Clusterwise Methods: a Synthesis and New Developments

Gilbert Saporta

Clusterwise Methods: a Synthesis and New Developments

Gilbert Saporta[1]

Clusterwise methods are a mix of cluster analysis and statistical modelling. When we look for some statistical modelization (regression, PCA, etc.) it would be unwise to fit a single model to the whole set of data if we know that the data set is heterogenous. The solution is simple: it consists to fit as many models as the number of clusters. When the clusters are not known beforehand, instead of finding first the clusters and after the models, clusterwise methods aim at finding simultaneously the clusters and the models, by optimizing some criterium.

There are two main approaches:

1. the least squares approach introduced by E.Diday in the 70's, derived from *k*-means;
2. mixture models with latent classes using maximum likelihood but only the first one easily enables prediction.

1.Typological PCA

In his pioneering paper, Diday (1974) proposed the simultaneous search of *k* subspaces with maximal inertia, ie local factorial planes. The algorithm, derived from *k*-means, is the following: after a first partition, units are reallocated (or not) to the nearest cluster according to their distances to the local plane. New local factorial planes are computed until convergence. Convergence is guaranteed since the sum of the inertia is increasing at each step. Note that there are two types of updating: after each reallocation (true *k*-means, or «stochastic»

[1]CNAM, Paris, *gilbert.saporta@cnam.fr*

algorithm) or after a « pass », or complete run of all observations which is the « batch » algorithm.

2. Clusterwise regression

Clusterwise regression simultaneously looks for a partition of the observations into clusters, and minimizes the sum of squared error computed over all the clusters. Starting from an initial partition Charles (1977) defined the reallocation step in order to get the smallest regression residual *ie* the best prediction. Since it could happen that a cluster might contain less observations than the number of predictors, a ridge or other kind of regularized regression should be used instead of OLS.

Esposito-Vinzi et al. (2005) studied clusterwise regression using common PLS components across clusters, while Niang *et al*. (2016) advocate the use of local PLS components.

Preda & Saporta (2005) proposed a clusterwise functional regression where for each cluster, one estimates the functional linear model $\hat{Y} = \int_0^T \beta_k(t) X_t dt$ by PLS regression since it is an ill-posed problem.

Carvalho *et al*. (2010) presented a clusterwise generalization of «center and range» regression for symbolic interval data. In this problem one predicts the center and the mid-ranges by two regressions.

Once each local model has been calibrated, a common issue is how to predict the response of a new unit whose only the predictors are known. The simplest way or “hard rule” is to allocate the new unit to the nearest cluster and apply the relevant model. This necessitates the choice of a relevant distance. A more flexible way is to use a weighted average of the *K* predictions; the weights being the posterior probabilities to belong to each cluster. A third solution is to pick at random, with unequal probabilities, one of the *K* models.

The three previous solutions are easy to implement in the framework of *k*-means like methods. But it is not the case for the mixture model, or latent class regression, since one needs to know the true cluster in order to compute the likelihood and get the posterior membership probabilities.

In order to find the adequate number of clusters, and prevent trivial solutions, it is necessary to use cross-validation.

3. Multiblock clusterwise methods

Bougeard *et al*. (2017, 2018) have extended clusterwise methods to the case where the variables are organized into blocks as illustrated in Figure 1.

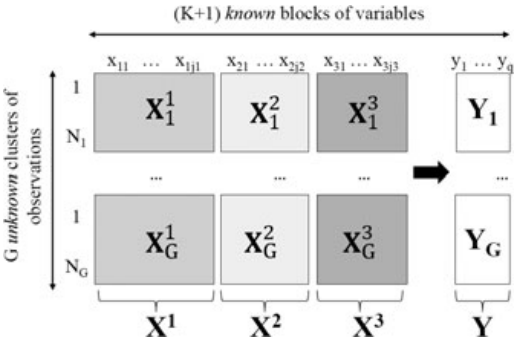

Figure 1: A *known* structure of variables into blocks and an *unknown* structure of observations into clusters.

In their first paper they “clusterize” PLS2 regression and multiblock redundancy analysis. The second paper introduces regularization within the specific goal of prediction.

The application shows that clusterwise regularized multiblock regression is a useful tool to analyze complex data as found, eg, in marketing, biology, social sciences, or many fields dealing with

population mixtures. The method and its interpretation tools are available for users though the R package mbclusterwise.

Keywords: Clusterwise regression, Mixture models, Dimension reduction, PLS regression, Multiblock data

References

BOUGEARD, S., ABDI, H., SAPORTA, G., NIANG KEITA, N. (2017): Clusterwise analysis for multiblock component methods, *Advances in Data Analysis and Classification.*

BOUGEARD, S., CARIOU,V., SAPORTA, G., NIANG KEITA, N. (2018): Prediction for regularized clusterwise multiblock regression, *Applied Stochastic Models in Business and Industry.*

CHARLES, C. (1977): *Régression Typologique et Reconnaissance des Formes.* Ph.D. Université Paris IX.

De CARVALHO, F., SAPORTA, G., QUEIROZ, D. (2010): A Clusterwise Center and Range Regression Model for Interval-Valued, *COMPSTAT'2010, 19th International Conference on Computational Statistics,* pp.461- 468,

DIDAY, E. (1974): Introduction à l'analyse factorielle typologique, *Revue de Statistique Appliquée*, 22, 4, pp.29-38.

ESPOSITO-VINZI, V. LAURO, C., AMATO, S. (2005): PLS Typological Regression: Algorithmic, Classification and Validation Issues, in Vichi M, Monari P, Mignani S, Montanari A, eds, *New Developments in Classification and Data Analysis,* pp.133-140, Springer.

NIANG, N., BOUGEARD, S., SAPORTA, G. (2016): Prédiction en régression clusterwise PLS, *48èmes Journées de Statistique*, Montpellier, France

PREDA, C., SAPORTA, G. (2005): Clusterwise PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 49, pp. 99–108.