

Some Simple Methods for Creating a Pooled Cluster Solution using K-Means Clustering in Multiply Imputed Data

Wednesday 23 July 2025 15:30 (15 minutes)

Oral presentation

Some Simple Methods for Creating a Pooled Cluster Solution using K-Means Clustering in Multiply Imputed Data

Author

Joost van Ginkel

Affiliation

Leiden University

Abstract

K-means clustering is a widely used technique to cluster cases in a dataset into a number of groups. When data are incomplete, missing data need to be treated prior to carrying out K-means clustering. Multiple imputation is a widely recommended procedure for dealing with missing data, which creates multiple plausible complete versions of the incomplete dataset. When applied to each of these imputed datasets, K-means clustering requires a method to combine the cluster solutions of the several imputed datasets into one overall cluster solution. Several combination methods have been proposed, such as majority vote, multiply imputed cluster analysis, and partitions pooling. These methods either come with practical problems, or try to resolve these problems using rather involved procedures. In the current presentation we propose two simple generalizations of the K-means clustering algorithm for complete data to multiply imputed datasets, which bypass all the problems that the other methods try to resolve. In a simulation study it is shown that the two newly proposed methods better recover the underlying cluster structure than the existing methods.

Primary author: VAN GINKEL, Joost (Leiden University)

Co-author: Dr LOVIK, Anikó (Leiden University)

Session Classification: Session 7 : "Clustering and classification methods in psychology"

Track Classification: Statistical analyses: Statistical analyses