Optimal Linear Projection based on Noise Variance Estimation - Application to Spectroscopic Modeling

Amaury Lendasse and Francesco Corona
Helsinki University of Technology, Laboratory of Computer and Information Science

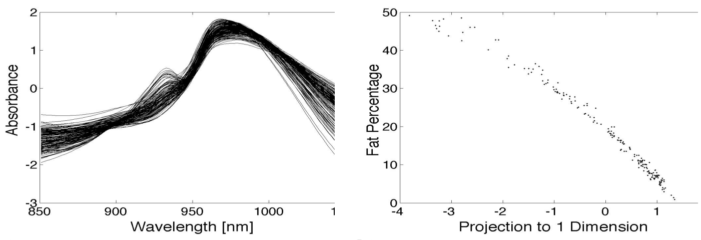
E-mail: lendasse@hut.fi, f.corona@hut.fi

Data from spectrophotometers form vectors with a large number of exploitable variables. Building quantitative models using these variables most often requires using a smaller set of variables than the initial one. Indeed, a too large number of input variables to a model results in a too large number of parameters, leading to overfitting and poor generalization abilities. Partial least squares regression (PLS-R) has been successfully used to deal with a large number of input variables. In PLS-R, the input variables are linearly projected to latent variables. The projection is done in order to keep the necessary information needed to build a linear model between the latent variables and the target variables. Even if PLS-R is providing good models in many practical situations, it can fail if the intrinsic relationship between input and output variables is nonlinear. Furthermore, the latent variables that are built by the PLS-R are optimized in order to provide the best input variables if a linear model is used. It is not straightforward that these latent variables are the optimal input variables for nonlinear models such Support Vector Machines (SVM).

In this work, we propose a method to build latent variables that are optimal if a nonlinear model is used. This method is based on Noise Variance Estimation (NNE). NNE is providing an estimate of the variance of the noise between input and output variables. The optimal set of inputs is the one that is minimizing the NNE. Furthermore, NNE is providing an estimate of the best performances that can be obtained without overfitting. The NNE used in this work is based on Delta Test. Consider a set of general input-output pairs (x_i, y_i) in $\Re^n \times \Re$, x_{NN_i} denotes the nearest neighbor of x_i and y_{NN_i} the output of x_{NN_i}. Then, the variance estimate δ provided by Delta Test is:

$$\delta = \lim_{N \to \infty} \sum_{i=1}^{N} \left(y_{NN_i} - y_i\right)^2 \Big/ 2 \approx \sum_{i=1}^{N} \left(y_{NN_i} - y_i\right)^2 \Big/ 2. \qquad (1)$$

In this work, the linear projection that builds latent variables is optimized by an iterative Forward-Backward Selection methodology in order to minimize the Delta Test. We successfully tested the proposed method with two data sets from the food industry (Wine and Tecator, [1]). The goal is to predict analytical values (such as the fat content) using NIR absorption spectra. The spectra of Tecator, the projections and the results obtained using LS-SVM models are summarized in Fig. 1.



(a) (b)

Method	Test MSE
PLS-R	4.45
Projection to $\Re$ + LSSVM	1.35
Projection to $\Re^2$ + LSSVM	0.85

(c)

Fig1: (a) Data set (b) Projection to 1-dimensional space (c) Comparison of the performances

References

[1] *Mutual information for the selection of relevant variables in spectrometric nonlinear modeling*, F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Chemometrics and Intelligent Laboratory Systems, Volume 80, Issue 2, 15 February 2006, pp. 215-226