Combined-information criterion for clusterwise elastic-net regression. Application to omic data.

Stéphanie Bougeard[1], Xavier Bry[2], Thomas Verron[3], Ndèye Niang[4]

[1] ANSES (French agency for food, environmental and occupational health safety), Ploufragan, France
[2] University of Montpellier, France
[3] SEITA, Paris, France
[4] CEDRIC CNAM, Paris, France

E-mail for correspondence: `stephanie.bougeard@anses.fr`

Abstract: Many research questions pertain to a regression problem assuming that the population under study is not homogeneous with respect to the underlying model. In this setting, we propose an original method called Combined Information criterion CLUSterwise elastic-net regression (CICLUS). This method handles several methodological and application-related challenges. It is derived from both the information theory and the microeconomic utility theory and maximizes a well-defined criterion combining three weighted sub-criteria, each being related to a specific aim: getting a parsimonious partition, compact clusters for a better prediction of cluster-membership and a good within-cluster regression fit. The solving algorithm is monotonously convergent under mild assumptions. The CICLUS method provides an innovative solution to two key issues: the automatic optimization of the number of clusters and the issue of a prediction model. We applied it to elastic-net regression in order to be able to manage high-dimensional data involving redundant explanatory variables. CICLUS is illustrated through a real example in the field of omic data, showing how it improves the quality of the prediction and facilitates the interpretation. It should therefore prove useful whenever the data involve a population mixture as for example in biology, social sciences, economics or marketing.

Key words: Clusterwise regression; Typological regression; Elastic-net regularization

Charles C (1977). Régression typologique et reconnaissance des formes. PhD, University of Paris IX, France.

Mortier F, Ouedraogo DY, . . . , and Picard N (2015). Mixture of inhomogeneous matrix models for species-rich ecosystems. Environmetrics, 26(1), 39–51.

Suk HW, and Hwang H (2010). Regularized fuzzy clusterwise ridge regression. Advances in Data Analysis and Classification, 4, 35–51.