Is that a cluster? A practical guide on how to avoid Type I and Type II error

Ambra Perugini, Enrico Toffalini, Tommaso Feraco, Filippo Gambarota,
Massimiliano Pastore, Gianmarco Altoè

Abstract:

Clustering, or cluster analysis, is a family of unsupervised machine learning methods that allow researchers to group sets of observations into smaller subsets (clusters) based on some sort of similarity. As cluster analysis is becoming popular in psychological and social sciences, it comes the need to point out some risks in performing cluster analysis. Even though these methods are mostly exploratory, they are more often used to make inference, therefore, inferential risks should be taken into consideration. Common risks include not only failing to detect existing clusters due to a lack of power but also revealing multiple clusters that do not exist in the population (Type I error).

Through data simulation we will go through a couple of examples highlighting these risks. And ultimately, we will introduce a tool developed to estimate Type I error and power which can be used a priori to determine whether the research design is respecting the assumptions. When assumptions in the data structure are not respected (skewness, kurtosis, correlation between indicators), there is a high risk of detecting clusters that are not there. Specifically, in psychological research those assumptions are rarely respected, and a wise choice of the technique used to perform cluster analysis is even more important, as different existing methods have different underlying assumptions (i.e. K-means, Gaussian Mixture Models).