Master thesis proposal
Czekanowski diagrams: simple visualization of distances

Krzysztof Bartoszek

November 6, 2018

Background—hierarchical clustering

The results of hierarchical clustering algorithms can be visualized using various methods. A common one is a tree representation of the data. Alternatively one uses only the distance information between objects and draws a heatmap. Unfortunately, these methods can suffer from an excess of information—it can be difficult for the human eye to spot the key characteristics of the hierarchy. Therefore, methods based on ordering/clustering and in turn reducing the amount presented information were developed. In fact the first such taxonomic based method was proposed in 1909 by Jan Czekanowski [2]. Due to the era of when it was proposed only black and white visualization was possible. Today, this is still relevant as sometimes only grayscale graphics are permissible (e.g. for printed publications). This visualization method is mostly used in anthropology, geography or econometrics.

Thesis project

Currently, black–and–white Czekanowski diagrams (see Fig. 1) are produced only by the program MaCzek [6]. The program runs only under Windows, does not seem to under development and is limited to 250 observations and 100 dimensions. Therefore, the first aim of the project is to build an `R`–package, publicly available on CRAN, that provides the functionality of producing Czekanowski diagrams for user defined distance functions. Of course the package should implement a number of typical distance functions. The diagram is created using the following steps

1. Calculate the distance between all objects.
2. Divide the calculated distances into k classes.
3. Assign a graphical symbol to each of the distances classes.
4. Graphically represent the distance matrix.

As a result of the above procedure one obtains an *unordered* Czekanowski diagram.

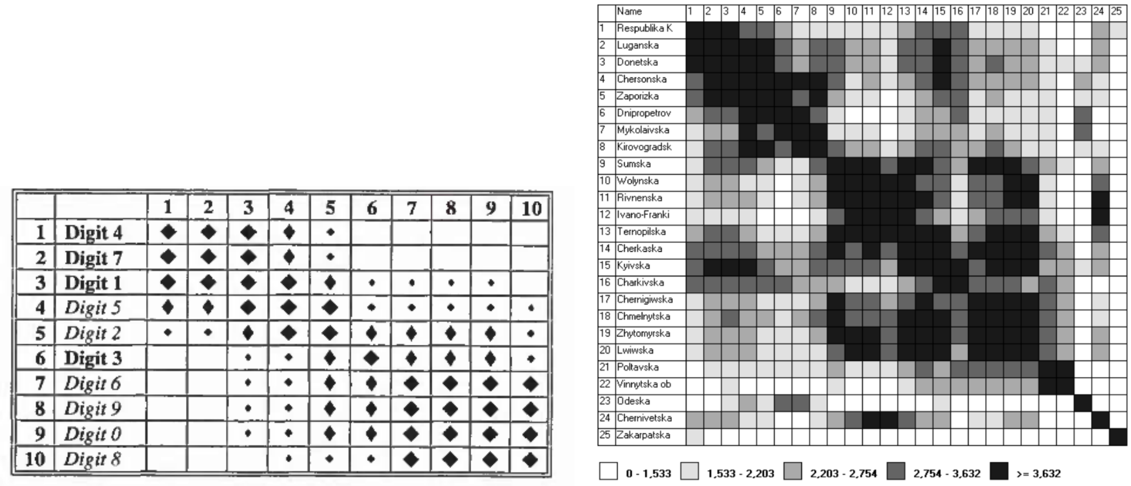

Figure 1: Example Czekanowski diagrams produced by the MaCzek program. Left: classification of Arabic numerals [6], right: similarities in milk production between different regions in Ukraine [4].

One key part that is missing is how to choose the number of distance classes, k, and then how to partition the distances into classes. The number of clusters and partition boundaries can be provided by the user but also automatic clustering functionality should be implemented. Furthermore, the ordering of the observations is not defined in the original formulation. In the beginning of the 20$^{\text{th}}$ century manual rearrangements of rows and columns was used. However, today this should be done algorithmically. In connection with Czekanowski diagrams, the Wrocław taxonomy method was proposed to both order and cluster the observations [3, 5]. When finding the best ordering one has to optimize over subsets of the permutations of the objects. This step can be implemented in various ways, e.g. genetic algorithms (as MaCzek does). After ordering and clustering one obtains *ordered* Czekanowski diagrams.

Following implementation the package should be tested on simulated and real data. The testing should be done also with respect to sensitivity of the choice of distance function and ordering methods. The simulated data should include both independent and dependent observations. The latter can be hierarchically correlated through phylogenetic trees [using e.g. mvSLOUCH 1]. The real data will be chosen according to the interests of the student and possibilities of obtaining the data. The methods should also be tested as the number of variables for each object increases and also for large amounts of data points. In the latter case it might be necessary to implement further visualization techniques.

Goals

1. An `R` package on CRAN that provides functionality for producing ordered Czekanowski diagrams.

2. A successful implementation of the package, should also hopefully result in a short

manuscript, announcing the package, to be submitted to a software oriented journal.

3. Test the sensitivity of the methods with respect to the choice of distance functions and ordering algorithms.

4. Explore the methods for data simulated under various scenarios.

5. Adjust the visualization for “Big Data”.

Data

The topic can be illustrated simulated and/or real data.

References

[1] K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, and T. F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *J. Theor. Biol.*, 314:204–215, 2012.

[2] J. Czekanowski. Zur Differentialdiagnose der Neandertalgruppe. *Korespondentblatt der Deutschen Gesellschaft für Anthropologie, Ethnologie und Urgeschichte*, XL(6/7):44–47, 1909.

[3] K. Florek, J. Łukasiewicz, J. Perkał H. Steinhaus, and S. Zubrzycki. Sur la liason et la division des points d’un ensemble fini. *Coll. Math.*, 2:282–285, 1951.

[4] M. Parlińska, Ł . Pietrych, and I. Petrovska. Evaulation of milk production diversification in Ukraine with using multidimensional statistical methods. *Metody Ilościowe w Badaniach Ekonomicznych*, XV(4):229–235, 2014.

[5] F. A. Szczotka. On a method of ordering and clustering of objects. *Applicationes Mathematicae (Zastosowania Matematyki)*, XIII(1):23–34, 1972.

[6] A Sołtysiak and P. Jaskulski. Czekanowski’s diagram. a method of multidimensional clustering. In *New Techniques for Old Times. CAA 98. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the* 26th *Conference, Barcelona, March 1998*, number 757 in BAR International Series, pages 175–184, Oxford, 1999.