JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; **18**: 253–264
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.868

Three-way component analysis of interval-valued data

Paolo Giordani[1]* and Henk A. L. Kiers[2]

[1]Department of Statistics, Probability and Applied Statistics, University of Rome 'La Sapienza', Ple Aldo Moro 5, I-00185 Rome, Italy
[2]Heymans Institute (DPMG), University of Groningen, Grote Kruisstraat 2/1, NL-9712 TS Groningen, The Netherlands

Received 26 July 2003; Revised 25 March 2004; Accepted 18 May 2004

Vertices Principal Component Analysis (V-PCA) and Centers Principal Component Analysis (C-PCA) are variants of Principal Component Analysis (PCA) to deal with two-way interval-valued data. In this case the observation units are represented as hyperrectangles instead of points. Tucker3 and CANDECOMP/PARAFAC are component analysis techniques to analyze the underlying structure of *three-way* data sets. In the present paper, after recalling the above mentioned methods, we extend the C-PCA and V-PCA methods to deal with three-way interval-valued data by means of Tucker3 and CANDECOMP/PARAFAC and we describe how to represent the observation units in the obtained low-dimensional space. Furthermore, an application of the extended methods—called Three-way Vertices Principal Component Analysis (3V-PCA) and Three-way Centers Principal Component Analysis (3C-PCA)—to three-way interval-valued air pollution data is described.

KEYWORDS: Principal Component Analysis; Tucker3, CANDECOMP/PARAFAC; interval-valued data sets; air pollution data

1. INTRODUCTION

In real life, many phenomena cannot be explained by using single-valued variables. For instance, we can think about the daily temperatures or the daily pollution levels registered in different places or about the mineral concentrations of food items. In the above examples a researcher can be more interested in the minimum and maximum values registered than in the average ones because they offer more detailed information about the examined phenomenon taking into account the variability of the features involved. Data in which each observation is given as an interval of values (indicated by a minimum and a maximum) are called 'interval-valued data'. Note that the intervals need not pertain to the actually observed maxima and minima but could also pertain to, for instance interquartile intervals or intervals relating to the middle 90% of the scores (e.g. in the presence of outliers).

In other real world applications the available information is vague and therefore cannot be revealed exactly by single numerical data. Again it can be useful to summarize the information using interval-valued data. We can define an interval data value as the score of the generic observation unit i on the generic variable j identified by a pair of values: a lower bound and an upper bound which enclose the exact observation or the registered observations in a specified range of time (or space).

Several techniques have been proposed for the analysis of multivariate interval-valued data (see Reference [1] for an overview), among which techniques for Principal Component Analysis (PCA) of interval-valued data. In this paper we propose a three-way extension of PCA of interval-valued data. The generic three-way interval-valued datum usually pertains to the score of the generic observation i on the generic interval-valued variable j at the occasion k. An occasion can be a specific time point or, in general, a specific measurement condition.

In the next section we review the available methods for two-way PCA of interval-valued data: Vertices Principal Component Analysis (V-PCA) and Centers Principal Component Analysis (C-PCA) [2]. These will be described in ample detail, because we need such details for the elaboration of our three-way generalizations. In Section 3 we present the two most popular three-way methods: Tucker3 [3] and CANDECOMP/PARAFAC [4,5]. These are methods for component analysis of three-way data sets. In Section 4 we propose how to apply the above methods to three-way interval-valued data sets. The new methods are called *Three-way Vertices Principal Component Analysis* (3V-PCA) and *Three-way Centers Principal Component Analysis* (3C-PCA). Finally, in Section 5 the results of an application of the three-way procedures to an empirical data set on air pollution will be presented.

2. PRINCIPAL COMPONENT ANALYSIS FOR TWO-WAY INTERVAL-VALUED DATA

In this section we describe two helpful methods in order to discover the underlying structure of a two-way interval-valued data set the V-PCA and C-PCA methods proposed by Cazes *et al.* [2].

*Correspondence to: P. Giordani, Department of Statistics, Probability and Applied Statistics, University of Rome 'La Sapienza', Ple Aldo Moro 5, I-00185 Rome, Italy.
E-mail: paolo.giordani@uniroma1.it

Prior to reviewing both methods, it is worth noticing that, when an observation unit is characterized by a score on a single interval-valued variable, the information can be represented as a segment in $\mathbb{R}^1$. Instead, in the general case of I observation units on J interval-valued variables, we have the data matrix

$$\mathbf{Y} = \begin{pmatrix} [y_{11}^-, y_{11}^+] & \cdots & [y_{1J}^-, y_{1J}^+] \\ \vdots & \ddots & \vdots \\ [y_{I1}^-, y_{I1}^+] & \cdots & [y_{IJ}^-, y_{IJ}^+] \end{pmatrix} \tag{1}$$

where y_{ij}^- and y_{ij}^+ denote, the minimum and maximum respectively of the interval of scores of observation unit i on variable j. Each row of the matrix in (1) pertains to a generic observation unit that can be represented by a hyperrectangle in $\mathbb{R}^J$ (a rectangle if $J = 2$) whose total number of vertices is 2^J.

2.1. Vertices Principal Component Analysis (V-PCA)

The first step in V-PCA consists of coding the interval-valued data matrix into a numerical one. For the sake of simplicity, let us consider that each observation unit has scores on $J = 2$ variables. Then the interval-valued data matrix is

$$\mathbf{Y} = \begin{pmatrix} [y_{11}^-, y_{11}^+] & [y_{12}^-, y_{12}^+] \\ \vdots & \vdots \\ [y_{i1}^-, y_{i1}^+] & [y_{i2}^-, y_{i2}^+] \\ \vdots & \vdots \\ [y_{I1}^-, y_{I1}^+] & [y_{I2}^-, y_{I2}^+] \end{pmatrix}. \tag{2}$$

Each row can be transformed into a specific numerical matrix representing the four vertices for the observation unit at hand. For the ith observation unit this matrix is

$${}_i\mathbf{Y} = \begin{pmatrix} y_{i1}^- & y_{i2}^- \\ y_{i1}^- & y_{i2}^+ \\ y_{i1}^+ & y_{i2}^- \\ y_{i1}^+ & y_{i2}^+ \end{pmatrix} \tag{3}$$

in which each row refers exactly to each vertex of the rectangle which represents the ith observation unit. ${}_i\mathbf{Y}$ is a matrix with $2^J = 2^2 = 4$ rows and $J = 2$ columns. For V-PCA the matrices ${}_i\mathbf{Y}$ are stacked below each other so as to obtain the supermatrix

$$\mathbf{Y}_{\text{V-PCA}} = \begin{pmatrix} {}_1\mathbf{Y} \\ \vdots \\ {}_I\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} y_{11}^- & y_{12}^- \\ y_{11}^- & y_{12}^+ \\ y_{11}^+ & y_{12}^- \\ y_{11}^+ & y_{12}^+ \end{pmatrix} \\ \vdots \\ \begin{pmatrix} y_{I1}^- & y_{I2}^- \\ y_{I1}^- & y_{I2}^+ \\ y_{I1}^+ & y_{I2}^- \\ y_{I1}^+ & y_{I2}^+ \end{pmatrix} \end{pmatrix}. \tag{4}$$

The matrix in (4) has $2^2 I$ rows and two columns. In the general case of J interval-valued variables, $\mathbf{Y}_{\text{V-PCA}}$ is constructed by stacking the $(2^J \times J)$-matrices ${}_i\mathbf{Y}$ (which represent all vertices) below each other; thus $\mathbf{Y}_{\text{V-PCA}}$ in general is a supermatrix of order $2^J I \times J$.

V-PCA is nothing but performing classical PCA on the possibly preprocessed (e.g. columnwise centered and/or scaled to sums of squares equal to the number of observation units) matrix $\mathbf{Y}_{\text{V-PCA}}$. To represent each observation unit, we project all the vertices of the concerned hyperrectangle in the obtained low-dimensional space. It would be insightful if, when we extract two components in V-PCA, the projected vertices could be represented by a rectangle. Unfortunately, the projected vertices do not define such a rectangle. This problem is solved, however, by representing the generic ith observation unit on each axis as the segment which includes all the projections. The rectangle or hyperrectangle (if more than two components are used) built from these segments will then fully comprise the projected hyperrectangle of the observation unit in the low-dimensional space. Thus, if for the ith observation unit we define ${}_ia_{ls}$, the component score of the vertex l on the sth component, we can find the lower bound of the segment including all the projections on the sth component as

$${}_ia_s^- = \min_{l \in L_i}({}_ia_{ls}) \tag{5}$$

and the upper bound as

$${}_ia_s^+ = \max_{l \in L_i}({}_ia_{ls}) \tag{6}$$

where L_i denotes the set of all vertices for observation unit i. It follows that, if we extract two components, we obtain a rectangle including all the projected vertices (see Figure 1).

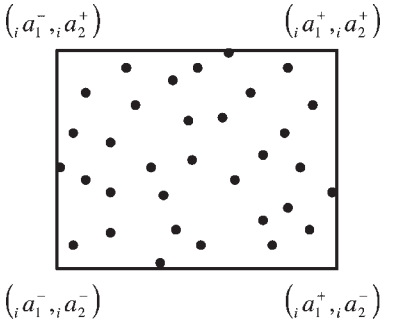

Figure 1. Example of the 32 vertices of a five-dimensional hyperrectangle projected on a two-dimensional space.

Remark 1

It was implicitly seen that the V-PCA method is computationally cumbersome because it requires the computation of the matrix $\mathbf{Y}_{\text{V-PCA}}$ whose number of rows is exponentially proportional to the number of variables. However, finding the loadings involves the computation of the cross-products matrix $\mathbf{Y}'_{\text{V-PCA}}\mathbf{Y}_{\text{V-PCA}}$. In fact, it is well known that the component loadings (in matrix $\mathbf{F}$) are obtained as eigenvectors of the cross-products matrix. Now $\mathbf{Y}_{\text{V-PCA}}$ has cross-products which are very simple to compute. We have [2]

$$\mathbf{Y}'_{\text{V-PCA}}\mathbf{Y}_{\text{V-PCA}} = 2^{J-2}\begin{pmatrix} 2\sum_{i=1}^I \left(y_{i1}^{-2} + y_{i1}^{+2}\right) & \sum_{i=1}^I \left(y_{i1}^- + y_{i1}^+\right)\left(y_{i2}^- + y_{i2}^+\right) & \cdots & \sum_{i=1}^I \left(y_{i1}^- + y_{i1}^+\right)\left(y_{ij}^- + y_{ij}^+\right) \\ \sum_{i=1}^I \left(y_{i2}^- + y_{i2}^+\right)\left(y_{i1}^- + y_{i1}^+\right) & 2\sum_{i=1}^I \left(y_{i2}^{-2} + y_{i2}^{+2}\right) & \cdots & \sum_{i=1}^I \left(y_{i2}^- + y_{i2}^+\right)\left(y_{ij}^- + y_{ij}^+\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^I \left(y_{ij}^- + y_{ij}^+\right)\left(y_{i1}^- + y_{i1}^+\right) & \sum_{i=1}^I \left(y_{ij}^- + y_{ij}^+\right)\left(y_{i2}^- + y_{i2}^+\right) & \cdots & 2\sum_{i=1}^I \left(y_{ij}^{-2} + y_{ij}^{+2}\right) \end{pmatrix}. \tag{7}$$

Thus such a computational short-cut to V-PCA gives an efficient way to find the component loadings matrix **F**. Next the component scores matrix can be obtained as $\mathbf{Y}_{\text{V-PCA}}\mathbf{F}$ taking into account that **F** is columnwise orthonormal. When the number of rows of $\mathbf{Y}_{\text{V-PCA}}$ becomes prohibitively large, the same holds for the component scores matrix that has the same number of rows. However, the representation of the hyperrectangles in the low-dimensional space according to Equations (5) and (6) can be computed easily using the equivalent formulae

$$ {}_{i}a_{s}^{-} = \sum_{j:f_{js}<0} y_{ij}^{+} f_{js} + \sum_{j:f_{js}>0} y_{ij}^{-} f_{js} \tag{8} $$

$$ {}_{i}a_{s}^{+} = \sum_{j:f_{js}<0} y_{ij}^{-} f_{js} + \sum_{j:f_{js}>0} y_{ij}^{+} f_{js} \tag{9} $$

where f_{js} is the generic element of the component loadings matrix **F** [2].

2.2. Centers Principal Component Analysis (C-PCA)

A different way to summarize two-way interval-valued data is centers principal component analysis (C-PCA), which does not involve the computation of huge matrices as $\mathbf{Y}_{\text{V-PCA}}$ was for the V-PCA method. Now the interval-valued data matrix in (1) is transformed into a new matrix of the same order whose generic element is the *midpoint* of each interval-valued variable. Thus, instead of (4), we obtain the transformed data matrix

$$ \mathbf{Y}_{\text{C-PCA}} = \begin{pmatrix} y_{11}^{c} & \cdots & y_{1J}^{c} \\ \vdots & \ddots & \vdots \\ y_{I1}^{c} & \cdots & y_{IJ}^{c} \end{pmatrix} \tag{10} $$

where

$$ y_{ij}^{c} = \frac{y_{ij}^{+} + y_{ij}^{-}}{2} \text{ for } i = 1, \ldots, I \quad \text{and } j = 1, \ldots, J \tag{11} $$

It follows that, in each row of (10), one can find the co-ordinates of the centers of the hyperrectangle in $\mathbb{R}^J$ associated with each observation unit. Therefore the data matrix in (10) is the matrix of the co-ordinates of the centers pertaining to the I observation units.

C-PCA consists of performing classical PCA on the possibly preprocessed version of the matrix in (10). Assuming that (10) is already preprocessed, the sth component score of the ith center is

$$ a_{is}^{c} = \sum_{j=1}^{J} y_{ij}^{c} f_{js} \tag{12} $$

where f_{js} is the component loading of the jth variable on the sth component, in which we use the property that **F** is columnwise orthonormal.

Just as with V-PCA, in C-PCA we wish to represent each observation unit as a hyperrectangle in a low-dimensional subspace (a rectangle in the principal plane extracting two components). We can do this in the same way as with V-PCA, i.e. by projecting all vertices on the principal plane and next, for each component, finding the segments that cover these scores, using Equations (8) and (9). It is helpful to note that the co-ordinates of the ith center, y_{ij}^{c}, $j = 1, \ldots, J$, are enclosed between y_{ij}^{-} and y_{ij}^{+}, $j = 1, \ldots, J$. In fact, it is worth noting that each such segment in the low-dimensional space will surely contain the projected center. Thus not only in V-PCA but also in C-PCA it is possible to represent each observation unit as a hyperrectangle in a low-dimensional subspace. In V-PCA the underlying structure is obtained by considering the bounds of the interval-valued data at hand. On the contrary, in C-PCA this is done by using only the midpoints of the data. Nonetheless, the size of the low-dimensional hyperrectangles is then obtained by considering the bounds of the data as a sort of supplementary information (using Equations (8) and (9)). The intuition behind C-PCA is based upon the assumption that, since the centers should be reasonable representations of the hyperrectangles, a subspace representing the centers well should not be too bad for the hyperrectangles either.

Remark 2

In Remark 1 we computed the cross-products matrix $\mathbf{Y}'_{\text{V-PCA}}\mathbf{Y}_{\text{V-PCA}}$. For the sake of completeness, we remark that the cross-products matrix obtained considering the C-PCA method is, with respect to the off-diagonal elements, proportional to that for V-PCA. Hence the only essential difference between the cross-products matrices is in the diagonal elements. See Reference [2] for further details.

In this section we have shown two methods to summarize two-way interval-valued data. Both methods can be extended to deal with three-way interval-valued data. In the case of V-PCA we will explicitly use the simple computation of the cross-products matrix (see (7)) for this purpose.

3. THREE-WAY COMPONENT ANALYSIS

Several methods are available in the literature for the analysis of three-way data (see Reference [6] for an overview). In this section we briefly present two of them. They are probably the two most popular ones: Tucker3 proposed by Tucker [3] and CANDECOMP/PARAFAC proposed independently by Carroll and Chang [4] (CANDECOMP) and Harshman [5] (PARAFAC). For more details about the above methods, see e.g. References [7,8].

3.1. The Tucker3 method

The Tucker3 method aims at analyzing the underlying structure of a three-way data matrix **Y** of order $I \times J \times K$ whose generic term is y_{ijk}, the score of observation unit i on variable j at occasion k; I, J and K are the total numbers of observation units, variables and occasions respectively. Using scalar notation, the Tucker3 model [3] can be written as

$$ y_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk} \tag{13} $$

where a_{ip}, b_{jq} and c_{kr} are the coefficients of observation unit i on the pth observation unit component, of variable j on the qth variable component and of occasion k on the rth occasion component respectively. They are also the generic elements

of the matrices **A**, **B** and **C** respectively, called component matrices of order $I \times P$, $J \times Q$ and $K \times R$ respectively. Apart from the generic error term e_{ijk}, g_{pqr} also appears in (13). It is the generic term for the so-called three-way core matrix **G**. This matrix plays the role of showing the contributions of the triple interactions among all the components of all the modes. The core can be considered a reduced version of the entire data set.

A matrix formulation of (13) is

$$\mathbf{Y}_a = \mathbf{A}\mathbf{G}_a(\mathbf{C}' \otimes \mathbf{B}') + \mathbf{E}_a \tag{14}$$

where $\mathbf{Y}_a$, $\mathbf{G}_a$ and $\mathbf{E}_a$ are two-way matrices obtained by juxtaposing the frontal slices of the three-way matrices **Y**, **G** and **E** (the error matrix) respectively. It follows that $\mathbf{Y}_a$ and $\mathbf{E}_a$ have order $I \times JK$ and $\mathbf{G}_a$ has order $P \times QR$.

Prior to fitting the model, it is often useful to preprocess the data by means of centering and scaling [9,10]. By centering, one aims at eliminating constants in the data that are not fitted by the method. These constants are often unknown, but one usually deletes them by subtracting the mean across particular modes. It is possible to center across one mode or a combination of modes. The aim of scaling is the elimination of undesirable differences in the importance of the entities of one or more modes. Rescaling the scores of the entities of the modes involved eliminates such unwanted effects.

The model is fitted to the data by minimizing the sum of squared residuals

$$\|\mathbf{Y}_a - \mathbf{A}\mathbf{G}_a(\mathbf{C}' \otimes \mathbf{B}')\|^2 \tag{15}$$

over **A**, **B**, **C** and **G**. Several Alternating Least Squares (ALS) algorithms are available to minimize (15) (see e.g. References [11,12]). We note that in the algorithm one can constrain the component matrices to columnwise orthonormality without loss of fit. The fit of the model can be calculated as

$$1 - \frac{\|\mathbf{Y}_a - \mathbf{A}\mathbf{G}_a(\mathbf{C}' \otimes \mathbf{B}')\|^2}{\|\mathbf{Y}_a\|^2} \tag{16}$$

which is usually multiplied by 100 to express the fit as a percentage.

The obtained solution is not unique. Equally well-fitting solutions can be obtained by arbitrary non-singular transformations of all component matrices, provided that these rotations are compensated in the core. In fact, we have

$$\mathbf{Y}_a = \mathbf{A}\mathbf{G}_a(\mathbf{C}' \otimes \mathbf{B}') + \mathbf{E}_a = \tilde{\mathbf{A}}\tilde{\mathbf{G}}_a(\tilde{\mathbf{C}}' \otimes \tilde{\mathbf{B}}') + \mathbf{E}_a \tag{17}$$

where $\tilde{\mathbf{A}} = \mathbf{AS}$, $\tilde{\mathbf{B}} = \mathbf{BT}$, $\tilde{\mathbf{C}} = \mathbf{CU}$ and $\tilde{\mathbf{G}}_a = \mathbf{S}^{-1}\mathbf{G}_a(\mathbf{U}^{-1\prime} \otimes \mathbf{T}^{-1\prime})$ for any non-singular square matrices **S**, **T** and **U**. The indeterminacy of the model can be exploited in order to search for simplicity of the component matrices or of the core matrix or of both (see Reference [13] for an overview). Finally, notice that in the Tucker3 analysis it can be very useful to plot the entities of one mode to better interpret the solution (see Reference [14] for details).

3.2. The CANDECOMP/PARAFAC model

Another well-known three-way method is CANDECOMP/PARAFAC [4,5], which is more parsimonious but less general than Tucker3, as will be made clear below. It is the straightforward generalization of two-way PCA by inserting component loadings for the occasion mode. Therefore the model is given by

$$y_{ijk} = \sum_{s=1}^{S} a_{is}b_{js}c_{ks} + e_{ijk} \tag{18}$$

where a_{is}, b_{js} and c_{ks} are the generic coefficients of the observation units, of the variables and of the occasions respectively on the sth component.

With respect to the formulation of the Tucker3 model in (13), at least two differences are evident. First, we find the same number of components for all the modes, and each component is related to exactly one component of the other modes. Thus we can interpret the extracted components across all the modes simultaneously. The second difference, related to the first, is that the core matrix disappears. In fact, we can still assume to have the core in CANDECOMP/PARAFAC after constraining it to be unit superdiagonal ($g_{pqr} = 1$ if $p = q = r = s$, $g_{pqr} = 0$ otherwise). Hence CANDECOMP/PARAFAC can be thought of as a constrained version of Tucker3. It also follows that we can express the model as

$$\mathbf{Y}_a = \mathbf{A}\mathbf{I}_a(\mathbf{C}' \otimes \mathbf{B}') + \mathbf{E}_a \tag{19}$$

which is the matrix formulation of Tucker3 in (14) imposing $\mathbf{G}_a = \mathbf{I}_a$, where $\mathbf{I}_a$ is the matrix of order $S \times S^2$ obtained by juxtaposing the S slices of the unit superdiagonal array.

Indeed, CANDECOMP/PARAFAC has a specific property also known as the intrinsic axis property: under mild assumptions the solution is unique up to scalar multiplications and permutations and hence cannot be rotated without loss of fit. It is well known that the uniqueness of the CANDECOMP/PARAFAC solution is a desirable property in chemical studies because it leads to unambiguous detection of the parameters underlying the chemical system (see, e.g. Reference [15]).

The problem of preprocessing is exactly the same as for Tucker3. For a detailed treatment of the preprocessing problem in three-way analysis we refer to Reference [10] and, especially in CANDECOMP/PARAFAC, to Reference [9].

The model is fitted to the data by minimizing the sum of squared residuals

$$\|\mathbf{Y}_a - \mathbf{A}\mathbf{I}_a(\mathbf{C}' \otimes \mathbf{B}')\|^2 \tag{20}$$

over **A**, **B** and **C** using an ALS algorithm [4,5]. Similarly to Tucker3, the fit of the model can be calculated (usually multiplying by 100) using (16), with the core matrix taken equal to $\mathbf{I}_a$. Again, as in Tucker3, also in the CANDECOMP/PARAFAC framework it could be interesting to plot the entities of a mode (see Reference [14] for more details).

4. THREE-WAY PRINCIPAL COMPONENT ANALYSIS OF INTERVAL-VALUED DATA

In the above sections we have introduced two techniques to analyze two-way interval-valued data sets, i.e. V-PCA and C-PCA, and two methods to analyze three-way numerical data sets, i.e. Tucker3 and CANDECOMP/PARAFAC.

In this section we extend C-PCA and V-PCA to deal with three-way interval-valued data by means of Tucker3 and CANDECOMP/PARAFAC. Thus we assume that the available information pertains to I observation units defined by J interval-valued variables, collected at K occasions. Let the data matrix at the generic occasion k, for $k = 1, \ldots, K$, be

$$\mathbf{Y}_k = \begin{pmatrix} \left[y_{11k}^-, y_{11k}^+\right] & \cdots & \left[y_{1Jk}^-, y_{1Jk}^+\right] \\ \vdots & \ddots & \vdots \\ \left[y_{I1k}^-, y_{I1k}^+\right] & \cdots & \left[y_{IJk}^-, y_{IJk}^+\right] \end{pmatrix} \tag{21}$$

whose generic element is the interval $\left[y_{ijk}^-, y_{ijk}^+\right]$, where y_{ijk}^- and y_{ijk}^+ are the lower bound and the upper bound respectively of the ith observation unit for the jth variable at the kth occasion. Then we can introduce the data matrix of the entire data set as

$$\mathbf{Y}_a = (\mathbf{Y}_1 \quad \cdots \quad \mathbf{Y}_K) = \left(\begin{pmatrix} \left[y_{111}^-, y_{111}^+\right] & \cdots & \left[y_{1J1}^-, y_{1J1}^+\right] \\ \vdots & \ddots & \vdots \\ \left[y_{I11}^-, y_{I11}^+\right] & \cdots & \left[y_{IJ1}^-, y_{IJ1}^+\right] \end{pmatrix} \cdots \begin{pmatrix} \left[y_{11K}^-, y_{11K}^+\right] & \cdots & \left[y_{1JK}^-, y_{1JK}^+\right] \\ \vdots & \ddots & \vdots \\ \left[y_{I1K}^-, y_{I1K}^+\right] & \cdots & \left[y_{IJK}^-, y_{IJK}^+\right] \end{pmatrix}\right). \tag{22}$$

The matrix in (22) is the two-dimensional matrix of order $I \times JK$ obtained by juxtaposing the data matrices defined in (21), for $k = 1, \ldots, K$. Hence the ith row of the matrix in (22) pertains to all scores of the ith observation unit. Let us now describe how to extend V-PCA and C-PCA to deal with three-way data. The basic idea of these three-way extensions is to first apply the same transformations as used in C-PCA and V-PCA to our three-way array and next analyze these transformed arrays by Tucker3 or CANDECOMP/PARAFAC analysis.

4.1. Three-way Vertices Principal Component Analysis (3V-PCA)

Now, if we apply the V-PCA transformation to one row of the supermatrix $\mathbf{Y}_a$, we obtain a matrix whose rows refer exactly to all vertices of the hyperrectangle in $\mathbb{R}^{JK}$ which represents each observation unit. Therefore its dimension is $2^{JK} \times JK$. It also follows that the transformed data matrix for the full matrix $\mathbf{Y}_a$ has order $I2^{JK} \times JK$, which, even for a small data array with, for instance, $I = 3$ observation units, $J = 3$ variables and $K = 4$ occasions, leads to a huge transformed data matrix $\mathbf{Y}_{3V\text{-}PCA}$ (of order $12\,288 \times 12$), where $\mathbf{Y}_{3V\text{-}PCA}$ is obtained by applying the V-PCA transformation on $\mathbf{Y}_a$. As noted above, the example is very small and, for increasing data sizes, the number of rows in $\mathbf{Y}_{3V\text{-}PCA}$ increases exponentially.

Clearly, analyzing such huge matrices by Tucker3 or CANDECOMP/PARAFAC analysis is unfeasible. However, for the analysis of $\mathbf{Y}_{3V\text{-}PCA}$ we can use modified algorithms that have been proposed for handling three-way arrays in which the number of entities of one of the modes is considerably higher than those of the remaining modes. Such algorithms have been proposed by Kiers and Krijnen [16], Kiers *et al.* [17], Alsberg and Kvalheim [18,19], Andersson and Bro [12], Bro and Andersson [20], Kiers and Harshman [21] and Carroll *et al.* [22] that, as shown by Kiers and Harshman [21], solve the same problem as Alsberg and Kvalheim [18,19] in a different context. The basis of all these procedures is that, in the iterative part of the procedure, only the cross-products of $\mathbf{Y}_{3V\text{-}PCA}$ are needed.

Analogously to the two-way case, this cross-products matrix can be easily computed as follows. The cross-products matrix $\mathbf{X}_{kl} = {}_k\mathbf{Y}'_{V\text{-}PCA}{}_l\mathbf{Y}_{V\text{-}PCA}$ at occasions k and l ($k \neq l$), where ${}_k\mathbf{Y}_{V\text{-}PCA}$ and ${}_l\mathbf{Y}_{V\text{-}PCA}$ denote the matrices resulting from the V-PCA transformation applied on $\mathbf{Y}_k$ and $\mathbf{Y}_l$ respectively, $k, l = 1, \ldots, K$, is

$$\mathbf{X}_{kl} = 2^{J-2}\begin{pmatrix} \sum_{i=1}^I \left(y_{i1k}^- + y_{i1k}^+\right)\left(y_{i1l}^- + y_{i1l}^+\right) & \sum_{i=1}^I \left(y_{i1k}^- + y_{i1k}^+\right)\left(y_{i2l}^- + y_{i2l}^+\right) & \cdots & \sum_{i=1}^I \left(y_{i1k}^- + y_{i1k}^+\right)\left(y_{ijl}^- + y_{ijl}^+\right) \\ \sum_{i=1}^I \left(y_{i2k}^- + y_{i2k}^+\right)\left(y_{i1l}^- + y_{i1l}^+\right) & \sum_{i=1}^I \left(y_{i2k}^- + y_{i2k}^+\right)\left(y_{i2l}^- + y_{i2l}^+\right) & \cdots & \sum_{i=1}^I \left(y_{i2k}^- + y_{i2k}^+\right)\left(y_{ijl}^- + y_{ijl}^+\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^I \left(y_{ijk}^- + y_{ijk}^+\right)\left(y_{i1l}^- + y_{i1l}^+\right) & \sum_{i=1}^I \left(y_{ijk}^- + y_{ijk}^+\right)\left(y_{i2l}^- + y_{i2l}^+\right) & \cdots & \sum_{i=1}^I \left(y_{ijk}^- + y_{ijk}^+\right)\left(y_{ijl}^- + y_{ijl}^+\right) \end{pmatrix} \tag{23}$$

whereas, if $k = l$, the diagonal elements are replaced by $2\sum_{i=1}^I \left(y_{ijk}^{-2} + y_{ijk}^{+2}\right)$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Then the cross-products matrix for all pairs of occasions is given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{11} & \cdots & \mathbf{X}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{K1} & \cdots & \mathbf{X}_{KK} \end{pmatrix} \tag{24}$$

the order of which is $JK \times JK$.

In our three-way extensions we now use the procedure proposed by Kiers and Harshman [21], which is based on the fact that the three-way analysis of the supermatrix $\mathbf{Y}_{3V\text{-}PCA}$ and that of a matrix $\mathbf{Z}$ which has exactly the same cross-products as $\mathbf{Y}_{3V\text{-}PCA}$ give essentially the same results. Specifically, if we find a matrix $\mathbf{Z}$ such that $\mathbf{Z}'\mathbf{Z} = \mathbf{X} = \mathbf{Y}'_{3V\text{-}PCA}$

$\mathbf{Y}_{3V\text{-}PCA}$, then applying Tucker3 or CANDECOMP/PARAFAC to $\mathbf{Z}$ or $\mathbf{Y}_{3V\text{-}PCA}$ will give exactly the same matrices $\mathbf{B}$ and $\mathbf{C}$ and also the same core. The only difference will be in the component scores for the A-mode (i.e. the vertices), but in practice one is not interested in all component scores for all vertices anyway. In fact, for each observation unit the main interest is in the component scores that allow one to determine the vertices and therefore the size of the low-dimensional hyperrectangle. Now the computational burden can be decreased enormously by finding a matrix $\mathbf{Z}$ that has far fewer rows than $\mathbf{Y}_{3V\text{-}PCA}$. To find such a small matrix $\mathbf{Z}$, we can use the following procedure. First we perform the eigendecomposition of $\mathbf{X}$ (which is decomposed as $\mathbf{K}\Lambda\mathbf{K}'$ with $\mathbf{K}'\mathbf{K} = \mathbf{K}\mathbf{K}' = \mathbf{I}_{JK}$ and Λ diagonal). We now set $\mathbf{Z}' = \mathbf{K}\Lambda^{1/2}$. It follows that $\mathbf{X} = \mathbf{Z}'\mathbf{Z}$. Hence we have found a matrix $\mathbf{Z}$ that has JK rows, which in practice is much smaller than $I2^{JK}$. In fact, other decompositions (e.g. the Cholesky one) can also be used such that the new 'data matrix' has the same cross-products as the original one (see Reference [21] for further details). Thus in this way the same component matrices for the second and third modes (and the core matrix if the Tucker3 model is performed) can be obtained. In principle, using the thus obtained results, we can now also compute all component scores, by using the full matrix $\mathbf{Y}_{3V\text{-}PCA}$ and defining $\mathbf{F} = (\mathbf{C} \otimes \mathbf{B})\mathbf{G}_a'$, as $\mathbf{A} = \mathbf{Y}_{3V\text{-}PCA}\,\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}$, as follows immediately from the fact that we minimize least squares function (15). Instead of computing all component scores for all vertices, however, we are mainly interested in finding hyperrectangles that contain these vertex component scores.

The plotting procedure based on (5) and (6) for the two-way case is no longer suitable for the three-way case because it requires the component matrix $\mathbf{A}$. In order to provide a low-dimensional representation of the observation units, similarly to Remark 1, the plotting procedure based on (8) and (9) can be considered. It can be adjusted to the three-way case as follows. Suppose that, depending on the features of the current data set, we have already performed the Tucker3 or CANDECOMP/PARAFAC analysis on the above mentioned matrix $\mathbf{Z}$. Now we would like to represent the observation units as hyperrectangles in the obtained low-dimensional subspace. To do so, we slightly modify Equations (8) and (9). Assuming, without loss of generality, that the data are already preprocessed (and that the upper and lower bounds are preprocessed in the same way as the centers matrix using the values pertaining to the centers), let us begin by considering the Tucker3 model. The co-ordinates of the vertices associated with each hyperrectangle on the first P principal axes are obtained as

$$\begin{aligned}\mathbf{A} &= \mathbf{Y}_{3V\text{-}PCA}\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\\ &= \mathbf{Y}_{3V\text{-}PCA}(\mathbf{C}\otimes\mathbf{B})\mathbf{G}_a'\left[\mathbf{G}_a(\mathbf{C}'\mathbf{C}\otimes\mathbf{B}'\mathbf{B})\mathbf{G}_a'\right]^{-1}\\ &= \mathbf{Y}_{3V\text{-}PCA}\mathbf{W}\end{aligned} \tag{25}$$

where $\mathbf{W}$ is defined implicitly in (25). If the component matrices are columnwise orthonormal (as is usually the case), we can simplify (25) as

$$\mathbf{A} = \mathbf{Y}_{3V\text{-}PCA}(\mathbf{C}\otimes\mathbf{B})\mathbf{G}_a'\left(\mathbf{G}_a\mathbf{G}_a'\right)^{-1} = \mathbf{Y}_{3V\text{-}PCA}\mathbf{W}. \tag{26}$$

Note that (25) and (26) offer an adequate plot of the observation units provided that $\mathbf{F}$ is columnwise orthonormal [14]. In (25) and (26) the matrix $\mathbf{W}$ plays the role of the component loadings matrix in the two-way case. Taking into account that the matrix $\mathbf{Y}_{3V\text{-}PCA}$ in (26) is usually huge and therefore not computed, the computation of the co-ordinates of the vertices is usually unfeasible as well. However, as for the two-way framework, we can still find, for each component, the bounds of the segment in which the projected vertices lie, by adjusting the formulae in (8) and (9) as

$${}_i a_p^- = \sum_{j,k:w_{pjk}<0} y_{ijk}^+ w_{pjk} + \sum_{j,k:w_{pjk}>0} y_{ijk}^- w_{pjk} \tag{27}$$

$${}_i a_p^+ = \sum_{j,k:w_{pjk}<0} y_{ijk}^- w_{pjk} + \sum_{j,k:w_{pjk}>0} y_{ijk}^+ w_{pjk} \tag{28}$$

where w_{pjk} is the generic element of $\mathbf{W}$. Using Equations (27) and (28), we can obtain a representation of the I observation units as hyperrectangles in the low-dimensional space spanned by $\mathbf{F}$.

Obviously, sometimes the matrix $\mathbf{F}$ is not columnwise orthonormal. Then the representation is distorted [14]. To solve this problem, we need a transformation matrix $\mathbf{V}$ for which $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{V}$ is columnwise orthonormal. The transformation must be compensated by postmultiplying the matrix $\mathbf{A}$ by $(\mathbf{V}')^{-1}$. The formulae in (26) and in (27) and (28) still hold, using $\tilde{\mathbf{W}} = \mathbf{W}(\mathbf{V}')^{-1}$ instead of $\mathbf{W}$. It can be verified that, in fact, $\tilde{\mathbf{W}} = \tilde{\mathbf{F}}$.

If we consider the CANDECOMP/PARAFAC model, the expression in (25) still holds but $\mathbf{F} = \mathbf{C}\odot\mathbf{B}$, where the symbol $\odot$ denotes the Khatri–Rao product, which is the columnwise Kronecker product (see e.g. Reference [23]). The bounds are defined again by (27) and (28) and should be slightly modified if $\mathbf{F}$ is not columnwise orthonormal, in the same way as above.

4.2. Three-way Centers Principal Component Analysis (3C-PCA)

In contrast to V-PCA, C-PCA can be generalized directly to the three-way situation. In fact, the transformation used in C-PCA does not modify the dimensionality of the data matrix. As a matter of fact, the three-way generalization of (10) is

$$\mathbf{Y}_{3C\text{-}PCA} = \begin{pmatrix} y_{111}^c & \cdots & y_{1J1}^c & \cdots & y_{11K}^c & \cdots & y_{1JK}^c \\ \vdots & \ddots & \vdots & \ldots & \vdots & \ddots & \vdots \\ y_{I11}^c & \cdots & y_{IJ1}^c & \cdots & y_{I1K}^c & \cdots & y_{IJK}^c \end{pmatrix} \tag{29}$$

where

$$y_{ijk}^c = \frac{y_{ijk}^+ + y_{ijk}^-}{2} \quad \text{for} \quad \begin{array}{l} i = 1,\ldots,I \\ j = 1,\ldots,J \quad \text{and} \quad k = 1,\ldots,K. \end{array} \tag{30}$$

Clearly, the matrix in (29) contains the co-ordinates of the centers of the hyperrectangles in $\mathbb{R}^{JK}$ pertaining to the I observation units. Note that in the three-way case, as for the two-way case, the cross-products matrices obtained by considering $\mathbf{Y}_{3V\text{-}PCA}$ or $\mathbf{Y}_{3C\text{-}PCA}$ in (29) differ only in the diagonal elements.

Assuming, without loss of generality, that (29) is already preprocessed, the co-ordinates of the I centers on the first P

principal axes of the observation units mode are obtained as (in the Tucker3 case)

$$\begin{aligned}\mathbf{A} &= \mathbf{Y}_{\text{3C-PCA}}\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1} \\ &= \mathbf{Y}_{\text{3C-PCA}}(\mathbf{C}\otimes\mathbf{B})\mathbf{G}_a'\left[\mathbf{G}_a(\mathbf{C}'\mathbf{C}\otimes\mathbf{B}'\mathbf{B})\mathbf{G}_a'\right]^{-1} \\ &= \mathbf{Y}_{\text{3C-PCA}}\mathbf{W}\end{aligned} \tag{31}$$

where $\mathbf{W}$ is defined implicitly in (31) and, of course, the component matrices are those obtained by performing the Tucker3 analysis on the matrix $\mathbf{Y}_{\text{3C-PCA}}$ in (29). The generic element of $\mathbf{A}$ in (31) is

$$a_{ip}^{c} = \sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^{c} w_{pjk}. \tag{32}$$

As (31) provides the projections of the centers on the obtained low-dimensional space, the bounds of the segments within which we are sure that the complete projected hyperrectangles are located can be found by means of Equations (27) and (28).

The same comments hold on performing the CANDECOMP/PARAFAC method, by suitably replacing the weights matrix $\mathbf{W}$.

To sum up, we have proposed two three-step procedures to perform three-way component analysis of three-way interval-valued data. We refer to these procedures as Three-way Vertices Principal Component Analysis (3V-PCA) and Three-way Centers Principal Component Analysis (3C-PCA). Briefly, the steps are:

- transforming the three-way interval-valued data set into the three-way matrix $\mathbf{Y}_{\text{3C–PCA}}$ in (30) for 3C-PCA or computing the cross-products matrix $\mathbf{X}$ in (24) in order to find the new 'data matrix' $\mathbf{Z}$ for 3V-PCA;
- performing Tucker3 or CANDECOMP/PARAFAC, depending on the features of the data set involved, on the matrix obtained in the previous step;
- plotting the I observation units by using the expressions in (27) and (28).

5. APPLICATION

The air pollution in Rome is monitored by 12 testing stations at various locations in Rome. In research by the municipality of Rome the testing stations are classified into four groups. The two groups denoted 'B' and 'C' can be considered as the urban groups. They are distinguished according to the traffic density of the areas involved. More specifically, group 'B' ('Arenula', 'Cinecittà', 'Magnagrecia', 'Preneste') refers to residential areas and group 'C' ('Francia', 'Fermi', 'Libia', 'Montezemolo', 'Tiburtina') refers to high traffic areas. Two testing stations ('Castel di Guido', 'Tenuta del Cavaliere') belong to group 'D'. They are areas indirectly exposed to vehicular pollution. Their utilization aims at understanding the complex photochemical phenomena that usually occur in Rome and in the suburban areas around Rome. Finally, just one testing station ('Ada') belongs to group 'A'. 'Ada' is located near a park and therefore will have lower pollution. This station should provide information about the lowest level of pollution in Rome. However, note that meteorological phenomena can lead to air pollution also in this area. The location of the testing stations is shown in Figure 2.

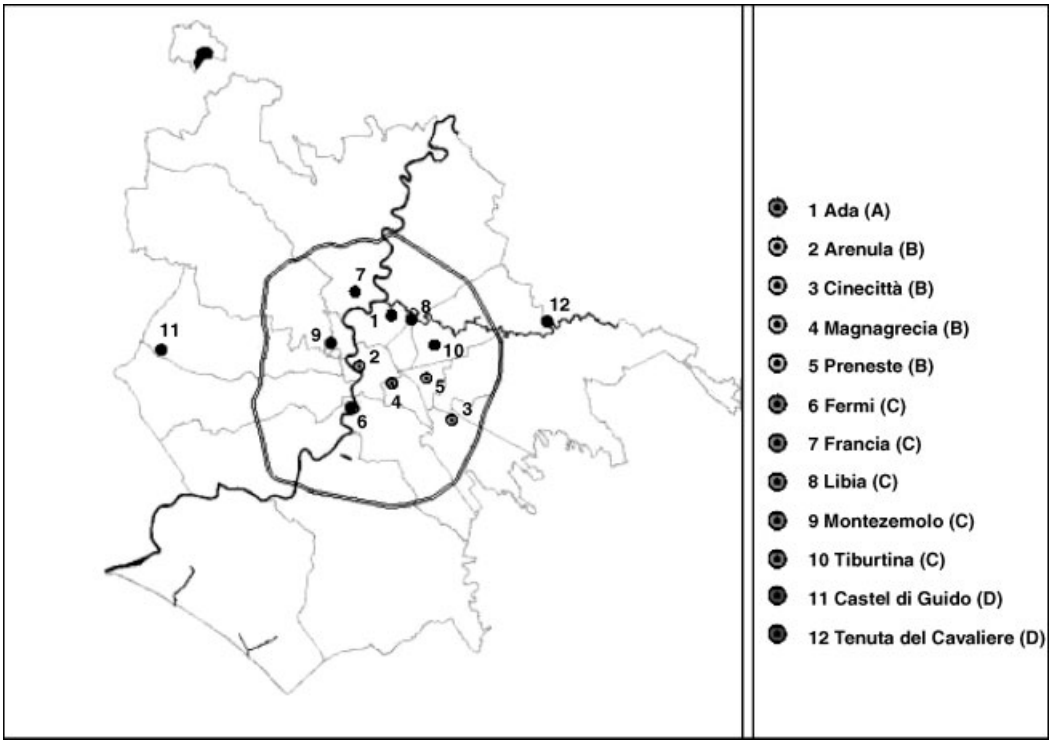

Figure 2. Map of the testing stations in Rome.
Note: The ring denotes a highway, which encloses the city of Rome. The dark grey line is the Tevere river. The light grey lines are the so-called 'consolar ways'. Source: http://www.comune.roma.it (official Rome web site).

The registered pollutants are CO, SO_2, O_3, NO, NO_2, NO_X, PM_{10} (amount of inhalable dust, i.e. atmospheric particles, the aerodynamic diameter of which is lower than 10 μm) and benzene. At the testing station 'Ada', additional pollutants and meteorological features (e.g. temperature, pressure, wind) are also considered.

In this application we consider a subset of testing stations such that all the testing stations record the same pollutants. Specifically, the testing stations considered here are 'Ada' (group 'A'), 'Magnagrecia' ('B'), 'Preneste' ('B'), 'Fermi' ('C'), 'Francia' ('C'), 'Castel di Guido' ('D') and 'Tenuta del Cavaliere' ('D'). The pollutants considered here are NO, NO_2 and O_3. We did not take into account SO_2, because only three testing stations register its concentration values and, above all, the emission of SO_2 has decreased dramatically in the past few years and continues to decrease owing to better fuel quality. Moreover, we did not use the available information on CO, PM_{10} and benzene, because they are registered only at a small number of testing stations. Finally, we did not consider NO_X to avoid linear dependences among the pollutants, since $NO_X = NO + NO_2$.

Among the remaining pollutants, we observe that they can be distinguished according to their origin. NO and NO_2 are emitted directly from a source (primary pollutants). Instead, O_3 (secondary pollutant) builds up in the atmosphere as a consequence of interactions between solar radiation and various pollutants.

In our analysis we consider measurements from three months (from 1 January to 31 March) in 1999. Specifically, we have a three-way interval-valued data set with $I = 7$ testing stations (observation units), $J = 3$ pollutants (variables) and $K = 90$ days (occasions) where the lower and upper bounds of the intervals are respectively the minimum and maximum values registered daily.

On this data set we applied the 3C-PCA and 3V-PCA methods using the Tucker3 model. Prior to fitting the data, we preprocessed them by centering across the observation units and scaling within the variables.

In order to choose the number of extracted components, we ran several Tucker3 analyses (in the 3C-PCA case) considering different values of P, Q and R. We decided to extract two components for the observation unit and variable modes ($P = Q = 2$) and one component for the occasion mode ($R = 1$). In fact, considering the 3C-PCA method, the goodness of fit according to (16) was 73.17%. This solution has been chosen as our favorite solution, because it had a good fit (even though, obviously, higher numbers of components led to a better fit) and is easily interpretable. Our simple structure rotation was performed as follows. Since $R = 1$, an equivalent formulation of the Tucker3 model as given in (14) is $\mathbf{Y}_a = (\mathbf{C}' \otimes \mathbf{A}\mathbf{G}_a\mathbf{B}') + \mathbf{E}_a$. Taking into account that $P = Q = 2$, the matrix $\mathbf{A}\mathbf{G}_a\mathbf{B}'$ has rank 2 and hence can be perfectly decomposed by PCA extracting two components. In order to find the PCA decomposition of $\mathbf{A}\mathbf{G}_a\mathbf{B}'$, we used the singular value decomposition of $\mathbf{A}\mathbf{G}_a\mathbf{B}' = \mathbf{P}\mathbf{D}\mathbf{Q}'$ with $\mathbf{P}'\mathbf{P} = \mathbf{I}_I$, $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_J$ and $\mathbf{D}$ diagonal. We obtained $\mathbf{A}^* = \mathbf{P}_2\mathbf{D}_2$ and $\mathbf{B}^* = \mathbf{Q}_2$ (where $\mathbf{P}_2$ and $\mathbf{Q}_2$ contain the columns of $\mathbf{P}$ and $\mathbf{Q}$ corresponding to the biggest two elements of $\mathbf{D}$, stored in the diagonal matrix $\mathbf{D}_2$). Thus we have replaced $\mathbf{A}\mathbf{G}_a\mathbf{B}'$ by the PCA decomposition $\mathbf{A}^*\mathbf{B}^{*\prime}$. Next we rotated the matrix $\mathbf{B}^*$ by means of a varimax rotation [24], while compensating for this rotation in $\mathbf{A}^*$, yielding $\mathbf{A}^v$ and $\mathbf{B}^v$. Thus we have finally replaced $\mathbf{A}\mathbf{G}_a\mathbf{B}'$ by the rotated PCA decomposition $\mathbf{A}^v\mathbf{B}^{v\prime}$ or equivalently by $\mathbf{A}^v\mathbf{I}_2\mathbf{B}^{v\prime}$. To sum up, $\mathbf{B}$ has been rotated by varimax, the core $\mathbf{G}_a$ has been replaced by the identity matrix and $\mathbf{A}$ has been rotated so as to compensate for the varimax rotation. The component matrices for the observation unit and variable modes are given in Tables I and II respectively. Since $R = 1$, the component matrix for the occasion mode reduces to a vector with 90 elements. This is displayed in Figure 3.

Table I. Component matrix for the observation units

Testing station	PC1	PC2
Ada ('A')	−3.41	−10.24
Fermi ('C')	−7.94	16.16
Francia ('C')	−9.30	6.04
Magnagrecia ('B')	−3.55	11.48
Preneste ('B')	4.59	2.74
Castel di Guido ('D')	16.41	−16.53
Tenuta del Cavaliere ('D')	3.20	−9.65

Table II. Component matrix for the variables

Pollutant	PC1	PC2
NO	−0.10	0.65
NO_2	0.08	0.76
O_3	1.00	0.00

By considering Table II, we can observe that the extracted components for the variables are strictly related to the kind of pollutants. As the core is an identity matrix, the components for the observation units and those for the variables are related one-to-one. Thus, when testing stations have high scores on a particular component, they have relatively high levels of the kinds of pollution represented by that particular component. For instance, when some testing stations are characterized by high first component scores, we can conclude that the involved testing stations have registered a high level of those pollutants that are associated most strongly with the first component for the variable mode.

The first component mainly reflects the presence of O_3. The two stations belonging to group 'C' assume the lowest first component scores (−9.30 for 'Francia' and −7.94 for 'Fermi'). 'Preneste' has the second highest component score. In particular, O_3 presence at 'Preneste' is greater than that at the remaining stations belonging to groups 'A', 'B' and 'C'. The score of 'Ada' (−3.41) is very similar to the one pertaining to 'Magnagrecia' (−3.55). Thus note that the amount of O_3 registered at the station 'Ada' is higher than those pertaining to 'Francia' and 'Fermi'. This can be explained by considering that O_3 values increase in suburban and rural areas where polluting issues are lower, as at 'Ada'. In fact, wind moves O_3 into areas in which lower air pollution conditions make it stabler than in areas with high air pollution. This comment holds even more strongly for the stations belonging to group 'D', which have high first-component scores. In particular, 'Castel di Guido' has the highest score

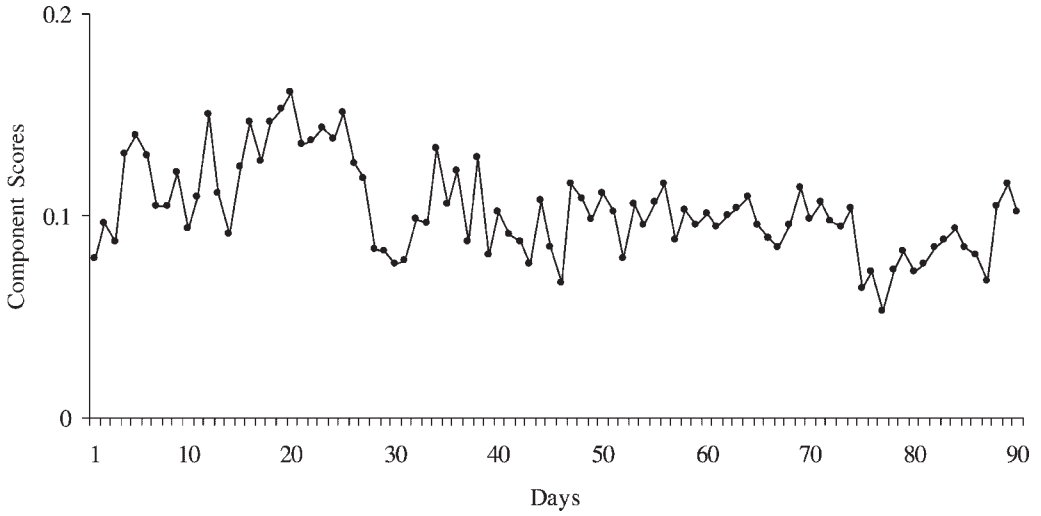

Figure 3. Occasion scores.

(16.41), considerably higher than those pertaining to the remaining stations.

The second component reflects the presence in the atmosphere of NO and NO_2. High component scores for the observation units denote testing stations in which the presence of the pollutants involved is high. As one might expect, the stations belonging to groups 'A' and 'D' have the lowest component scores. Thus 'Ada' has the lowest component score (−10.24) among the urban stations. This is consistent with its status of 'background' area whose registered air pollution values indicate the characteristics of Roman air without polluting sources. At 'Castel di Guido' and 'Tenuta del Cavaliere' the levels of primary pollutants are very low (−16.53, which is the lowest score among all the stations, and −9.65 respectively). Hence the application confirms the status of the stations belonging to group 'D'. Their positions are not directly exposed to air pollution (low level of primary pollutants). Nevertheless, the presence of O_3 implies that these areas are more polluted than town areas because of the role played by atmospheric agents such as wind or sun.

With respect to the primary pollutants, 'Fermi' is the most polluted station (16.16) among those belonging to groups 'B' and 'C', while 'Preneste' is the least polluted (2.74). The component scores of the remaining two stations do not emphasize differences between the two groups. Clearly, with respect to NO and NO_2 the results show that groups 'B' and 'C' are more polluted than groups 'A' and 'D'. By considering simultaneously both component scores, the differences between groups 'B' and 'C' become clearer. Group 'C' is characterized by very high primary pollutant levels and low O_3 levels. This is consistent with the status of high traffic areas in which the direct emission of pollutants is high. The stations of group 'B' (residential areas) are polluted by both primary and secondary pollutants, but the intensities are not high for either kind of pollutants.

Taking into account the structure of the core matrix, the component for the occasion mode provides a measure of the air pollution during each day of the recording time. Thus Figure 3 shows that the level of pollution took the highest values during the first month. Stable air pollution levels can be observed during February and March, although they permanently decreased during the second half of March (end of the recording time).

Finally, in Figure 4 we have plotted the testing stations using the component scores as co-ordinates. In fact, we used the short-cut offered by Equations (27) and (28), which allow us to find the bounds of the low-dimensional hyperrectangles. We got the low-dimensional representation in the space spanned by the columnwise orthonormal matrix $\mathbf{F} = (\mathbf{C} \otimes \mathbf{B})\mathbf{G}_a'$, shown in Figure 4. Figure 4 represents the stations as low-dimensional hyperrectangles, which, because $P = 2$, here reduce to rectangles.

By considering the position of the rectangles in the principal plane, realizing that the first axis represents the presence of O_3 and the second the presence of primary pollutants, we can quickly observe that the most polluted stations (groups 'B' and 'C') are displayed on the high left side whereas the less polluted ones (groups 'A' and 'D') are on the low right side. This obviously confirms our earlier interpretation using the component matrices. In the plot now, from right to left, the groups are approximately ordered as 'D', 'B', 'A', 'C'.

In Figure 4 the stations are distinguished not only with respect to the position of the rectangles but also with respect to their size. The fluctuations of the registered values during each reference day yield a somewhat cluttered representation. In fact, the size of the rectangles reflects how much the pollutant values vary during each day. Thus the size of the rectangles indicates the variation in levels of air pollution at each station: the bigger a rectangle is, the more the levels of air pollution vary. 'Preneste' is represented by the biggest rectangle, 'Ada' and the stations belonging to group 'D' by the smallest (in particular, 'Castel di Guido' is the smallest one). This can be explained by noting that 'Ada', 'Castel di Guido' and 'Tenuta del Cavaliere' registered the lowest pollutant values and that the differences between the minimum and maximum daily values for these stations are noticeably smaller than those pertaining to the remaining stations belonging to groups 'B' and 'C'. In particular, note that the interval widths with respect to the second component at 'Ada', 'Castel di Guido' and 'Tenuta del Cavaliere'

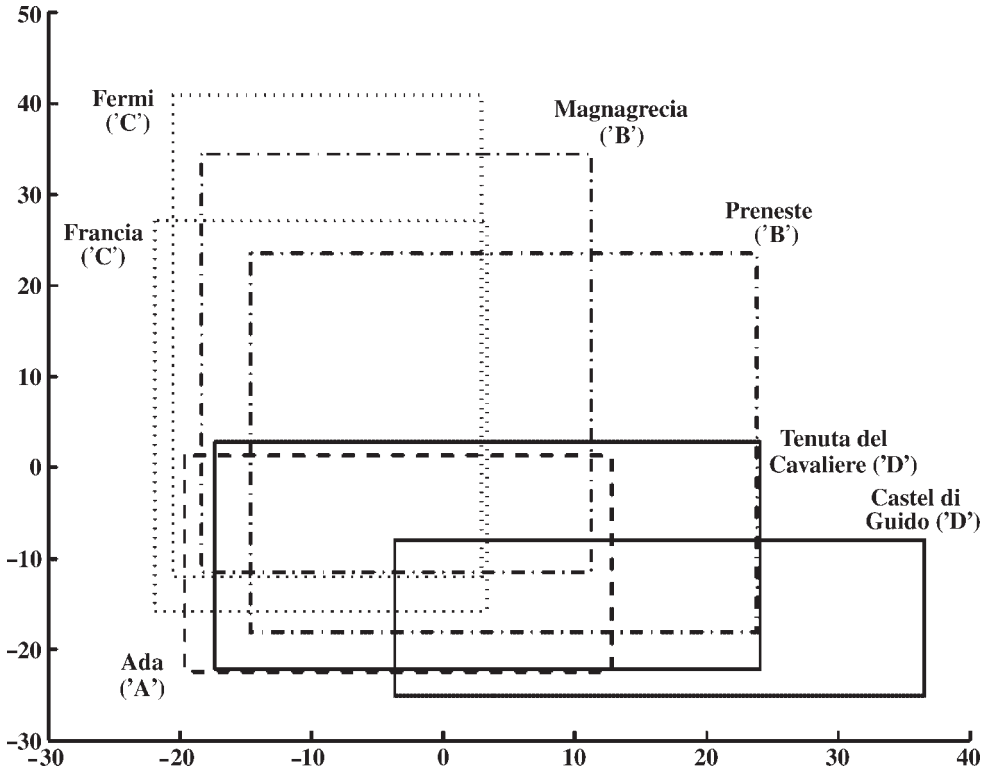

Figure 4. Low-dimensional representation of the testing stations.

are small. This shows that, during the entire day, the levels of air pollution given by primary pollutants are very low and stable. On the contrary, at 'Preneste' the differences between the minimum and maximum daily values are highest, because the pollution increases dramatically during the rush hours in comparison with the remaining time. Note that the second biggest rectangle pertains to 'Magnagrecia'. This seems consistent with the status of residential area (belonging to group 'B'). In fact, at these stations the levels of pollution are high in specific time bands and decrease strongly in others.

Further information can be extracted by observing the overlaps between pairs of rectangles. In fact, such overlaps show the similarities among the stations. 'Tenuta del Cavaliere' is in the middle between 'Castel di Guido' and 'Ada', whose rectangle is very overlapped with that pertaining to 'Tenuta del Cavaliere' with respect to the second axis. It follows that the levels of NO and NO_2 are very similar for these two stations. On the contrary, the presence of O_3 discriminates the air pollution at 'Tenuta del Cavaliere' and 'Ada'. Among the remaining rectangles, the one of 'Francia' is almost completely enclosed in that of 'Fermi' (both stations belong to group 'C'). The differences between these two stations can mainly be found in the maximal level of the primary pollutants at 'Fermi'. The rectangle pertaining to 'Magnagrecia' overlaps those of both 'Fermi' and 'Preneste'. Hence, as already noted, some features of 'Magnagrecia' are consistent with those of 'Preneste' (both stations belong to group 'B') while other features identify it as a compromise between groups 'B' and 'C'.

For the sake of completeness, we performed the 3V-PCA method on the same data set, taking into account the cross-products matrix $\mathbf{X}$ in (24). We decomposed the obtained matrix whose order is 270×270 by means of the eigendecomposition of $\mathbf{X}$. The new 'data matrix' is $\mathbf{Z} = \mathbf{\Lambda}^{1/2}\mathbf{K}'$, where $\mathbf{\Lambda}$ and $\mathbf{K}$ are the matrices containing the eigenvalues and the eigenvectors respectively. The solution is attained by running the '*N*-way toolbox for MATLAB' [25]. In order to compare the results obtained performing 3V-PCA with those obtained performing 3C-PCA (extracting the same number of components), we worked as follows. We rotated the component matrices for the variable mode and rescaled the component scores for the occasion mode so that they were as similar as possible to the ones obtained by 3C-PCA (given in Table II and Figure 2). Moreover, we transformed the core to an identity matrix so that it coincided with that obtained performing the 3C-PCA method. All these transformations were compensated for in the observation unit mode component matrix. As a result, we found the component matrix for the variables given in Table III, which can be seen to be very similar to that resulting from 3C-PCA (in Table II). The occasion component scores from 3V-PCA were also very similar to those from 3C-PCA (in Figure 3) and are therefore not given here. Finally, if we plotted the testing stations in the obtained low-dimensional space spanned by $\mathbf{F} = (\mathbf{C} \otimes \mathbf{B})\mathbf{G}_a'$ resulting from 3V-PCA, the configuration would be similar to that displayed in Figure 4. This strong similarity between the 3V-PCA and 3C-PCA results was to be expected because of the high number of occasions. Specifically, we showed

Table III. Component matrix for the variables (3V-PCA)

Pollutant	PC1	PC2
NO	−0.07	0.69
NO_2	0.04	0.72
O_3	1.00	0.01

earlier that the cross-products matrices given by the 3C-PCA and 3V-PCA methods differ only in their diagonal elements. As the number of lines of the cross-products matrix increases, the number of different elements decreases proportionally. In the application, only 0.4% of the elements (270 main diagonal elements out of 270^2 elements) differ.

6. CONCLUSION

The present paper has described methods to recover the underlying structure of three-way interval-valued data. Our methods are based on the Tucker3 and CANDECOMP/PARAFAC methods, useful to analyze three-way numerical data, and upon C-PCA and V-PCA, useful to analyze two-way interval-valued data. The proposed methods first transform the (three-way) interval-valued data into numerical data. After that, the Tucker3 or CANDECOMP/PARAFAC method, depending on the features of the data set involved, is fitted to the obtained numerical matrix. However, the three-way extension of the V-PCA method involves the computation of the cross-products matrix, which can be easily obtained. The interpretation of the solution is exactly the same as for the ordinary C-PCA and V-PCA methods, except that now we deal with three component matrices and the core matrix instead of the component scores and component loadings matrices.

An advantage of analyzing interval-valued data over analyzing single-valued data is that the missing data problem is less complex to manage in the former case. The reason is that, if some original data values are missing, an interval can still be defined. In cases where the complete interval (e.g. measurements registered during an entire day) is missing, the problem is similar to that in PCA, and we could use well-known procedures such as listwise deletion to handle such data.

Like PCA, Tucker3 and CANDECOMP/PARAFAC offer the chance to better understand the similarities among observation units by projecting them on the obtained low-dimensional space. The entities are represented as points in this subspace. The three-way extensions of C-PCA and V-PCA offer the same opportunity, but now the observation units are represented as hyperrectangles. In fact, to achieve this, we have extended in the three-way context the existing procedure for the two-way methods by finding, for each component, the smallest segment in the low-dimensional space that encloses all the projected vertices. Finally, in order to show how the proposed methods work, they are fitted to an empirical three-way data set.

In the case of interval-valued data, preprocessing reduces to preprocessing centers (as in C-PCA) or vertices (as in V-PCA). Although these preprocessing procedures can be carried out technically without any problem, their implications do not seem to have been studied yet in depth. Such an in depth study should focus on which procedure is the most convenient and meaningful way to preprocess this kind of data and on how such procedures affect the model parameters.

Finally, it should be noted that in several situations the information about the lower and the upper bound might be supplemented by means of membership functions, thus yielding 'fuzzy data'. It follows that points do not lie uniformly distributed in the hyperrectangles, but the distribution of the points depends on the membership functions of the fuzzy variables. For an overview of fuzziness, see References [26,27]. Giordani and Kiers [28] suitably extended two-way PCA to deal with fuzzy data. A generalization of three-way component methods for fuzzy data still remains to be developed.

Acknowledgements

We would like to thank the anonymous referees for their comments on an earlier version of this paper, and Giovanna Jona Lasinio Regione Lazio for making available the data used in Section 5.

REFERENCES

1. Bock HH, Diday E. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer: Heidelberg, 2000.
2. Cazes P, Chouakria A, Diday E, Schektman Y. Extension de l'analyse en composantes principales à des données de type intervalle. *Rev. Statist. Appl.* 1997; **45**: 5–24.
3. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279–311.
4. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an *n*-way generalization of 'Eckart–Young' decomposition. *Psychometrika* 1970; **35**: 283–319.
5. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an 'exploratory' multimode factor analysis. *UCLA Working Papers Phonet.* 1970; **16**: 1–84.
6. Kiers HAL. Hierarchical relations among three-way methods. *Psychometrika* 1991; **56**: 449–470.
7. Kroonenberg PM. *Three-mode Principal Component Analysis: Theory and Applications*. DSWO Press: Leiden, 1983.
8. Kiers HAL, Van Mechelen I. Three-way component analysis: principles and illustrative application. *Psychol. Methods* 2001; **6**: 84–110.
9. Harshman RA, Lundy ME. Data preprocessing and the extended PARAFAC model. In *Research Methods for Multimode Data Analysis*, Law HG, Snyder CW, Hattie JA, McDonald RP (eds). Praeger: New York, 1984; 602–642.
10. Bro R, Smilde AK. Centering and scaling in component analysis. *J. Chemometrics* 2003; **17**: 16–33.
11. Kroonenberg PM, de Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* 1980; **45**: 69–97.
12. Andersson CA, Bro R. Improving the speed of multi-way algorithms. Part I: Tucker3. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 93–103.
13. Kiers HAL. Recent developments in three-mode factor analysis: constrained three-mode factor analysis and core rotations. In *Data Science, Classification and Related Methods*, Hayashi C, Ohsumi N, Yajima K, Tanaka Y, Bock HH, Baba Y (eds). Springer-Verlag: Tokyo, 1998; 563–574.
14. Kiers HAL. Some procedures for displaying results from three-way methods. *J. Chemometrics* 2000; **14**: 151–170.
15. Appellof CJ, Davidson ER. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Anal. Chem.* 1981; **53**: 2053–2056.

16. Kiers HAL, Krijnen WP. An efficient algorithm for PARAFAC of three-way data with large numbers of observation units. *Psychometrika* 1991; **56**: 147–152.
17. Kiers HAL, Kroonenberg PM, ten Berge JMF. An efficient algorithm for TUCKALS3 with large numbers of observation units. *Psychometrika* 1992; **57**: 415–422.
18. Alsberg BK, Kvalheim OM. Speed improvement of multivariate algorithms by the postponing of basis matrix multiplication method. Part I. Principal component analysis. *Chemometrics Intell. Lab. Syst.* 1994; **24**: 31–42.
19. Alsberg BK, Kvalheim OM. Speed improvement of multivariate algorithms by the postponing of basis matrix multiplication method. Part II. Three-mode principal component analysis. *Chemometrics Intell. Lab. Syst.* 1994; **24**: 43–54.
20. Bro R, Andersson CA. Improving the speed of multi-way algorithms. Part II: Compression. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 105–113.
21. Kiers HAL, Harshman RA. Relating two proposed methods for speedup of algorithms for fitting two- and three-way principal component and related multilinear models. *Chemometrics Intell. Lab. Syst.* 1997; **36**: 31–40.
22. Carroll JD, Pruzansky S, Kruskal JB. CANDELINC: a general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika* 1980; **45**: 3–24.
23. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**: 105–122.
24. Kaiser HF. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 1958; **23**: 187–200.
25. Andersson CA, Bro R. The *N*-way toolbox for MATLAB. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 1–4.
26. Dubois D, Prade H. *Fuzzy Sets and Systems: Theory and Applications*. Academic Press: New York, 1980.
27. Zimmermann HJ. *Fuzzy Set Theory and its Applications*. Kluwer Academic: Dordrecht, 2001.
28. Giordani P, Kiers HAL. Principal component analysis of symmetric fuzzy data. *Comput. Statist. Data Anal.* 2004; **45**: 519–548.

