Optimal Gaussian Basis Functions for Chemometrics

Tuomas Kärnä and Amaury Lendasse
Helsinki University of Technology, Laboratory of Computer and Information Science

E-mail: tuomas.karna@hut.fi, lendasse@hut.fi

Spectral data are often high dimensional which causes problems in data analysis. Computational complexity of many analysis methods grows exponentially with respect to the number of variables. Furthermore, the analysis suffers from the curse of dimensionality, which states that the theoretical lower bound of error increases with data dimensionality. In chemometrics, the curse of dimensionality is especially problematic because the available data sets are often small: in some cases the number of variables exceeds the number of training examples. This is a poor starting point for machine learning and it very likely leads into poor generalization performance.

To overcome the curse of dimensionality, one can focus on studying only a small subset of the data or cast the data into a smaller dimensional space. Although the first alternative is often effective, it is not efficient: finding a relevant subset can be very time consuming. On the other hand, since spectral curves are relatively smooth, function fitting provides a straight forward way for dimension reduction. Often standard function bases, such as the B-splines or wavelets, are used for the fitting [1]. However, it seems appealing to tune the basis according to the data at hand so that minimal number of coefficients (or weights) is needed for representing the data.

We propose that Gauss-Newton optimized Gaussian functions are a good choice for the basis. The locations and widths of the Gaussian functions are optimized for an accurate fit in the entire data set. Consequently, the basis follows the statistical nature of the data and a good representation is obtained with a very small number of basis functions. Furthermore, there is only one unknown parameter to select: the number of Gaussian functions. The propose methodology is summarized in Fig. 1.

We tested the proposed method with two data sets from the food industry [2]. The goal was to predict analytical values (such as the fat content) using NIR absorption spectra. The Gaussian basis was optimized for the data sets and the obtained fitting coefficients were used to train a Least-Squares Support Vector Machine (LS-SVM) model for the final prediction. The results show that the reduced number of variables (obtained by the fitting) gives better prediction performance than LS-SVM or PLS with the original data.

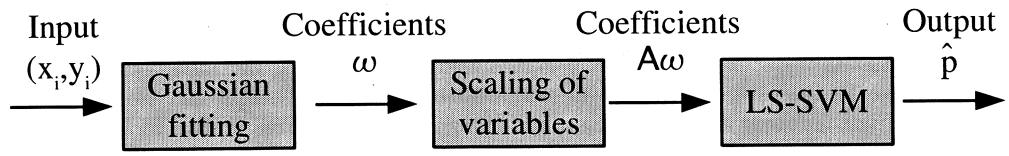

Fig1: The proposed methodology.

References
[1] Functional Data Analysis, J. O. Ramsay and B. W. Silverman., second edition, Springer, New York, 2005.
[2] 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.