PLS regression for functional data

Gilbert Saporta

Conservatoire National des Arts et Métiers
292 rue Saint Martin, F- 75141 Paris cedex 03
gilbert.saporta@cnam.fr

Abstract

Functional data occurs when we observe curves or paths from a stochastic process X_t.
If for each curve or path we have a single response variable Y, we have a regression problem when Y is numerical, a classification problem when Y is categorical.

Linear methods look for predictors which may be expressed as an integral sum $\int_0^T \beta(t)X(t)dt$

The problem is not new and comes back to Fisher (1924).
When *t* takes continuously its values in an interval [0;*T*], there is severe multicollinearity that leads to inconsistent estimation of the parameters.
Since the works of Ramsay & Silverman (1997), many techniques have been applied to solve these kinds of problems, mostly by using explicit regularization techniques.

In this communication we will focus on linear methods based on an orthogonal decomposition of the predictors. The use of components derived from the Karhunen-Loeve decomposition is, for functional data, the equivalent of principal components regression (PCR).
We show that partial least squares (PLS) regression performs better than PCR since PCR components are obtained irrespective of the response.
Clusterwise PLS regression may be used when heterogeneity in the data is present. This corresponds to a mixture of several regression models. Clusters and local models are found by an extension of k-means clustering.

References

Fisher R.A. (1924) The Influence of Rainfall on the Yield of Wheat at Rothamsted.
Philosophical Transactions of the Royal Society, B: 213: 89-142

Preda C., Saporta G. (2005a): PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, 48, 149-158.

Preda C., Saporta G. (2005b) Clusterwise PLS regression on a stochastic process . *Computational Statistics and Data Analysis*, 49(1): 99-108, 2005.

Ramsay J.O., Silverman B. (1997), *Functional data analysis*, Springer