Supervised and Unsupervised Linear Methods for Functional Data

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

Functional data occurs when we observe curves or paths from a stochastic process $X_{t\,.}$
We will assume that *t* takes continuously its values in an interval [0;*T*].

Karhunen-Loeve decomposition provides then an unsupervised method for describing this kind of data and is the continuous counterpart of principal component analysis.

If for each curve or path we have in addition a single response variable Y, we have a regression problem when Y is numerical, a supervised classification problem when Y is categorical.
Linear methods looks 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).
The strong autocorrelations between predictors leads to inconsistent estimation of the parameters. Since the works of Ramsay & Silverman (1997), many techniques have been applied to solve these kind of problems, mostly by using explicit regularization techniques.

We will focus here 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) but 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 (Preda *et al.*, 2005). This corresponds to a mixture of several regression models. Clusters and local models are found by an extension of *k*-means clustering.

Previous methods are easily generalized to binary classification, since Fisher's linear discriminant function is equivalent to a multiple regression (Preda *et al.*, 2007) .
Logistic regression has also been generalized to functional predictors (Escabias *et al.*, 2007)
In many real time applications like industrial process, it is of the highest interest to make anticipated predictions. A method based on a bootstrap test for comparing areas under ROC curves helps to determine an optimal time *t**<*T* giving a prediction based on [0; *t**] almost as good as the prediction using all the data. (Costanzo *et al*., 2006)
Instead of using the same anticipated decision time *t** for all data, we may adapt *t** to each new trajectory given its incoming measurements. The procedure is similar in its spirit to a on-line sequential test (Costanzo *et al*., 2008).

Applications on real and simulated data will be presented.

References

Costanzo D., Preda C. , Saporta G. (2006). Anticipated prediction in discriminant analysis on functional data for binary response . In *COMPSTAT2006,* 821-828, Physica-Verlag

Escabias M., Aguilera A.M. , Valderrama M.J. (2007) Functional PLS logit regression model. *Computational Statistics & Data Analysis*, 51, 10, 4891-4902

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. (2005) Clusterwise PLS regression on a stochastic process . *Computational Statistics & Data Analysis,* 49(1): 99-108, 2005.

Preda C. , Saporta G., Lévéder C., (2007) PLS classification of functional data, *Computational Statistics*, 22(2), 223-235

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