Clusterwise methods, past and present

Gilbert Saporta
Conservatoire National des Arts et Métiers, Paris, France – gilbert.saporta@cnam.fr

Abstract

Instead of fitting a single and global model (regression, PCA, etc.) to a set of observations, clusterwise methods look simultaneously for a partition into *k* clusters and *k* local models optimizing some criterion. 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 using maximum likelihood
but only the first one easily enables prediction.
After a survey of classical methods, we will present recent extensions to functional, symbolic and multiblock data.

Keywords: clusterwise regression; mixture models; dimension reduction; PLS regression

1. Introduction

When a data set in partitioned into *k* heterogenous clusters, it would be unwise to fit a single model to the whole set of data without taking into account this information, as shown in Figure 1 for a simple regression.

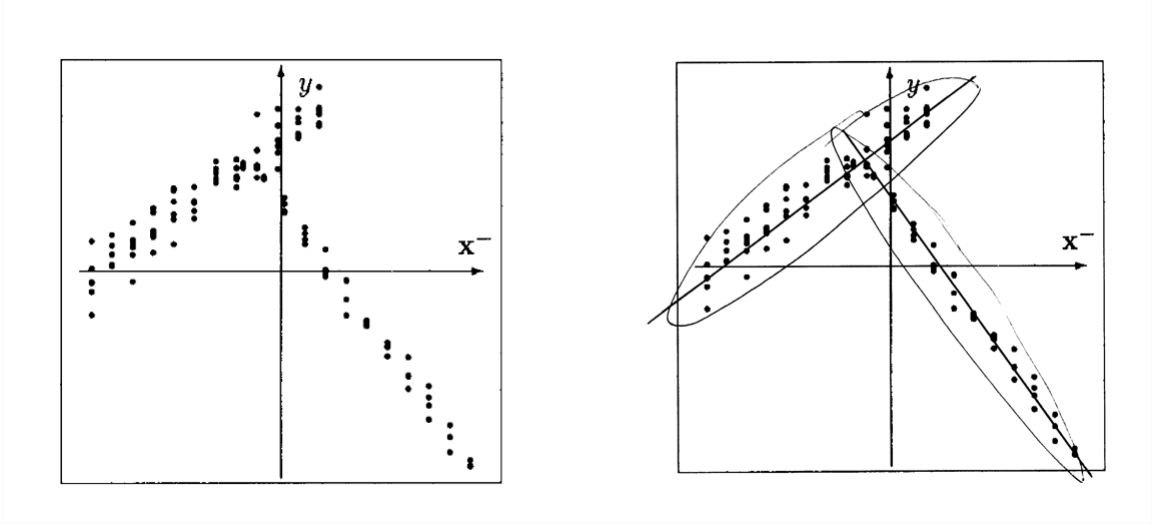

Figure 1 from Hennig (2000).

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 in two steps 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 ways for performing clusterwise analysis: *k*-means methods optimizing a least-squares criterium and mixture models with latent classes.

2. 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. As in the following methods, there are two types of updating: update criterium and partition after each reallocation (true *k*-means, or « stochastic » algorithm) or update criterium and partition after a « pass », or complete run of all observations which is the « batch » algorithm.

3. Clusterwise regression

The principle is for each class, to fit a linear model maximizing the global criterium

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbf{1}_k(i)\left(y_i-(\alpha_k+\boldsymbol{\beta}_k\mathbf{x}_i)\right)^2$$

Where $\mathbf{1}_k(i)$ is equal to 1 if unit i belongs to cluster k , otherwise 0. 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. Späth (1979) did not noticed this problem, when he coined the expression « clusterwise regression » together with a Fortran package.

Esposito-Vinzi *et al.* (2003, 2005) studied clusterwise regression using common PLS components across clusters, while Niang *et al.* (2016) advocate the use of local PLS component.

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.

More recently 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.

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

4. Latent class regression

DeSarbo & Cron (1988) assumed that the response is distributed as a finite sum, or mixture, of conditional univariate normal densities:

$$\sum_{k=1}^{K}\lambda_k\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(\frac{-\left(y_i-\boldsymbol{\beta}_k\mathbf{x}_i\right)^2}{2\pi\sigma_k^2}\right)$$

The weights and the parameters are estimated by maximum likelihood through EM algorithm.

Wedel & al. (1993) extended the method to Poisson regression, and Wedel & DeSarbo (1995) to generalized local linear models.

Mixture models are not always identifiable and Hennig (2000) gave the following result: mixtures of linear regression models with Gaussian noise are identifiable, if the number of components K is smaller than the minimal number of (feasible) hyperplanes necessary to cover all covariate points (without intercept).

5. The prediction problem

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 3 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. In the FlexMix package, which is widely used,

one needs to know the true cluster in order to compute the likelihood and get the posterior membership probabilities. As F.Leisch wrote (personal communication) "*Without y you cannot determine the likelihood and hence not into which cluster the observation belongs. You could calculate predictions for each cluster, but then you have K answers, not one*."

6. Software

Latent class regression is implemented in commercial software such as GLIMMIX http://www.scienceplus.com/glimmix, LatentGold 5.0 and XLSTAT-LG. Free packages are easily available: Flexmix by Leisch (2004), Mixmod and many others.

Until very recently (Bougeard, 2016) there was no available software for clusterwise regression using least squares, despite several publications dealing with global optimization techniques like simulated annealing (DeSarbo *et al*., 1989), Variable Neighbour Search metaheuristics (Caporossi & Hansen, 2005) and Carbonneau *et al*., 2014 which have not been compared between them.

Numerical experiments will be presented which prove the efficiency of *k*-means-like algorithms, coupled with PLS regression. Figure 2 from Niang *et al* (2016) shows the performance of clusterwise PLS regression with 3 components on an example derived from Späth's data:

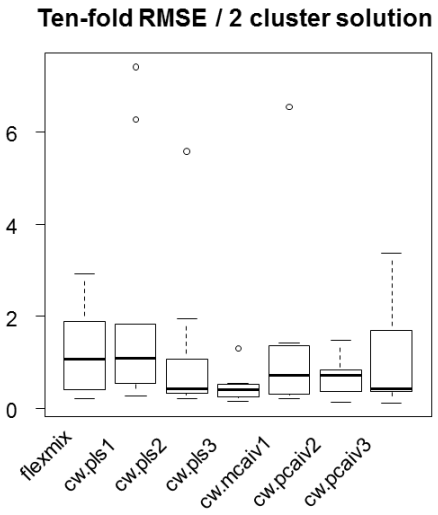

Figure 2: comparison of the prediction accuracy of 7 clusterwise methods

7. Conclusions

The publication bias towards mixture models may be a reflect of the long controversy between explaining and predicting, see Breiman (2001), Saporta (2008). As long as prediction is the main goal, methods based on least-squares and regularized regression are efficient and need less assumptions. *K*-means algorithms may be applied in various situations, like multiblock data Bougeard *et al*., (2016). Works in progress include clusterwise sparse PLS.

References

Bougeard, S. (2016): Clusterwise Multiblock Analyses, https://cran.r-project.org/package=mbclusterwise

Bougeard, S., Niang, N., Saporta, G. (2016): Regularized clusterwise multiblock regression, Compstat 2016, Oviedo, Spain

Breiman, L. (2001): Statistical Modeling: The Two Cultures, *Statistical Science*, 16, 3, 199–231

Caporossi, G. , Hansen, P. (2005): Variable Neighborhood Search for Least Squares Clusterwise Regression, *Cahiers du GERAD*, HEC Montréal

Carbonneau, R.A., Caporossi, G., and Hansen, P. (2014) : Globally Optimal Clusterwise Regression By Column Generation Enhanced with Heuristics, Sequencing and Ending Subset Optimization, *Journal of Classification*, 31, pp.219-241
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,
DeSarbo, W.S. and Cron, W.L. (1988): A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5, pp.249-282.
DeSarbo, W.S , Kamakura W.A. , Wedel M. (2005): Latent Structure Regression, In: *Handbook of Marketing Research* , R. Grover & M. Vriens, (eds), London, Sage, 394-417.
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 *New Developments in Classification and Data Analysis*, pp.133-140, Springer
Esposito-Vinzi, V. Lauro, C. (2003): PLS Regression and Classification, In: *Proceedings of the PLS'03 International Symposium*, DECISIA, pp. 45-56
Hennig, C. (1999): Models and methods for clusterwise linear regression. In*:Classification in the Information Age*, Springer, pp.179-187.
Hennig, C. (2000): Identifiability of models for Clusterwise linear regression. *Journal of Classification*, 17, pp.273-296.
Leisch, F. (2004) : FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(8).
Niang, N., Bougeard, S., Saporta, G. (2016): Prédiction en régression clusterwise PLS, *48 èmes 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.
Saporta, G. (2008): Models for Understanding versus Models for Prediction, in *Proceedings COMPSTAT'08*, Brito, P. (ed.), Springer, pp.315-322
Späth, H. (1979): Clusterwise linear regression, *Computing*, 22, pp.367-373
Wedel M., DeSarbo W.S. (1995): A Mixture Likelihood Approach for Generalized Linear Models, *Journal of Classification*, 12, 21–55.