Supplementary material

We clustered patients using the approximate singular value-based tensor decomposition (ASVTD) method described in Ruffini et al 2017[15]. The method takes as input a table where each row corresponds to a patient and each column to an observed variable on patients, such as a diagnostic, a clinical result, demographics such sex and age, etc. plus a number k of desired clusters. It returns the description of the k clusters found, where each cluster is described by the average value of each observed variable in it.

Most methods used for clustering require a notion of similarity (equivalently, a distance) among patients, and then define clusters so that the intra-cluster similarity is high and the inter-cluster similarity is low; this is the case, for example, for k-means, k-medoids (also known as PAM), and dendogram or hierarchical methods. In contrast, ASVTD takes a probabilistic approach: It assumes that data is generated as a mixture of k unknown populations (the clusters), and finds the descriptions of the k populations whose mixture makes the observed data most likely.

More precisely, ASVTD assumes that there is a set of N observed variables, and a single unobserved or latent variable that takes k possible values. Each unobserved value creates a cluster. Naturally, the observed variables are correlated in arbitrary ways in the data. The main assumption is that, when one fixes the value of the hidden variable (= fixes a cluster), the observed variables become all independent within each cluster, and in particular uncorrelated. The method then finds the partitioning of the instances in k clusters that follows this assumption most closely, that is that makes the observed variables (almost) uncorrelated within each cluster. Then an instance, in the dataset or out of it, can be assigned to the cluster that generates it with the highest probability.

An intuitive explanation of why this strategy makes sense is as follows. Suppose that after partitioning the instances in clusters, two of the observed variables (say, two diagnostics) are still correlated within a cluster. This means that patients in this cluster tend to either have both diagnostics, or to have neither of the two. This in turns means that this cluster can be reasonably subdivided into two clusters: That with patients that have both diagnostics, and that with patients that have neither. Therefore, the clustering is not optimal. Only when all observed variables are independent within every cluster, there is no way of further splitting the clusters more finely.

Compared to similarity-based methods, ASVTD avoids the complicated decision of which distance or similarity function to use, which risks adding a-priori assumptions on the relevant variables. Often, one uses by default the Euclidean distance; this considers equally all attributes, and works badly in high-dimensional data, and especially in the presence of noisy or irrelevant attributes. This does not happen in ASVTD: Irrelevant attributes are not helpful to explain any separation of the data, and are therefore not used in the assignment of instances to clusters. Finally, ASVTD has the potential of creating “all the rest” clusters, collecting the instances that do not fit any of the clear patterns captured by other clusters; this is difficult to do with similarity-based methods.

In ASVTD, the task of choosing a final number of clusters is left to the user. There are mathematical approaches to choosing an optimal number, such as BIC or AIC criteria. In this paper we have used clinical relevance and interpretability as a less formal criterion.

The MATE tool of Amalfi Analytics (www.amalfianalytics.com) has been used in this paper. It implements an extended version of the method in [15] that in particular that can deal with continuous and categorical variables in addition to binary ones.