Collaborative Sliced Inverse Regression

A. Chiancone, S. Girard, J. Chanussot

In multidimensional data analysis, one has to deal with a dataset *X* made of *n* points in dimension *p*. When *n* and *p* are simultaneously large, classical statistical analysis methods and models fail. Supervised and unsupervised dimensionality reduction techniques are widely used to preprocess high dimensional data retaining the information useful to solve the original problem.

In regression context Sliced Inverse Regression [1] has proven to achieve good results retrieving a base of the so called *effective dimension reduction* (*e.d.r.*) space i.e. the smallest space containing the information needed to correctly regress the function. Recently, many papers focused on the complex structure of real data showing that often the data is organized in subspaces. Kuentz & Saracco (2009) proposed to clusterize *X* and use SIR in each cluster to better fit the so called linearity condition.

Our hypothesis is that the *e.d.r.* space is not unique all over the data and that the different clusters can be assigned to different *e.d.r.* spaces. We introduce a novel technique to identify the number of *e.d.r.* spaces based on a weighted distance between the different spaces. First we clusterize the data (in our simulation study we used the standard k-means) then we apply SIR independently in each cluster. A greedy merging algorithm is proposed to assign each cluster to its *e.d.r* space taking into account the size of the cluster on which SIR is performed.

Our approach is illustrated on simulated data from a Gaussian mixture model.

This work is founded by LabEx Persyval.

[1] Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association, 86, 316-342.

[2] Kuentz, V., & Saracco, J. (2010). Cluster-based sliced inverse regression. Journal of the Korean Statistical Society, 39(2), 251-267.