       0.20596       0.103104
10 vs111 2599 12    9     162.1045 158215.21      16.60893       8.425900
10 vs111 2599 12   10      20.1820 158215.21       0.25744       0.130603
vs222              11     276.2334 122895.32      62.08935      24.466893
2599 vs222 12 10   13      28.2110  77264.61       1.03005       0.255191
2599 vs222 12 10   14      10.7468  77264.61       0.14948       0.037033
12 vs222 2599 10   16     169.1854 118622.16      24.13016       9.178092
12 vs222 2599 10   17      99.8820 118622.16       8.41025       3.198903
10 vs222 2599 12   19      15.6388  76876.78       0.31813       0.078421
10 vs222 2599 12   20      12.4100  76876.78       0.20033       0.049383
vs333              21      28.1106   2741.56      28.82325       0.253377
2599 vs333 12 10   23       6.5390    853.89       5.00744       0.013710
12 vs333 2599 10   26       8.7500    958.79       7.98538       0.024550
12 vs333 2599 10   27       5.7349    958.79       3.43029       0.010546
10 vs333 2599 12   29      21.2609   1792.86      25.21260       0.144941
10 vs333 2599 12   30      16.5225   1792.86      15.22673       0.087534
 ++++                ++++
 Shown are selected over  21  PT  with var > 0.01 % total
```

Some other preprocessing transformations or pre-model examples can be seen with the EEG data in Leibovici (2000), exploiting the ANOVA interpretation of the factors involved. The EEG data consisted of a repeated cross-over design on 12 subjects with 3 *verum* doses ($10mg$, $30mg$ and $90mg$) and 1 placebo. The EEG bands quantification was available spatially at 28 electrodes with repetitions over 10 times of measurements (before drug administration and every hour or so after until 6 hours post-dosing). For this dataset the preprocessing aim was pretty much towards minimizing *subject* variability. Part of Figure 4 in Leibovici (2000) is reproduced here in Figure 3, showing a comparison of the first principal tensors obtained from PTA4-modes of the $subject \times dose \times electrode \times time \times EEG - band$ with different centering and scaling. So the best result (c) was obtained after scaling each subject and then removing the *subject* effect and all two-way interactions with the *subject* factor. Would it have been better to remove subject 11 from the analysis?

4.4. Summary method on "PTAk" objects

Figure 2 and the previous output are examples of output from the `summary()` method of "`PTAk`" objects generated by calling `PTAk()` and the other methods in the R package. The identifier for singular values and principal tensors corresponds to the column `-no-` which is the order number in the processing history. The first column of the output table from `summary()`

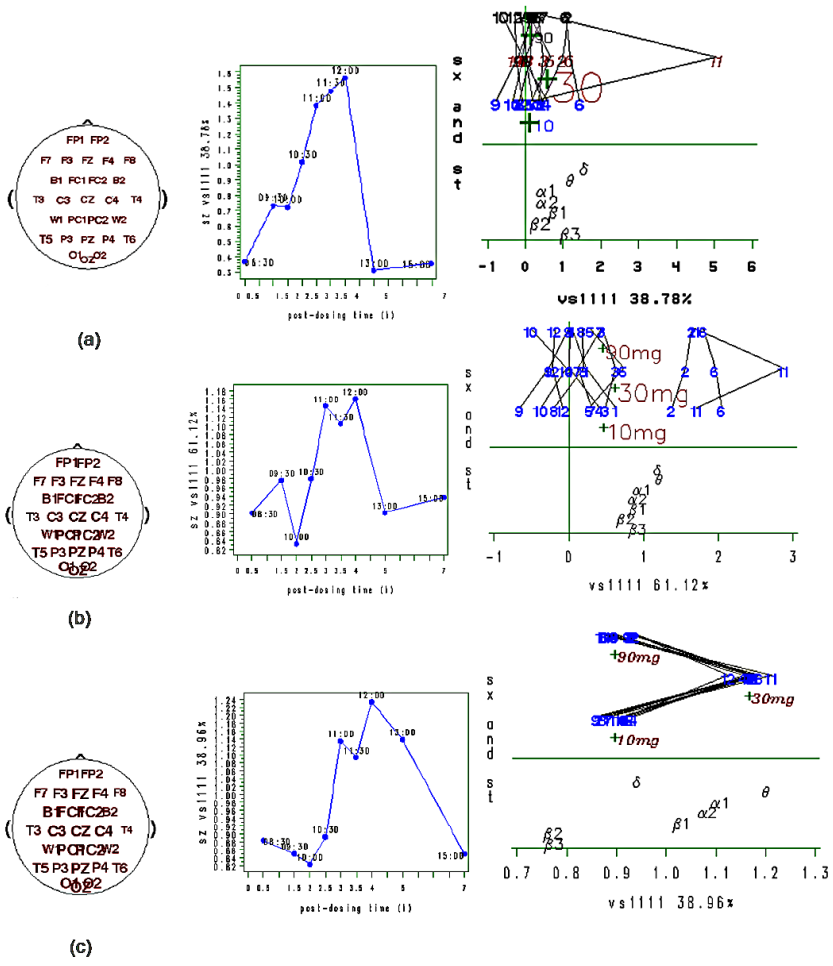

Figure 3: 1st principal tensor from PTA4-modes of the EEG data *versus* baseline *versus* placebo with different centering and scaling: (a) raw data, (b) subject scaled, (c) ANOVA residuals from *subject factor* (main effect and other two-ways interactions with it) —the subject— doses and bands plots are artificially scattered vertically for better reading of the dispersion on the horizontal axe (means by dose have been overlaid).

qualifies the identifier with a leading name. For main principal tensors the name is unique, like `vs111` or `vs222` where the number corresponds to an order from the top level hierarchy (the repetition of the number is to emphasize the level of the hierarchy corresponding to the order of the current tensor analyzed). The others names are qualifying associated principal tensors, expressing the dimension of the mode from which the association is made (contraction by the corresponding component vector). In 3-modes analysis, they also show the last two dimensions (on which a PCA is done).

For PTA4-modes the schema starts with `vs1111`, the second level corresponds to 3-modes analyses with names such as `12-vs222` where here the component contracting the data is explicit for the dimension, 12, but implicit for the principal tensor it comes from, as here `vs222` expresses the current 3-modes analysis. The third level brings in names such as `12-300 vs222 10 7` with the same meaning as in 3-modes analysis: associated with the PCA on the table 10×7. This schema is then similar for all other higher modes. Notice for the 4-

modes analysis there is no `12-vs111` as in fact this `SINGVA()` optimization is redundant with the `vs1111` solution. In the same manner, the first principal component for every 2-modes analysis is redundant with the solution just optimized within the 3-modes analysis. This comes from the fact that in order to really take the benefit of the recursivity it is easier in the implementation of Equation 8 to perform the PTA$(k-1)$-modes analysis just onto the contracted tensor and not onto the projection of it on the tensorial orthogonal of the rest of the principal tensor. Computationally it is then easier to let the recursive algorithm perform all the solutions associated and discarding the redundant ones. That is why there are gaps in the number for the column `-no-`, e.g., after the principal tensor `-no- 1` there will never be a principal tensor `-no- 2`. The generic form of the `PTAk()` algorithm which is implemented in the R package is then:

$$\begin{aligned} \mathrm{PTA}k(X) &= \sigma_1(\psi_1^1 \otimes \psi_1^2 \otimes ... \otimes \psi_1^k) \\ &+ \psi_1^1 \otimes_1 \mathrm{PTA(k-1)}^*(X..\psi_1^1) \\ &+ \psi_1^2 \otimes_2 \mathrm{PTA(k-1)}^*(X..\psi_1^2) \\ &+ ... \\ &+ \psi_1^k \otimes_k \mathrm{PTA(k-1)}^*(X..\psi_1^k) \\ &+ \mathrm{PTA}k(P(\psi_1^{1\perp} \otimes \psi_1^{2\perp} \otimes ... \otimes \psi_1^{k\perp})X) \end{aligned} \tag{11}$$

in where * means that the "top" solutions of each PTA$(k-1)$-modes have to be discarded as redundant from previous optimization.

5. Plotting and interpreting

A class "`PTAk`" object is a `list` of lists. For each mode of the tensor, the list, reachable by its mode number in the dimension of the array, contains few descriptors of the mode and the components stored in a matrix `$v` where the row numbers match the order (`-no-`) given in the `summary()`, while the columns are the names list given in `$n` (taken from the `dimnames()` of the array) for the corresponding mode. The list for last mode has extra summaries describing the analysis (applicable to all modes or the whole analysis), such as the singular values stored in `$d`:

```
R> Zone3w.PTA3[[3]]$v[c(1, 9, 11), ]

           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
[1,]  0.2406405 -0.4552913 -0.4382538  0.2727417 -0.4560138 -0.4447370
[2,]  0.2406405 -0.4552913 -0.4382538  0.2727417 -0.4560138 -0.4447370
[3,] -0.2976787 -0.1066016 -0.0960886 -0.1850355 -0.1017793 -0.1103574
           [,7]         [,8]         [,9]        [,10]
[1,]  0.2501758  0.001285919 -0.003664209 -0.002239094
[2,]  0.2501758  0.001285919 -0.003664209 -0.002239094
[3,] -0.2404791 -0.297859857  0.610534012  0.560992015

R> Zone3w.PTA3[[3]]$d[c(1, 9, 11)]

[1] 362.1039 162.1045 276.2334
```

```
R> Zone3w.PTA3[[3]]$n

[1] "P" "Tave" "ETo" "PETo" "Tmax" "Tmin" "Q3" "Alt" "dM2T" "dMETo"
```

Notice that here, looking at the components for mode 3 of the tensors 1, 9, and 11, the principal tensor 9 is an associated principal tensor to the first principal tensor via the *indicator* mode, its component for this mode is therefore equal to the one from the first principal tensor.

Interpretations of the extracted features of the dataset expressed in the principal tensors can be derived from various plots of components which can be read simultaneously. For the ecoclimatic analysis one would read and interpret a map configuration (*spatial-location* component), an annual pattern (*month* component) and an axis describing associations and oppositions of the variables (*indicator* component), together expressing the monthly dynamic of the ecoclimatic characteristics. **PTAk** provides a `plot()` method for "`PTAk`" objects which basically overlays the scattering plots of components for the asked modes. For the spatial component here we used a modified version of the `plot.Map()` (given in the file "`v34i10-additions.R`"). The following `plot()` calls can be seen on Figure 4 which gathers the basic 3 plots related to the principal tensors 1 and 11: the first plot is a joint plot of their modes 2 and 3 (time and indicators), the other plots are their respective spatial modes (mode 1).

```
R> plot(Zone3w.PTA3-modes, mod = c(2, 3), nb1 = 1, nb2 = 11,
+    lengthlabels = 5, coefi = list(c(1, 1, 1), c(1, -1, -1)))
R> plot(Zone_climTUN, ol = NA, auxvar = Zone3w.PTA3-modes[[1]]$v[1, ],
+    nclass = 30, colrmp = colorRampPalette(Yl)(31), mult = 100)
R> plot(Zone_climTUN, ol = NA, auxvar = Zone3w.PTA3-modes[[1]]$v[11, ],
+    nclass = 30, colrmp = colorRampPalette(Yl)(31), mult = 100)
```

Looking closely at the given outputs, one sees that the principal tensor `-no- 11`, `vs222`, makes an opposition between "number of dry months" `dM2T`, `dMETo` (two different ways deriving this ecoclimatic indicator) with positive weightings and the other indicators with negative weights. Nonetheless the `plot()` on Figure 4 shows the opposite signs and as a matter of fact the argument `coefi`, in the call of the `plot()`, is indicating this change for the tensor `nb2 = 11` on mode 2 and 3. The reason for this ad-hoc change can be understood for example from the fact:

$$\psi_i \otimes \varphi_i \otimes \phi_i \otimes \xi_i = \psi_i \otimes (-\varphi_i) \otimes \phi_i \otimes (-\xi_i) \tag{12}$$

So as in PCA where a principal component and the corresponding variable loadings can be arbitrarily multiplied by (-1), k-modes analysis having $k > 2$ set of components will show different ways of distributing this changes. Therefore simultaneous or joint interpretation has to be cautious about this fact. Interpretation has to look at either components separately giving a within-component description (association, opposition, etc.) or all the component scores giving a whole principal tensor interpretation, but not reading only 2 out of 3 components for example.

Using the associativity of the tensor product, a theoretical example of reading associations and oppositions for different components is given in Table 1, the interpretations are all compatible, and also identical for any other tensor equivalence transformation.

In PCA, examining correlations of variables with the principal components is the traditional way of having interpretations across components. The duality in 2-modes analysis implies

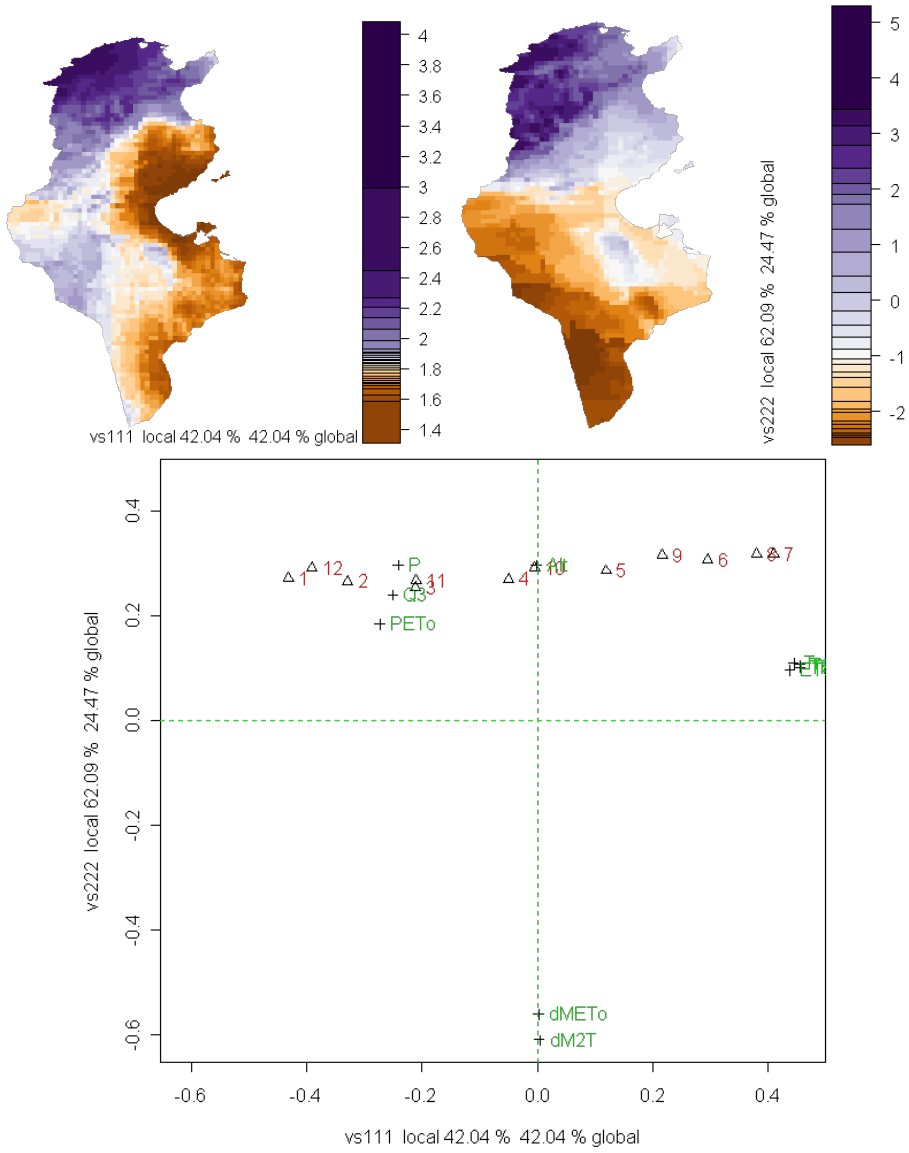

Figure 4: Principal tensors 1 and 11 from PTA3-modes of the ecoclimatic Tunisian data.

	−	+
ψ	A	B
φ	t_1	t_2
ϕ	n	m

ψ	$\otimes$	φ	$\otimes$	ϕ
ψ	$\otimes$	$(-\varphi)$	$\otimes$	$(-\phi)$
$(-\psi)$	$\otimes$	φ	$\otimes$	$(-\phi)$
$(-\psi)$	$\otimes$	$(-\varphi)$	$\otimes$	ϕ

Oppositions 2 by 2
$(Bt_2\mathrm{m})$, $(At_1\mathrm{m})$,
$(Bt_1\mathrm{n})$, $(At_2\mathrm{n})$

Table 1: Linked information: Tensor Components | Tensor Equivalences | Interpretations across components: (xyz) meaning semantic association.

full equivalence as reading interpreting directly the factor loadings. With 3-modes, theses correlations have to be between the variables represented by the vectors of the tensor unfolded into a matrix and the $\psi_i \otimes \phi_i$ (for the 2-modes variables). Similar considerations occur with $k > 3$ modes.

6. PTAk-modes with non identity metrics

As mentioned in the introduction, all multiway decomposition methods in **PTAk** allow us to use non-identity metrics for every space involved in the tensorial space, i.e., symmetric positive definite matrices used in the inner products within each space. (The canonical inner product —sum of cross-products— corresponds to the identity metric and the metric on the tensorial space the tensorial product of the individual metrics, see for example Leibovici 1993; Dauxois *et al.* 1994). Instead of feeding the methods with an "`array`" object `X`, one uses a `list` where `$data` contains the "`array`" object and `$met` is a `list` of "`matrix`" objects or "`vector`" object objects (diagonal metrics) representing the metrics associated with the inner products in each space. Algebraically within the tensor framework this has an effect on the contracted product and on the norms of the vectors, therefore on the optimization of Equations 6 and its equivalent for any k. Going back to one of the definitions of the tensorial product gives a hint for this natural extension:

$$\langle a \otimes b \otimes c, \psi \otimes \varphi \otimes \phi \rangle_{D \otimes Q \otimes M} = \langle a, \psi \rangle_D \langle b, \varphi \rangle_Q \langle c, \phi \rangle_M = {}^t a D \psi \, {}^t b Q \varphi \, {}^t c M \phi \tag{13}$$

where a and ψ belong to the metric space $\mathbb{R}^s$ with as metric the symmetric positive definite matrix D, s being the length of vector a, and similar definitions for the other elements of Equation 13. If a, b, and c cover the canonical basis elements in each space and X being expressed naturally in this basis, it is possible to grasp *via* the contraction of elementary tensors (rank one tensors), the utilisation of metrics in the contraction of any two tensors. Computationally one could also write:

$$X_{..m}(\psi \otimes \varphi \otimes \phi) = {}^t({}^t({}^t X_D D \psi)_Q Q \varphi)_M M \phi \tag{14}$$

in which the use of non-identity metrics in the contracted product has been emphasized by a subscript $..._m$; the subscript with a metric means rearranging the given tensor in a matrix where the columns lie in that metric space, i.e., the number of rows is equal to the dimension of the given metric space.

Like in PCA or SVD, where one analyses a triplet (X, Q, D), data and metrics for the two modes, it is convenient to refer to $(k+1)$-uples of objects defining the analysis, that is the tensor of order k as the data, and the k metrics associated to each mode of the tensor:

$$(X, M_1, M_2, ..., M_k) \tag{15}$$

6.1. Use of metrics in PCA

A PCA of a triplet (X, Q, D) with X a data matrix $n \times p$, Q a $p \times p$ metric on the rows (or in the column-space) and similarly D a $n \times n$ metric on the columns (or in the row-space), is generalizing a standard PCA by diagonalizing with Q-normed vectors the matrix ${}^t X D X Q$ equivalent, to the covariance matrix if X is column-centered, $D = 1/n \mathrm{Id}_n$ and $Q = \mathrm{Id}_p$, or to the correlation matrix if instead of the identity metric $Q = \mathrm{diag}(1/\mathrm{var}_1 \ldots 1/\mathrm{var}_p)$ (see also Dray and Dufour 2007, for more details).

6.2. Choice of metrics for spatial data

A classical use of metrics is in discriminant analysis when a known group structure is part of the design experiment, and either assessing or minimizing the impact of this structure on

the variability is the goal of the analysis. For example it would be possible to a perform (i) a PTAIVk-modes to assess the known structure and (ii) an orthogonal-PTAIVk-modes (that is a residual analysis) to minimize the structure in the analysis, where using as metric the inverse of within-group variations would improve both analyses. When the structure is unknown, as is often the case, the goal of the analysis becomes to look for what is structuring the data. As mentioned by Caussinus and Ruiz (1990) a good strategy to reveal dense groups with generalized PCA would be to reveal outliers first using the metric W_o^{-1} given in Equation 17 with $0.05 \leq \beta \leq 0.03$ and remove them before using the metric W_l^{-1}, Equation 16, using a $\beta \geq 1$. W_l is a robust estimate of the within covariance of the unknown structure given by the function `CauRuimet`:

$$W_l = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} G_{ij} ker(d^2_{S^-}(Z_i, Z_j))(Z_i - Z_j)\ {}^t(Z_i - Z_j)}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} G_{ij} ker(d^2_{S^-}(Z_i, Z_j))} \tag{16}$$

where Z is a data matrix, $d^2_{S^-}(.,.)$ is the squared Euclidean distance with S^- the inverse of a robust sample covariance and G is a graph structure expressing some known proximity (when no knowledge of proximity is chosen $G_{ij} = 1$ and one just has to put `mo = 1` in the function). *ker* is a positive decreasing function which would provide a kernel function for the different weighting: by default $e^{-\beta u}$ with choices on β. W_l corresponds to the definition of *local variance* (Lebart 1969; Caussinus and Ruiz 1990; Faraj 1994).

$$W_o = \frac{\sum_{i=1}^{n} ker(d^2_{S^-}(Z_i, \tilde{Z}))(Z_i - \tilde{Z})\ {}^t(Z_i - \tilde{Z})}{\sum_{i=1}^{n} ker(d^2_{S^-}(Z_i, \tilde{Z}))} \tag{17}$$

To complete the different metrics or semi-metrics associated to known or unknown structure variances, the function can give also something analog to a *global variance*:

$$W_g = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} G_{ij} ker(d^2_{S^-}(Z_i, Z_j))(Z_i - \tilde{Z})\ {}^t(Z_j - \tilde{Z})}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} G_{ij} ker(d^2_{S^-}(Z_i, Z_j))} \tag{18}$$

where $\tilde{Z}$ is the *location* vector of the multivariate distribution, i.e., a robust estimate of the mean. In the case of some known structure the semi-metric $W_l^{g\alpha-} = W_g^{1/\alpha} W_l^{-1} W g^{1/\alpha}$ may be used to extract robust sub-structure, in the sense that the analysis will tend to minimize the local variance and be consistent at $1/\alpha$ with the global variance, with convergence towards W_l^{-1} as α increases.

So far, using these metrics is not specific to spatial data as even the graph structure can reflect any proximity in attribute space, but it is also reminiscent of neighborhood graphs as in Lebart (1969); Faraj (1994). In a spatial context, $W_l^{g\alpha-}$ is particularly interesting when looking for finer-scale structures compatible with some coarser scale relationships represented by the G_{ij} (G taken into account in W_g but in W_l computations).

With the Tunisian climatic data example, we used $W_l y^{-1}$ as metric in the indicators space computed on the yearly average across spatial-location but also the same metric computed on the concatenated array across spatial-location and month mode (noted below W_l^{-1}).

```
R> Zvm <- CONTRACTION(Zone3w, rep(1/12, 12), Xwiz = 2, zwiX = 1)
R> Wly <- CauRuimet(Zvm, ker = 2, m0 = 1, withingroup = TRUE,
+    loc = substitute(apply(Z, 2, mean, trim = 0.1)), matrixmethod = TRUE)
```

```
R> Wlv <-  Powmat(Wly, -1)
R> Zone3w <- list("data" = Zone3w, "met" = list(1, 1, Wlv))
R> Zone3w.PTA3-modesm3 <- PTA3-modes(Zone3w, nbPT = 3, nbPT2 = 3,
+    modesnam = c("carte", "mois", "var"),
+    addedcomment = "PTA3-modes metric Wly var yearly")
R> summary(Zone3w.PTA3-modesm3, testvar = 1e-2)

++++ PTA3-modes ++++
    PTA3-modes metric Wly var yearly           data =   Zone3w   2599 12 10
                ------Percent Rebuilt---- 96.56307 %
                ------Percent Rebuilt from Selected ---- 96.46785 %
                 -no- --Sing Val--  --ssX-- --local Pct-- --Global Pct--
vs111               1     801.918 1112488.0     57.8049      57.80492+
2599 vs111 12 10    3     308.729  750707.7     12.6965       8.56761+
2599 vs111 12 10    4     104.273  750707.7      1.4483       0.97734
12 vs111 2599 10    6     404.862  821714.2     19.9477      14.73391+
12 vs111 2599 10    7     101.507  821714.2      1.2539       0.92618
10 vs111 2599 12    9     100.671  658692.8      1.5386       0.91099
10 vs111 2599 12   10      51.986  658692.8      0.4103       0.24293
vs222              11     287.342  167518.8     49.2871       7.42167-
2599 vs222 12 10   13     138.938  102851.1     18.7685       1.73518
...
```

The summary description of the analysis shows a "concentration" of explained variability onto the first principal tensors (marked with a +) with a decrease for example of `vs222`. Nonetheless using W_{ly}^{-1}, little differences were seen for the spatial components. When using the W_l^{-1} metric instead, taking into account *month* differences, results are completely redistributed in the principal tensors as for example the principal tensor `vs222` of the analysis without metric now becomes the main features extracted by `vs111`, see and compare Figures 5 and 4.

```
R> Zv <- matrix(as.vector(aperm(Zone3w$data, c(1, 2, 3))), c(2599*12, 10))
R> Wlv <- Powmat(CauRuimet(Zv, ker = 1, m0 = 1, withingroup = TRUE,
+    loc = substitute(apply(Z, 2, mean, trim = 0.1)),
+    matrixmethod = TRUE), -1)
R> Zone3w.PTA3-modesm3all <- PTA3-modes(list("data" = Zone3w$data,
+    "met" = list(1, 1, Wlv)),
+    nbPT = 3, nbPT2 = 3, modesnam = c("carte", "mois", "var"),
+    addedcomment = "PTA3-modes metric Wlv var")
R> summary(Zone3w.PTA3-modesm3all, testvar = 1e-2)

++++ PTA3-modes ++++
    PTA3-modes metric Wlv var               data =   Zone3w   2599 12 10
                ------Percent Rebuilt---- 82.45044 %
                ------Percent Rebuilt from Selected ---- 79.82162 %
                 -no- --Sing Val-- --ssX-- --local Pct-- --Global Pct--
vs111               1       411.92  493139        34.408        34.4080
2599 vs111 12 10    3       210.18  233375        18.929         8.9582
```

```
2599 vs111 12 10    4       127.69  233375        6.986         3.3061
12 vs111 2599 10    6       157.71  247705       10.041         5.0435
12 vs111 2599 10    7       151.75  247705        9.297         4.6699
10 vs111 2599 12    9       107.83  184021        6.319         2.3580
vs222              11       191.73  167397       21.959         7.4540
12 vs222 2599 10   16       156.67  100306       24.470         4.9772
12 vs222 2599 10   17       129.00  100306       16.591         3.3746
vs333              21       121.81   60459       24.543         3.0089
...
```

Figure 5 is realised from the three different plots given below, the scatter plot being done using the `plot.PTAk()` method, and spatial components plotted using `plot.Map()` from the **maptools** R package:

```
R> plot(Zone3w.PTA3-modesm3all, mod = c(2, 3), nb1 = 1, nb2 = 11,
+    lengthlabels = 5)
R> plot(Zone_climTUN, ol = NA, auxvar = Zone3w.PTA3-modesm3all[[1]]$v[1, ],
+    nclass = 30, colrmp = colorRampPalette(Yl)(31), mult = 100)
R> plot(Zone_climTUN, ol = NA, auxvar = Zone3w.PTA3-modesm3all[[1]]$v[11, ],
+    nclass = 30, colrmp = colorRampPalette(Yl)(31), mult = 100)
```

Now using this metric combining smoothed global variance constraint with inverse within local structure, given by $W_l^{g\alpha-}$, the results demonstrate a more interesting spatial pattern than in previous analysis, Figure 6. Note as in the analysis with W_l^{-1}, `vs222` of the previous analysis (without metric) now becomes the main features extracted by `vs111`, elsewhere different in a compatible description of the spatio-temporal patterns as expected by the link between the metrics.

```
R> DD <-  round(100*t(Zone3w.PTA3-modes[[1]]$v[c(1, 6, 9, 11, 16, 17), ]))
R> ddd2 <- exp(as.matrix(round(dist(DD))))
R> ddd2 <- 1/(1 + (ddd2-min(ddd2)))
R> ddd2[ddd2 <= quantile(ddd2, probs = 0.25)] <- 0
R> Zv1 <- CONTRACTION(Zone3w$data, rep(1, 12), zwiX = 1, Xwiz = 2)
R> Wov <- CauRuimet(Zv1, ker = 2, m0 = ddd2, withingroup = FALSE,
+    loc = substitute(apply(Z, 2, mean, trim = 0.1)), matrixmethod = FALSE)
R> Wglv <- Powmat(Wov, 1/9) %*% Wlv %*% Powmat(Wov, 1/9)
R> Zone3wgl.PTA3-modes <- PTA3-modes(list("data" = Zone3w$data,
+    "met" = list(1, 1, Wglv)), nbPT = 3, nbPT2 = 3,
+    modesnam = c("carte", "mois","var"),
+    addedcomment = "PTA3-modes metric Wglv")
R> summary(Zone3wgl.PTA3-modes, testvar = 1e-2)

++++ PTA3-modes ++++
data =   list(data  =  Zone3w$data, met  =  list(1, 1, Wglv))   2599 12 10
    PTA3-modes metric Wglv
                ------Percent Rebuilt---- 87.57398 %
                ------Percent Rebuilt from Selected ---- 87.56613 %
```

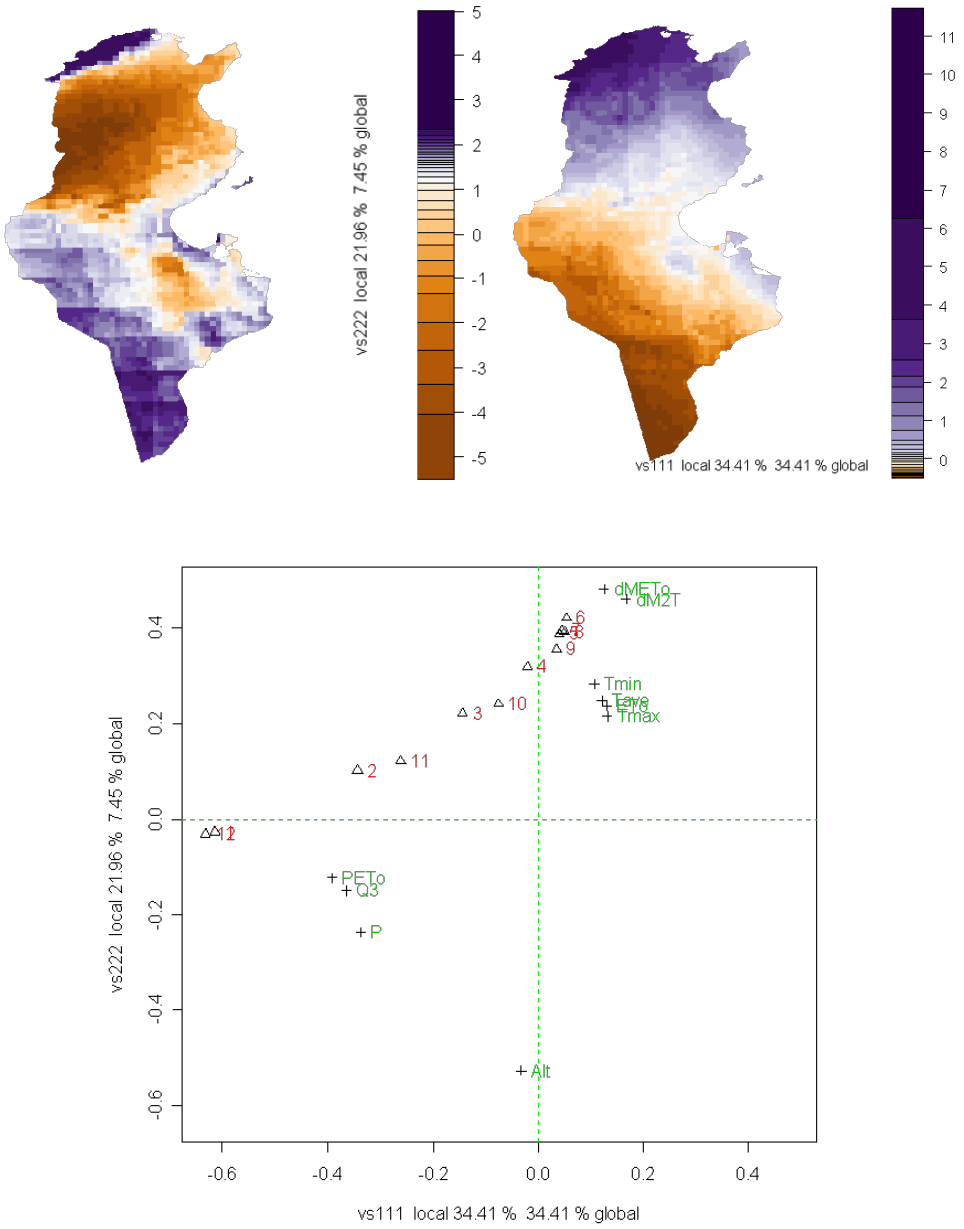

Figure 5: Principal tensor 1 and 11 (`vs111` and `vs222`)) using W_l^{-1} metric for the indicator mode.

	-no-	--Sing Val--	--ssX--	--local Pct--	--Global Pct--	
vs111	1	1233.432	3609793	42.14520	42.145198	a
2599 vs111 12 10	3	590.828	1913730	18.24071	9.670305	a
2599 vs111 12 10	4	194.028	1913730	1.96720	1.042910	a
12 vs111 2599 10	6	584.066	2348657	14.52462	9.450226	b
12 vs111 2599 10	7	480.787	2348657	9.84207	6.403597	c
10 vs111 2599 12	9	429.092	1731332	10.63456	5.100556	d
10 vs111 2599 12	10	97.945	1731332	0.55409	0.265754	

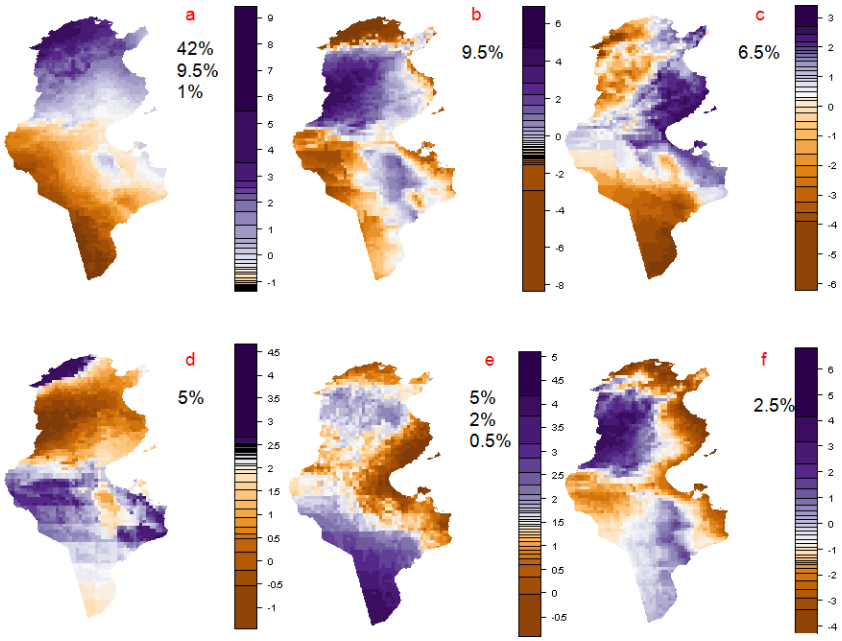

Figure 6: Main spatial components from principal tensors of Tunisian climatic data with metric $W_l^{g\alpha-}$: letters correspond to the output listing with (`metric Wglv`).

```
vs222                 11      412.755  658783      25.86084     4.719574 e
2599 vs222 12 10      13      265.521  267878      26.31853     1.953064 e
2599 vs222 12 10      14      150.321  267878       8.43528     0.625971 e
12 vs222 2599 10      16      295.822  370631      23.61118     2.424251 f
...
```

The coarse structure, used here for the $W_l^{g\alpha-}$ metric, is a non-linear transformation of distances (see `ddd2` in the above code chunk) based on components extracted from the PTA3-modes without metric (from which global spatial effects were expected to emerge). This particular example, with this choice of coarse structure, could be seen as as an iterative process, assimilating non-linear estimation of smoothed spatial pattern, from and within constrained spatio-temporal decomposition of multivariate dynamics.

In a similar context, geographically weighted discriminant analysis (Brunsdon *et al.* 2007), takes a similar kernel approach with geographic distances, to take into account spatial proximity in the estimation of variance-covariance matrix playing also the role of a metric. The approach of Borcard and Legendre (2002), claiming to account for a range of scales by using an eigenvalue decomposition of a truncated matrix of distances between sampling sites, is also applicable here to build a metric depending on spatial proximity.

7. Correspondence Analysis on k-way tables

The tensorial framework developed previously, using different metrics and particular datasets to perform a PTAk-modes, extends the framework of multidimensional analysis using PCA as a generic method (Escoufier 1987; Dray and Dufour 2007). A method of particular interest is

correspondence analysis; a generalization to multiple contingency table (Leibovici 1993, 2000), has been used to analyze spatial patterns of attributes of categorical variables (Leibovici *et al.* 2008).

7.1. 2-way correspondence analysis

Correspondence analysis (FCA) of a two-way contingency table with cells n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ can be described as follows. The usual notations are:

$$n_{i.} = \sum_j n_{ij}, \quad n_{.j} = \sum_i n_{ij}, \quad n_{..} = N = \sum_{ij} n_{ij}$$

and then the observed proportions are defined as $p_{ij} = n_{ij}/N$. Diagonal metrics containing vector margins $P_{I.} = {}^t(\cdots p_{i.} \cdots)$ and $P_{.J}$ used thereafter, are noted D_I and D_J. Correspondence analysis provides a decomposition of the measure of lack of independence between the two categorical variables indexed respectively by i and j in performing the principal component analysis, PCA or generalized PCA, of the following triplet (Escoufier 1987):

$$(D_I^{-1} P D_J^{-1} \; D_I, \; D_J) \tag{19}$$

The measure of lack of independence is linked to the analysis by:

$$1 + \frac{\chi^2}{N} = 1 + \sum_{ij} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_{ij} p_{i.}p_{.j} (\frac{p_{ij}}{p_{i.}p_{.j}})^2 = \sum_s \sigma_s^2 \tag{20}$$

where the σ_s are the singular values from the PCA of the triplet given which has $\sigma_0 = 1$ with components equal to unit vectors in their respective spaces.

7.2. k-way correspondence analysis

As FCA is a particular PCA, we proposed a generalization of correspondence analysis to k-way tables, FCAk-modes, as particular PTAk-modes (Leibovici 1993). Presenting here only the case $k = 3$, and using similar notations as in the previous section, the three-way table $I \times J \times K$, is analyzed by the PTA3-modes of the quadruple:

$$((D_I^{-1} \otimes D_J^{-1} \otimes D_K^{-1})..P, \quad D_I, \quad D_J, \quad D_K) \tag{21}$$

If one notes: $\Pi_{ijk} = \Pi_{.jk} + \Pi_{i.k} + \Pi_{ij.} + \Delta_{ijk}$ for

$$(\frac{p_{ijk} - p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}) = (\frac{p_{.jk} - p_{.j.}p_{..k}}{p_{.j.}p_{..k}}) + (\frac{p_{i.k} - p_{i..}p_{..k}}{p_{i..}p_{..k}}) + (\frac{p_{ij.} - p_{i..}p_{.j.}}{p_{i..}p_{.j.}}) + (\frac{p_{ijk} - \delta_{ijk}}{p_{i..}p_{.j.}p_{..k}}),$$

where $\delta_{ijk} = p_{ij.}p_{..k} + p_{i.k}p_{.j.} + p_{.jk}p_{i..} - 2p_{i..}p_{.j.}p_{..k}$, one has the following property:

$$\|\Pi_{ijk}\|^2 = \|\Pi_{.jk}\|^2 + \|\Pi_{i.k}\|^2 + \|\Pi_{ij.}\|^2 + \|\Delta_{ijk}\|^2 \tag{22}$$

where $\|.\|$ is the norm on the tensor space, i.e., using the metric $D_I \otimes D_J \otimes D_K$. This result dates from Lancaster in 1951, reported more recently in Carlier and Kroonenberg (1996) where another generalization of correspondence analysis is derived. Equation 22 means that deviation from three-way independence can be orthogonally decomposed into deviations from

independence for the two-way margins of the three-way table, and a three-way interaction term. Each two-way margin deviation from independence is reminiscent of (simple) correspondence analysis:

$$\begin{aligned}\frac{\chi^2}{N} &= \sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk}-p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}\right)^2 \\ &= \sum_{jk} p_{.j.}p_{..k}\left(\frac{p_{.jk}-p_{.j.}p_{..k}}{p_{.j.}p_{..k}}\right)^2 + \sum_{ik} p_{i..}p_{..k}\left(\frac{p_{i.k}-p_{i..}p_{..k}}{p_{i..}p_{..k}}\right)^2 + \sum_{ij} p_{i..}p_{.j.}\left(\frac{p_{ij.}-p_{i..}p_{.j.}}{p_{i..}p_{.j.}}\right)^2 \\ &+ \sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk}-\delta_{ijk}}{p_{i..}p_{.j.}p_{..k}}\right)^2.\end{aligned} \tag{23}$$

showing that, the PTA3-modes (Equation 21) simply retrieves the two-way lacks of marginal independence, and this, in a natural way according to the algorithm schema 11. The inertia or sum of squares is:

$$\sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk}}{p_{i..}p_{.j.}p_{..k}}\right)^2 = \sum_{s=0}^{r}\sigma_s = 1 + \sum_{s=1}^{r}\sigma_s = 1 + \frac{\chi^2}{N} \tag{24}$$

where the first ($s = 0$) principal tensor being $\mathbb{1}_I \otimes \mathbb{1}_J \otimes \mathbb{1}_K$ with $\sigma_0 = 1$, its *associated principal tensors* relate to two-way margins decompositions, i.e., each term of the second row of the Equation 23. The use of the `FCAk()` function, implementing this particular `PTAk()`, is described in the next section for a particular spatial analysis.

7.3. Analyzing multiple collocations

Looking at collocations of attributes with $v \geq 1$ categorical variables issued from spatial-location processes (such as point processes) leads to analyse multiway tables of spatial cooccurrences (Leibovici *et al.* 2008). Order-two cooccurrences have now a long history within spatial pattern analysis (Ripley 1981; Diggle 2003) and can be used within R for example with the R package **spatstat** (Baddeley and Turner 2005).

They can also be analyzed using correspondence analysis whilst higher order cooccurrences revealing spatial patterns can be extracted via FCAk-modes performed within **PTAk** using the method `FCAk()`(respectively the CAOO and CAkOO statistical methods defined in Leibovici *et al.* 2008). This is illustrated here using the dataset `Lansing` from **spatstat**. This dataset consist of an ecological study where categories of trees and their positions in the studied area are recorded. Analyzing the pattern of categories will help the ecologist to study tree associations in the development of the forest. A collocation of order 3 at distance 0.1 unit (square window of 1 unit × 1 unit) on the single point process marked with a categorical variable describing the tree species was used. The collocation function `cOO3d1S()` is given in the file "`v34i10-additions.R`". Other functions for other applications of the generic PTAk method can be found at Leibovici (2004).

The function `cOO3d1S()` in the following code computes the collocation of order 3 for one marked point process, keeping the "locations of collocations" in one entry of the output array (instead of marginalizing for each category). Each cell $n_{s_i,j,k}$ contains the number of collocations, up to distance $d = 0.1$ unit, of categories i, j and k at location s_i of category i.

```
R> lansing.1 <- cOO3d1S(lansing, 0.1)
R> lansing.1.FCAk  <- FCAk(lansing.1, nbPT = 3, nbPT2 = 3, minpct = 0.01,
+    modesnam = c("2251points", "6catTree", "6catTree"),
+    addedcomment = "S I I", chi2 = FALSE, E = NULL)
R> summary(lansing.1.FCAk)

++++ FCA3-modes++++    d = 0.1 unit
  ++ collocation Table  lansing.1   2251 6 6  ++     S I I
               -----Total Percent Rebuilt---- 69.68052 %
  ++ Percent of lack of complete independence rebuilt  ++ 42.70666 %
            selected pctoafc >  0.1 %  total =   42.61869
              -no- --Sing Val--  --ssX-- --Global Pct--  --FCA--
vs111             1     1.000000 2.124032       47.08027       NA
2251 vs111 6 6    3     0.256438 1.074332        3.09603  5.85042
2251 vs111 6 6    4     0.062902 1.074332        0.18628  0.35201
6 vs111 2251 6    6     0.493220 1.429735       11.45302 21.64225
6 vs111 2251 6    7     0.250328 1.429735        2.95025  5.57495
vs222            11     0.238690 0.190229        2.68231  5.06863
6 vs222 2251 6   16     0.115381 0.092153        0.62677  1.18437
6 vs222 2251 6   17     0.103075 0.092153        0.50020  0.94521
vs333            21     0.132362 0.062104        0.82483  1.55865
6 vs333 2251 6   26     0.051125 0.024298        0.12306  0.23254
6 vs333 2251 6   27     0.048546 0.024298        0.11096  0.20967
...
```

The `FCAk()` listing results obtained by the `summary()` function on a `FCAk` class object (inheritance from "`PTAk`" class object) differs from the `PTAk()` listing multi-hierarchical tree of singular values by the extra column `-FCA-` percentage which is merely a percentage of variability without regard to `vs111` the trivial singular value 1. Associated with this trivial principal tensor we get the marginal FCA's (or generally the marginal FCA($k-1$)-modes).

```
R> plot(lansing.1.FCAk, mod = c(2, 3), nb1 = 6, nb2 = 11, lengthlabels = 5)
R> lansing$marks <- lansing.1.FCAk[[1]]$v[6, ]
R> plot(lansing[lansing$marks< = 0], cols = "blue",
+    main = "6 vs111 2251 6 FCAk 21%")
R> par(new = TRUE)
R> plot(lansing[lansing$marks>0], cols = "red", main = "")
R> lansing$marks <- lansing.1.FCAk[[1]]$v[11, ]
R> plot(lansing[lansing$marks< = 0], cols = "blue", main = "vs222 FCAk 5%")
R> par(new = TRUE)
R> plot(lansing[lansing$marks>0], cols = "red", main = "")
```

Figure 7 displays some effects associated with spatial dependencies of the categories of trees as measured by the lack of independence for the collocation counts of order 3. The principal tensors 6 and 11 are indeed showing the same thing: collocations with *Blackoak* and *Hickory* in the top left corner opposed to the collocations with *Maple* and *Miscellaneous* trees, and the reverse in the big blue squares areas. Nonetheless principal tensor 11 is a real third order

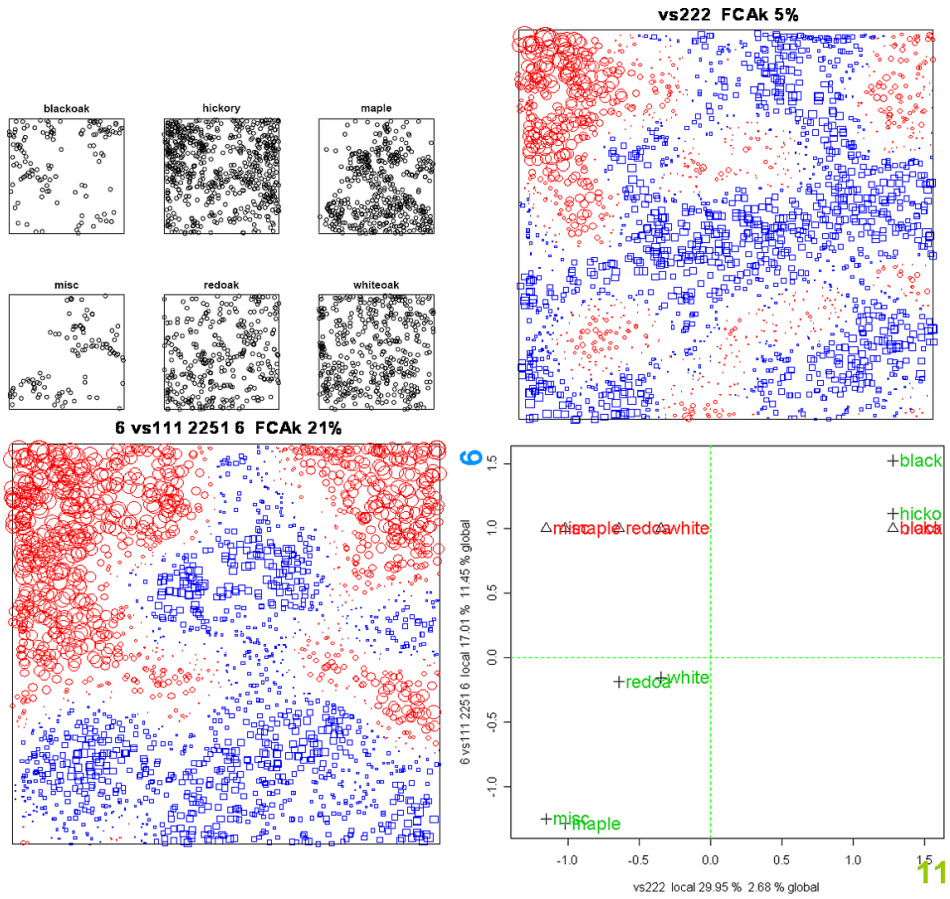

Figure 7: Main effects. Principal tensors 6 and 11 (see `summary(lansing.1.FCAk)`) from CAkOO analysis for collocation counts of order 3 of the Lansing data: blackoak, hickory, maple, misc, redoak and whiteoak trees.

discrepancy effect, and principal tensor 6 is like a second order *a posteriori* as it is associated with a marginal effect (an FCA2-modes after marginal contraction, i.e., associated with `vs111` which is all components values equal to 1 (the $s = 0$ of Equation 24).

When analyzing cooccurrences of high orders on the same categorical variable, such as the `lansing` data, one gets lots of symmetries within the tensor to be analyzed by `FCAk()`. A fully symmetrical tensor can cause problems to obtain convergence for the actual algorithm within `SINGVA()`. It is expected that the renewed interest in multiway tensor analysis, particularly for symmetric tensors (Ni and Wang 2007; Comon *et al.* 2008) will bring new algorithms for rank-one approximation (as in `SINGVA()` with the RPVSCC algorithm): see also Faber *et al.* (2003) for a discussion on recent PARAFAC algorithms and de Silva and Lim (2008) for a general discussion on lower-rank approximation in general. Other multiway methods linked to higher order cooccurrences, spatially or non-spatially, are focusing on generalizing dissimilarities or similarities for multidimensional scaling. For example Bennani's work from his thesis, published in Heiser and Bennani (1997), was looking at Euclidean approximation of 3-way dissimilarities using an approach generalizing the unfolding metric multidimensional scaling.

8. Penalized optimization

The optimization algorithm within `PTA3()` or `PTAk()`, can be seen as alternating unconstrained least squares. Utilizing metrics could be seen as introducing linear constraints or linear smoothing, which was pushed a bit further by allowing any smoothing of the components within each iteration step. This very simple constraint makes a panelization algorithm which will be equivalent to using metrics or semi-metrics if the penalizing operators are linear in a separable Hilbert space, e.g., polynomial smoothing, spline smoothing (Leibovici and El Maâche 1997; Leibovici 2008; Besse *et al.* 2005). Nonetheless even in the previous case, the structure of the algorithm is more flexible, as for a particular `mode`, the smoothing parameter is a list of smoothers which can be different along the decomposition process. However, the increased flexibility of optimization may lead to an invalid or even non-convergent `PTAk()`, but nearly orthogonal decompositions may be intersting. Work to fully describe the properties of such smoothing parameterization for the multiway methods in **PTAk** analysis is still underway by the current author.

Below is a comparison of two different PTA3 for a verbal study on 12 subjects in which brain activity is measured by fMRI during the verbal paradigm on/off showed by the square curves on Figure 8. The first analysis (top of Figure 8) is a standard PTA3-modes on the *brain* $\times$ *time* $\times$ *subject* data, and the analysis at the bottom is using panelization for time and space: `Wav2D()` a 2-D wavelet smoothing for *brain* adapted from **wavethresh** and for *time* a double kernel smoother `Susan1D()` provided in **PTAk**, which additionally to traditional kernel smoothing does preserve high peaks. The smoothing arguments in `PTAk()` or `PTA3()` are then included: `PTA3(..., smoothing = TRUE, smoo = list(Wav2D, Susan1D, NA), ...)`. Beforehand, the data was detrended, using `Detren` on the time mode and scaled for each subject using `Multcent()`. Only the first principal tensors are shown here, some more results especially combining metrics and penalization can be seen at Leibovici (2004).

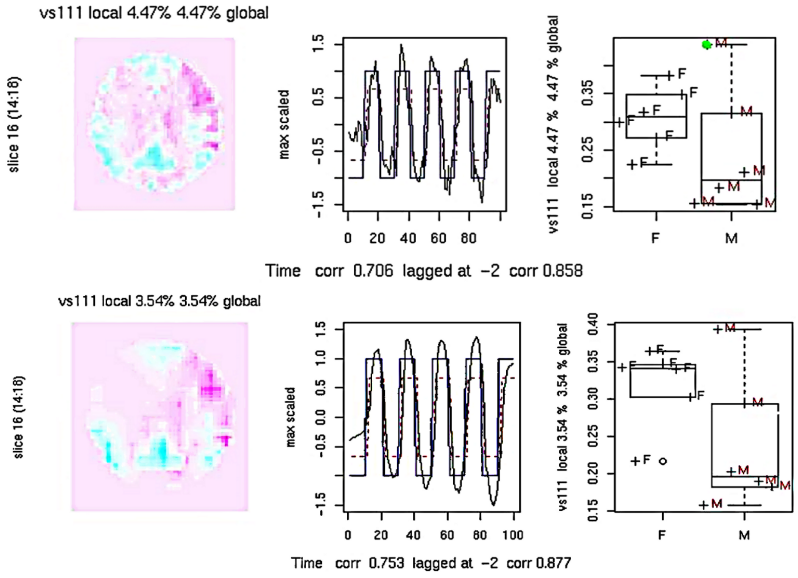

Figure 8: Verbal study data *brain* $\times$ *time* $\times$ *subject* with: canonical `PTA3-modes` (top), penalised `PTA3-modes` with smooth constraints on *brain* and *time* (bottom).

9. Conclusions and perspectives

As a possible way of extending multidimensional analysis on tables with 2 entries to multi-entries data analysis, **PTAk** allows multiway decompositions of high order interactions and can describe efficiently multiple domain interactions including the spatio-temporal domain. Using specific metrics, linked to the input data or to a covariate structure, leads also to extend well known 2-way analysis or principles such as discriminant analysis and correspondence analysis, which have been shown to be of interest for contiguity analysis, or spatial data defined by collocation events. Metrics, linear constraints and finally penalizing components during optimization are made possible by the framework developed; they are efficient ways to take into account spatio-temporal intrinsic properties within multiway analysis. Even though we presented mainly the `PTAk()` generic method, two other multiway methods, `CANDPARA()` and `PCAn()`, implemented in the R package possess these generalized features.

Non-linear adjustment is desirable in multidimensional analysis and also in multiway analysis, **PTAk** gives some answers but remains fundamentally multilinear. Non-linear objective functions used with multilinear models, can nonetheless be very powerful such as in independent component analysis (ICA) for noisy data (Leibovici and Beckmann 2001; Beckmann and Smith 2005).

Acknowledgments

Thanks go to the data providers: Dr. Jane Adcock who let me play with her verbal study experiment at fMRIB laboratory, University of Oxford (United Kingdom); Dr. Gérard Derzko from SANOFI-Recherche, Montpellier (France) for the EEG data and to Gilles Quillevere at IRD, La Maison de La Télédétection, Montpellier (France) who showed interest in using **PTAk** for climatic GIS data for circum Saharan Africa.

References

Baddeley AJ, Turner R (2005). "**spatstat**: An R Package for Analyzing Spatial Point Patterns." *Journal of Statistical Software*, **12**(6), 1–42. URL http://www.jstatsoft.org/v12/i06/.

Beckmann CF, Smith SM (2005). "Tensorial Extensions of Independent Component Analysis for Multi-Subject fMRI Analysis." *NeuroImage*, **25**(1), 294–311.

Besse PC, Cardot H, Faivre R, Goulard M (2005). "Statistical Modelling of Functional Data: Research Articles." *Applied Stochastic Models in Business and Industry*, **21**(2), 165–173.

Borcard D, Legendre P (2002). "All-Scale Spatial Analysis of Ecological Data by Means of Principal Coordinates of Neighbour Matrices." *Ecological Modelling*, **153**, 51–68.

Borg I, Groenen P (2005). *Modern Multidimensional Scaling: Theory and Applications*. 2nd edition. Springer-Verlag, New York, NY.

Brunsdon C, Fotheringham S, Charlton M (2007). "Geographically Weighted Discriminant Analysis." *Geographical Analysis*, **39**(4), 376–396.

Cailliez F, Pagès JP (1976). *Introduction à l'Analyse Des Données*. Société de Mathématiques Appliquées et de Sciences Humaines (SMASH), Paris, France.

Carlier A, Kroonenberg PM (1996). "Decompositions and Biplots in Three-Way Correspondence Analysis." *Psychometrika*, **61**(2), 355–373.

Carroll JD, Chang JJ (1970). "Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of 'Eckart-Young' Decomposition." *Psychometrika*, **35**, 283–319.

Caussinus H, Ruiz A (1990). "Interesting Projections of Multidimensional Data by Means of Generalized Principal Components Analysis." In *COMPSTAT90*, pp. 121–126. Physica-Verlag, Heidelberg, Germany.

Chessel D, Dufour AB, Dray S, Lobry JR, Ollier S, Pavoine S, Thioulouse J (2007). ***ade4**: Exploratory and Euclidean Methods in Environmental Sciences*. R package version 1.4-5, URL http://CRAN.R-project.org/package=ade4.

Comon P, Golub G, Lim LH, Mourrain B (2008). "Symmetric Tensors and Symmetric Tensor Rank." *SIAM Journal on Matrix Analysis and Applications*, **30**(3), 1254–1279.

Dauxois J, Romain Y, Viguier-Pla S (1994). "Tensor Products and Statistics." *Linear Algebra and Its Applications*, **210**, 59–88.

De Leeuw J, Mair P (2009). "Multidimensional Scaling Using Majorization: **SMACOF** in R." *Journal of Statistical Software*, **31**(3), 1–30. URL http://www.jstatsoft.org/v31/i03/.

de Silva V, Lim LH (2008). "Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem." *SIAM Journal on Matrix Analysis and Applications*, **30**(3), 1084–1127.

Diggle PJ (2003). *Statistical Analysis of Spatial Point Patterns*. Hodder Arnold, London, UK.

Dray S, Dufour AB (2007). "The **ade4** Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software*, **22**(4), 1–20. URL http://www.jstatsoft.org/v22/i04/.

Escoufier Y (1987). "The Duality Diagram: A Means of Better Practical Applications." In P Legendre, L Legendre (eds.), *Developments in Numerical Ecology*, Serie G, pp. 139–156. NATO Advanced Institute, Springer-Verlag, Berlin, Germany.

Faber NM, Bro R, Hopke PK (2003). "Recent Developments in CANDECOMP/PARAFAC Algorithms: A Critical Review." *Chemometrics*, **65**(1), 119–137.

Faraj A (1994). "Interpretation Tools for Generalized Discriminant Analysis." In *New Approaches in Classification and Data Analysis*, pp. 286–291. Springer-Verlag, Heidelberg, Germany.

Franc A (1992). *Etude Algébrique des Multitableaux: Apports de l'Algèbre Tensorielle*. Ph.D. thesis, Université de Montpellier II, Montpellier, France.

Gollob HF (1968). "A Statistical Model which Combines Features of Factor Analytic and Analysis of Variance Techniques." *Psychometrika*, **33**, 73–116.

Harshman RA (1970). "Foundations of the PARAFAC Procedure: Models and Conditions for 'an Explanatory' Multi-Modal Factor Analysis." *UCLA Working Papers in Phonetics 16*, UCLA. UMI Serials in Microform, No. 10,085.

Heiser WJ, Bennani M (1997). "Triadic Distance Models: Axiomatization and Least Squares Representation." *Journal of Mathematical Psychology*, **41**(2), 189–206.

Hill MO (1973). "Reciprocal Averaging: An Eigenvector Method of Ordination." *Journal of Ecology*, **61**, 237–249.

Kaptein A, Neudecker H, Wansbeek T (1986). "An Approach to n-Mode Component Analysis." *Psychometrika*, **51**(2), 269–275.

Kolda TG (2003). "A Counterexample to the Possibility of an Extension of the Eckart-Young Low-Rank Approximation Theorem for the Orthogonal Rank Tensor Decomposition." *SIAM Journal on Matrix Analysis and Applications*, **24**(3), 762–767.

Kroonenberg PM (1983). *Three-Mode Principal Component Analysis: Theory and Applications*. DWO Press, Leiden, Netherlands.

Kroonenberg PM, De Leeuw J (1980). "Principal Component Analysis of Three-Mode Data by Means of Alternating Least Squares Algorithms." *Psychometrika*, **45**, 69–97.

Lang S (1984). *Algebra*. 2nd edition. Addison-Wesley, Reading, Massachusetts.

Le Roux B, Rouanet H (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Lebart L (1969). "Analyse Statistique de la Contiguitée." *Publication de l'Institut de Statistiques Universitaire de Paris*, **XVIII**, 81–112.

Leibovici DG (1993). *Facteurs à Mesures Répétées et Analyses Factorielles: Applications à un Suivi Épidémiologique*. Phd thesis, Université de Montpellier II, Montpellier, France.

Leibovici DG (2000). "Multiway Multidimensional Analysis for Pharmaco-EEG Studies." *Report initiated at SANOFI-RECHERCHE TR00DL2*, FMRIB Centre, University of Oxford, UK. URL http://www.fmrib.ox.ac.uk/analysis/techrep/tr00dl2/tr00dl2.pdf.

Leibovici DG (2001). ***PTAk**: Principal Tensor Analysis on k Modes*. R package version 1.1-4, URL http://CRAN.R-project.org/package=PTAk.

Leibovici DG (2004). "c3s2i: Conseils Services Statistique SIG Ingénieurie Informatique." URL http://c3s2i.free.fr/.

Leibovici DG (2008). "A Simple Penalised Algorithm for SVD and Multiway Functional Methods." *Technical report*, Centre for Geospatial Science, University of Nottingham, UK. URL http://cgs.nottingham.ac.uk/~lgzdl/.

Leibovici DG (2009). ***PTAk**: Principal Tensor Analysis on k Modes*. R package version 1.2-0, URL http://CRAN.R-project.org/package=PTAk.

Leibovici DG, Bastin L, Jackson M (2008). "Discovering Spatially Multiway Collocations." In *GISRUK08 Conference*, pp. 66–71. GIS Research UK, Manchester, UK.

Leibovici DG, Beckmann CF (2001). "An Introduction to Multiway Methods for Multi-Subject fMRI Experiments." *Technical Report TR01DL1*, FMRIB Centre, University of Oxford, UK. URL http://www.fmrib.ox.ac.uk/analysis/techrep/tr01dl1/tr01dl1.pdf.

Leibovici DG, El Maâche H (1997). "Une Décomposition en Valeurs Singulières d'un Elément d'un Produit Tensoriel de k Espaces de Hilbert Séparables." *Compte Rendus de l'Académie des Sciences I*, **325**(7), 779–782.

Leibovici DG, Quillevere G, Desconnets JC (2007). "A Method to Classify Ecoclimatic Arid and Semi-Arid Zones in Circum-Saharan Africa Using Monthly Dynamics of Multiple Indicators." *IEEE Transactions on Geoscience and Remote Sensing*, **45**(12), 4000–4007.

Leibovici DG, Sabatier R (1998). "A Singular Value Decomposition of k-Way Array for a Principal Component Analysis of Multiway Data, PTA-k." *Linear Algebra and Its Applications*, **269**, 307–329.

Ni G, Wang Y (2007). "On the Best Rank-1 Approximation to Higher-Order Symmetric Tensors." *Mathematical and Computer Modelling*, **46**, 1345–1352.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ripley BD (1981). *Spatial Statistics*. John Wiley & Sons, New York, NY.

Rougier JC (2002). ***tensor**: Tensor Product of Arrays*. R package version 1.4, URL http://CRAN.R-project.org/package=tensor.

van den Boogaart KG (2007). ***tensorA**: Advanced Tensors Arithmetic with Named Indices*. R package version 0.31, URL http://CRAN.R-project.org/package=tensorA.

Affiliation:

Didier G. Leibovici
Centre for Geospatial Science
University of Nottingham
Sir Clive Granger Building
Nottingham NG7 2RD, United Kingdom
E-mail: didier.leibovici@nottingham.ac.uk
URL: http://www.nottingham.ac.uk/cgs/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 34, Issue 10 *Submitted:* 2008-05-08
May 2010 *Accepted:* 2009-08-19