The Journal of Open Source Software

Akmedoids R package for generating directionally-homogeneous clusters of longitudinal data sets

Monsuru Adepeju[1], **Sam Langton**[1], and **Jon Bannister**[1]

1 Crime and Well-being Big Data Centre, Manchester Metropolitan University

DOI: 10.21105/joss.02019

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 13 January 2020
Published: 17 January 2020

In partnership with

AMERICAN ASTRONOMICAL SOCIETY

This article and software are linked with research article DOI 10.3847/xxxxx, published in the The Journal of Open Source Software.

Summary

In social and behavioural sciences, longitudinal clustering is widely used for identifying groups of individual trends that correspond to certain developmental processes over time. Whilst popular clustering techniques, such as k-means, are suited for identifying spherical clusters (Curman, Andresen, & Brantingham, 2015; Genolini & Falissard, 2011), there has been little attempt to modify such methods to identify alternative forms of cluster, such as those that represent linear growth over time (i.e. directionally-homogeneous clusters). To address this shortcoming, we introduce `Anchored k-medoids`, a package referred to as `Ak-medoids`, which implements a medoid-based expectation maximisation (MEM) procedure within a classical k-means clustering framework. The package includes functions to assist in the manipulation of longitudinal data sets prior to the clustering procedure, and the visualisation of solutions post-procedure. The potential application areas of `Ak-medoids` include criminology, transport, epidemiology and brain imaging.

Design and implementation

Previous studies have taken advantage of the various functional characteristics of longitudinal data in order to extract theoretically or empirically interesting clusters of subjects. Examples include using the Fourier basis (Tarpey & Kinateder, 2003) or the coefficients of the B-spline derivative estimates (De Boor, 1978; Schumaker, 2007) which anchor clustering routines to better capture a presumed developmental process. Here, we develop an `Anchored k-medoids` (`Akmedoids`) clustering package, which employs the ordinary least square (OLS) trend lines of subjects, and a bespoke expectation-maximisation procedure, specifically to capture long-term linear growth. In criminology, identifying such slow-changing trends helps to unravel place-based characteristics that drive crime-related events, such as street and gang violence, across a geographical space (Griffiths & Chavez, 2004). To date, explorations of these trends have deployed existing techniques, namely k-means (Andresen, Curman, & Linning, 2017; Curman et al., 2015) and group-based trajectory modelling (Bannister, Bates, & Kearns, 2017; Chavez & Griffiths, 2009; Weisburd, Bushway, Lum, & Yang, 2004), which are suited for spherical clusters (Genolini & Falissard, 2011). The sensitivity of such techniques to short-term fluctuations and outliers in longitudinal datasets makes it more difficult to extract clusters based on the underlying long-term trends. `Akmdeoids` is tailored for such a scenario. The main clustering function in the `Akmedoids` package implements a medoid-based expectation maximisation (MEM) procedure by integrating certain key modifications into the classical k-means routine. First, it approximates longitudinal trajectories using OLS regression and second, anchors the initialisation process with medoid observations. It then deploys the medoid observations as new anchors for each iteration of the expectation-maximisation procedure (Celeux & Govaert, 1992), until convergence. In a similar fashion to classical k-means, the routine relies on distance-based similarity between vectors of observations and is scale invariant. This implementation ensures that the impact of short-term fluctuations and outliers are minimised. The final groupings are augmented with the raw trajectories, and visualised, in order to provide a clearer delineation of the long-term linear trends of subject trajectories. Given an `l` number of iterations, the computational complexity of the clustering routine is the same as that of a classical k-means algorithm, i.e. `O(lkn)`, where `k` is the specified number of clusters and `n`, the number of individual trajectories. The optimal number of clusters for a given data may be determined using the average silhouette (Rousseeuw, 1987) or the Calinski and Harabasz criterion (Calinski & Harabasz, 1974) or. A full demonstration is provided in the package vignette of how to deploy `Akmedoids` to examine long-term relative exposure to crime in `R`. We encourage the use of the package outside of criminology.

Clustering and cluster representations

The main clustering function of akmedoids is `akmedoids.clust`. The function captures directionally homogeneous clusters within any given longitudinal dataset using the procedure detailed above. For crime inequality studies, the package includes the `props` function for converting the absolute (or rate) measures of individual trajectories into a relative measure over time. The `statPrint` function draws from the `ggplot2` library (Wickham, 2016) in order to visualise the resulting clusters in either a line or an areal-stacked graph format, alongside descriptive cluster statistics.

Acknowledgment

We gratefully acknowledge the Economic and Social Research Council (ESRC), who funded the Understanding Inequalities project (Grant Reference ES/P009301/1) through which this research was conducted.

References

Andresen, M. A., Curman, A. S., & Linning, S. J. (2017). The trajectories of crime at places: Understanding the patterns of disaggregated crime types. *Journal of Quantitative Criminology*, *33*, 427–449.

Bannister, J., Bates, E., & Kearns, A. (2017). Local variance in the crime drop: A longitudinal study of neighbourhoods in greater glasgow, scotland. *British Journal of Criminology*, *58*, 177–199.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*, 1–27.

Celeux, G., & Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Journal of Computational Statistics and Data Analysis*, *14*, 315–332.

Chavez, J. M., & Griffiths, E. (2009). Neighborhood dynamics of urban violence: Understanding the immigration connection. *Homicide Studies*, *13*, 261–273.

Curman, A. S. N., Andresen, M. A., & Brantingham, P. J. (2015). Crime and place: A longitudinal examination of street segment patterns in vancouver, bc. *Journal of Quantitative Criminology*, *31*, 127–147.

De Boor, C. (1978). *A practical guide to splines*. Springer-Verlag New York. Retrieved from https://www.springer.com/gp/book/9780387953663

Genolini, C., & Falissard, B. (2011). KmL: K-means for longitudinal data. *Computational Statistics*, *2*, 317–328.

Griffiths, E., & Chavez, J. M. (2004). Communities, street guns and homicide trajectories in chicago, 1980-1995: Merging methods for examining homicide trends across space and time. *Criminology*, *42*, 941–978.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 53–65.

Schumaker, L. L. (2007). *Spline functions: Basic theory*. Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9780511618994

Tarpey, T., & Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification*, *20*, 93–114.

Weisburd, D., Bushway, S., Lum, C., & Yang, S. (2004). Trajectories of crime at places: A longitudinal study of street segments in the city of seattle. *Criminology*, *42*, 283–322.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org