Special Issue

Journal of CHEMOMETRICS

Received: 17 July 2007, Revised: 6 April 2008, Accepted: 16 May 2008, Published online in Wiley Interscience: 3 September 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1168

Relevance vector machines for multivariate calibration purposes

Noslen Hernández[a]*, Isneri Talavera[a], Angel Dago[b], Rolando J. Biscay[c], Marcia M. Castro Ferreira[d] and Diana Porro[a]

The introduction of support vector regression (SVR) and least square support vector machines (LS-SVM) methods for regression purposes in the field of chemometrics has provided advantageous alternatives to the existing linear and nonlinear multivariate calibration (MVC) approaches. Relevance vector machines (RVMs) claim the advantages attributed to all the SVM-based methods over many other regression methods. Additionally, it also exhibits advantages over the standard SVM-based ones since: it is not necessary to estimate the error/margin trade-off parameter *C* and the insensitivity parameter ϵ in regression tasks, it is applicable to arbitrary basis functions, the algorithm gives probability estimates seamlessly and offer, additionally, excellent sparseness capabilities, which can result in a simple and robust model for the estimation of different properties. This paper presents the use of RVMs as a nonlinear MVC method capable of dealing with ill-posed problems. To study its behavior, three different chemometric benchmark datasets are considered, including both linear and non-linear solutions. RVM was compared with other calibration approaches reported in the literature. Although RVM performance is comparable with the best results obtained by LS-SVM, the final model achieved is sparser, so the prediction process is faster. Taking into account the other advantages attributed to RVMs, it can be concluded that this technique can be seen as a very promising option to solve nonlinear problems in MVC. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: relevance vector machines; multivariate calibration; Bayesian learning; Kernel methods

1. INTRODUCTION

The goal of multivariate calibration (MVC) is to model the relation between a univariate chemical or physical property and a multivariate set of predictor variables in many cases contained in easily measured spectra. The property of interest is usually determined by using a reference method that is often time consuming, expensive and laborious. A good calibration model should be able to replace the reference method. The most commonly used MVC technique is partial least squares (PLS). This method has the advantage that it deals with the so-called ill-posed problems often produced as a consequence of strong correlation between the measured variables in a spectral data or large number of measurements in comparison with the number of recorded spectra [1–3]. However with PLS, nonlinear relations can only be modeled in a limited way by taking into account more latent variables.

Neural Networks are probably the most popular way of doing high dimensional regression estimation. They are considered to be good to deal with nonlinear spectral data. Nevertheless, they have two main drawbacks [4]. Their architecture has to be determined *a priori* or modified while training by some heuristic which results in a non-necessarily optimal structure of the network and can become a difficult combinatorial problem from multilayer networks. Unfortunately, Neural networks can get stuck in local minima while training and in many cases it is not even desired for the network to obtain its minimum.

The introduction in the field of chemometrics of SVM-based regression methods [5–7] offers advantageous alternatives to the existing linear and nonlinear MVC approaches, due to their capabilities to solve ill-posed problems and lead to global models that are often unique, and exhibit good prediction abilities and good performance when dealing with nonlinear MVC problems, where the common methods are weak.

The excellent performance of SVR and LS-SVM as MVC techniques solving ill-posed problems in a selected case has been demonstrated by Thissen and co-workers [8,9]. These papers show that both SVM and LS-SVM outperform the other models (PLS, two-dimensional penalized signal regression methods (TPSR)), and it can be seen that LS-SVM also performs better than its predecessor SVM. The authors attribute this behavior to the fact that probably LS-SVM can be optimized much more accurately due to its computational simplicity (less parameters and much faster). However, they presented as a possible advantage of SVMs over LS-SVM the fact that usually less

* *Correspondence to: N. Hernández, Advanced Technologies Applications Center, MINBAS, Havana, Cuba.*
E-mail: nhernandez@cenatav.co.cu

a *N. Hernández, I. Talavera, D. Porro*
Advanced Technologies Applications Center, MINBAS, Havana, Cuba

b *A. Dago*
Research Petroleum Center, MINBAS, Havana, Cuba

c *R. J. Biscay*
Institute of Cybernetics, Mathematics and Physics, CITMA, Cuba

d *M. M. C. Ferreira*
Instituto de Química, Universidad Estadual de Campinas (UNICAMP),13083-970, Campinas, SP, Brazil

Copyright © 2008 John Wiley & Sons, Ltd.

support vectors than training objects are required in SVMs models (sparseness). LS-SVM uses all training objects in their final models; hence, no sparseness is obtained, only using pruning techniques applied to the Lagrange multipliers [10] sparse models can be obtained.

In spite of the advantages attributed to the SVMs, the support vector methodology does exhibit significant drawbacks. It does not allow for the liberal use of an arbitrary kernel function $K(\cdot,\cdot)$, because this function must satisfy Mercer's conditions. Furthermore, SVMs require to determine the error/margin trade-off parameter C and the insensitivity parameter ϵ, which generally entails a cross-validation procedure, which is wasteful both of data and computation. Finally, its output is a point estimate instead of the conditional distribution $p(t|\mathbf{x})$ in order to capture uncertainty in the prediction, therefore its predictions are not probabilistic. Taking into consideration these facts, it is possible to achieve improvements in relation with the other SVM-based methods if we are able to solve these disadvantages.

One possible way to approach this objective is the use of relevance vector machine for regression (RVMR), originally introduced by Tipping [11]. RVMR is a probabilistic sparse kernel model identical in functional form to the SVM, where a Bayesian approach to learning is adopted [12,13], introducing a prior over the weights governed by a set of hyperparameters, one associated with each weight whose most probable values are iteratively estimated from the data. Sparsity is achieved because in practice the posterior distributions of many of the weights are sharply peaked around zero [14].

One of the main advantages of the RVMR is its capability to obtain a generalization performance comparable to SVM but using dramatically fewer kernel functions. Furthermore, the RVMR suffers from none of the other limitations of SVM outlined above.

The main goal of this paper is to demonstrate the use of RVMR as a MVC technique capable of dealing with ill-posed problems, specially in dealing with nonlinear data.

2. RELEVANCE VECTOR MACHINES IN REGRESSION

Tipping [11] proposed the relevance vector machine (RVM) to recast the main ideas behind SVMs in a Bayesian context. For a regression problem, given a training dataset $\{\mathbf{x}_n, t_n\}_{n=1}^N$, the following generalized linear regression model can be used to describe the mapping relation between the input pattern vector $\mathbf{x}$ and the scalar target t:

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \quad \mathbf{t} = \mathbf{y} + \boldsymbol{\epsilon} \tag{1}$$

where the 'errors' $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ are modeled probabilistically as independent zero-mean Gaussian, with variance σ^2, so $p(\epsilon) = \prod_{n=1}^N \mathcal{N}(\epsilon_n|0, \sigma^2)$; $\mathbf{w} = (w_1, \ldots w_M)$ is the parameter vector and $y(\mathbf{x}_n, \mathbf{w})$ can be expressed as a linearly weighted sum of some basis functions $\phi(\mathbf{x})$:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) + w_0, \quad \mathbf{y} = \Phi\mathbf{w} \tag{2}$$

Here $\Phi = [\phi_1, \ldots, \phi_M]$ is the $N \times M$ 'design' matrix whose columns comprise the complete set of M 'basis vectors'. Note that the form of the function (2) is equal to the form of the function for an SVM, where we identify our general basis functions with the kernel as parameterized with the training vectors: $\phi_m(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_m)$ and $\phi(\mathbf{x}_n) = [1, K(\mathbf{x}_1, \mathbf{x}_n), \ldots, K(\mathbf{x}_N, \mathbf{x}_n)]$. If bias is not included, Φ will be an $N \times N$ matrix, just eliminating the first column in which all elements are 1.

The error model assumed implied $p(t_n|\mathbf{x}_n) = \mathcal{N}(t_n|y(\mathbf{x}_n), \sigma^2)$, where the notation specifies a Gaussian distribution over t_n with mean $y(\mathbf{x}_n)$ and variance σ^2. Due to the assumption of independence of the t_n, the likelihood of the complete dataset can be written as

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \Phi\mathbf{w}\|^2\right\} \tag{3}$$

Maximum-likelihood estimation of $\mathbf{w}$ and σ^2 from the above equation will generally lead to overfitting problem. Here, though, RVM adopts a Bayesian perspective, and 'constrains' the parameter by defining an explicit prior probability distribution over them, encoding a preference for smoother (less complex) functions by making the popular choice of a zero-mean Gaussian prior distribution:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} \mathcal{N}\left(w_i|0, \alpha_i^{-1}\right) \tag{4}$$

$$= (2\pi)^{-M/2} \prod_{m=1}^{M} \alpha_m^{1/2} \exp\left(-\frac{\alpha_m w_m^2}{2}\right) \tag{5}$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ a vector of *M hyperparameters*. It must be stressed that there is an individual hyperparameter associated independently with every weight, moderating the strength of the prior over its associated weight. It is this form of prior that is ultimately responsible for the sparsity properties of the model [15].

Given $\boldsymbol{\alpha}$, the *posterior* parameter distribution conditioned on the data is given by combining the likelihood and prior within Bayes's rule:

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)} \tag{6}$$

is a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with

$$\Sigma = (\mathbf{A} + \sigma^{-2}\Phi^T\Phi)^{-1} \quad \boldsymbol{\mu} = \sigma^{-2}\Sigma\Phi^T\mathbf{t} \tag{7}$$

and $\mathbf{A}$ is defined as $diag(\alpha_1, \ldots, \alpha_M)$. Rather than extending the model to include Bayesian inference over these hyperparameters (which is analytically intractable), Sparse Bayesian learning can be formulated as a *type-II maximum likelihood* procedure; that is a most probable point estimate $\boldsymbol{\alpha}_{MP}$ may be found throughout the maximization of the *marginal likelihood*, or equivalently, its logarithm $\mathcal{L}(\boldsymbol{\alpha})$ with respect to the hyperparameters $\boldsymbol{\alpha}$ (the same can be done to estimate the hyperparameter σ):

$$\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \log \int_{-\infty}^{\infty} p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})\, d\mathbf{w}$$

$$= -\frac{1}{2}\left[N \log 2\pi + \log|\mathbf{C}| + \mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}\right] \tag{8}$$

Copyright © 2008 John Wiley & Sons, Ltd.

with

$$\mathbf{C} = \sigma^2\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^{\mathrm{T}} \quad (9)$$

Once most probable values $\boldsymbol{\alpha}_{\mathrm{MP}}$ (and σ_{MP}) have been found, a point estimate $\boldsymbol{\mu}_{\mathrm{MP}}$ for the parameters is then obtained by evaluating Equation (7) with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{MP}}$ and $\sigma = \sigma_{MP}$. The crucial observation is that typically the optimal values of many hyperparameters are infinite [15]. From Equation (7), this leads to a parameter posterior infinitely peaked at zero for many weights w_m with the consequence that $\boldsymbol{\mu}_{\mathrm{MP}}$ correspondingly comprises very few non-zero elements. Those training vectors associated with the remaining non-zero weights are called 'relevance' vectors, in deference to the principle of *automatic relevance determination* [15].

Predictions are made based on the posterior distribution over the weights, conditioned on the maximized values $\boldsymbol{\alpha}_{\mathrm{MP}}$ and σ^2_{MP}. The predictive distribution for a new datum $\mathbf{x}_*$, using Equation (6) is defined as follows:

$$p\left(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}\right) = \int p\left(t_*|\mathbf{w}, \sigma^2_{MP}\right) p\left(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}\right) \mathrm{d}\mathbf{w} \quad (10)$$

which is easily computed due to the fact that both terms in the integrand are Gaussian, resulting in a Gaussian too $p(t_*|\mathbf{t}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) = \mathcal{N}(t_*|y_*, \sigma^2_*)$ with:

$$y_* = \boldsymbol{\mu}^T\phi(\mathbf{x}_*) \quad (11)$$

$$\sigma^2_* = \sigma^2_{MP} + \phi(\mathbf{x}_*)^T\Sigma\phi(\mathbf{x}_*) \quad (12)$$

It can be seen that the predictive mean is intuitively $y(\mathbf{x}_*, \boldsymbol{\mu})$ giving a final (posterior mean) approximator $\mathbf{y} = \Phi\boldsymbol{\mu}_{\mathrm{MP}}$

3. EXPERIMENTAL SECTION

In our experiments, RVMR method is compared with other regression methods like PLS, SVR and LS-SVM, whose performance have been studied in the literature [9].

3.1. Datasets

The proposed method is evaluated in three different datasets. The first dataset is related to NIR spectra of ternary mixtures of ethanol, water and 2-propanol, originally measured and described by Wülfert *et al*. [16]. A mixture design of 19 different combinations of mole fractions are analyzed in a wavelength range of 850–1049 nm with a resolution of 1 nm (200 wavelengths). Each mixture is measured at five different temperatures 30, 40, 50, 60, 70°*C*. These data are 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, relations between spectra from different temperatures cannot be made straightforward.

In order to compare our results with previous works presented by Thissen *et al*. [9], we maintain the same test set, containing the mixtures 5, 6, 9, 11, 14, 15 per temperatures and the other 13 mixtures per sample making up the training set (65 objects). In the same way, pretreatment of the spectra has been performed according to Wülfert's paper (baseline corrected and mean-center).

Taking into account the advantages of SVM-based global models to sell-off the prediction for all the temperatures with a unique model, we set up the RVMR global model with a training set data from all the temperatures.

The second dataset named Tecator, comes from the food industry [17]. 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-channel absorbance spectrum in the 850–1050 nm wavelength range and is associated with a content description of meat sample, obtained by analytic chemistry; the percentage of fat, water and protein are reported. The regression problem consists in the prediction of the fat percentage from the spectrum. From the 215 spectra, 43 are kept aside as a testing set and the 172 remaining samples are used for model estimation (training set). Original spectra are preprocessed, each spectrum is reduced to zero mean and unit variance.

We have utilized as third dataset a pharmaceutical tablets dataset [18,19]. This dataset is a near infrared (NIR) transmittance spectra for pharmaceutical tablets with 310 spectra and 404 variables or wavelengths from 7400 to 10507 cm^{-1}. Calibration models in this paper were carried out with a relative small dataset defined as 'preliminary calibration set' in the original paper [18], consisting of 120 samples from the pilot scale. The goal in this dataset is to predict the active substance content (w/w) of a pharmaceutical tablet. Dyrby *et al*. [18] reported that PLS was capable of achieving acceptable performance, indicating that the inherent data structure is approximately linear. We have selected this dataset to investigate if the proposed method can deal with linear problems. We have divided the complete dataset in 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 have been performed using Matlab. SVR was performed using a toolbox for Matlab called spider [20]. For LS-SVM, the Matlab/C toolbox [21] was used. For RVMR calculations, an implementation described by Tipping [15] and developed in Matlab [22] was used.

RVMR models have been created using a Gaussian (RBF) or a linear spline kernel. Its kernel-specific parameter has been tuned using *k*-fold cross-validation. Since the cross-validation procedure was performed over just one parameter (σ or η), the optimization error for each of the parameter values can be visualized in a line plot, so the lowest prediction error can be seen and then the optimal parameter value can be picked.

It is important to notice that while doing the cross-validation, we cannot run RVMR setting an arbitrary kernel-specific parameter. For example for RBF kernel, when the width is too big, the Hessian will become ill-conditioned easily. Also, if there are some repeated input vectors or some of the input vectors are very close to each other, the Hessian will become semi-positive definite and it would lead to numerical troubles when computing its inverse using Cholesky decomposition. So it is difficult to implement the cross-validation for RVMR in practice.

Optimization of the SVR (C, ϵ and specific kernel parameter) and LS-SVM (γ and specific kernel parameter) parameters have been done by a grid search based on *k*-fold cross-validation. Here, a range of parameter values is specified and each combination of parameters values is cross-validated. The combination with the lowest error is selected for training the algorithm.

 Copyright © 2008 John Wiley & Sons, Ltd.

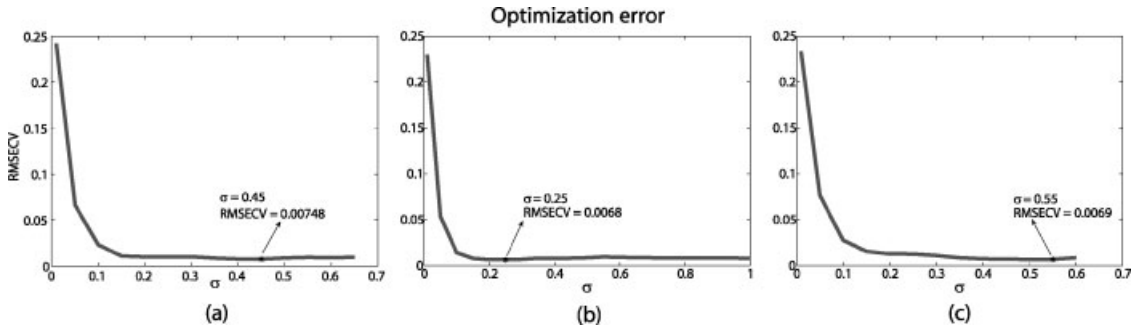

Figure 1. Root mean square errors in cross-validation for (a) ethanol, (b) water and (c) 2-propanol. Indicated with an asterisk is the optimal value of σ.

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

$$\text{RMSEP} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{t}_n - t_n)^2} \tag{13}$$

4. RESULTS

4.1. Temperature influenced near-infrared spectra dataset

Some methods reported in the literature were applied to this dataset with the aim of comparing the performance of RVMR with alternative approaches. Those with best results were selected: SVR global model [8] and LS-SVM [9] global pruning model for nonlinear approaches. Furthermore, a local PLS model [8] (linear regression) was chosen because it is the best PLS approach reported. The global PLS method was not included due to its poor performance, with a reported RMSEP of 0.0195 for ethanol, 0.085 for water and 0.023 for 2-propanol.

In order to obtain the prediction error (RMSEP) values of these methods reported in the literature, we reproduced them running the different algorithms with the parameter values specified in the original papers.

For RVMR, a Gaussian (RBF) kernel was utilized and its single input scale parameter was tuned using fivefold cross-validation on the training set for each one of the three compounds.

Different range of values have been specified for each compound: 0.01–0.65 for ethanol, 0.01–1 for water and 0.01–0.6 for 2-propanol all in steps of 0.05.

The corresponding (σ) values obtained for ethanol, water and 2-propanol as shown in Figure 1 are $\sigma = 0.45$, $\sigma = 0.25$ and $\sigma = 0.55$, respectively. With these σ values, the RVMR was trained again with the complete training set.

A comparison of the mole fraction prediction errors of different methods for ternary mixtures of ethanol, water and 2-propanol are shown in Figure 2.

Figure 2 shows that RVMR, like the other nonlinear methods, performs better than the PLS-based models. It can be noticed that LS-SVM is slightly more accurate than RVMR (in a factor of 1.16) but the latter has better performance than the original SVR (in a factor of 1.08), except for the case of water.

A graphical representation of real values versus predicted values of the dependent variable in the test set for each component using the RVMR method is presented in Figure 3. It confirms that prediction accuracies (generalization capabilities) are quite good for all of them.

However, the RVMR method exhibits some important advantages over the LS-SVM and the SVR methods. Table I presents the numbers of support vectors for SVM and LS-SVM [8,9] and the relevance vectors for RVMR needed to build each model for the prediction of the three chemical components.

It is evident that RVMR method is drastically the most sparse; a few relevant vectors are required to build a good model. In all cases, the sparsity is reduced more than a half with respect to the other methods implying that less than a quarter of the training set objects contribute to the fitted model.

LS-SVM normally uses all training objects in the final model. To achieve some sparsity in the model, it is necessary to apply pruning techniques to the Lagrange multipliers [10], which implies an increase of both final error and computational cost. Nevertheless, neither the SVR nor the pruned LS-SVM can reach the levels of sparsity that are obtained by the RVMR method.

It is interesting to notice that, in the case of water, even though more relevance vectors were required by the RVMR method in comparison with the other components, it is precisely in this case where it does not outperform the other nonlinear approaches. When Thissen *et. al.* [9] pruned the water LS-SVM model, they could not reach the low number of support vectors

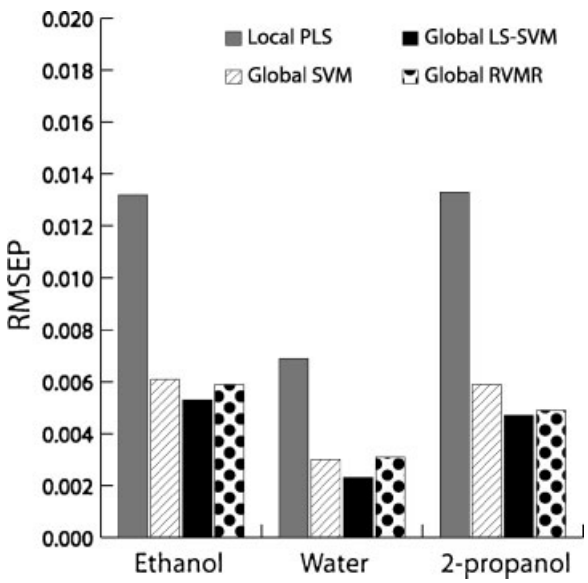

Figure 2. Performances of different approaches together with the newly presented global model based on RVMR.

Copyright © 2008 John Wiley & Sons, Ltd.

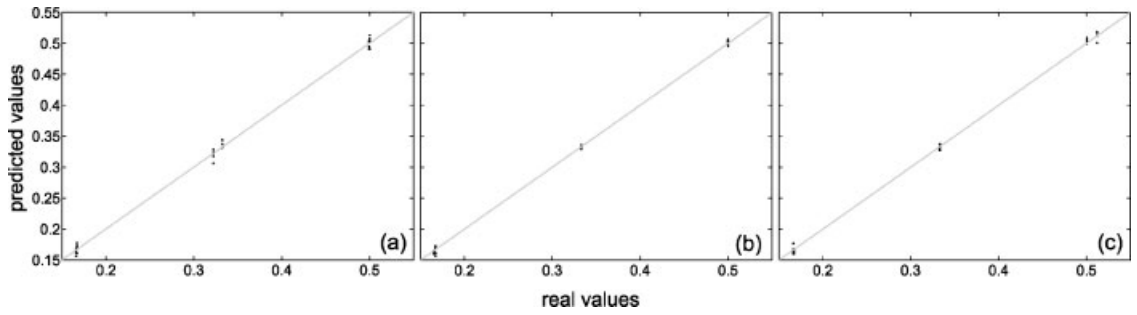

Figure 3. Real values versus predicted values for (a) ethanol, (b) water and (c) 2-propanol for RVMR models.

Table I. Number of vectors used and the percent of the training set that they represented for each method

	Ethanol		Water		2-propanol	
	# SV	%	# SV	%	# SV	%
SVR	39	60	27	41.5	36	55
LS-SVR	37	57	57	88	35	54
	# RV	%	# RV	%	# RV	%
RVMR	13	20	16	24	14	21

as SVR (41.5%), because the prediction error increased up to 0.0071. Hence, to obtain results comparable with SVR, they had to maintain the number of support vectors equal to 57(88%). However, RVMR achieves in this case much more sparsity (almost half of the number of vectors used in SVR), although not better RMSEP value. It is possible that such high level of sparsity shown by RVMR is attempting against improvement of accuracy for the case of water, although this is the best predicted compound.

To exhibit the influence of each training object to the final solution, the estimated inverse variances values of the models were used. Notice that when we talk in terms of RVMR, the estimation of the inverse variance parameter α_i according to the maximum likelihood II method can be considered as a weighting of the objects 'importance' in the regression. According to this, the highest alpha values (lowest α_i^{-1}) are assigned to the least relevant objects for the model fitting.

Figure 4 shows the most important training objects for each model (relevance vector), as well as those who do not contribute to the solution ($\alpha_i \to \inf$ or $\alpha_i^{-1} = 0$).

Due to the fact that the inverse variance values differ in magnitude for each model (notice that the scales are different), we decided to study the contribution of the training objects in each model separately.

In order to show the location of the most important training objects in each mixture design (proposed by Wülfert *et al*. [16]), we follow the procedure explained by Thissen *et al*. [9]. Hence, the importance of each mixture point in a design has been obtained by taking the mean of the individual five-mixture design corresponding to the five different temperatures.

Mixture design representing the model for predicting each compound shows different distributions of the important training objects. The relative importance of each object into a mixture design is different too (Figure 5). Nevertheless there is a common tendency in all the designs. It can be seen that all of them

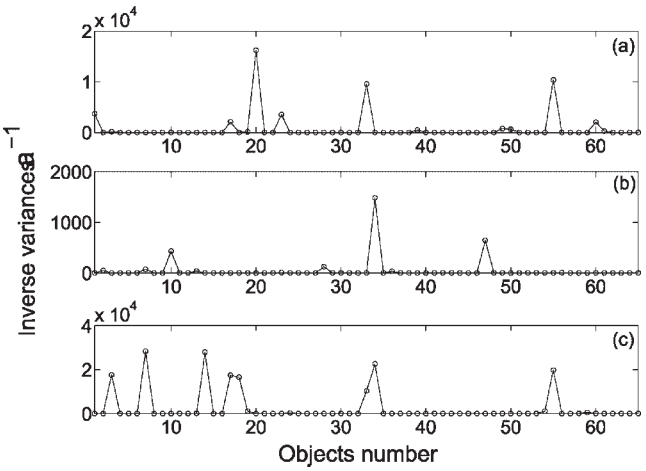

Figure 4. Inverse variances of the objects for (a) ethanol, (b) water and (c) 2-propanol models.

uses training objects with a high mole fraction of ethanol and 2-propanol and a low fraction of water. Similar behavior was obtained by Thissen *et al*. [9] using the traditional SVM. This was explained from the fact that the NIR spectra of ethanol and 2-propanol are similar while the one from water deviates much more. Consequently, it is possible to conclude that it is more difficult to distinguish ethanol from 2-propanol than ethanol from water.

4.2. Tecator meat sample dataset

For this dataset, SVR and LS-SVM were carried out in order to compare their performance with the new methodology proposed. We also compare RVMR in this dataset with an approach proposed by Rossi *et al*. [23] that combines variable selection based on mutual information and LS-SVM (MI-LSVM). Performance of different regression methods on this traditional benchmark are reported in Rossi's paper. The RMSEP for PLS was gathered from this work.

A Gaussian (RBF) kernel was used for both SVR and LS-SVM. A grid search based on k-fold cross-validation have been done with $k = 4$ for selecting the hyperparameter values which result in ($C = 1000$, $\sigma = 0.97$, $\epsilon = 0.5$) for SVR and ($\gamma = 2989.21$, $\sigma = 2.13$) for LS-SVM.

For RVMR, a linear spline kernel was used and its single scale parameter ($\eta = 3.52$) was chosen with fourfold cross-validation.

A comparison of fat content prediction errors for the different methods are shown in Figure 6.

 Copyright © 2008 John Wiley & Sons, Ltd.

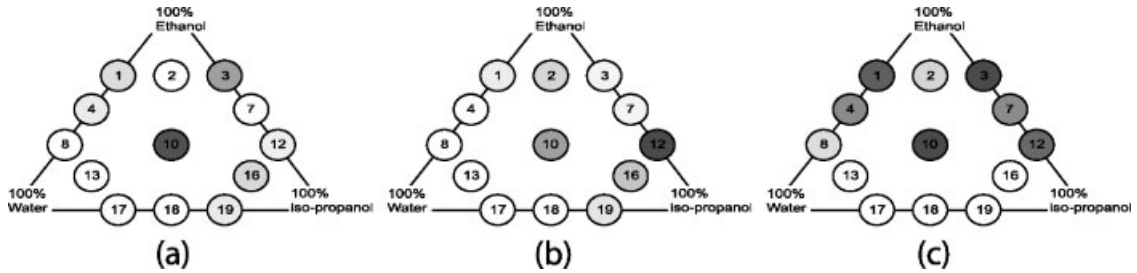

Figure 5. Relative importance of training objects for (a) ethanol, (b) water and (c) 2-propanol.

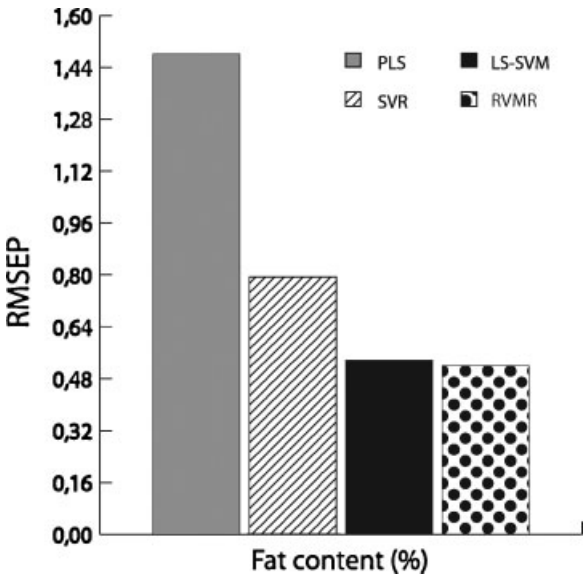

Figure 6. Performances of different approaches together with the newly presented model based on RVMR for Tecator dataset.

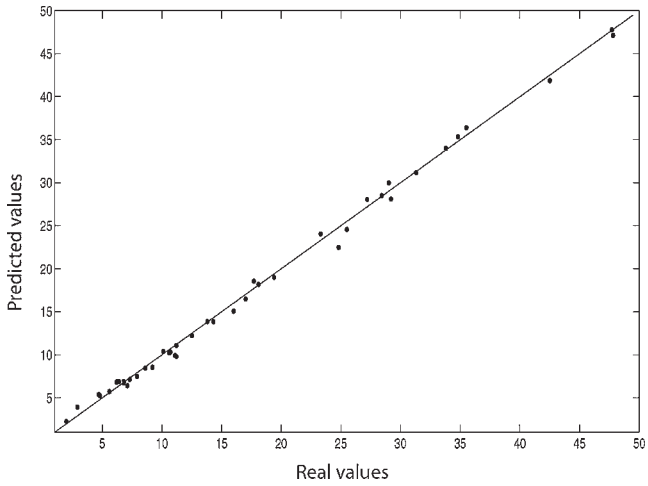

Figure 7. Real values versus predicted values for RVMR model.

Table II. Number of vector used and the percent of the training set that they represented for each method			
	SVR	LS-SVM	RVMR
# SV	147	148	35
%	85	86	20

Figure 6 shows almost the same behavior than in the previous dataset. All nonlinear methods outperform the traditional PLS. The only difference appears that in this dataset RVMR is slightly more accurate than LS-SVM. Real values versus predicted values on the test set using RVMR model are graphically represented in Figure 7.

Interesting results for this dataset was reported by Rossi *et al*. [23]. He compared combination of lineal and nonlineal methods with dimensionality reduction techniques and variable selection. His better results on this dataset were obtained with MI-LSVM, which present an RMSEP of 0.66 (the NMSE reported by Rossi was converted to RMSE for comparing). Predictions achieved with the RVMR model built with all original variables are better than predictions achieved by MI-LSSVM model built on a subset of selected variables. Although Rossi in his paper obtained the best results with MI-LSSVM, he did not include in his comparisons any model built on original variables. But, instead of thinking in RVMR as an alternative to variable selection, it would be interesting to investigate in future works if RVMR combined with variable selection, in the same manner as Rossi did with LS-SVM and RBF, improve RVMR results obtained here.

Table II confirms that RVMR is drastically more sparse than SVR and LS-SVM. RVMR model required just 20% of the training set, less than a quarter of the training samples required by the other methods. In the case of LS-SVM, prunning techniques was applied for obtaining a sparse model.

Figure 8 shows the influence of each sample in the training set for SVR, LS-SVM and RVMR. The estimated inverse variances were taken into account for RVMR model and estimated Lagrange multipliers for SVR and LS-SVM models. There are samples that are used by the three methods as well as others that are not significant for any of them. This is the dataset in which RVMR results less sparse, but it can be noticed that there are a great difference with the other two methods.

4.3. Tablet dataset

SVR, LS-SVM and RVMR models were constructed for this dataset. The PLS results reported in the literature were reproduced [18,24].

For SVR and LS-SVM parameters selection, the grid search was carried out based on 10-fold cross-validation and a Gaussian (RBF) kernel was used. The resulting hyperparameters values were ($C = 10$, $\epsilon = 0.2$, $\sigma = 0.61$) for SVR and ($\gamma = 46.261$, $\sigma = 0.65$) for LS-SVM.

For RVMR, a Gaussian (RBF) kernel was used and 10-fold cross-validation to select its scale parameter value ($\eta = 0.7$).

Copyright © 2008 John Wiley & Sons, Ltd.

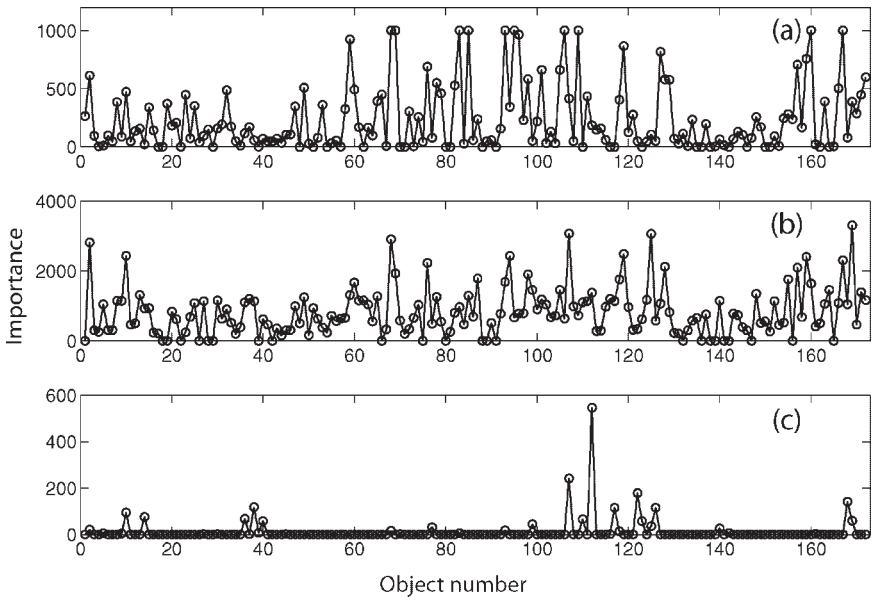

Figure 8. Importance of the objects for (a) SVR, (b) LS-SVM and (c) RVMR.

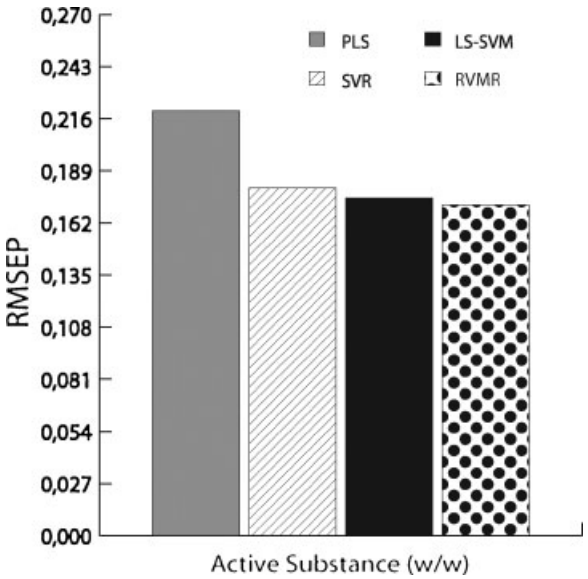

Figure 9. Performances of different approaches together with the newly presented model based on RVMR for Tablet dataset.

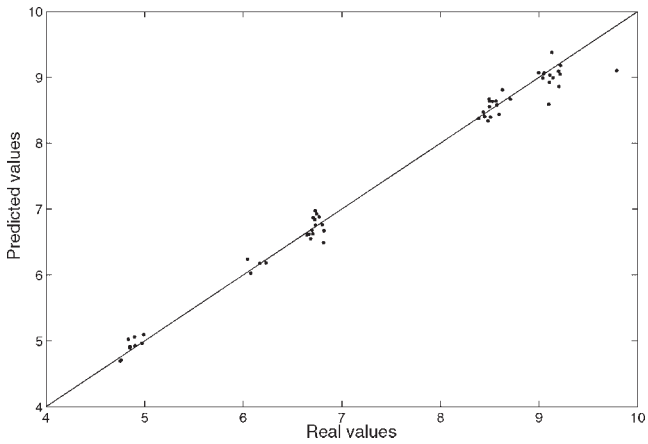

Figure 10. Real values versus predicted values of w/w for RVMR model.

The active substance content prediction errors for the different approaches are shown in Figure 9.

It can be noticed that in this dataset, prediction errors of all nonlinear approaches are very similar (SVR, LS-SVM and RVMR). RVMR is slightly more accurate than SVR and LS-SVM. Also, the differences between prediction errors of these methods and traditional PLS is not so significant as in the other datasets. PLS performs reasonably well on this dataset, corroborating the strong linear relationship between the spectra and the desired property; so RVMR, as well as the other nonlinear methods, have demonstrated its good performance solving linear problems.

This dataset was also studied by Chen *et al*. [24]. They proposed the use of Gaussian process (GP) for MVC. GP and RVMR have many things in common. Indeed, RVMR can be interpreted as a special case of a GP, just that RVMR presents the sparsity property. Support vector-based methods and nonlinear methods reported by Chen *et al*. (artificial neural network (ANN) and GP) present very similar performance on this dataset; especially for the case of GP. Chen *et al*. stated that maybe a weak nonlinearity is present in this dataset and for that reason nonlinear methods are more accurate. Support vector-based methods and particulary RVMR works very well in this dataset, achieving an estimate function that does not overfit this possible weak nonlinearity present in this dataset and generalizes well for unseen data.

The same number of support vector was selected by SVR and LS-SVM (after applying prunning techniques). RVMR again is much more sparse, retaining only a 6% of the training data. These results are shown in Table III

The influence of samples in the training set for each method are shown in Figure 11. Those objects which do not influence in any of the models were marked in black. The great sparsity of RVMR can be seen; all samples used by this method were used by the others too. Also, there are many samples in common used by SVR and LS-SVM.

 Copyright © 2008 John Wiley & Sons, Ltd.

Table III. Number of vector used and the percent of the training set that they represented for each method

	SVR	LS-SVM	RVMR
# SV	34	34	4
%	52	52	6

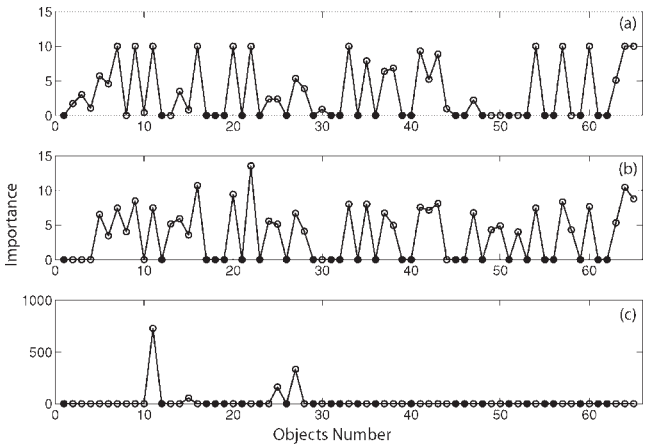

Figure 11. Importance of the objects for (a) SVR, (b) LS-SVM and (c) RVMR.

It should be pointed out that the prediction process of new objects using the RVMR models obtained in this paper, compared to the previously applied models is much faster, due to sparsity. This is an important advantage if prediction has to be performed online. The RVMR algorithm used in this paper makes no use of specially developed efficient algorithms for training. More efficient training approaches for RVMR are discussed in Faul *et al*. [25].

5. CONCLUSIONS

In this paper, a new method for nonlinear MVC based on RVMR was presented. Good results obtained with RVMR suggest that they constitute an effective tool that could lead to improvements in other physical–chemical properties estimation from NIR spectra. RVMR also demonstrates excellent sparseness capabilities, which can result in simple and accurate models for the estimation of these properties.

RVMR was compared with other methods: PLS, SVR and LS-SVM. Two nonlinear dataset were used, one of them with NIR spectra affected by nonlinear temperature-induced variation in which Global models were applied, by implicitly taking the temperature into account. One linear dataset was studied for testing the performance of RVMR in a linear problem, giving excellent results. For all these dataset, RVMR prediction accuracy was comparable to those reached by SVR and LS-SVM, obtaining in some cases the best results and achieving the greatest levels of sparsity. However, this does not allow to absolutely conclude that RVMR is always the best method to be used. This should be determined for each concrete problem.

The inverse variance in the RVMR model and Lagrange multipliers in SVR and LS-SVM were used to indicate the relative importance of each training object which aids in the interpretation of the results.

Although RVMR has the advantages mentioned above, it also has some limitations in practical applications. The learning procedure of RVMR is usually much slower than the SVR and LS-SVM. It is a problem that requires $O(N * M^2)$ operations and $O(M^2)$ storages due to the necessity of repeatedly computing and inverting the Hessian matrix. Then the training set cannot be too large. But for small dataset, as those used in this paper, its computational cost is comparable with SVR and LS-SVM, especially due to the fact that in RVMR just one parameter has to be tuned, unlike SVR and LS-SVM which have to adjust three and two parameters, respectively. Also in the case of LS-SVM, for obtaining sparse models, it is necessary to do an extra prunning process, that also consumes computational time.

Acknowledgements

The authors thank the reviewers for their detailed and constructive comments.

REFERENCES

1. Swierenga H. Robust multivariate calibration models in vibrational spectroscopic applications. *Ph.D. Thesis*, University of Nijmegen, 2000.
2. Hageman JA, Streppel M, Wehrens R, Bydens LMC. Wave length selection with Tabu search. *J. Chemom.* 2003; **17**: 427–437.
3. Estienne F, Massart DL. Multivariate calibration with Raman data using fast principal components and partial least square method. *Anal. Chim. Acta* 2001; **450**: 123–129.
4. Bishop CM. *Neural Networks for Pattern Recognition*. Clarendon Press: Oxford, 1995.
5. Belousov AI, Versakov SA, Von Frese J. A flexible classification approach with optimal generalization performance: support vector machines. *Chemon. Intell. Lab. Syst.* 2002; **64**: 15–25.
6. Belousov AI, Versakov SA, Von Frese J. Application aspects of supports vector machines. *J. Chemom.* 2002; **16**: 482–489.
7. Thissen U, van Brakel R, de Weijer AP, Melssen WJ, Bydens LMC. Using support vector machines for time series prediction. *Chemon. Intell. Lab. Syst.* 2003; **69**: 35–49.
8. Thissen U, Pepers M, Üstun B, Melssen WJ, Bydens LMC. Comparing support vector machines to PLS for spectral regression applications. *Chemon. Intell. Lab. Syst.* 2004; **73**: 169–179.
9. Thissen U, Üstun B, Melssen WJ, Bydens LMC. Multivariate calibration with least-squares support vector machines. *Anal. Chem.* 2004, **76**: 3099–3105.
10. Suykens JAK, Van Gestel T, De Branmaster J, De Moor B, Vandewalle J. Least Square Support Vector Machine. World Scientific: Singapore, 2002.
11. Tipping ME. The relevance vector machines. In *Advances in Neural Information Processing Systems 12*, Solla SA, Leen TK, Muller K-R (eds). MIT Press: Cambridge, MA, 2000; 652–658.
12. Mackay DJC. Bayesian interpolation. *Neural Comput.* 1992; **4**(3): 415–447.
13. Berger JO. *Statistical Decision Theory and Bayesian Analysis* (2nd edn). Springer: New York, 1985.
14. Faul AC, Tipping ME. Analysis of sparse Bayesian learning. In *Advances in Neural Information Processing Systems 14*, Dietterich TG, Becker S, Ghahramani Z (eds). MIT Press: Cambridge, MA, 2002; 383–389.
15. Tipping ME. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 2001; **1**: 211–244.
16. Wülfert F, Kok W. Th, Smilde AK. *Anal. Chem.* 1998; **70**: 1761–1767.
17. Tecator dataset. Available at Statlib: http://lib.stat.cmu.edu/datasets/tecator
18. Dyrby M, Engelsen SB, Nrgaard L, Bruhn M, Lundsberg-Nielsen L. Chemometric quantitation of the active substance in a pharmaceutical tablet using near infrared (NIR) transmittance and NIR FT Raman spectra. *Appl. Spectrosc.* 2002; **56**(5): 579–585.
19. Tablet Dataset. Available at http://www.models.kvl.dk.research/data/Tablets/
20. Matlab Toolbox for Kernel Methods: The Spider. Available at http://www.kyb.tuebingen.mpg.de/bs/people/spider/

Copyright © 2008 John Wiley & Sons, Ltd.

21. LS-SVMlab: a Matlab/C toolbox for least squares support vector machines; 2002. Available at http://www.esat.kuleuven.ac.be/sista/lssvmlab/
22. Tipping ME. "SparseBayes" a Matlab implementation of sparse Bayesian learning, 2001. Available at http://research.microsoft.com/mlp/rvm/
23. Rossi F, Lendasse A, Francois D, Wertz V, Verleysen M. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometr. Intell. Lab. Syst.* 2006; **80**(2): 215–226
24. Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. *Chemon. Intell. Lab. Syst.* 2007; **87**: 59–71.
25. Tipping M, Faul A. Fast marginal likelihood maximization for sparse Bayesian models. In Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics, 2003.

 Copyright © 2008 John Wiley & Sons, Ltd.