COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 131 (2016) 97–110

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Multidimensional scaling analysis of virus diseases

António M. Lopes [a,*], José P. Andrade [b], J.A. Tenreiro Machado [c]

[a] UISPA—LAETA/INEGI, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[b] Department of Anatomy, Faculty of Medicine, University of Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal

[c] Institute of Engineering, Department of Electrical Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal

ARTICLE INFO

Article history:
Received 20 August 2015
Received in revised form
26 February 2016
Accepted 30 March 2016

Keywords:
Multidimensional scaling
Clustering
Virus diseases

ABSTRACT

Background and Objective: Viruses are infectious agents that replicate inside organisms and reveal a plethora of distinct characteristics. Viral infections spread in many ways, but often have devastating consequences and represent a huge danger for public health. It is important to design statistical and computational techniques capable of handling the available data and highlighting the most important features.

Methods: This paper reviews the quantitative and qualitative behaviour of 22 infectious diseases caused by viruses. The information is compared and visualized by means of the multidimensional scaling technique.

Results: The results are robust to uncertainties in the data and revealed to be consistent with clinical practice.

Conclusions: The paper shows that the proposed methodology may represent a solid mathematical tool to tackle a larger number of virus and additional information about these infectious agents.

1. Introduction

Viruses exert enormous damage on humans worldwide and are the single most important cause of infectious morbidity and mortality. History was, and still is, shaped since ancient times by viral diseases. These diseases began to be characterized in the 19th century leading to the identification and differentiation of many viral illnesses [1]. The first viruses were identified at the end of the 19th century and since then the process of discovery has continued steadily with a growing momentum in these years. In fact, in recent years it is possible to visualize viral structure at an atomic level of resolution, nucleotide sequences of viral genomes are known, and functional domains of numerous viruses and enzymes have been established [1,2]. This information is now being applied to the development of diagnostic tools and effective antiviral therapies.

The classification of viruses has also evolved. Firstly, subclassifications were based on pathologic features such as the preference of a specific organ (for example, the liver in viral hepatitis). Secondly, some epidemiologic characteristics were defined as the transmission by arthropods (arbovirus, for example) [1]. The current classifications are based on the type and structure of the viral nucleic acid and its replication strategy, the symmetry type of the capsid of the virus, and the presence or absence of a lipid envelop [1,2].

* *Corresponding author.* UISPA—LAETA/INEGI, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal. Tel.: +351 220413486; fax. +351 225081445.
E-mail address: aml@fe.up.pt (A.M. Lopes).

http://dx.doi.org/10.1016/j.cmpb.2016.03.029
0169-2607/

More than 2000 species of viruses have been identified and approximately 650 are capable of infecting humans and animals [2]. Diseases can range from the common cold to fatal events such as Ebola, Smallpox or Rabies [2]. Globally, viral diseases are very diverse and present several degrees of complexity.

In this study we will adopt multidimensional scaling (MDS) to visualize the relationships between 22 selected human viral infectious diseases. Some viruses were selected based on recent viral outbreaks and presence in the media (for example, Influenza A virus subtype H5N1, Ebola and Chikungunya), others were chosen due to historical reasons (for example, Rabies, Poliomyelitis and Smallpox), and still others due to their prevalence and incidence in human populations (for example, Influenza, Rhinovirus and Norovirus). In two viral diseases (Human Immunodeficiency Virus and Rabies) we consider both the treated and untreated paradigms of the disease due to the huge discrepancy in mortality.

MDS is proven to obtain a new perspective on visualizing global data associated with human pathologies. MDS is a set of techniques used to analyse similarities in data that produce spatial or geometric representations of complex objects [3–5]. MDS had its origin in behavioural sciences for its help in understanding judgements of individuals (as preference, or relatedness) concerning elements in a set of objects [6–8]. Nowadays, MDS is used with a large variety of real data, such as biological taxonomy [9–12], finance [13,14], marketing [15], sociology [16], physics [17], geophysics [18–20], communication networks [21,22], biology and biomedics [23,24], among others [25,26].

Bearing these ideas in mind, the paper is organized as follows. In Section 2 we present the MDS technique. In Section 3 we study and compare data regarding 22 virus diseases. Finally, in Section 4 we draw the main conclusions.

2. Multidimensional scaling

Given s objects in a m-dimensional space and a measure of proximity, δ_{ij}, between objects i and j, a symmetric $s \times s$ matrix, $\mathbf{C} = [\delta_{ij}]$, of item to item (dis)similarities is calculated in a first step. The MDS algorithm produces a $s \times q$ ($q < m$) configuration, $\mathbf{X}$, representing point coordinates (items), where q is specified by the user. Thus, row i from matrix $\mathbf{X}$ gives the coordinates of object i in the q-dimensional embedding space. Configuration $\mathbf{X}$ preserves, as best as possible, the proximities between pairwise elements in the higher m-dimensional space and unveils the underlying data structure. MDS is, consequently, different from other similar techniques, such as factor and cluster analysis, because there are no assumptions concerning which factors might drive each dimension. Additionally, MDS is able to treat distinct types of data, has better convergence rates, and is less complex than other methods [3,27].

In order to arrive at the best configuration $\mathbf{X}$, MDS evaluates different alternative configurations while minimizing a goodness-of-fit function. This problem, equivalent to minimizing the raw stress function, σ^2, can be formulated as [28]:

$$\sigma^2 = \sum_{i=2}^{s} \sum_{j=1}^{i-1} z_{ij} \left(\delta_{ij} - d_{ij}\right)^2, \tag{1}$$

where z_{ij} is a user chosen non-negative weight and d_{ij} is a measure of the (dis)similarities among the items in the embedding space. Therefore, d_{ij} is usually a distance measure. Smaller (larger) distances between two objects translate into more (less) similarities between them. For example, the Minkowski distance provides a general way to specify distance for quantitative data in a multidimensional space:

$$d_{ij} = \left(\sum_{k=1}^{q} \alpha_k \left| x_{ik} - x_{jk} \right|^r \right)^{1/r}, \quad r \geq 1, \tag{2}$$

where x_{ik} is the value of dimension k for object i and α_k is a weight factor. When $\alpha_k = 1$, the Euclidean and the city-block distances are obtained for $r = 2$ and $r = 1$, respectively. Nevertheless, the MDS technique allows users to choose other metrics for the comparison of objects that can be better adequate for their data. In the sequel we will adopt the Canberra distance and the cosine correlation.

There are different stress measures, such as the normalized raw stress, which is σ^2 divided by the sum of squared dissimilarities. Possible alternatives are Kruskal's stress-1 and Kruskal's stress-2, which divide σ by the sum of squared distances, or by a function of the variances of distances, respectively. Another example is the S-stress measure given by the sum of squared errors between squared distances and squared dissimilarities [29,30].

The Shepard diagram is used to infer the quality of the MDS solution. Let p_{ij} denote the similarities between objects i and j. A Shepard diagram consists of pairs (p_{ij}, p_{ij}) and (p_{ij}, δ_{ij}). If a line connecting the pairs (p_{ij}, δ_{ij}) is drawn, then the approximation error, concerning dissimilarities of each object, is given by $d_{ij} - \delta_{ij}$. The Shepard diagram is thus useful for visualizing the residuals and outliers resulting from the MDS application to the data. A narrow scatter around the 45 degree line indicates a good fit between d_{ij} and δ_{ij}.

The stress plot represents σ^2 versus the number of dimensions q of the MDS maps. Usually, we get a monotonic decreasing chart and we choose q as a compromise between reducing σ^2 and having a low dimension for the MDS charts.

MDS can be divided according to the classification of data similarities, the number of similarity matrices and the nature of the MDS model. We thus have the non-metric, or metric MDS, if similarity data are qualitative or quantitative. In what concerns the number of similarity matrices and nature of the model we have classical MDS (i.e., with one matrix and unweighted models), replicated MDS (i.e., with several matrices and unweighted models) and weighted MDS (i.e., with several matrices and weighted models).

The MDS interpretation is based on the emerging clusters and distances between points in the map, rather than on their absolute coordinates, or the geometrical form of the locus. Thus, we can rotate or translate the MDS chart since the distances between points remain identical. Usually, two or three dimensional charts are selected, because they allow a direct graphical representation.

MDS has advantages over other methods, such as principal component analysis (PCA), since MDS can follow similarity/dissimilarity matrices based on several distinct metrics. MDS uses the inter-object distances rather than the coordinates of