A functional approach to variable selection in spectrometric problems*

Fabrice Rossi[1], Damien François[2], Vincent Wertz[2], and Michel Verleysen[3]

[1] Projet AxIS, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France
[2] Université catholique de Louvain - Machine Learning Group, CESAME, 4 av. G. Lemaître, 1348 Louvain-la-Neuve, Belgium
[3] Université catholique de Louvain - Machine Learning Group, DICE, 3 place du Levant, 1348 Louvain-la-Neuve, Belgium

Abstract. In spectrometric problems, objects are characterized by high-resolution spectra that correspond to hundreds to thousands of variables. In this context, even fast variable selection methods lead to high computational load. However, spectra are generally smooth and can therefore be accurately approximated by splines. In this paper, we propose to use a B-spline expansion as a pre-processing step before variable selection, in which original variables are replaced by coefficients of the B-spline expansions. Using a simple leave-one-out procedure, the optimal number of B-spline coefficients can be found efficiently. As there is generally an order of magnitude less coefficients than original spectral variables, selecting optimal coefficients is faster than selecting variables. Moreover, a B-spline coefficient depends only on a limited range of original variables: this preserves interpretability of the selected variables. We demonstrate the interest of the proposed method on real-world data.

1 Introduction

In many real-world problems, objects are described by sampled functions rather than by vectors. In the simplest case, an object is given by a function f, from $\mathbb{R}$ to $\mathbb{R}$, specified by a list of m input/output pairs, $((x_j, f(x_j)))_{1\leq j\leq m}$ (m and $(x_j)_{1\leq j\leq m}$ depend on the object). Examples of such situation include applications in which temporal evolution of objects is monitored (and therefore where each object is described by one or several time series) and others in which objects are characterized by spectra (near infrared transmittance for instance).

One of the main problems in spectrometry is regression: one wants to predict physical or chemical properties of a sample via its spectrum. Because chemical and physical analyses are long, difficult and expensive, we have generally

* M. Verleysen is Research Director of the Belgian F.N.R.S. (National Fund for Scientific Research). D. François is funded by a grant from the Belgian F.R.I.A. Parts of this research result from the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

a low number of examples, typically a few hundreds. On the contrary, spectra are generally sampled at a high resolution, up to thousands of wavelengths per spectrum. It is therefore quite common to have more sampling points (spectral variables) than spectra.

As the sampling is generally fixed (i.e. m and $(x_j)_{1 \leq j \leq m}$ are fixed for the whole data set), each spectrum can be considered as a high-dimensional vector. However, because of the low number of spectra, even simple linear methods are difficult to apply directly to many spectrometric problems. In practice, the standard solution is to rely on dimension reduction methods coupled with linear regression, mainly Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). PCR for instance consists in a simple linear model constructed on a few Principal Components of the original data. PLSR consists in finding linear projections that have maximum correlation with the target variable; a linear regression is then built on the projected coordinates. While those methods give generally satisfactory results, they are unfortunately difficult to interpret: the linear model is constructed on projected features whose dependency to the original spectral variables, albeit linear, can be quite complex. In general, one cannot determine from the model which spectral range is useful for the regression problem under focus. Moreover, PCR and PLSR are intrinsically limited by their linear nature.

Another solution consists in using variable selection methods to keep only a small number of the original spectral variables and then to build a nonlinear model on those data [1–3]. When a small number of spectral variables are selected, this approach both avoids overfitting and eases interpretation. Even if the processing of the selected variables is nonlinear, the model emphasizes generally the dependency of the target variable to a small number of original spectral variables. However, those methods suffer from two problems. They are generally quite slow, even when filter selection methods are used (i.e., when the relevance of a group of variables is estimated via a simpler model than the nonlinear model). Moreover, while filter methods tend to be less sensitive to overfitting than the nonlinear models used for the second part of the analysis, they nevertheless face the difficulty of dealing with high-dimensional data and can select redundant or useless variables.

This paper proposes to use a functional representation approach as a preprocessing step before variable selection for regression problems. The main idea is to leverage the functional nature of spectra to replace high-resolution representations by a low number of variables that keep almost all the original information and still allow one to assess the importance of some wavelength ranges in the regression task. This prior reduction of the data dimensionality eases the variable selection both in terms of computational requirement and in terms of statistical significance.

2 Functional Data Analysis

The idea is this paper is based on the concepts of Functional Data Analysis (FDA [4]). The main idea of FDA is to adapt standard data analysis methods such that they make explicit use of the functional aspect of their inputs. There are two standard approaches in FDA: *regularization* and *filtering*.

The regularization approach addresses the overfitting problem via complexity control. Let us consider for example the problem of linear regression on functional data. In a theoretical and perfect setting, we have a random variable X with values in L^2 (the Hilbert space of square integrable functions from $\mathbb{R}$ to $\mathbb{R}$) and a target random variable Y with values in $\mathbb{R}$. The functional linear model is given by $Y = \langle h, X \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in L^2, i.e.

$$Y = \int hX \mathrm{d}\lambda.$$

Finding h via observations, i.e. via realizations of the pair (X, Y), is difficult: because of the infinite dimension of L^2, the problem is ill-posed. It has generally an infinite number of solutions and they are difficult to estimate. The problem is solved by looking for smooth candidates for the function h, for instance twice differentiable functions with minimal curvature, see e.g. [5, 6]. This can be considered as a functional version of the standard ridge regression [7]. The regularization approach has been applied to other data analysis problems, such as Principal Component Analysis [8]. As shown in [9] in the case of discriminant analysis, when the data are sampled functions, ridge regularization leads to worse solutions than a functional regularization.

In the filtering approach, each list of input/output pairs is considered as a function approximation problem for which a simple truncated basis solution is chosen: the list $((x_j, f(x_j)))_{1 \leq j \leq m}$ is replaced by the vector $(u_1, \ldots, u_p)$ obtained as the minimizer of

$$\sum_{j=1}^{m} \left(f(x_j) - \sum_{k=1}^{p} u_k \phi_k(x_j) \right)^2,$$

i.e., the square reconstruction error of f by the basis functions $(\phi_k)_{1 \leq k \leq p}$. Filtering can therefore be considered as a preprocessing step in which functional data are consistently transformed into vector data. It has been used as a simple way to adapt many data analysis methods to functional data, see for instance [10] for linear regression, [11, 12] for Multi-Layer Perceptrons and Radial Basis Function Networks, [13] for k-nearest neighbors and [14] for Support Vector Machines.

3 B-spline representation

3.1 B-splines

Obviously, the filtering approach of FDA can be used to reduce the dimensionality of spectra: one can freely choose p (the number of basis functions) and

therefore greatly reduce the number of spectral variables. In fact, this idea of using function representation for spectra has been used in the field of chemometrics since [15]. The idea of this early work was to compress the spectra in order to speed up linear methods (PCR and PLSR for instance). We use the same idea to speed up variable selection. As in [16, 17] we use B-splines, however, our additional motivation is the locality of B-splines, that allows us to maintain interpretability.

Let us consider an interval $[a, b]$ and p sub-intervals, defined by the $p + 1$ values, $t_0, \ldots, t_p$, called *knots*, such that $t_j < t_{j+1}$, $t_0 = a$ and $t_p = b$. We recall that splines of order d are C^{d-2} piecewise polynomial functions given by a polynomial of degree $d - 1$ on each interval $[t_j, t_{j+1}[$ (the last interval is $[t_{p-1}, t_p]$). The vector space of such functions has a basis of $p - 1 + d$ *B-splines*, $B_1^d, \ldots, B_{p-1+d}^d$ (see [18] for details).

We consider n spectra, $(s_i)_{1 \leq i \leq n}$ which are functions from $\mathbb{R}$ to $\mathbb{R}$, observed at m wavelengths, $(w_j)_{1 \leq j \leq m}$. We denote a the smallest wavelength and b the largest. Given $p+1$ knots as above, we associate to a spectrum s_i the coordinates $\mathbf{c}(s_i)$ of its best approximation by a spline of order d on the associated B-spline basis. The proposed method consists in replacing the m original variables by the $p - d + 1$ B-spline coordinates.

Those coordinates are the solution of the standard least square optimization problem:

$$\mathbf{c}(s_i) = \arg \min_{\mathbf{c} \in \mathbb{R}^{p-1+d}} \sum_{j=1}^{m} \left(s_i(w_j) - \sum_{k=1}^{p-1+d} c_k B_k^d(w_j) \right)^2 . \tag{1}$$

This quadratic problem leads to a linear solution, i.e. there is a $(p - 1 + d) \times m$ matrix R, that depends only on d, p and $(w_j)_{1 \leq j \leq m}$, such that

$$\mathbf{c}(s_i) = R\mathbf{s}_i, \tag{2}$$

where $\mathbf{s}_i$ denotes the vector representation of s_i, i.e. $\mathbf{s}_i = (s_i(w_1), \ldots, s_i(w_m))$.

3.2 Interpretability

Of course, this type of linear relationship applies to any filtering approach that consists in projecting the considered functions on the sub-vector space spanned by some well-chosen basis functions. An interesting and quite specific aspect of B-splines however it that R is approximately localized. Let us consider more precisely the case of a single spectrum s. The coordinates of its projection are given by:

$$c(s_i)_l = \sum_{j=1}^{m} R_{lj} s_i(w_j).$$

A remarkable property of B-splines is that most coefficients in R have a very small magnitude. In practice, this means that the value of a new variable depends only on some of the original variables. Moreover, dependencies are localized: the

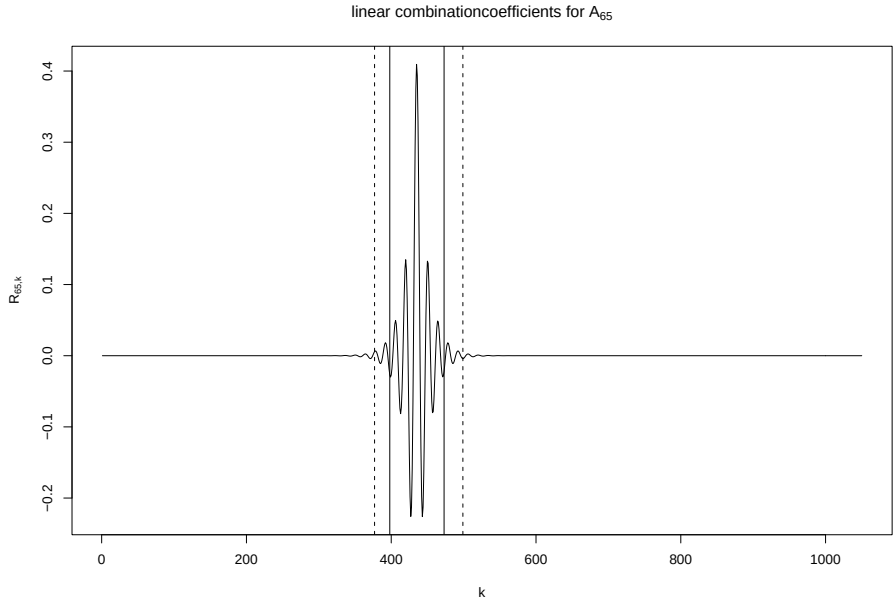

Fig. 1. Graphical representation of $R_{65,k}$ for splines of order 5 with 155 B-splines calculated for 1050 original spectral variables

coordinate of s_i on the B-spline B_k^d depends only on a sub-interval of the original wavelength interval $[a, b]$, as shown on an example in Figure 1.

In theory, values in R are nonzero because when we construct an orthogonal basis of the space of splines of order d, the functions of this basis do not have a smaller support than interval $[a, b]$. Compactly supported wavelet bases [19] provide alternative solutions in which some lines of R have actual zero entries. However, in this case, the low-resolution wavelet spans the full original interval and therefore some lines of R do not have any negligible coefficient. Up to a small approximation, B-splines offer on the contrary a localized basis.

In practice, a wavelength range can be associated to each new variable $c(s)_l$. If we assume the list of wavelengths $(w_j)_{1\leq j\leq m}$ to be in increasing order, and given a precision ratio $\epsilon > 0$, the indexes of the bounds of the interval are

$$l_i = \max\left\{1 \leq j \leq m \;\middle|\; \max_{1\leq k<j} |R_{ik}| < \epsilon \max_{1\leq k\leq m} |R_{ik}|\right\}, \tag{3}$$

$$u_i = \min\left\{1 \leq j \leq m \;\middle|\; \max_{j<k\leq m} |R_{ik}| < \epsilon \max_{1\leq k\leq m} |R_{ik}|\right\}, \tag{4}$$

with the convention that $\max_{1\leq k<1} |R_{ik}| = \max_{m<k\leq m} |R_{ik}| = 0$. The lower bound w_{l_i} corresponds to the largest index j such that all coefficients R_{ik} for $k < j$ are smaller than ϵ times the maximal coefficient. The upper bound w_{u_i}

is defined in a symmetric way. Figure 1 displays two wavelength intervals: the vertical solid lines give the bounds of the interval calculated for $\epsilon = 0.05$ and the dashed lines correspond to $\epsilon = 0.01$.

3.3 Optimal B-splines basis

Obviously, the quality of the new variables depends both on d and on p. For instance $d = 1$ corresponds to a piecewise constant approximation that has generally a low quality. In order to compare possible choices for d and p, we use the leave-one-out error estimate described in [11]. This estimate is based only on the spectra themselves and does not take into account the regression task. It can be implemented very efficiently : for a single spectrum, the cost is $O(p^2 + pm)$. Moreover, because most of the calculation does not depend on the spectrum, the cost for n spectra is $O(p^2 + pmn)$. It should be noted that in spectrometric applications, spectra do not exhibit very strong differences and it is therefore possible to select the optimal basis by using only a small subset of the original data set.

4 Experimental results

4.1 Methodology

In this section we apply the general idea of using a B-spline representation to a spectrometric regression problem. The actual method consists in the following steps:

1. Extraction of the B-spline coefficients for each spectrum. The number and the order of the B-splines is chosen by the leave-one-out error estimate.
2. Selection of the B-spline coefficients through mutual information (MI) maximization with a forward-backward search (as in [3]). Any other variable selection method could be used.
3. Calculation of the wavelength ranges associated to the selected variables, as explained in Section 3.2, with $\epsilon = 0.01$.
4. Construction of a nonlinear model (Radial Basis Function Network, RBFN) on the coefficient selected by the previous step. The meta-parameters of the RBFN are chosen by a 3-fold cross validation technique.

In order to assess the performances of the proposed method, its results are compared to the performances of linear models namely a principal component regression (PCR) and a partial least square regression (PLSR). The numbers of components in the PLSR and in the PCR model are chosen with the same 3-fold cross-validation method used to choose the meta-parameters of the nonlinear model. To motivate the use of a nonlinear model, we also include the results of a standard linear regression (LR) built on the selected variables.

The comparison of the models is done according to the Normalized Mean Square Error (NMSE) they reach on an independent test set.

Finally, we use for comparison a simple method to extract the wavelengths that play a significant role in the prediction of the target variable by the best linear model obtained with PCR or PLSR. The output of such a model can be written

$$y = \alpha_0 + \sum_{j=1}^{m} \alpha_i X(w_j), \tag{5}$$

where $X(w_j)$ is a scaled version of the original input variable $s(w_j)$ (i.e., $X(w_j)$ has zero mean and unit variance). As in section 3.2, we consider that wavelength w_j is important if $|\alpha_j| > \epsilon \max_{1 \leq l \leq m} |\alpha_l|$.

4.2 Results

We use the data set from the software contest organized at the International Display Research Conference held in 1998. It consists of scans and chemistry gathered from fescue grass (*Festuca elatior*). The grass was bred on soil medium with several nitrogen fertilization levels. The aim of the experiments was to try to find the optimum fertilization level to maximize production and to minimize the consequences on the environment. In this context, the problem to address is the following: can NIR spectrometry measure the nitrogen content of the plants?

Although the scans were performed on both wet and dry grass samples, we only consider wet samples here (i.e., the scans were performed directly after harvesting). The dataset contains 141 spectra discretized to 1050 different wavelengths, from 400nm to 2498nm. The nitrogen level goes from 0.8 to 1.7 approximately. The data can be obtained from the Analytical Spectroscopy Research Group of the University of Kentucky[4].

We have split randomly the dataset into a test set containing 36 spectra and a training set with the remaining 105 spectra. The random split has been done in a way that preserve roughly the distribution of the target variable (the nitrogen level).

The leave-one-out error calculation leads to the selection of an optimal basis of 155 B-splines of order 5 (the optimal number of B-splines is chosen in $[50, 500]$) and achieves therefore a good compression ratio. The forward-backward mutual information procedure selects ten coordinates. Both phases take a few minutes on a personal computer, whereas the same variable selection procedure would have taken several hours on the original variables.

The results on the test set (NMSE) for the studied methods are given in Table 1. The 10 variables selected by maximizing the mutual information cannot be used to construct a linear model with performances comparable to the ones of the optimal linear models. The nonlinear model constructed on those variables has clearly the best performances.

While the mutual information maximization leads to the selection of 10 variables, they are calculated using only three intervals of the original wavelength range: $[400, 816]$, $[874, 1118]$ and $[2002, 2478]$. Figure 2 represents the normalized

[4] `http://kerouac.pharm.uky.edu/asrg/cnirs/shoot_out_1998/`

Table 1. Normalized mean square error on the test set for the nitrogen content prediction problem

Method	Variables	NMSE (test)
PCR	10	$1.57\,10^{-1}$
PLSR	9	$1.51\,10^{-1}$
MI + RBFN	10	$1.21\,10^{-1}$
MI + LR	10	$2.59\,10^{-1}$

coefficients used to compute the new variables. It appears clearly that only some original wavelengths are used.

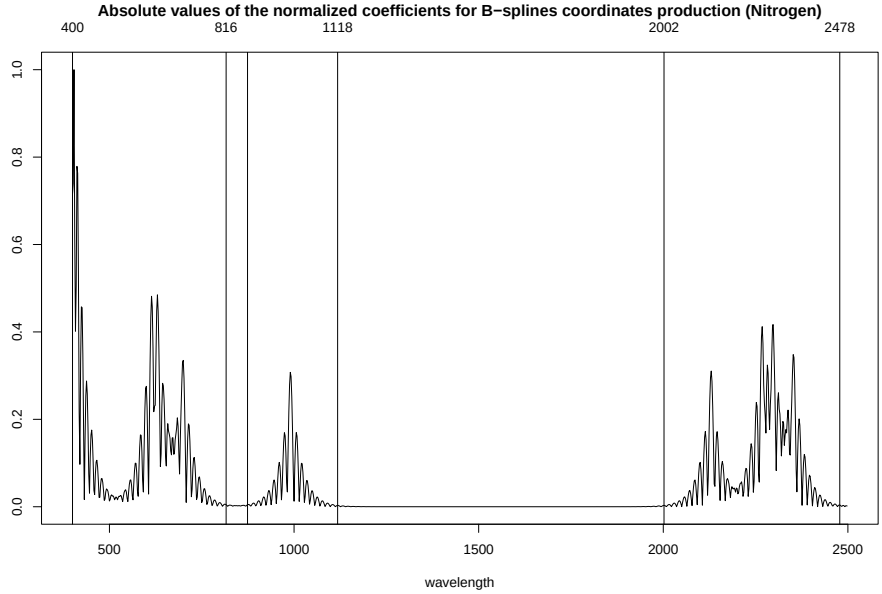

Fig. 2. Normalized absolute value of the coefficients used to compute the selected variables from the original spectral variables

It is not possible to select a few wavelength ranges from the linear model induced by the PLSR: only 17 weights out of 1050 are smaller than $\epsilon = 0.01$ times the higher one in this linear model. As illustrated by Figure 3, the PLSR uses almost the full wavelength range.

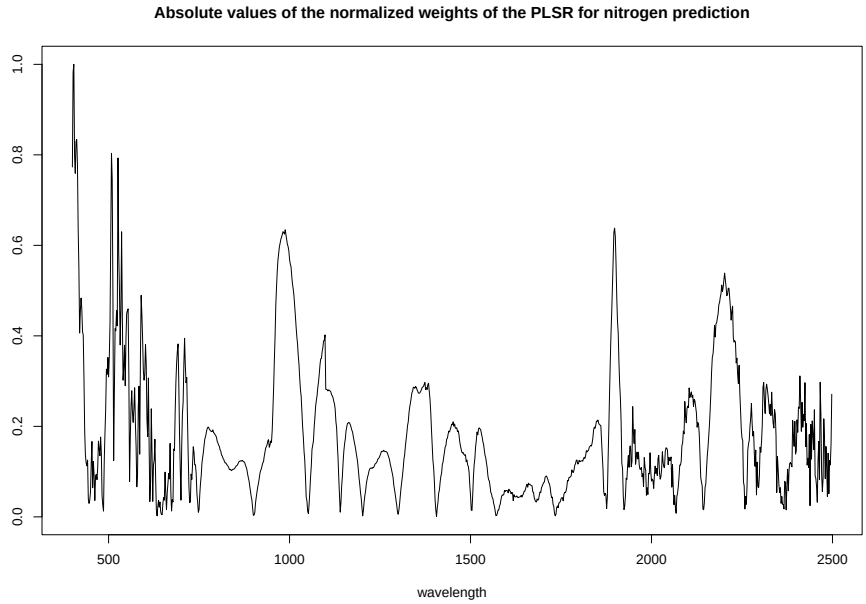

Fig. 3. Normalized absolute values of the coefficients of the linear model induced by PLSR

5 Conclusion

This paper proposes a simple and generic approach for variable selection in spectrometric regression problems. The method is based on the standard filtering approach of Functional Data Analysis: the original spectral variables are replaced by coordinates of the corresponding functions on a B-spline basis. The new variables are still interpretable as each of them depends only on a limited sub-range of the original wavelength interval. The optimal basis is selected by a fast leave-one-out procedure. The prior reduction of the number of variables allows one to use time consuming variable selection methods such as the forward-backward mutual information maximization used in the present paper. The performances obtained on a real-world benchmark are very good. Constructing the full regression model takes a few minutes, compared to several hours that would be needed to run the chosen variable selection procedure on the original spectral variables. On the chosen benchmark, the obtained model outperforms linear models. While these linear techniques are faster, they induce complex dependencies between the target variable and almost all considered wavelengths, and provide therefore no insight on the data. On the contrary, the selected variables are based on only 3 interpretable sub-intervals of the initial wavelength range.

References

1. Benoudjit, N., Cools, E., Meurens, M., Verleysen, M.: Chemometric calibration of infrared spectrometers: Selection and validation of variables by non-linear models. Chemometrics and Intelligent Laboratory Systems **70**(1) (2004) 47–53
2. Benoudjit, N., François, D., Meurens, M., Verleysen, M.: Spectrophotometric variable selection by mutual information. Chemometrics and Intelligent Laboratory Systems **74**(2) (2004) 243–251
3. Rossi, F., Lendasse, A., François, D., Wertz, V., Verleysen, M.: Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. Chemometrics and Intelligent Laboratory Systems **80**(2) (2006) 215–226
4. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer Verlag (1997)
5. Hastie, T., Mallows, C.: A discussion of "A statistical view of some chemometrics regression tools" by I.E. Frank and J.H. Friedman. Technometrics **35** (1993) 140–143
6. Marx, B.D., Eilers, P.H.: Generalized linear regression on sampled signals with penalized likelihood. In A. Forcina, G. M. Marchetti, R.H., Galmacci, G., eds.: Statistical Modelling. Proceedings of the 11th International workshop on Statistical Modelling, Orvietto (1996)
7. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1) (1970) 55–67
8. Pezzulli, S., Silverman, B.: On smoothed principal components analysis. Computational Statistics **8** (1993) 1–16
9. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. Annals of Statistics **23** (1995) 73–102
10. Cardot, H., Ferraty, F., Sarda, P.: Functional linear model. Statist. & Prob. Letters **45** (1999) 11–22
11. Rossi, F., Delannay, N., Conan-Guez, B., Verleysen, M.: Representation of functional data in neural networks. Neurocomputing **64** (2005) 183–210
12. Rossi, F., Conan-Guez, B.: Theoretical properties of projection based multilayer perceptrons with functional inputs. Neural Processing Letters **23**(1) (2006) 55–70
13. Biau, G., Bunea, F., Wegkamp, M.: Functional classification in Hilbert spaces. IEEE Transactions on Information Theory **51** (2005) 2163–2172
14. Rossi, F., Villa, N.: Support vector machine for functional data classification. Neurocomputing **69**(7–9) (2006) 730–742
15. Alsberg, B.K.: Representation of spectra by continuous functions. Journal of Chemometrics **7** (1993) 177–193
16. Alsberg, B.K., Kvalheim, O.M.: Compression of nth-order data arrays by b-splines. part 1: Theory. Journal of Chemometrics **7**(1) (1993) 61–73
17. Olsson, R.J.O., Karlsson, M., Moberg, L.: Compression of first-order spectral data using the b-spline zero compression method. Journal of Chemometrics **10**(5–6) (1996) 399–410
18. de Boor, C.: A Practical Guide to Splines. Volume 27 of Applied Mathematical Sciences. Springer (1978)
19. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Communications in Pure & Applied Mathematics **41** (1988) 909–996