Topics for Master theses: Anticlustering

Advisor: Martin Papenberg

If you are interested in pursuing one of the topics below for your master thesis, please contact me (martin.papenberg@hhu.de); this document only provides very basic information as a starting point. Basic knowledge with the statistical programming language R is required (e.g., by taking the class „Testtheorie mit R") as well as the willingness to become more proficient with R.

Anticlustering is a method to divide a set of elements into parts in such a way that the parts are similar to each other. For example, in psychological research, anticlustering is useful to divide a set of experimental stimuli into groups that are assigned to different experimental conditions in a within-subjects design. The topics below are research projects investigating the anticlustering method, e.g., by identifying best practices or improvements.

The following sources contain additional information on the anticlustering method:

- Papenberg, M., & Klau, G. W. (2021). Using anticlustering to partition data sets into equivalent parts. *Psychological Methods, 26*(2), 161-174. https://doi.org/10.1037/met0000301 (Preprint)
- Papenberg, M. (2023). K-plus Anticlustering: An Improved k-means Criterion for Maximizing Between-Group Similarity. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.12315 (Preprint)
- Homepage of the R package anticlust: https://m-py.github.io/anticlust/
- A video on using the R package anticlust (in German): https://youtu.be/YGrhSmi1oA8

1. Anticlustering with categorical variables

Investigate which strategy of including categorical variables with anticlustering works best:

1. Stratification: Include categorical variables as a „hard constraint", by using the argument „categories" in anticlustering().
2. Use categorical variables as part of the optimization criterion: *K-Means* anticlustering.
3. Use categorical variables as part of the optimization criterion: Use an appropriate distance measure for categorical data and optimize the *diversity* criterion.

Method: Program a simulation study in R, which compares different approaches under varying conditions.

- See „Using categorical variables with anticlustering" (https://m-py.github.io/anticlust/articles/Categorical_vars.html)

2. Number of exchange partners in anticlustering optimization

Investigate how many exchange partners are optimal / required during the anticlustering optimization process. It is also of interest to investigate the influence of the method of selecting exchange partners (e.g. random selection; nearest neighbours).

Method: Program a simulation study in R, which compares different approaches under varying conditions.

- See „Speeding up anticlustering“ (https://m-py.github.io/anticlust/articles/Speeding_up_anticlustering.html)

3. Anticlustering with measurement error

Anticlustering is conducted on the basis of observed data and optimizes between-group similarity with regard to the attributes that are available. The data is taken „as is“. Previous comparisons of different anticlustering objectives (Papenberg & Klau, 2021; Papenberg, 2023) (implicitly) assumed that the attributes contained no measurement error, but real data is always subject to unreliability. It is of interest to investigate if the conclusions in previous simulation studies also hold up when attributes are measured with error.

Method: Program a simulation study in R (e.g. by replicating / extending the simulation study in Papenberg, 2023), where the numeric attributes are subject to measurement error, and compare different anticlustering objectives.

- See https://github.com/m-Py/k-plus-anticlustering/tree/master/Simulation_2