Multivariate Analysis of Data on Migraine Treatment

Agostino Tarsitano and Ilaria L. Amerise*

Dipartimento di Economia, Statistica e Finanza, Università della Calabria, Via Pietro Bucci, Cubo 1c, 87036 Rende (CS), Italy

Abstract: Migraineur constitutes a multidimensional model of health disorder involving a complex combination of genetic, psychological, demographic, enviromental and economic factors. This model provides a framework to describe limitations of an individual functional ability and quality of life, and to aid in the elaboration of more adequate intervention programs for each patient. Our primary objective in this paper is a data-driven profiling of patients.

The approach followed consists of examining affinity/dissimilarity between sufferers on the basis of different family of indicators and then aggregating multiple partial matrices, where each matrix expresses a particular notion of the dissimilarity of one patient from another. One important particularity of our method is the notion of multi-dimensional dissimilarity for static and dynamic indicators, without ignoring any portion of data.

The partial dissimilarity matrices are assembled in the form of a global matrix, which is used as input of subsequent calculations, such as multidimensional scaling and cluster analysis. Our main contribution is to show how multi-scale, cross-section and longitudinal data from individuals involved in a migraine treatment program may optimally be combined to allow profiling migraine-affected patients.

Keywords: Kostecki-Dillon, General dissimilarity coefficient, Cluster analysis, Multi-dimensional scaling.

1. INTRODUCTION

The study presented here is part of an industrial research and experimental development project. Its general objective is to design, develop and apply an innovative technological services platform to support the effectiveness and efficacy of the integrated clinical management of the cephalalgic patients. The partners in of the project have chosen to start a vast data collection on headaches in Calabria passing from a conventional on-demand healthcare approach centered on communication between patients and healthcare professionals, and by providing a means to match level of medical care with disease severity.

The portion of the study that we present here is a pilot research that aims to provide practical indications for designing and improving the questionnaire which will be circulated among patients, physician and healthcare professionals. To this end, we analyze the well-known data set on the severity and frequency of migraine headaches data set described by Kostecki-Dillon *et al*., 1999 [1]. This data set contains indicators measured on different scales and units of measurement. The various indicators can be partitioned into groups of variables. Our point of departure is the computation of a partial dissimilarity matrix for each group of indicators taken separately ("partial" because each of them is linked to a specific group of indicators and not to the globality of the issues reported in the data set). Successively, we use an optimization-based procedure to build a global dissimilarity matrix in the form of a weighted average of the partial dissimilarity matrices. At this point, the global dissimilarity matrix becomes the basis for metric and semi-metric scaling techniques that can be used for visual exploration of data. Additionally, it is possible to assess the presence of a group structure on the migraine sufferers, who are enrolled in the program. In this regard, both the global original dissimilarity matrix and the global dissimilarity matrices derived from a synthesis of the scaling methods are submitted to a partitioning around medoids algorithm in an attempt to identify profiles or clusters of patients that share similar needs.

The remainder of the article is organized as follows. In the next section, we address the problem of specifying differential weights for each group of indicators in order to reflect their significance, reliability and statistical adequacy. In Section 3, the global dissimilarity matrix acts as a starting point for the partitioning around medoid algorithm to cluster the patients on the basis of their static and dynamic characteristics. In Section 4, the global dissimilarity matrix between migraineurs is subjected to a multidimensional scaling (MDS) method to settle possible disagreements and reveal underlying relationships among the results. Finally, the last section provides a summary and identifies topics for future study.

*Address correspondence to this author at the Dipartimento di Economia, Statistica e Finanza, Università della Calabria, Via Pietro Bucci, Cubo 1c, 87036 Rende (CS), Italy; Tel: +39098492455; Fax: +390984492421;
E-mail: ilaria.amerise@unical.it

E-ISSN: 1929-6029/19

2. COMPUTING DISSIMILARITY MEASURES FOR MIXED DATA

Let us suppose that p different indicators regarding n patients are organized into m non-overlapping sets including m_j indicators with $\sum m_j = p$. In general, dissimilarity-based methods can properly handle problems with multiple scales including categorical, ordinal, and time-series data. Our idea is that analogies and differences between migraine sufferers can be condensed into m partial matrices $\mathbf{D}_h, h = 1,2,\ldots,m, m > 2$ of $n \times n$ order

$$\mathbf{D}_h = \begin{bmatrix} d_{1,1,h} & d_{1,2,h} & \ldots & d_{1,n,h} \\ d_{2,1,h} & d_{2,2,h} & \ldots & d_{2,n,h} \\ \ldots & \ldots & \ldots & \ldots \\ d_{n,1,m} & d_{n,2,h} & \ldots & d_{n,n,h} \end{bmatrix}, \qquad h = 1,2,\ldots,m. \qquad (1)$$

In more formal terms, we assume that each matrix is real, non negative and symmetrical, with zero diagonal and positive row sums regardless of what, and how many, indicators have been used in each group or how the matrices have been constructed.

With the usage of (1), we realize a transition from an information basis represented by a data matrix (consisting of a number of indicators taking values over a number of patients) to an information basis in terms of a dissimilarity matrix whose generic element $d_{i,j,h}$ expresses the dissimilarity between patients i and j with regard the h-th set of indicators. The main advantage of this approach is that it makes data measured with heterogenous methods and instruments comparable for multi-dimensional analysis such as clustering and scaling methods.

It is commonly observed that cluster analysis according to different variables can produce different classifications even when they are applied to the same set of subjects. Similar statement holds for multi-dimensional scaling, which can substantially differ according to the variables used. Consequently, combining multiple dissimilarity measures $\mathbf{D}_h, h = 1,2,\cdots,m$ in a global dissimilarity matrix $\mathbf{D}$ can be beneficial, provided that the mixing is done in an optimal way.

In the present paper, the global dissimilarity matrix is made up of a weighted average of the of the partial dissimilarity matrices

$$\mathbf{D} = \sum_{h=1}^{m} w_h \mathbf{D}_h \qquad with \quad w_h \geq 0; \ \sum_{h=1}^{m} w_h = 1 \qquad (2)$$

The positive sign of the weights and the linearity of (2) ensures that every variation in $d_{i,j}$ corresponds to an increase or decrease in at least one of the $d_{i,j,h}, h = 1,2,\ldots,p$. On the other hand, gain with respect to one or more partial matrix could compensate loss with respect to another group of partial matrices by the same amount of variation. We must assume that such behavior is desired or at least is not detrimental to the problem being worked; otherwise averaging over component data sources would be inappropriate.

2.1. Data Description

In this study we use the subset of data on migraine treatments collected by Kostecki-Dillon *et al*., 1999 [1] and freely available in the *KosteckiDillon* data set from the R package *carData*. The data consists of headache logs kept by n=133 patients in a treatment program in which bio-feedback was used to reduce migraine frequency and severity. Patients entered the program at different times over a period of about 3 years. Patients were encouraged to begin their logs four weeks before the onset of treatment and to continue for one month afterwards, but only 55 patients have data preceding the onset of treatment. On average, patients recorded information on 31.2 days, with the number of days ranging from 7 to 121. The variables involved are time:

- time in days relative to the onset of treatment, which occurs at time 0.
- dos: time in days from the start of the study, January 1 of the first year of the study. Note that there is a perfect correlation between “time" and “dos" (except for patient 126 were time 28 is missing). Consequently, we decided to ignore information derived from the variable “dos".
- pretreatment: a binary variable representing dichotomization of time: as 0 if time is positive and 1 if the patient began its logs before the onset of treatment.
- compliance: a binary variable coded 1 if the duration of treatment is greater than the average and 0 otherwise.
- hatype: an ordered factor with three levels: “no aura", “mixed", “aura" describing the type of migraine experienced by a subject.
- age: at onset of treatment, in years.
- airq: a measure of air quality.
- medication: an ordered factor with levels “none" (recorded as 1), “reduced" (recorded as 2), “continuing" (recorded as 3) representing

subjects who discontinued their medication, who continued but at a reduced dose, or who continued at the previous dose.

- headache: all patients were assigned a pain score recorded as 0 or 1, whereby the former indicates low pain and the latter high pain. The cutoff point is not indicated.

- gender: a factor with levels “female" and “male".

In summary, there are two type of indicators: dynamic variables: “time" (X_1), “airq" (X_2) and “headache" (X_3); static variables where only one value is observed for each subject: “age" (X_4), “pretreatment" (X_5), “compliance" (X_6), “gender" (X_7), “hatype" (X_8), “medication" (X_9)

2.2. Dissimilarity for Mixed Data

The definition of dissimilarity is crucial for the effectiveness of our approach. The dissimilarity functions proposed to compute the degree of closeness between patients are as follows

Time series data. We adopted the dynamic time warping (dtw) distance as dissimilarity function between time series of different lengths such as those involved in this study. This distance tries to find a natural peak-to-peak, valley-to-valley alignment between a pair of sequences by warping them such that, for example, the Manhattam or city-block distance between the warped time series is minimal [2].

Let $\delta^*_{i,j,h}$ be the dtw distance between patient i and j with respect to the variables “time" ($h=1$), “airq" ($h=2$) and “headache" ($h=3$). If the sequences are of different lengths, the dtw distance is not symmetrical, but a symmetrized version can be defined as $\delta_{i,j,h}=(\delta^*_{i,j,h}+\delta^*_{j,i,h})/2, h=1,2,3$. Furthermore, the dtw distance does not satisfy the triangle inequality. Ratio variables. The age of the patients yields $\delta_{i,j,4}=|x_{i,4}-x_{j,4}|$. Binary symmetric. In this case 0/0 and 1/1 matches are treated as equally indicative of similarity. We have treated together the binary indicators of the *Kostecki-Dillon* data set: “pretreat", “compliance" and “gender". The dissimilarity due to these variables is measured by the Rogers-Tanimoto coefficient, which, using the symbolism of the tetrachoric contingency table, is given by

$$\delta_{i,j,5}=\left[\frac{b_{i,j}+c_{i,j}}{0.5\left(a_{i,j}+e_{i,j}\right)+b_{i,j}+c_{i,j}}\right]^{0.5}$$

where

$$a_{i,j}=\sum_{h=5}^{7} i(x_{i,h}=1\cap x_{j,h}=1) \quad e_{i,j}=\sum_{h=5}^{7} i(x_{i,h}=1\cap x_{j,h}=0)$$

$$b_{i,j}=\sum_{h=5}^{7} i(x_{i,h}=0\cap x_{j,h}=1) \quad c_{i,j}=\sum_{h=5}^{7} i(x_{i,h}=1\cap x_{j,h}=0)$$

for $h=5,6,7$ and with $i(x)=1$ if x is true and 0 otherwise. The Rogers-Tanimoto coefficient takes into consideration the existing agreement among the subjects when factors coincide and when they are absent. Concordances receive half weights. By doing so, the coefficient actually gives slightly less emphasis to the positive matches.

Ordinal scale. The dissimilarity concerning “medication" is measured by considering quadratic contrasts *i.e.* the factor enters not only linearly but also quadratically. So, not only linear, but also quadratic effects can be captured. In particular, $\delta_{i,j,6}=P_{x_{i,9},x_{j,9}}$ where

$$P=\begin{bmatrix} 0 & 11 & 6.171573 \\ 11 & 0 & 5 \\ 6.171573 & 5 & 0 \end{bmatrix}$$

Nominal. Since only one indicator fell in this group we pose $\delta_{i,j,7}=1$ if $x_{i,8}=x_{j,8}$ and $\delta_{i,j,7}=0$ otherwise. From the $p=9$ indicators reported in the Kostecki-Dillon data set, we have derived $m=7$ set of variables, which serve to generate dissimilarity summary in form of symmetric matrices with non-negative off-diagonal elements and zero diagonal elements.

2.3. Weighting Matrices

In constructing the global dissimilarity matrix, decisions must be made about the weight to be given to each set of indicators. To this end, we need an expression for how much a certain variable affects the global dissimilarity. This can be derived from the total sum of the squares of the elements in the matrices $\mathbf{D}_h, h=1,2,\ldots,m$ choosing the weights so as to maximize the variance of the elements in $\mathbf{D}$.

Initially, each matrix $\mathbf{D}_h$ is transformed into a cross-product matrix $\mathbf{B}_h$ by an operation called double-centering

$$\mathbf{B}_h=-0.5\mathbf{C}\mathbf{D}_h^2\mathbf{C}^t, \qquad with \quad \mathbf{C}=\mathbf{I}_n-n^{-1}\mathbf{u}_n\mathbf{u}_n^t \tag{3}$$

where $\mathbf{u}_n$ is an $n\times 1$ vector of 1’s and $\mathbf{I}_n$ be the identity matrix of order n. The notation $\mathbf{D}_h^2$ stands for the matrix whose (i,j)-th element is the square of

$d_{i,j,h}$. Note that $\mathbf{B}_h\mathbf{u}_n = 0$. Therefore, double-centering reduces the rank of the original matrix by one because one of the eigenvalues of $\mathbf{B}_h$ is forced to be zero.

As a preliminary step we define the global cross-product matrix $\mathbf{B}$

$$\mathbf{B} = \sum_{h=1}^{m} w_h \mathbf{B}_h \quad with \quad w_h \geq 0, \quad and \quad \sum_{h=1}^{m} w_h = 1. \tag{4}$$

Let $\beta_h = Vec(\mathbf{B}_h), h = 1,2,\ldots,m$ be the column vector obtained by stacking the columns of $\mathbf{B}_h$ on top of one another. A standard method to measure how the $\mathbf{B}_1,\ldots,\mathbf{B}_m$ matrices resemble each other is the vector correlation

$$\frac{\beta_r^t \beta_s}{\| \beta_r \| \| \beta_s \|} = \frac{Tr\left[\left(\mathbf{B}_r^t \mathbf{B}_s\right)^t \left(\mathbf{B}_r^t \mathbf{B}_s\right)\right]}{\| \mathbf{B}_r^t \mathbf{B}_r \|_F \| \mathbf{B}_s^t \mathbf{B}_s \|_F} = a_{r,s}. \tag{5}$$

where $\| \mathbf{B}_h \|_F^2 = Tr\left(\mathbf{B}_h^t \mathbf{B}_h\right)$ is the square of the Frobenius norm of $\mathbf{B}_h$. Naturally, $a_{r,s} = a_{s,r}$ and $a_{r,r} = 1, r = 1,2,\ldots,m$ and $0 \leq a_{r,s} \leq 1$. In general, we have

$$a_{r,s} = \frac{\sum_{i=1}^{m} r^2\left(\mathbf{b}_{i,r}, \mathbf{b}_{i,s}\right) + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left[r^2\left(\mathbf{b}_{i,r}, \mathbf{b}_{j,s}\right) + r^2\left(\mathbf{b}_{i,s}, \mathbf{b}_{j,r}\right)\right]}{\sqrt{\left[m + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} r^2\left(\mathbf{b}_{i,r}, \mathbf{b}_{j,r}\right)\right]\left[m + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} r^2\left(\mathbf{b}_{i,s}, \mathbf{b}_{j,s}\right)\right]}} \tag{6}$$

where $\mathbf{b}_{i,r}, \mathbf{b}_{j,s}$ are, respectively, the i-th and the j-th column of $\mathbf{B}_r$ and $\mathbf{B}_s$ and $r^2\left(\mathbf{b}_{i,r}, \mathbf{b}_{j,s}\right)$ is the square of the Pearson correlation between them [3]. A value of $a_{r,s}$ equal zero or near zero implies that the two matrices are not linearly related, whereas a value close to one indicates a strong linear relationship between $\mathbf{B}_s$ and $\mathbf{B}_r$. If $a_{r,s} = 1$ then $\mathbf{B}_r$ can be derived from $\mathbf{B}_s$ through a homothetic transformation. It must be pointed out that $a_{r,s}$ is invariant to a linear transformations of the matrices. We remark that an exact value of zero is virtually precluded because the matrices $\mathbf{B}_h, h = 1,2,\cdots,m$ are all symmetrical with zero row and column sums, have a rank less than or equal to $(n-1)$ and, above all, are derived from distance matrices computed on the same set of subjects. Actually, since $a_{r,s}$ is positive, there will be generally an appearance of correlation produced in this way, even if no such relation really exists.

Let the vector correlations (5) be placed in a matrix $\mathbf{A} = (\beta_1, \beta_2, \cdots, \beta_m)$. The principal component analysis of $\mathbf{A}$ yields a set of m orthogonal eigenvectors, $\mathbf{q}_h, h = 1,2,\cdots,m$ and a vector of eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ corresponding to each eigenvector. In particular, the first principal component is that linear combination of the columns of $\beta_1, \beta_2, \cdots, \beta_m$ of $\mathbf{A}$ which describes the greatest amount of total variance between dissimilarities.

We can assume, without being too restrictive, that $\mathbf{A}$ is positive. Under such condition, the Perron-Frobenius theorem (see, for example, [4]) ensures that there is a single eigenvalue, say λ_1, that is positive and greater than or equal to all other eigenvalues in modulus and that there is a strictly positive eigenvector $\mathbf{q}_1$ corresponding to λ_1. A principal component with such characteristics represents some overall “profileÓ or “incidence" of a type of variables on the patients. The global cross-product matrix can now be computed by using (4) with weights $\mathbf{w} = \left(\mathbf{u}_m^t \mathbf{q}_1\right)^{-1} \mathbf{q}_1$.

Now, we can construct the related global dissimilarity matrix, but before we have to consider the role of the Euclidity of a matrix. An $(n \times n)$ matrix is said to be Euclidean if its entries reproduce exactly the distances between n points in a Euclidean space, that is, $d_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \| \; for \; 1 \leq i, j \leq n$. It has been shown that $\mathbf{D}_h$ is Euclidean if and only if the matrix $\mathbf{B}_h$ is positive semi-definite (see, for example, [5,6]). Moreover, we know that the largest eigenvalue of $\mathbf{B}_h$ is positive (see, [7]), but the others may be negative, zero or positive. In most real-data applications, the dtw function will not provide a Euclidean dissimilarity matrix, and so a correction will be needed. There exist a great number of approaches for different approximation criteria and some algebraic results (see, [8,9]), which have been widely discussed in MDS and we need not consider them here.

Let us suppose, also in reason and on the basis of the transformation referred to above, that $\mathbf{B}_h, h = 1,2,\cdots,m$, are all $n \times n$ positive semi-definite matrices with rank greater than or equal to m. Owing to this fact, a linear combination of positive semi-definite matrices is also positive semi-definite (see, for example, [10]). It follows that the global matrix of cross-products can be written as

$$\mathbf{B} = \sum_{h=1}^{m} w_h \tilde{\mathbf{B}}_h \tag{7}$$

According, for example, to [11], the corresponding global dissimilarity matrix can be defined as

Table 1: Weights for the Kostecki-Dillon Datas

Correction	time	airq	headache	age	binaries	hatype	medication
None	0.2300	0.2132	0.2103	0.0188	0.2234	0.0598	0.0445
Quasi	0.1336	0.1830	0.2583	0.0207	0.2495	0.1109	0.0440
Lingoes	0.2459	0.2496	0.2727	0.0398	0.0959	0.0478	0.0482
Cailliez	0.1477	0.2216	0.2591	0.0473	0.1895	0.0721	0.0628

$$\tilde{\mathbf{D}} = \sqrt{\sum_{h=1}^{m} w_h \tilde{\mathbf{D}}_h} \qquad with \quad \tilde{\mathbf{D}}_h = \mathbf{g}_h \mathbf{u}_m^t + \mathbf{u}_m \mathbf{g}_h^t - 2\mathbf{G}_h \qquad (8)$$

where $\mathbf{G}_h = \left(\| \mathbf{B}_h \|_F\right)^{-1} \mathbf{B}_h$ and $\mathbf{g}_h$ is composed of the elements in the diagonal of $\mathbf{G}_h$. The normalization of the cross-product matrices is necessary to make $\mathbf{B}_h$ invariant under change of scale.

The matrix $\tilde{\mathbf{D}}$ depends on the additive transformation adopted to make Euclidean some of the partial dissimilarity matrices. Three recommended correction methods are “cailliez", "lingoes" and “quasi-Euclidean", whose application yields the results reported in Table **1**.

There are no large differences between the various systems of weights. Scarce importance is given to “age", “hatype" and “medication" and this is presumably attributable to the reduced variability of the dissimilarities concerning these variables. When no correction is applied, we obtain four larger almost equal weights and weights related to “age", “hatype" and “medication" much smaller than the average of the other in the same row. The additive corrections “quasi-Euclidean" and "Cailliez" confirm the irrelevance of ‘age", “hatype" and “medication", but “time" and “binaries" see their role radically changed with respect to those reported in the first row. In view of these controversial findings, it is encouraging to see that the weight concerning “headache" in one of the largest in all the systems and this is of extreme importance because this indicator is of major interest in the current study.

The fact is that a transformation of the dissimilarity matrices is nevertheless required if we want to use relation (8). The Lingoes additive correction gives weights that are very similar to the natural weights reported in the first row, but with an important distinction: the set of indicators called "binaries" loses part of its importance in favor of “airq" and “headache", which could also be a good thing due to quantity of information being conveyed by these indicators. To simplify matters, the weights shown th the row headed “lingoes" will be those chosen for computing the global dissimilarity matrix.

3. CLUSTER ANALYSIS

Over the years, many methods have been used to find groups in data, but here we will concentrate on the partitioning around medoids (PAM) method. The PAM algorithm searches between the units of the available data $U_1, U_2, \cdots, U_n$ (which in our case are patients) for k representative units, called medoids, among the subjects of the data set. Medoids are computed such that the total dissimilarity of all subjects to the nearest medoid is minimal. In short, the goal is to select k medoids $R_1, R_2, \cdots, R_k$ that minimize the objective function

$$T\left(R_1, R_2, \cdots, R_k\right) = \sum_{i=1}^{n} \min_{j=1,2,\cdots,k} d\left(U_i, R_j\right) \qquad (9)$$

The number of possible choices for the medoids ranges between $(n/k)^k$ and $(ne/k)^k$ where e is the base of natural logarithms, which is so large that we must rely on optimization techniques.

Initially, k medoids are chosen at random from the set of n units. Each remaining unit is added to the cluster corresponding to the closest medoid: $U_i \in C_j \Rightarrow d\left(U_i, R_j\right) \le d\left(U_i, R_l\right), l = 1,2,\cdots,k$. In case of ties, the unit is assigned to one of the clusters according to their order of presentation. Each medoid is re-determined as the units for which the sum of dissimilarities to all the other units in the cluster is as small as possible.

$$R_j \Rightarrow \sum_{U_t \in C_j} d\left(U_t, R_j\right) = \min_{U_t \in C_j} \sum_{U_i \in C_j} d\left(U_t, U_i\right), \quad j = 1,2,\cdots,k \qquad (10)$$

This step, known as the “build step", is repeated until a satisfactory set of initial medoids has been found.

In a successive step, transfers of a non-medoid units are attempted. Let $e_{s,i,j} = T_s(\hat{R}_1, \hat{R}_2, \cdots, \hat{R}_k) - T_s(R_1, R_2, \cdots, R_k)$ be the amount of change that occurs in the objective function (9) if $U_s \in C_i$ is placed in C_j for $i \neq j$, $j = 1,2,\cdots,k$; $i = 1,2,\cdots,n$ and where $\hat{R}_1, \hat{R}_2, \cdots, \hat{R}_k$ are the medoids

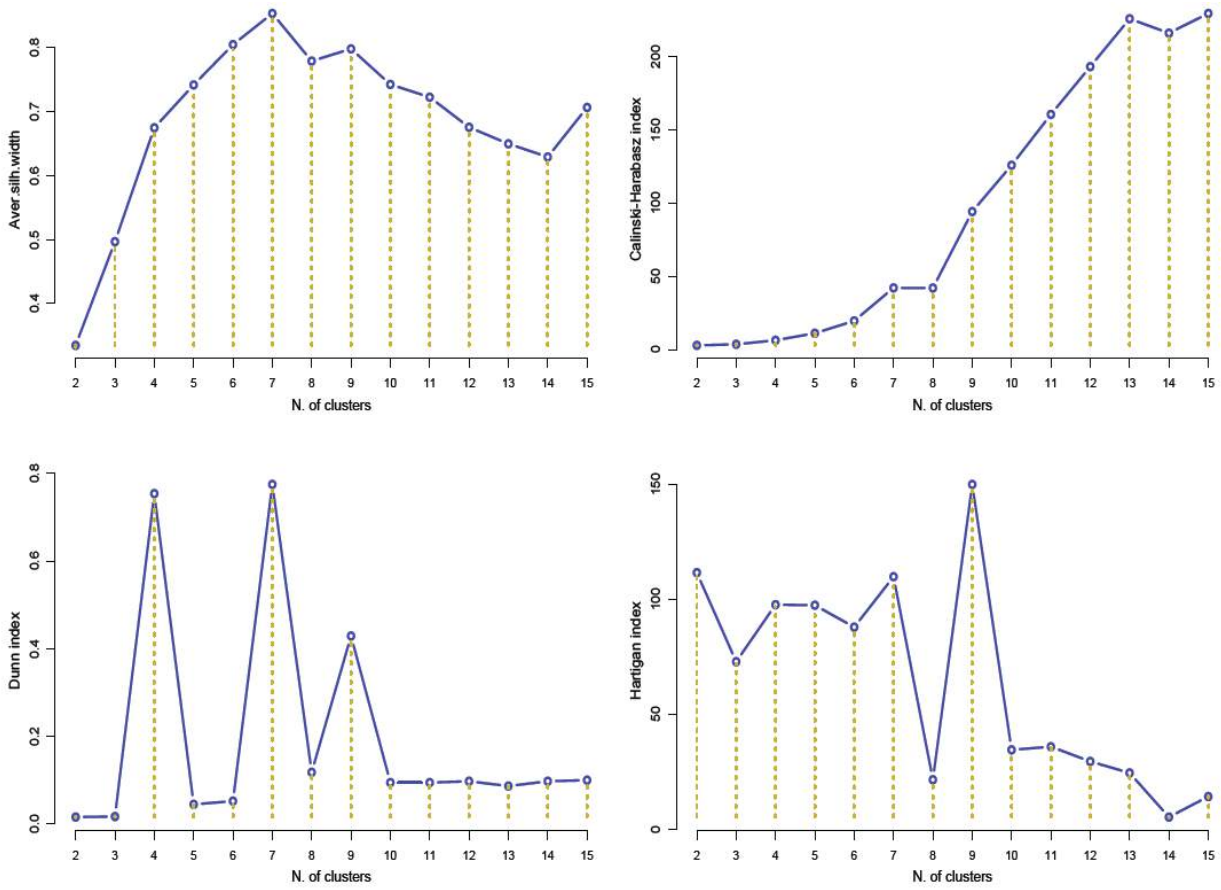

Figure 1: Choice of optimal number of clusters.

that would result after the realization of the change. If all the $e_{s,i,j} \geq 0$ then PAM algorithm stops otherwise it carries out the transfers associated with negative values of $e_{s,i,j}$ according to their order of magnitude while avoiding exchanges between clusters already involved in a transfer and leaving at least one unit in each cluster.

The number of clusters is assumed to be known. In the absence of such a priori information, a procedure is needed to find a suitable number of clusters. Therefore, it is common for the algorithm to be applied with different k and then the best solution so far obtained is selected by using a validity index. We find the optimal number of clusters by choosing the k for which a convergence is achieved between several criteria.

To help the choice of k, Figure **1** shows the plot of four indices commonly used to select the optimal number of clusters: average silhouette width, Calinski and Harabasz, Dunn and Hartigan index (see [12]). A peak in the graph of the estimated index values versus the number of clusters, indicates a candidate as the best number of clusters. The graphs in Figure **1** suggest $k = 7$ clusters, that is, the value toward which all criteria converge.

Table **2** gives the results of the PAM for a partition of the Kostechi-Dillon data set in $k = 7$ clusters.

Firstly, we note that the patients enrolled in the Kostechi-Dillon experiment are divided into two types of groups. One consists of clusters C_2, C_3, C_6, and C_7 differentiated from all remaining clusters by the gender (female) of the patient, that is, the females' behavior shows more observable instances, while for males the effects of headache illness are less straightforward and thus less observable. This is to be expected because, unlike many other chronic diseases, the morbidity attributable to these disorders is largely concentrated in otherwise healthy young and middle-aged people, particularly women.

Two clusters of female patients had a treatment duration longer than the average: C_2 and C_3. Cluster C_3 is distinct from C_2 because all the patients included in C_3 were subjected to pretreatment. The same demarcation line can be set between C_6 and C_7. The clusters C_1, C_4, C_5 are formed with male migraineurs. Those classified in C_5 had a duration of treatment for a period longer than the average including a period of pretreatment. The duration of the treatment is short than the average in the case of patients classified in

Table 2: Results of the PAM Clustering

C	Typical	Membership						
1	M63+C*0	M49+N*0 M26+C*0	M55oN*0	M32oC*0	M58-R*0	M52oC*0	M59+C*0	M46oN*0
2	F33+C*1	F36+C*1 F34oR*1 F37-R*1	F28+C*1 F39+R*1 F54oC*1	F63+R*1 F28-R*1	F62+C*1 F53+N*1	F42+C*1 F50oC*1	F48oN*1 F29oC*1	F33oR*1 F57-R*1
3	F49oC●1	F30+C●1 F50+N●1 F46+C●1 F49-C●1	F49oC●1 F21+C●1 F53oC●1 F56+C●1	F49oC●1 F52+C●1 F50oR●1 F52+N●1	F35+N●1 F60oC●1 F44oN●1 F18oN●1	F62oR●1 F66oR●1 F40oC●1 F52oC●1	F50oR●1 F46+N●1 F43oC●1 F33-C●1	F28oC●1 F35+C●1 F21+C●1 F35oC●1
4	M41oC●0	M24+R●0	M18+N●0	M44oC●0	M48oR●0	M56oC●0		
5	M33-N●1	M45oR●1	M53-N*1	M45oC●1	M43-C*1			
6	F40oC●0	F30+C●0 F46oN●0 F32oR●0 F47oR●0	F44oC●0 F32oC●0 F35+N●0	F52oC●0 F46oC●0 F23oC●0	F36+C●0 F46+R●0 F43+C●0	F42oC●0 F53oC●0 F32oC●0	F43+R●0 F27+R●0 F29+N●0	F33+C●0 F51+C●0 F60oC●0
7	F46+C*0	F43+C*0 F50oN*0 F36oC*0 F47+C*0 F48+C*0 F54oC*0 F41oC*0	F43+C*0 F33oC*0 F46oC*0 F36+R*0 F40+C*0 F29+C*0	F47+C*0 F60oN*0 F32+C*0 F51oC*0 F55+R*0 F43oN*0	F24oR*0 F46oC*0 F45+C*0 F34oC*0 F51+C*0 F42oC*0	F27oN*0 F46oC*0 F55oN*0 F41+C*0 F62+C*0 F46+R*0	F21+N*0 F54oC*0 F18-C*0 F24oN*0 F35+C*0 F24oN*0	F38+C*0 F30+R*0 F45+C*0 F51oR*0 F36-C*0 F20-C*0

Legend: hatype: "+" aura, "-" no aura, "o" mixed; medication: "C" continuing, "R" reduced, "N" none; pretreatment: "¥" yes, "*" no; compliance: "1" duration of treatment greater than the average, "0" duration lesser than the average.

clusters C_4, even though all have had a phase of pretreatment. Patients belonging to C_1 are male patients showing low compliance.

The PAM group structure is not confirmed by the best hierarchical partition obtained with the Ward link. The dendrogram in Figure **2** indicates clusters which have little in common with the clustering in Table **2**.

Although the data lend themselves to a tree description, the separation of the groups is not as clear as that achieved with the PAM clustering. The considerable divergence between iterative partitioning and hierarchical clustering can be explained by the dominance of the dissimilarities across the time series that describe the patient pathways in the current study. In effect, the weights of the dynamic indicators are much more bigger than the others and the proximity measure by symmetrized dtw might have been influenced by the very fact that the patient records follow affine periods and duration so they are near but not similar.

4. CLASSICAL MULTIDIMENSIONAL SCALING

The underlying hypothesis of Section 4 is that any set of data, even time-sequences, can be utilized for building partial dissimilarity matrices that are successively aggregated in a global dissimilarity matrix by a trade-off strategy. The resulting matrix should be less sensitive to distortion and noise in the input data than single partial matrices. We have seen, however, that hierarchical and non-hierarchical cluster analysis can lead to surprisingly different solutions even when using exactly the same data. The existing incongruences suggest that some improvement of the analysis of the global dissimilarity matrix is desirable.

The multi-dimensional Euclidean coordinate space offers an appropriate environment for a broad variety of different standard and special tools for classification, but such a geometric representation requires the availability of variables measured on a ratio scale. This is not possible in the case of the Kostecki-Dillon data set, not only because of the involvement of nominal, binary and ordinal indicators, but also because of time series data, which are difficult to reconcile with the

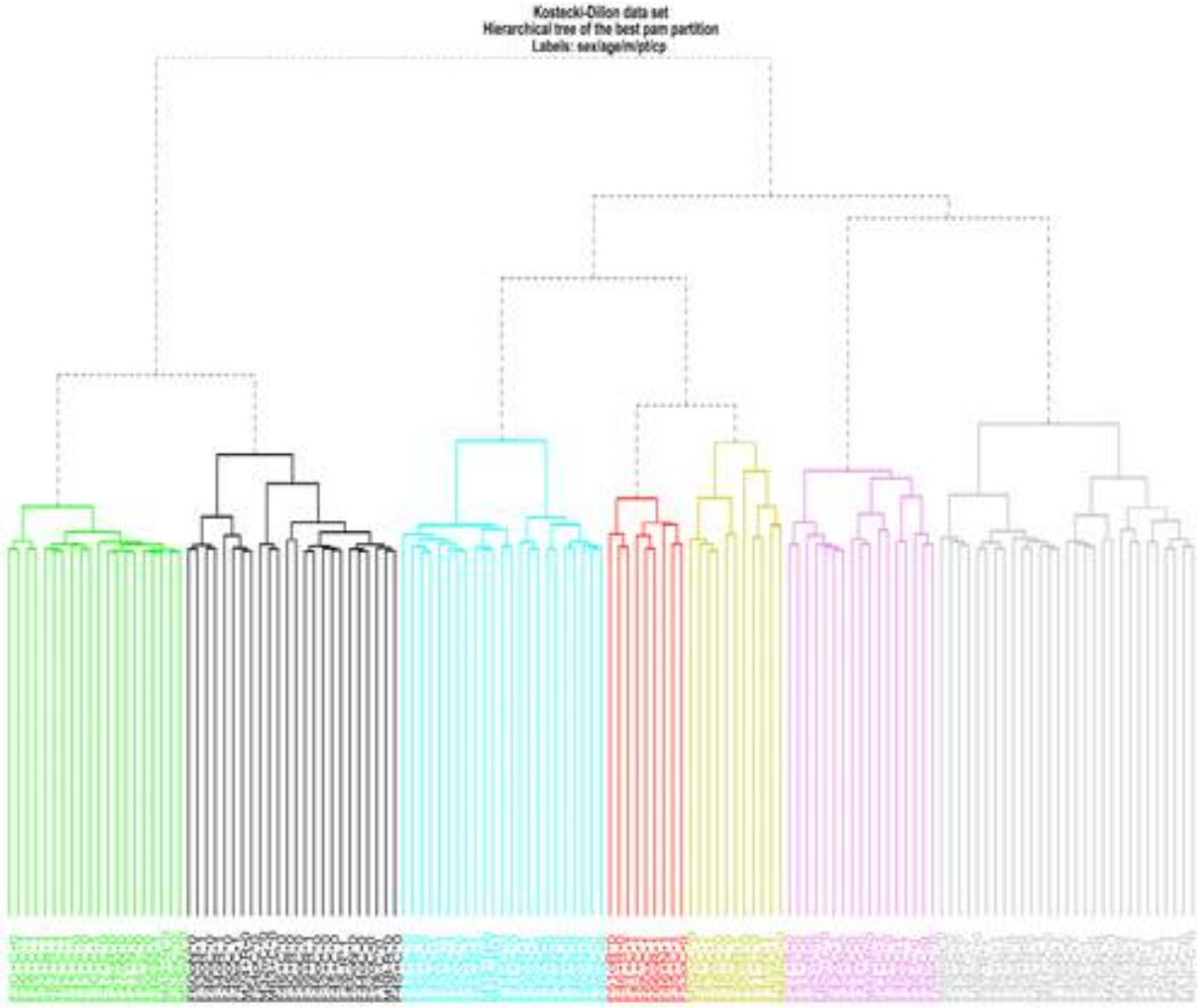

Figure 2: Hierarchical group structure of the Kostecki-Dillon data set.

concept of a data matrix in which the rows correspond to the patients in the experiment and the columns to the variables measured on them. We conjecture that the findings in Section 4 may not be entirely satisfactory due to redundancy of information about not significant aspects of patients' experience. Thus an application of a dimension reduction technique is proposed to convert a high dimensional data in form of dissimilarities between patients to two- or three-dimensional scatter plot of points, which are easily understandable.

In this section, we first embed the data into a Euclidean space, then applies the clustering algorithms to the pseudo-coordinates obtained by a classical multi-dimensional scaling (MDS) method [13]. Preliminarily, we note that the following ordering exists on the observed global dissimilarities

$$d_{1,2} \leq \cdots \leq d_{1,n} \leq \cdots \leq d_{2,n} \leq \cdots \leq d_{n-1,n} \tag{11}$$

Whenever any set of $n(n-1)/2$ non-negative numbers satisfy these inequalities, we shall say that they are monotonically related to the observed dissimilarities. The purpose of MDS is to derive a set of distances between points in a space of dimensionality $\nu < n$ from information about the dissimilarity between patients. A possible choice is $\Delta_{r,s} = f(d_{r,s}), r,s = 1,2,\cdots,n$ where f satisfies the monotonicity constraint

$$\Delta_{r,s} \leq \Delta_{r',s'} \quad \textit{if and only if} \quad d_{r,s} \leq d_{r',s'} \tag{12}$$

Relation (12) means that if r and s are less dissimilar than r' and s', then the points representing r and s must be nearer than the points representing r' and s'. Lingoes, 1971 [14] gives a rigorous and greatly simplified proof that at most $(n-2)$ dimensions are required to reproduce the order information for both the distance and dissimilarity matrices. Rivas Moya, 2000 [15] observes that, even though it is always possible to obtain a perfect reproduction of all dissimilarities in an Euclidean space of $(n-2)$ dimensions (provided that $\mathbf{D}$ is an Euclidean distance matrix), in practice, a representation in a low-

dimensional space, $\nu < n$, is ordinarily desired. Thus it is not always guaranteed that a perfectly monotonic dissimilarity-distance relation can be found. In other words, constraint (12) may be sometimes violated.

Let $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n$ be vectors of hypothetical observations on ν pseudo-indicators and let the distance function between points r and s be the Euclidean metric

$$\Delta_{i,j} = \left[\sum_{i=1}^{\nu} | y_{r,i} - y_{s,i} |^2 \right]^{0.5} \tag{13}$$

The coordinates of the points $\mathbf{Y}$ are the unknowns of the problem. Without loss of generality we set the center of the coordinate system at $(0,0,\cdots,0)$. If the coordinates are determined such that the corresponding distances $\Delta_{r,s}$ satisfy (12), then the multi-dimensional scaling problem is solved. If not, we need an index that quantifies the degree in which the relation between $\Delta_{r,s}$ and $d_{r,s}$ is not monotone.

A very commonly used index is the standardized and unit-free Stress (STandardized REsidual Sum of Squares) function

$$S(\mathbf{Y}) = \frac{\sum_{r<s}^{n} \left(\Delta_{r,s}(\mathbf{Y}) - d_{r,s} \right)^2}{\sum_{r<s}^{n} d_{r,s}^2} \tag{14}$$

where $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n$. The values of (14) are always between zero and one. Lower values of $S(\mathbf{Y})$ denote better fit in the lower-dimensional space. Any value less than 0.10 is typically taken to mean that there is a good representation of the dissimilarities by the points in the given configuration. The $\Delta_{r,s}, r,s = 1,2,\cdots,n$ must be calculated, in such a way that they verify (12) and minimize the Stress (14). Given the global dissimilarities $\tilde{d}_{i,j}, i,j = 12,\cdots,n$ in (8), the denominator of $S(\mathbf{Y})$ is fixed and, therefore, the Stress is minimized by solving the following optimization problem

$$\min_{\mathbf{Y}} \sum_{r<s}^{n} \left(\Delta_{r,s}(\mathbf{Y}) - \tilde{d}_{r,s} \right)^2 \quad subject\ to \quad \tilde{d}_{r,s} \le \tilde{d}_{r',s'} \rightarrow \Delta_{r,s}(\mathbf{Y}) \le \Delta_{r',s'}(\mathbf{Y}) \tag{15}$$

A reasonable way to solve (15) is to exploit the euclidity of $\mathbf{D}$ in (8) and apply formula (3)

$$\hat{\mathbf{B}} = -0.5\mathbf{C}\tilde{\mathbf{D}}^2\mathbf{C}^t, \qquad with \quad \mathbf{C} = \mathbf{I}_n - n^{-1}\mathbf{u}_n\mathbf{u}_n^t \tag{16}$$

Now $\hat{\mathbf{B}}$ is symmetric, positive semi-definite and of rank $(n-1)$ and hence can be written in terms of its singular value decomposition $\hat{\mathbf{B}} = \mathbf{V}\mathbf{L}^{0.5}\mathbf{V}^t$ where $\mathbf{L} = diag(l_1 \ge l_2 \ge \cdots \ge l_{n-1})$ is the diagonal matrix of the eigenvalues of $\hat{\mathbf{B}}$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{n-1}]$ is the matrix of corresponding eigenvectors, normalized such that $\mathbf{v}_i^t\mathbf{v}_i = 1$. The eigenvectors $\hat{\mathbf{Y}} = \mathbf{L}^{0.5}\mathbf{V}^t$ arranged in decreasing order of their corresponding eigenvalues and standardized so that the sum of squares of their elements equals the relevant eigenvalue, provide the solution of (15). The rows of $\hat{\mathbf{Y}}$ are called the principal coordinates of the pseudo-variables in ν dimensions.

We have performed MDS by using the built-in function “cmdscaleÓ for classical metric scaling. in the “R" environment. The usual controls on the eigenvalues suggest $\nu = 5$ principal coordinates. Having converted the global dissimilarity matrix $\mathbf{D}$ to an $n \times 5$ matrix of pseudo coordinates $\hat{\mathbf{Y}}$, it is possible to treat $\hat{\mathbf{Y}}$ as a “data matrix" for input to cluster analysis. Table **3** reports the cluster membership derived from PAM clustering with $k = 9$ groups as indicated by the same criteria used in Figure **1**. We note that the main changes are the splits of two of the old clusters described in Table **3**. Specifically, the old cluster C_3 is subdivided into two new clusters $C_3^{'}$ and $C_4^{'}$, the first focusing on female patients, who undergo a treatment duration longer than the average, had a pretreatment history and a migraine of the mixed aura subtype; and, the second including patients reporting migraine with aura.

Another major change is the split of the old cluster C_7 into two new clusters: $C_8^{'}$ containing female patients, who undergo a treatment duration shorter than the average, had not had a pretreatment history and reported a migraine of the aura subtype; the other $C_9^{'}$, grouping together patients who are distinct from those in $C_8^{'}$ as the migraine type, which is, prevalently, of the mixed aura subtype. Other minor changes are observed in the membership of the old clusters C_1, C_2, C_4, C_5, C_7. Figure **3** shows the projections of the principal coordinates in a bi-dimensional space so that the configuration can be plotted easily. The passage from a 7-cluster solution to a 9-cluster solution resulted in a configuration slightly more confused in its inability fully to separate between the patients according to gender, pretreatment and duration of the treatment. For example, patient M52oC*0 from the old cluster C_1 to the new cluster $C_5^{'}$ and patient F23oC¥0 from C_6 to $C_2^{'}$. These individuals are not consistent with the cluster to which they belong. On the

Table 3: Results of the PAM Clustering of Principal Coordinates

C'	Typical	Membership						
1	M63+C*0	M49+N*0	M55oN*0	M32oC*0	M58-R*0	M26+C*0	M59+C*0	M46oN*0
2	F33+C*1	F36+C*1	F28+C*1	F63+R*1	F62+C*1	F42+C*1	F48oN*1	F33oR*1
		F34oR*1	F39+R*1	F28-R*1	F53+N*1	F50oC*1	F29oC*1	F57-R*1
		F23oC●0	F37-R*1	F54oC*1				
3	F21+C●1	F49oC●1	F49oC●1	F62oR●1	F50oR●1	F28oC●1	F60oC●1	F66oR●1
		F53oC●1	F50oR●1	F44oN●1	F40oC●1	F43oC●1	F49oC●1	
		F18oN●1	F52oC●1	F35oC●1				
4	F49oC●1	F30+C●1	F35+N●1	F50+N●1	F21+C●1	F52+C●1	F46+N●1	F21+C●1
		F49-C●1	F56+C●1	F52+N●1	F35+C●1	F33+C●0	F33-C●1	F46+C●1
5	M41oC●0	M24+R●0	M18+N●0	M44oC●0	M48oR●0	M56oC●0	M52oC*0	
6	M33-N●1	M45oR●1	M53-N*1	M45oC●1	M43-C*1			
7	F40oC●0	F30+C●0	F44oC●0	F52oC●0	F36+C●0	F42oC●0	F43+R●0	F51+C●0
		F46oN●0	F32oC●0	F46oC●0	F46+R●0	F53oC●0	F27+R●0	F60oC●0
		F32oC●0	F35+N●0	F29+N●0	F46+R*0	F32oR●0	F47oR●0	
8	F48+C*0	F43+C*0	F43+C*0	F47+C*0	F21+N*0	F38+C*0	F46+C*0	F30+R*0
		F32+C*0	F45+C*0	F45+C*0	F47+C*0	F36+R*0	F41+C*0	F48+C*0
		F40+C*0	F55+R*0	F51+C*0	F62+C*0	F35+C*0	F29+C*0	F43+C●0
9	F24oN*0	F24oR*0	F27oN*0	F50oN*0	F33oC*0	F60oN*0	F46oC*0	F46oC*0
		F54oC*0	F36oC*0	F46oC*0	F55oN*0	F18-C*0	F51oC*0	F34oC*0
		F24oN*0	F51oR*0	F36-C*0	F54oC*0	F43oN*0	F42oC*0	F24oN*0
		F20-C*0	F41oC*0					

Legend: hatype: "+" aura, "-" no aura, "o" mixed; medication: "C" continuing, "R" reduced, "N" none; pretreatment: "¥" yes, "*" no; compliance: "1" duration of treatment greater than the average, "0" duration lesser than the average.

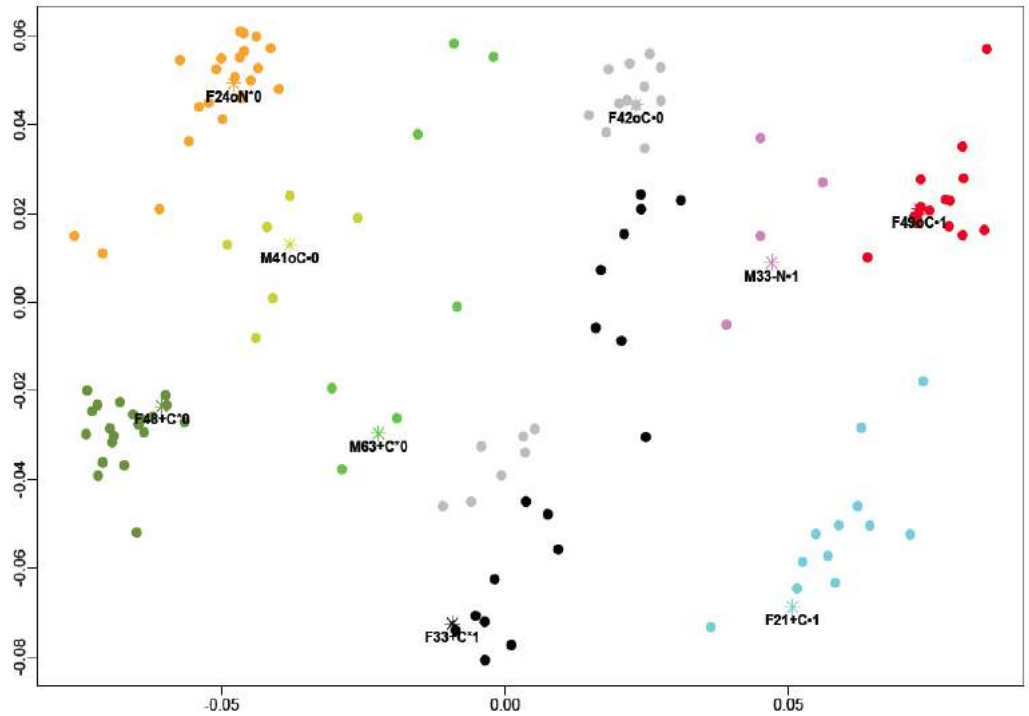

Figure 3: Scatter plot of patients plotted against the first principal coordinate by the second principal coordinate.

other hand, the splits of the two old clusters C_3 into $(C_3^{'}, C_4^{'})$ and C_7 into $(C_8^{'}, C_9^{'})$ follow along the migraine type thus involving a more specific indicator.

5. CONCLUDING REMARKS

Many research problems require statistical methods that must consider multiple dissimilarity matrices each of which offers a distinct point of view of the pairwise dissimilarities between the same set of subjects. This paper presents an approach aimed at the design of a procedure for aggregating the various sources of information for overcoming the discrepancies between different sources, particularly when the data set indicators have mixed types. This becomes a particular challenge in the case of longitudinal data concerning migraine-affected patients. We were given a foretaste of this by analyzing the well-known [1] data set in which time-series data co-exist with a mixture of numeric, ordinal, binary and nominal indicators.

In this work, we have devised a new global distance function based on the partial distance matrices between individuals obtained from various types of indicators observed on each patient. The partial distance matrices are combined as a weighted average and the resulting global distance matrix is then used for profiling patients by subgroups to enable personalized treatment. Classification of patients belonging to the Kostechi-Dillon data set has been obtained by combining PAM clustering and classical multidimensional scaling. The two methods have identified three clusters that deserve to be taken seriously into consideration. 1) Female patients, who undergo a treatment duration longer than the average, had a pretreatment history and a migraine of the mixed aura subtype. 2) Female patients, who undergo a treatment duration shorter than the average, had not had a pretreatment history and reported a migraine of the aura subtype. 3) Patients reporting migraine with aura.

The achievements of the research project will highlight the degree to which and in what ways the preliminary multivariate statistical analysis carried out in this paper can be of help for effectiveness and efficacy of data collection and the integrated clinical management of the cefalalgic patients.

REFERENCES

[1] Kostecki-Dillon T, Monette G, Wong P. Pine trees, comas and migraines. Newsletter. York University Institute for Social Research 2018; 14: 1-4.

[2] Giorgino T. Computing and visualizing dynamic time warping alignments in R: the dtw package. Journal of Statistical Software 2009; 31: 1-24. https://doi.org/10.18637/jss.v031.i07

[3] Amerise IL, Tarsitano A. Combining dissimilarity matrices by using rank correlations. Computational Statistics 2016; 31: 353-367. https://doi.org/10.1007/s00180-015-0590-x

[4] Lin KY. An elementary proof of the Perron-Frobenius theorem for non-negative symmetric matrices. Chinese Journal of Phisics 1977; 15: 283-285.

[5] Schoenberg IJ. Metric spaces and positive definite functions. Transactions of the American Mathematical Society 1938; 44: 522-536. https://doi.org/10.1090/S0002-9947-1938-1501980-0

[6] Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 1966; 5: 325-338. https://doi.org/10.1093/biomet/53.3-4.325

[7] Zhang F, Zhang Q. Eigenvalue inequalities for matrix product. IEEE Transactions on Automatic Control 2006; 51: 1506-1509. https://doi.org/10.1109/TAC.2006.880787

[8] Mardia KV. Some properties of classical multi-dimesional scaling. Communications in Statistics-Theory and Methods 1978; 13: 1233-1241. https://doi.org/10.1080/03610927808827707

[9] Bénasséni J, Dosse MB, Joly S. On a general transformation making a dissimilarity matrix Euclidean. Journal of Classification 2007; 24: 33-51. https://doi.org/10.1007/s00357-007-0005-y

[10] Caillez F, Kuntz P. A contribution to the study of the metric and Euclidean structures of dissimilarities. Psychometrika 1996; 61: 24-253. https://doi.org/10.1007/BF02294337

[11] Gower JC. Euclidean distance geometry. The Mathematical Scientist 1982; 7: 1-14.

[12] Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software 2014; 61: 1-36. https://doi.org/10.18637/jss.v061.i06

[13] Cox TF, Cox MAA. Multidimensional Scaling. 2nd Edition. Chaoman & Hall, Boca Raton FL, USA 2001. https://doi.org/10.1201/9781420036121

[14] Lingoes JC. Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika 1971; 36: 195-203. https://doi.org/10.1007/BF02291398

[15] Rivas Moya T. Calculating isotonic regression of the distance function in nonmetric multidimensional scaling model. Methods of Psychological Research Online 2000; 5. http://www.mpr-online.de

Received on 20-06-2019 Accepted on 12-07-2019 Published on 26-08-2019

https://doi.org/10.6000/1929-6029.2019.08.06