KmLcov: k-means for longitudinal data with covariates

Mickaël Canouil[1,2,3], **Christophe Genolini**[4,5,6,*], **René Ecochard**[1,2,3]

1. Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France.
2. Université de Lyon, F-69000, Lyon, France.
3. CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biotatistique-Santé, F-69100, Villeurbanne, France.
4. Université Paris Sud, F-91400, Orsay, France.
5. INSERM, UMR-S 669, F- 75014, Paris, France.
6. ModalX, Université Paris Nanterre, F-92000, Nanterre, France.

*Contact author: christophe.genolini@u-paris10.fr

Keywords: k-means, longitudinal data, covariates, clustering.

Many packages offer various solutions to cluster longitudinal data (Latent Class Analysis). They range from non-parametric algorithms, such as k-means (Genolini and Falissard (2010)), to model-based approaches. The former identify clusters of response variable trajectories without allowing for additional covariates. The latter offer several options to cluster longitudinal data and take into account the natures of the trajectories (continuous, censored, etc.) and of the covariates (continuous, binary, time-dependant, etc.). Some model-based packages are devoted to specific aspects: they accept unequal within-group variances or use splines to model the evolutions of the trajectories.

Mixing the advantages of both approaches might be interesting: the use of the k-means approach to cluster trajectories in a non-parametric way and the use of regression models to study the relationship between a response variable and one or more covariates.

KmLcov is a new package devoted to semi-parametric Latent Class Analysis: it is non-parametric regarding the evolution of Latent Class trajectories and parametric regarding the adjustment on the covariates. The main idea is to identify clusters of individual trajectories according to their similar evolutions taking into account the effect of one or more covariates on the outcome.

KmLcov allows studying variability at different levels using regression models: between individual measures, between individual trajectories, and between Latent Class trajectories.

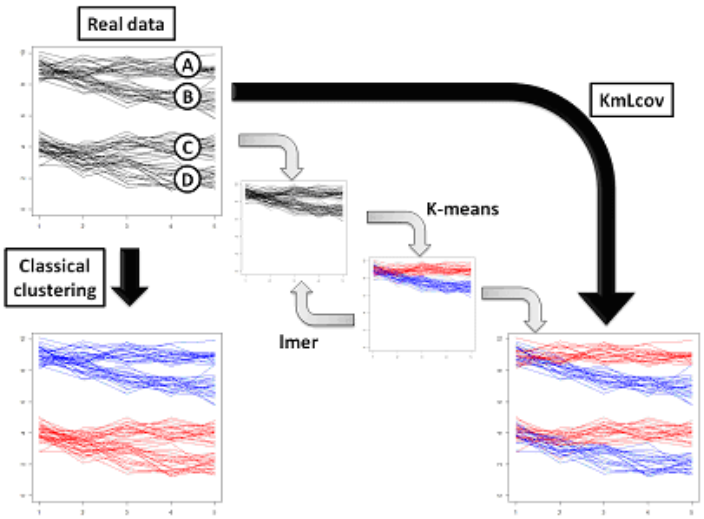

Figure 1: We want to identify healthy and sick groups in real data (with A=Healthy male; B=Sick male; C=Healthy Female; D=Sick Female.) Classical algorithms find male/female. `KmLcov` adjusts on male/female then finds healthy/sick clusters.

References

Genolini, C. and B. Falissard (2010). KmL: k-means for longitudinal data. *Computational Statistics 25*(2), 317–328.