Analytica Chimica Acta 642 (2009) 110–116

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Support vector regression for functional data in multivariate calibration problems

Noslen Hernández[a,*], Isneri Talavera[a], Rolando J. Biscay[b], Diana Porro[a], Marcia M.C. Ferreira[c]

[a] Advanced Technologies Application Center, Havana 12200, Cuba
[b] Institute of Mathematics, Physics and Computation, Cuba
[c] Instituto de Química, Universidade Estadual de Campinas, Brazil

ARTICLE INFO

Article history:
Received 31 July 2008
Received in revised form 14 October 2008
Accepted 15 October 2008
Available online 5 November 2008

Keywords:
Support vector regression
Functional Data Analysis
Multivariate calibration

ABSTRACT

Quantitative analyses involving instrumental signals, such as chromatograms, NIR, and MIR spectra have been successfully applied nowadays for the solution of important chemical tasks. Multivariate calibration is very useful for such purposes and the commonly used methods in chemometrics consider each sample spectrum as a sequence of discrete data points. An alternative way to analyze spectral data is to consider each sample as a function, in which a functional data is obtained. Concerning regression, some linear and nonparametric regression methods have been generalized to functional data. This paper proposes the use of the recently introduced method, support vector regression for functional data (FDA-SVR) for the solution of linear and nonlinear multivariate calibration problems. Three different spectral datasets were analyzed and a comparative study was carried out to test its performance with respect to some traditional calibration methods used in chemometrics such as PLS, SVR and LS-SVR. The satisfactory results obtained with FDA-SVR suggest that it can be an effective and promising tool for multivariate calibration tasks.

1. Introduction

With the fast development of instrumental analysis equipments, such as spectrophotometers, chromatographers, signal analyzers and other modern measurement devices, huge amounts of data, as high-resolution digitized functions, are generated nowadays. As a consequence, regression tasks in which a predictor variable is some type of functional data (FD), and not of a low-dimensional vector, are quite common. For example, the prediction of chemical physical properties of an industrial product from its spectral function is very important nowadays [1,2].

The direct application of classical multivariate regression methods for this type of data exhibits serious limitations. Indeed, digitized functions (e.g., spectral data) are generally represented by high-dimensional vectors whose coordinates are strongly correlated. Furthermore, usually the dimension of such vectors greatly exceeds the number of independent observations (e.g., the number of measured spectra). In such situations, standard regression analysis leads to ill-posed inverse problems, which causes a number of difficulties. In practice, there are a number of approaches to tackle these deficiencies. One of the commonest solution is to use dimension reduction methods together with linear regression, mainly Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) [3]. Whereas those methods generally give satisfactory results, nonlinear relations can only be modeled in a limited way by taking into account more latent variables.

Another way to deal with these problems consists in using variable selection methods to keep only a small number of relevant spectral variables [4,5]. Those methods are less sensitive to overfitting and lead to an easy interpretation of the results, but they are generally quite slow. Although there are some works [6] aimed at improving the computational time of the variable selection methods.

Functional Data Analysis (FDA) is an extension of traditional multivariate analysis that is specifically oriented to deal with observations of functional nature [7]. For this, each object is characterized by one or more continuous real-valued functions, rather than by a finite-dimensional vector. This allows applying functional processing techniques such as derivation, integration, etc. On this basis, several classical multivariate statistical methods have been extended to functional data (FD). Concerning regression, the first works in this direction were focused on linear models [7,8]. Also, some dimensionality reduction approaches for linear regression, such as PCR and PLSR have been generalized to FD [9–12]. More recently, a number of estimation methods for functional nonparametric regression models have also been introduced, namely, estimators based on functional data adaptations of classical neural networks [13], Naradaya-Watson Kernel (NWK) estimators [14,15]

* Corresponding author. Tel.: +53 727 21670; fax: +53 727 30045.
E-mail address: nhernandez@cenatav.co.cu (N. Hernández).

0003-2670/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2008.10.063

and regularization in Reproducing Kernel Hilbert Spaces (RKHS) [16]. These nonparametric techniques require a high amount of computer memory to encode the estimates in order to make future predictions. To avoid this drawback Hernández et al. [17] introduced support vector (SV) regression methods for functional data.

The main goal of this paper is to demonstrate the feasibility and practical performance of Support Vector Regression for functional data (FDA-SVR) in the solution of both linear and nonlinear multivariate calibration chemical problems. Additionally, the performance of this new methodology regarding other well known multivariate calibration methods used in chemometrics such as the traditional PLS and others based on support vectors such as support vector regression (SVR) [18,19] and Least Square Support Vector Machines (LSSVM) [20], is shown.

2. Support vector estimators for functional nonparametric regression

Support vector Regression methods for Functional Data have been introduced recently by Hernández et al. [17]. It is known that estimation methods for very general regression models have been elaborated on the basis of regularization in RKHS [18,19,21].

Let $\mathcal{X}$ be a linear space with norm $\|\|_{\mathcal{X}}$, and $\mathbb{R}^{\mathcal{X}}$ be the set of functions from $\mathcal{X}$ into $\mathbb{R}$. Suppose it is given some positive definite (pd) function (or kernel) $\kappa : \mathcal{X} \times X \to \mathbb{R}$. It is known that there exists a RKHS $\mathcal{H}_\kappa \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel κ. The norm on $\mathcal{H}_\kappa$ will be denoted by $\|\|_{\mathcal{H}_\kappa}$.

Consider the abstract nonparametric regression model:

$$Y = \Psi(X) + e,$$

where (X, Y) is a random variable on some probability space $(\Omega, \mathcal{F}, P)$ with values in $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} \subset \mathbb{R}$, e is a real-valued random variable with zero mean, which is assumed to be independent from (X, Y), and Ψ is an unknown mapping $\mathcal{X} \to \mathbb{R}$.

The problem of interest is to estimate the regression mapping Ψ on the basis of data (X_i, Y_i), $1 \le i \le n$, formed by independent and identically distributed observations of (X, Y).

For this, let $\mathcal{H}_0$ be a given finite-dimensional linear subspace of $\mathbb{R}^{\mathcal{X}}$ with basis $G_1, \ldots, G_m$. Denote by $\mathcal{H} = \mathcal{H}_\kappa + \mathcal{H}_0$ the space of functions $F = F_\kappa + F_0$ with $F_\kappa \in \mathcal{H}_\kappa$ and $F_0 \in \mathcal{H}_0$. Henceforth, it will be assumed that $\Psi \in \mathcal{H}$. Within this framework the regression problem is formulated as a variational problem of finding the function F that minimized the Regularized Empirical Risk, $R_\lambda(F)$

$$\hat{\Psi}_\lambda = arg\min_{F \in \mathcal{H}} R_\lambda(F).$$

The Regularized Empirical Risk is defined by:

$$R_\lambda(F) = R_{\text{emp}}(F) + \lambda \|F_\kappa\|_{\mathcal{H}_\kappa},$$

where the data dependent term $R_{\text{emp}}(F) = (1/n)\sum_{i=1}^{n} c(X_i, Y_i, F(X_i))$ is called Empirical Risk and the second term, called stabilizer, is a norm $\|F_\kappa\|_{\mathcal{H}_\kappa}$ in $\mathcal{H}_\kappa$ and λ is the regularization parameter.

In the case of support vector regression methods the cost function c within the Empirical Risk is the so-called ϵ-insensitive cost function

$$c(X, Y, F(X)) = |Y - F(X)|_\epsilon = \max(|Y - F(X)| - \epsilon, 0),$$

where $\epsilon \ge 0$ is some given constant.

It can be noticed that the general framework for regression estimation through regularization in RKHSs, can be applied for any specification of the space $\mathcal{X}$, the pd kernel κ on $\mathcal{X}$, the contrast function c, and the finite-dimensional subspace $\mathcal{H}_0$ of $\mathbb{R}^{\mathcal{X}}$. Functional nonparametric regression models deal with cases in which $\mathcal{X} \subset \mathbb{R}^{\mathcal{T}}$ is a set of functions $x : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is an infinite-dimensional set. Thus, in regression models with FD, the unknown regression mapping Ψ is a *functional* defined on a normed space $\mathcal{X}$ of real-valued functions.

It is known that under quite general conditions on the cost function c (which are satisfied by the ϵ-insensitive contrast), the regularized estimate $\hat{\Psi}_\lambda$ in any abstract regression model has the following explicit form: for all $x \in \mathcal{X}$,

$$\hat{\Psi}_\lambda(x) = \sum_{i=1}^{n} a_i \kappa(x_i, x) + \sum_{j=1}^{m} b_j G_j(x)$$

for some a_i, $b_j \in \mathbb{R}$ that depend only on the hyperparameter λ and the matrices $K = (\kappa(x_i, x_j))$, $G = (G_j(x_i))$.

Therefore, once the kernel κ is given, the numerical computation of RKHS-based regularized estimators $\hat{\Psi}_\lambda$ in regression for FD input (i.e., when $\mathcal{X}$ is a space of functions) is exactly the same as in regression for multivariate input (i.e., when $\mathcal{X}$ is a subspace of some finite-dimensional linear space $\mathbb{R}^d$). Some results that allow for the construction of some classes of nnd kernels on functional spaces were given in Ref. [17], where the Gaussian kernel appears as a particular case of these class of kernel.

3. Experimental

In order to assess the performance of the proposed method, FDA-SVR is applied to some datasets and the results are compared with other calibration methods widely used in chemometrics like traditional PLS, SVR and LS-SVM.

3.1. Datasets

The FDA-SVR method was applied to three different spectral datasets. The first one, named Tecator dataset, is from the food industry [22]. It consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food and Feed Analyzer. Each observation consists of a 100-chanel absorbance spectrum in the 850–1050 nm wavelength range. Each spectrum in the data base was associated with the percentage of fat, water and protein, determined according to reference methods as described in Ref. [22]. The regression problem in this work consists of predicting the fat content from a meat sample, based on its spectral data. From the 215 spectra, 43 were kept aside as a testing set and the 172 remaining samples are used for model estimation (training set). The raw data was preprocessed, each spectrum was reduced to zero mean and unit variance.

The second dataset contains NIR spectra of ternary mixtures of ethanol, water, and 2-propanol, originally measured and described by Wülfert et al. [23]. A mixture design of 19 different combinations of mole fractions was analyzed in a wavelength range of 850–1049 nm with a resolution of 1 nm (200 wavelengths). The spectral data for each mixture was measured at five different temperatures 30, 40, 50, 60, and 70 °C. This dataset is representative of a well known analytical chemical problem in which NIR spectra of a ternary mixture are nonlinearly affected by temperature-induced spectral variations. As a result, the relations among spectra measured at different temperatures are not straightforward. The same training and test set proposed in Ref. [23] for the construction of global models, were maintained. The test set contains the mixtures 5, 6, 9, 11, 14 and 15 per temperature and the other 13 mixtures per temperature make up the training set (65 objects). In the same way, data pretreatment was performed according to Wülfert's paper (baseline correction and mean-center).

The third and last dataset originated from pharmaceutical tablets [24,25]. This dataset consist of near infrared (NIR) transmit-

tance spectra of pharmaceutical tablets with 310 spectra and 404 variables or wavelengths from 7400 cm^{-1} to 10507 cm^{-1}. Calibration models in this paper were carried out with a relative small data set defined as "preliminary calibration set" in the original paper [24], consisting of 120 samples from the pilot scale. The goal of this analysis set is to predict the active substance content (w/w) of a pharmaceutical tablet. According to the authors [24], the PLS model was capable of achieving acceptable performance, indicating that the inherent data structure was approximately linear. This dataset was selected to investigate if the proposed method can deal with linear problems. The complete dataset was split into 65 samples for training and 55 samples for testing. Multiplicative scatter correction (MSC) was used as preprocessing method.

3.2. Software and optimization

All calculations were carried using Matlab software [26]. SVR calculations were performed using a toolbox for Matlab called spider [27]. For LS-SVM, the Matlab/C toolbox [28] was used. PLS model were built using the PLS Toolbox 3.5 [29]. The calibration method proposed was implemented using the SVM toolbox [30] and the functions implemented in Matlab by Ramsay [31] for Functional Data Analysis.

For the application of FDA-SVR each function (spectra) was represented by a n th order B-spline approximation with p basis functions. The optimal values for the number of B-spline coefficients (p) and the order of the spline basis (n) were selected following a leave-one-out (LOO) cross-validation procedure described in Ref. [32]. The kernel functions used within the kernel-based algorithms (SVR, LS-SVM, FDA-SVR) were the Gaussian kernel or the linear kernel.

The optimal hyperparameters values for training FDA-SVR (λ, ϵ and specific kernel parameter), SVR (λ, ϵ and specific kernel parameter) and LS-SVM (λ and specific kernel parameter) have been tuned using a grid search based on k-fold cross-validation. The number of latent variables used in the PLS models was selected with k-fold cross-validation.

In order to obtain the final prediction error, an independent test set was used. The comparison of the accuracy among the different models was done using RMSEP, defined by:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)} \quad (1)$$

where y_n and $\hat{y}_n$ are the measured and estimated values of the studied property for a sample, respectively, and N is the number of samples in the prediction set.

4. Results and discussion

In this section results of applying FDA-SVR on the three datasets described above are presented as well as the comparison with the other reference methods selected.

4.1. Tecator meat sample dataset

PLS, SVR and LS-SVM models were built for this data set in order to compare their performance with the new methodology proposed.

A Gaussian (RBF) kernel was used in all the kernel-based methods (SVR, LS-SVM, FDA-SVR).

For FDA-SVR, the leave-one-out (LOO) cross-validation procedure, followed by selecting the parameters of the B-spline basis to represent each spectra as a function, leads to the selection of a B-spline basis of order 5 and 48 basis functions. Previous studies on this standard dataset [4,32] have pointed out the relevance of the spectral shape and, for that reason, the L_2 norm of the second derivative of each spectrum was used within the Gaussian kernel used in FDA-SVR.

Table 1
Optimal hyperparameter values

	FDA-SVR	SVR	LS-SVM	PLS
λ	1000	1000	2989	×
rbf (σ)	0.1	0.97	2.13	×
ϵ	0.1	0.5	×	×
# PC	×	×	×	14

Table 1 shows the hyperparameter values resulting from the grid search based on 4-fold cross-validation for FDA-SVR, SVR and LS-SVM. The number of latent variables for PLS was selected by 4-fold cross-validation and appears in this table also.

With this hyperparameter setting all the algorithms were trained and the obtained models were applied to the test set. A comparison of fat content prediction errors for the different methods is shown in Fig. 1.

Fig. 1 shows that FDA-SVR, like the other nonlinear methods, outperforms the traditional PLS (by a factor of 3.5). This behavior is a consequence of the nonlinear relationship between the analyte (fat content) and the spectra, which is being better explained by the nonlinear methods rather than by linear PLS, despite the great number of latent variables in the PLS model. FDA-SVR is more accurate (by a factor of 1.51) than the original SVR, which appears as an evidence that the new method is taking advantage of the functional information provided. It is visible that the results from FDA-SVR and LS-SVM are practically the same differing just by a factor of 1.03. LS-SVM is a simpler method and requires the optimization of fewer parameters than the other support vector-based methods, so the models could be optimized more accurately and probably this might be the reason for its good performance.

The standardized residuals versus predicted values for the test set using FDA-SVR model are graphically represented in Fig. 2.

The data points in the plot seem to be randomly distributed around zero and also, there are no samples with high residuals, suggesting that FDA-SVR model is statistically well posed.

Regarding sparsity, all sparse methods use similar percent of the training data as support vectors. In order to achieve a sparse model with LS-SVM, it is necessary to apply pruning techniques to the

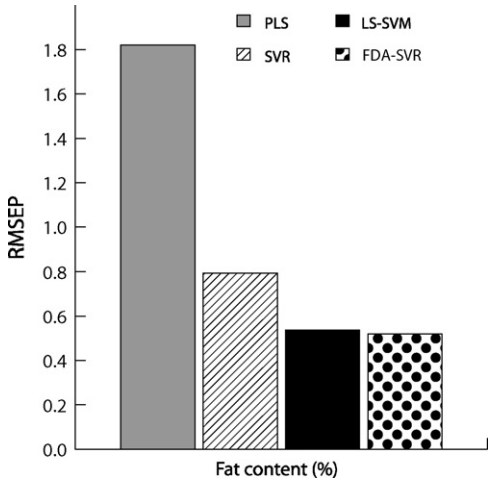

Fig. 1. Performances of different approaches together with the newly presented model based on FDA-SVR for Tecator dataset.

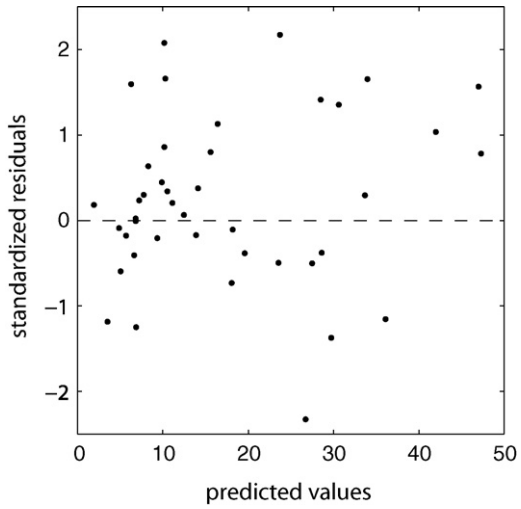

Fig. 2. Standardized residuals versus predicted values of fat content for FDA-SVR model.

Table 2
Number of support vectors for Tecator dataset

	FDA-SVR	SVR	LS-SVM
# SV	150	147	148
%	87	85	86

Lagrange multipliers [20], which implies in an increase of both final error and computational cost. The number of support vectors and the percent of the training set that they represent for each method are shown in Table 2.

4.2. Temperature influenced near-infrared spectra dataset

This dataset has been studied in previous works [33,34], so the results for PLS, SVR and LS-SVM models were gathered from the literature.

FDA-SVR global models were constructed for each compound in the mixture (ethanol, water and iso-propanol). For this, each spectra in this dataset was approximated by a B-spline basis of order 6 and 48 basis functions (the LOO cross-validation procedure mentioned in previous section was used to select these parameter values). The Gaussian (RBF) kernel function was utilized, but for this data, the actual values of spectral variables appeared to be as important as the shape of the spectra. For that reason, not only the information gathered from the second derivative of the spectrum was used but also the values of the spectrum itself. Then, the norm used within the Gaussian kernel was a linear combination of the L_2 norm of the second derivatives of the spectrum and the L_2 norm of the spectrum itself, that is $\|x\| = \|x\|_{L_2} + \|x''\|_{L_2}$.

The corresponding hyperparameter values, resulting from the grid search based on 10-fold cross-validation, for ethanol, water and iso-propanol models, are shown in Table 3.

A comparison of the mole fraction prediction errors of different methods for ternary mixtures of ethanol, water and 2-propanol is shown in Fig. 3. In order to achieve greater clarity, the numerical results for RMSEP are presented in Table 4.

Table 3
Optimal hyperparameter values of FDA-SVR model for each compound

	Ethanol	Water	2-Propanol
λ	500	500	500
rbf (σ)	0.5	0.5	0.5
ϵ	0.003	0.003	0.003

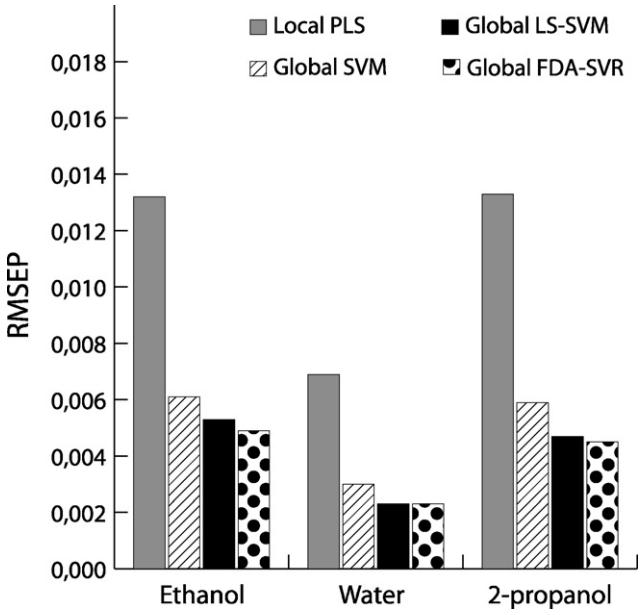

Fig. 3. Performance of different approaches together with the newly presented model based on FDA-SVR, for temperature influenced near-infrared dataset.

It can be seen in Fig. 3 a similar behavior to that of the previous dataset analyzed. Again, the nonlinear approaches perform better than PLS for all the compounds. This result is not surprising if we remember that in this dataset temperature appears as a nonlinear interference and global models are built, for which spectra measured at different temperatures are being used. So, nonlinear methods can explained better than PLS this nonlinearity in the dataset. FDA-SVR is more accurate than the other support vector-based methods for the case of ethanol and iso-propanol, achieving equal performance to LS-SVM in the case of water.

A graphical representation of standardized residuals versus fitted values for each compound is shown in Fig. 4.

Most of the residuals are randomly scattered around zero in Fig. 4. However, there are some important aspects that are worth to comment. It can be noticed that the variance of residuals for low concentration of ethanol is slightly lower than for other concentration values. While the opposite occurs for high concentration of water. Also, the residuals for high concentration of water and iso-propanol are biased, i.e., the residuals are not randomly distributed. These facts indicate that the regression function has uncertainty at the boundaries.

Levels of sparsity achieved by support vector-based methods for ethanol, water and iso-propanol are shown in Table 5. There are not great differences between the percent of training objects used for the different approaches to construct the models. The only notable difference appears in the case of water, where LS-SVM requires almost twice the amount of SVs used by FDA-SVR.

Table 4
RMSEP for all the methods

	RMSEP		
	Ethanol	Water	2-Propanol
PLS	0.0132	0.0069	0.0133
SVR	0.0061	0.0030	0.0059
LS-SVM	0.0053	0.0023	0.0047
FDA-SVR	0.0049	0.0023	0.0045

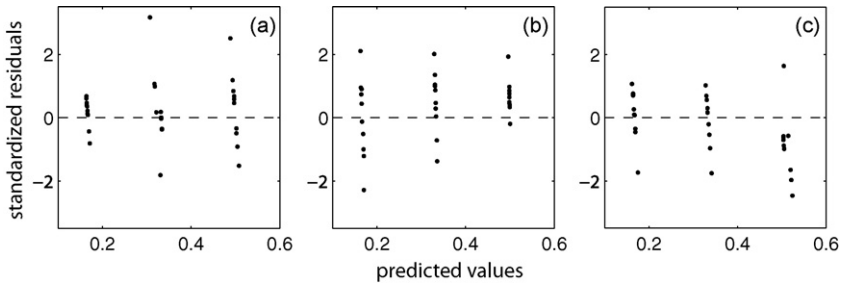

Fig. 4. Standardized residuals versus predicted values of (a) ethanol (b) water and (c) iso-propanol for FDA-SVR model.

Table 5
Number of support vectors used in all analysis for all compounds

	Ethanol		Water		2-Propanol	
	#sv	%	#dsv	%	# sv	%
SVR	39	60	27	41.5	36	55
LS-SVR	37	57	57	88	35	54
FDA-SVR	36	55	30	46	35	54

Because the exact composition of each sample is known in this dataset, the interpretation of the models from the chemical point of view turns to be easier. To exhibit the influence of each training object to the final FDA-SVR solution, the estimated Lagrange multipliers of the models were used. The greater the absolute value of the Lagrange multiplier associated with each sample in the training set, greater its influence in the model. Fig. 5 shows the most important training objects for each model (support vectors), as well as those which do not contribute to the solution ($\alpha_i = 0$). In order to known the relation between the importance of objects and its chemical structure, we follow the procedure explained by Thissen et al. [34], to show its importance in the mixture design proposed by Wülfert et al. [23]. Here, we will study the contribution of the training objects in each model separately. The importance of each mixture point in a design of an specific model, was obtained by taking the mean of the individual five mixture design corresponding to the five different temperatures at which each mixture was measured, and represented with a gray scale color (the greater the influence, darker the color).

Notice that Lagrange multipliers for the three models are in the same range of values (Fig. 5), due to the regularization parameter λ which has the same value in all the models. Despite of we are analyzing the influence of each object for each model separately, comparison between models can be done.

At first sight it can be seen from Fig. 6 that objects with a high mole fraction of ethanol and iso-propanol are the most important in all the models. However, their influence varies from design to design. For example, they seem to have a stronger influence in ethanol and iso-propanol than in water model. In fact, the design corresponding to water reaffirms the fact that it has the greatest sparsity level as was presented in Table 5. Especially the zone of high water concentration, in which the amount of objects practically not used, is greater than in the other designs.

Similar behavior was obtained by Thissen et. al [34] interpreting traditional SVM. According to chemical structures (two homologous alcohols versus water), the NIR spectra of ethanol and 2-propanol are similar whereas the one from water is significantly different. Consequently, it is more difficult for this methods to distinguish ethanol from 2-propanol than ethanol from water, being the mixtures with high concentration of water the best predicted and thus non support vectors objects.

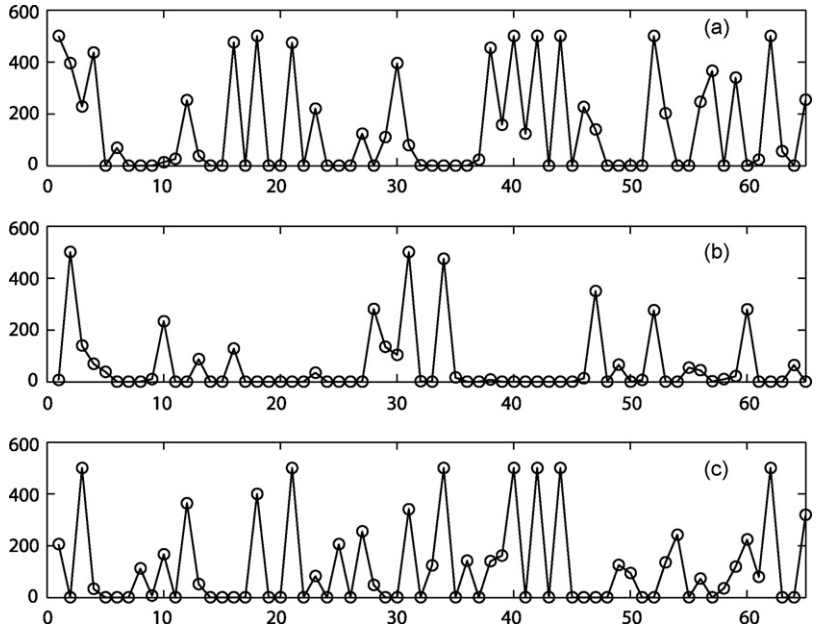

Fig. 5. Lagrange multipliers versus sample number showing the influence of training samples in (a) ethanol (b) water and (c) iso-propanol models.

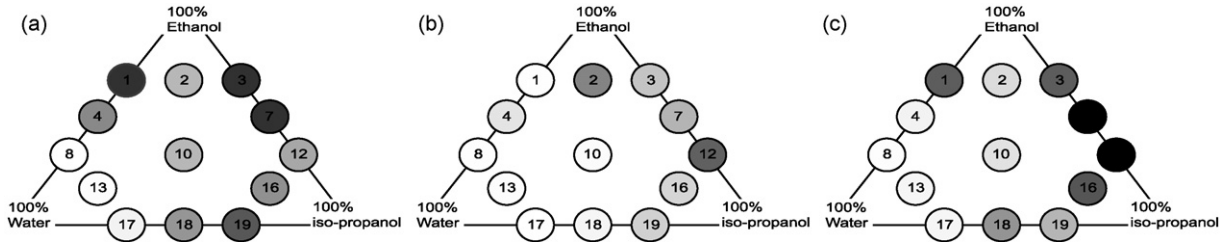

Fig. 6. Importance of training objects for (a) ethanol, (b) water and (c) 2-propanol FDA-SVR models.

Table 6
Optimal hyperparameter values

	FDA-SVR	SVR	LS-SVM	PLS
λ	100	10	46.26	×
rbf (σ)	×	0.61	0.93	×
ϵ	0.08	0.2	×	×
# PC	×	×	×	3

4.3. Tablet dataset

PLS, SVR, LS-SVM and FDA-SVR models were constructed for this dataset. For SVR and LS-SVM, the Gaussian (RBF) kernel was used.

For FDA-SVR, each spectrum is approximated by a B-spline basis of order 4 and 128 basis functions. The kernel function used within this algorithm is a linear kernel, which means that essentially the inner product between functions is taken into account.

The hyperparameter values shown in Table 6 were selected by performing a grid search based on 10-fold cross-validation for support vector-based methods and 10-fold cross-validation for PLS model.

The prediction errors for the active compounds content obtained from different approaches are shown in Fig. 7. It can be noticed that for this dataset, the prediction errors of all nonlinear approaches are very similar (SVR, LS-SVM and FDA-SVR). FDA-SVR is slightly less accurate than SVR and LS-SVM. This may be a consequence of the few information, from the functional point of view, that could be extracted from the data and supplied to the method. For example, information could has been lost by using only the coefficients resulting from the projection of the function in the B-spline basis.

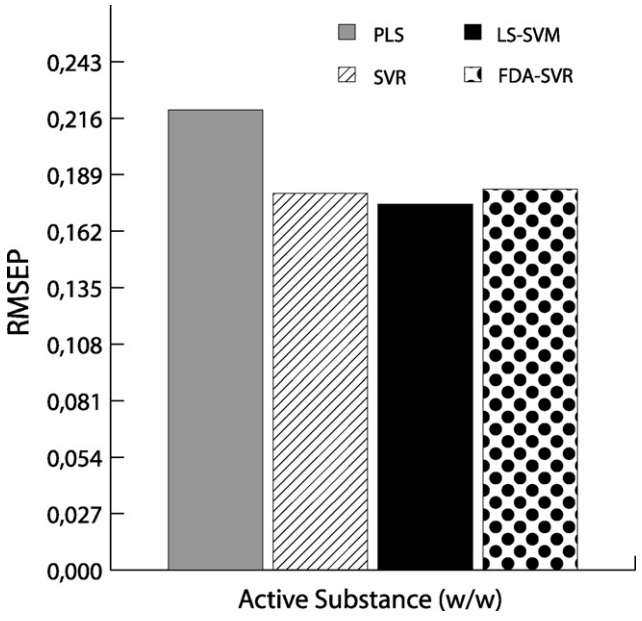

Fig. 7. Performances of different approaches together with the newly presented model based on FDA-SVR for Tablet dataset.

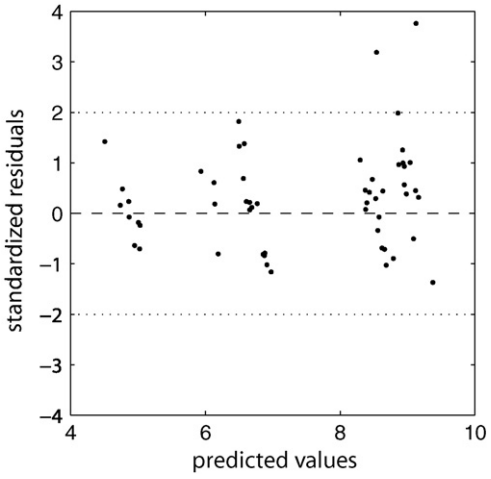

Fig. 8. Standardized residuals versus predicted values of active compound content for FDA-SVR model.

Table 7
Number of support vectors for Tablet dataset

	FDA-SVR	SVR	LS-SVM
# SV	48	34	34
%	73	52	52

Differences between prediction errors from these nonlinear methods and traditional PLS are not so significant as in the other datasets. PLS performs reasonably well, corroborating the strong linear relationship between the spectra and the desired property. So FDA-SVR, as well as the other nonlinear methods, have demonstrated good performance in solving linear problems.

The standardized residuals versus predicted values by FDA-SVR model of the active compound content are graphically represented in Fig. 8.

The residuals are randomly scattered around zero, within the 95% confidence interval given by the dashed lines, except by two samples with high residuals (probably two outliers).

The same number of support vector was selected by SVR and LS-SVM (after applying prunning techniques). FDA-SVR requires more training objects to build the model (see Table 7).

5. Conclusions

This paper proposes a new method based on support vector regression and functional data analysis (FDA-SVR) for multivariate calibration. Its practical feasibility for the solution of both linear

and nonlinear multivariate calibration problems has been shown. The new method was compared with other calibration approaches: PLS, SVR and LS-SVM, showing good performance, what can be explained by the possibility of capturing more information from the shape of the spectrum, due to the functional preprocessing techniques. Although FDA-SVR outperformed the other support vector-based methods in some of the examples studied in this work, it requires more decision to be made, like the choice of the base functions, the choice of internal knots and the number of knots, among others. This implies that no method should be rejected, before investigating which is the most appropriated method for the solution of a given problem. For example, this drawback of FDA-SVR, just mentioned, can be compensated when working with high dimensional spectra, due to the advantage of this method for dealing with high dimensional data. The satisfactory results obtained by using FDA-SVR method suggest that it can be an effective tool for the estimation of physical-chemical properties from NIR spectra, emerging as a very promising tool for multivariate calibration tasks in practice.

References

[1] P. Geladi, Chemometrics in spectroscopy: Part 1, classical chemometrics, Spectrochem. Acta, Part B 58 (2003) 767–782.
[2] H. Swierenga, Multivariate calibration model in vibrational spectroscopic applications, Ph.D. thesis, Univ. of Nijmegen, (2000).
[3] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–148.
[4] F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometr. Intell. Lab. Syst. 80 (2006) 215–226.
[5] R.F. Teófilo, J.P.A. Martins, M.M. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemometr., in press.
[6] F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Fast selection of spectral variables with B-spline compression, Chemometr. Intell. Lab. Syst. 86 (2007) 208–218.
[7] J. Ramsay, B. Silverman, Functional Data Analysis, Springer, 1997.
[8] J. Ramsay, Some tools for functional data analysis, J. Roy. Stat. Soc. Ser. B 53 (1991) 539–572.
[9] A.M. Aguilera, F.A. Ocana, M.J. Valderrama, An approximated principal component prediction model for continuous-time stochastic processes, Appl. Stoch. Models Data Anal. 13 (1997) 61–72.
[10] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statist. Probab. Lett. 45 (1999) 11–22.
[11] C. Preda, G. Saporta, PLS regression on stochastic processes, Comput. Stat. Data Anal. 48 (2005) 149–158.
[12] M.M.C. Ferreira, W.C. Ferreira, B.R. Kowalski, Rank determination and analysis of non-linear processes by global linearizing transformation, J. Chemometr. 10 (1996) 11–30.
[13] F. Rossi, B. Conan-Guez, Functional multi-layer perceptron: a non-linear tool for functional data analysis, Neural Networks 18 (2005) 45–60.
[14] F. Ferraty, P. Vieu, Nonparametric models for functional data with application in regression, time series prediction and curve discrimination, J. Nonparametr. Statist. 16 (2004) 11–125.
[15] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
[16] C. Preda, Regression models for functional data by reproducing kernel hilbert space methods, J. Statist. Plan. Infer. 137 (2007) 829–840.
[17] N. Hernández, R.J. Biscay, I. Talavera, Support vector regression methods for functional data, Lecture Notes Comput. Sci. 4756 (2008) 564–573.
[18] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, 2001.
[19] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, in: Advances in Computational Mathematics, MIT Press, 2000, pp. 1–50.
[20] J. Suykens, T. Gestel, J. Brabanter, B. Moor, J. Vandewalle, Least Square Support Vector Machine, World Scientific, Singapore, 2002.
[21] O. Bousquet, A. Elisseeff, Stability and generalization, J. Mach. Learn. Res. 2 (2002) 499–526.
[22] Tecator dataset, Available at Statlib: http://lib.stat.cmu.edu/datasets/tecator (2007).
[23] F. Wülfert, W.Th. Kok, A.K. Smilde, Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models, Anal. Chem. 70 (1998) 1761–1767.
[24] M. Dyrby, S. Engelsen, L. Nørgaard, M. Bruhn, L.L. Nielsen, Chemometric quantitation of the active substance (containing c n) in a pharmaceutical tablet using near infrared (nir) transmittance and nir ft raman spectra, Appl. Spectrosc. 56 (2002) 579–585.
[25] Tablet dataset, Available at http://www.models.kvl.dk.research/data/Tablets/ (2008).
[26] Matlab 7.3.0, MathWorks, Inc., Natick, MA (2006).
[27] Matlab toolbox for kernel methods: the spider, Available at http://www.kyb.tuebingen.mpg.de/bs/people/spider/ (2006).
[28] Ls-svmlab: a matlab/c toolbox for least squares support vector machines, Available at http://www.esat.kuleuven.ac.be/sista/lssvmlab/ (2007).
[29] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, Pls toolbox 3.5, Eigenvector Research, Inc. (2005).
[30] Support vector machines for classification and regression, Available at http://www.isis.ecs.soton.ac.uk/isystems/kernel/ (2006).
[31] FDA functions, Available at ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns (2007).
[32] F. Rossi, N. Delannay, B. Conan-Guez, M. Verleysen, Representation of functional data in neural networks, Neurocomputing 64 (2005) 183–210.
[33] N. Hernández, I. Talavera, A. Dago, R.J. Biscay, M.M.C. Ferreira, Relevance vector machines for multivariate calibration purposes, J. Chemometr. 22 (2008) 686–694.
[34] U. Thissen, B. Üstun, W. Melssen, L. Bydens, Multivariate calibration with least squares support vector machines, Anal. Chem. 76 (2004) 3099–3105.