Analyst

Cite this: *Analyst*, 2011, **136**, 1703

www.rsc.org/analyst

PAPER

Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data

Roman M. Balabin*[a] and Ekaterina I. Lomakina[b]

Received 8th June 2010, Accepted 31st January 2011
DOI: 10.1039/c0an00387e

In this study, we make a general comparison of the accuracy and robustness of five multivariate calibration models: partial least squares (PLS) regression or projection to latent structures, polynomial partial least squares (Poly-PLS) regression, artificial neural networks (ANNs), and two novel techniques based on support vector machines (SVMs) for multivariate data analysis: support vector regression (SVR) and least-squares support vector machines (LS-SVMs). The comparison is based on fourteen (14) different datasets: seven sets of gasoline data (density, benzene content, and fractional composition/boiling points), two sets of ethanol gasoline fuel data (density and ethanol content), one set of diesel fuel data (total sulfur content), three sets of petroleum (crude oil) macromolecules data (weight percentages of asphaltenes, resins, and paraffins), and one set of petroleum resins data (resins content). Vibrational (near-infrared, NIR) spectroscopic data are used to predict the properties and quality coefficients of gasoline, biofuel/biodiesel, diesel fuel, and other samples of interest. The four systems presented here range greatly in composition, properties, strength of intermolecular interactions (*e.g.*, van der Waals forces, H-bonds), colloid structure, and phase behavior. Due to the high diversity of chemical systems studied, general conclusions about SVM regression methods can be made. We try to answer the following question: to what extent can SVM-based techniques replace ANN-based approaches in real-world (industrial/scientific) applications? The results show that both SVR and LS-SVM methods are comparable to ANNs in accuracy. Due to the much higher robustness of the former, the SVM-based approaches are recommended for practical (industrial) application. This has been shown to be especially true for complicated, highly nonlinear objects.

1. Introduction

Modern quality control of industrial products, such as food products, pharmaceuticals, and petroleum products, is in need of rapid, robust, and cheap analytical methods to continuously monitor product quality parameters.[1–9] Ideally, product quality parameters would be measured in real time, online, which would reduce the amount of waste or production of defective goods, minimize the amount of raw materials and energy consumption required, optimize product quality (*e.g.*, maximizes gasoline octane number during fraction mixing and compounding), and minimize environmental impact.[1–11] These factors are especially important when dealing with the multi-trillion (US) dollar, environmentally unfriendly petroleum industry.[12,13] For example, the ability to increase the gasoline octane number by one–two units or the yield of diesel fuel from crude oil by 1% could lead to enormous financial and environmental benefits.

To control the quality of industrial products in an online regime, spectroscopic methods are often used.[1–9] Vibrational spectroscopy[14–17] (mid-infrared (MIR), Raman, and near infrared (NIR)) is one of the best ways to obtain information about chemical structure and quality coefficients of different mixtures, even multicomponent mixtures. Alternative analytical methods include ultraviolet-visible (UV-Vis) absorption spectroscopy,[18] nuclear magnetic resonance (NMR) spectroscopy,[19] gas or high pressure liquid chromatography (GC/HPLC),[20,21] and mass spectrometry.[22] The latter is frequently combined with a soft ionization technique, such as matrix-assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI).[23]

The relatively low cost of modern MIR/NIR/Raman spectrometers compared to mass spectrometers or NMR spectrometers makes vibrational spectroscopy the technique of choice for real-world applications. The possibility of remote quality control *via* fiber optics, which is easily achievable in the NIR spectrum, makes NIR spectroscopy one of the most promising analytical techniques for industrial applications.[10,11,24,25]

[a]*Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland. E-mail: balabin@org.chem.ethz.ch; Tel: +41-44-632-4783*
[b]*Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 119992 Moscow, Russia*

The combination of an information-rich analytical technique, such as NIR spectroscopy, with efficient regression tools, provided by modern mathematics, makes the creation of accurate and robust methods for prediction of object properties possible.[26–28] The analysis of such sophisticated, multicomponent, and "dirty" samples as petroleum (whose composition, properties, and even structure[29,30] can vary greatly over time or by oil source) is almost impossible without multivariate data analysis (MDA) techniques. The progress in chemometrics has a direct influence on the field of analytical chemistry.[4–7] The modern petroleum industry is in need of accurate and reliable calibration methods.[4–7] The same can be said about the modern and rapidly growing biofuel industry.[31] Note that Geladi[32] has provided a general overview of the subject, including a description of how chemometrics can be used for data analysis, classification, curve resolution, and multivariate calibration with spectroscopic data.

The partial least squares (PLS) or projection to latent structures regression method appeared many years ago and has become extremely popular.[33] Together with its variants and modifications, the PLS calibration model is the most widely used regression technique for spectroscopic data analysis.[33] The greatest problem in PLS methodology is that the spectrum–property relationship is assumed to be linear. This assumption is not always valid for industrial samples, and it is completely unacceptable for systems with strong intermolecular or intramolecular interactions, including π-stacking[29,34,35] and hydrogen bonding.[35–38] The shifts in positions of vibrational bands[15–17,35–38] and non-fulfillment of the Beer–Lambert–Bouguer law[35] lead to intrinsic nonlinearity of the spectrum–property relationship in these systems. Examples of such systems include crude oil, black oil, ethanol–gasoline fuel mixtures, and solutions of petroleum macromolecules.[4–7,13,29,34] Even relatively weak van der Waals intermolecular forces[35,39–41] in chemical systems like gasoline, biodiesel, paraffin wax, or aromatic hydrocarbons can influence the accuracy of the PLS model. Note that nonlinear relations can only be modeled by PLS in a limited way by considering more latent variables.[26,27,42] Exactly the same can be said about the principal component regression (PCR) technique.[4,26,27]

It should be stated separately that the degree of nonlinearity can be rather different for different properties of the same chemical system. However, one can sometimes make general conclusions about object nonlinearity, or system nonlinearity,[4] based on a number of system properties or rather general characteristic behavior.

One should note that a number of nonlinear PLS-based approaches exist, such as Poly-PLS[43,44] and Spline-PLS.[45] The only difference between these two algorithms and (linear) PLS is one step in which the linear function is changed into a polynomial function (for Poly-PLS) or a piecewise polynomial function called a spline function (for Spline-PLS). These two techniques are referred to as "quasi-nonlinear" calibration methods.[4,43–45]

Although partial least squares regression has been a cornerstone of MDA of chemical data for many years, it is neither perfect nor complete.[4,6,19,20,26–28,33] Since the assumption about the linearity of the input–output dependence is a rough approximation for most chemical systems, usually only valid within a small interval of input/output values, alternative regression tools are needed.[4,6,43–45]

Modern applied mathematics offers a wide variety of nonlinear methods, and artificial neural networks, or ANNs, are among the most effective and popular methods.[46] Based on Kolmogorov's theorem,[46–48] one can claim that the standard multilayer feed-forward neural network with a single hidden layer that contains a finite number of neurons (see Fig. 1 in ref. 28) can be regarded as a universal approximator; that is, ANN can approximate any linear or nonlinear dependence between the input and output values with an appropriate choice of free parameters or weights.[28,46] This background makes ANN one of the most pervasive nonlinear data analysis techniques in almost all fields of chemistry, from quantitative structure–property relationship studies (QSPR/QSAR)[49] to quantum chemistry (QC)[28,50,51] to petroleum studies.[4–7]

The disadvantages of the ANN approach to spectroscopic data analysis are[1–9,46] as follows:

(i) the stochastic nature of the ANN training (model building) process;

(ii) the dependence of the final result on the initial parameters;

(iii) the need to repeat network training many (hundreds of) times;

(iv) the non-uniqueness of the final solution, or ANN weights, that produces the best result, given that many networks with completely different sets of free parameters can produce very similar results;

(v) the available sample set should be relatively large for effective ANN training;

(vi) the tendency to overfitting; and

(vii) the training time and computational resources: ANN training can take many hours, and even days, of CPU time even with modern computers (as of mid-2010).

Note that techniques such as clamping and analysis of weights can provide detailed insights into how an ANN functions.

Does any alternative to these ANN-based methods exist? Support vector machines (SVMs) might be regarded as the perfect candidate for spectral regression purposes.[1–3,52] SVM-based techniques are very interesting methods, simple in their theoretical background and very powerful in model and real-world applications. A large advantage of SVM-based techniques is their ability to model nonlinear relationships.[24,52–54] Compared to neural networks, SVM has the advantage of leading to a global model that is capable of efficiently dealing with high dimensional input vectors.[1–3] SVMs have the additional advantage of being able to handle ill-posed problems and lead to global models that are often unique.[3] Furthermore, due to their specific formulation, sparse solutions can be found in many cases. However, finding the final SVM model can be very difficult computationally because it requires quadratic programming and the solution to a set of nonlinear equations.[3]

First used as a classification methodology,[55] SVM has been extended to regression tasks *via* two approaches: support vector regression (SVR)[1] and least-squares support vector machines (LS-SVMs).[3] Both will be discussed in our current study. See Section 3 for the basic theoretical concepts of the both methods. It should be noted that support vector machines, unlike PLS and ANN regression methods, are still relatively unknown to scientists in the field of chemometrics.[1–3]

A number of studies dealing with SVM-based approaches for solving chemically or industrially important problems have been

published in recent years.[1–3,8,9,56–63] Unfortunately, none of them are sufficiently general; only a few sets of spectra (at most) are usually used in each case. So, it is currently difficult to draw any definite conclusions about the efficiency of SVR or LS-SVM and the potential for the application of these approaches in spectroscopic data analyses. Different studies report different accuracies for SVM- and ANN-based approaches that cannot be compared because of differences in experimental or computational methodologies.[1–3,8,9,56–63] The role of SVM-based regression in the area of chemometrics and multivariate data analysis is still unclear.

In the current study, we try to make a rather general comparison of SVM-based regression models, SVR and LS-SVM, with linear (PLS), "quasi-nonlinear" (Poly-PLS), and nonlinear (ANN) regression methods. Due to our previous experiences[4–7,53,54] and the great importance of this particular field, petroleum systems were chosen as a representative example of real-world samples. Five very different chemical systems were studied, differing in complexity, composition, structure, and properties; these systems are gasoline, ethanol–gasoline biofuel, diesel fuel, aromatic solutions of petroleum macromolecules, and petroleum resins in benzene. Fourteen different sample sets ("NIR spectrum—sample property", see below) were used in total. We try to rule out factors that influence SVR/LS-SVM behavior (relative to PLS, Poly-PLS, and ANN) when dealing with spectroscopic data. General conclusions are made about the applicability of SVM-based regression tools in the modern analytical chemistry of petroleum and its products.

2. Experimental

2.1. Sample sets

Fourteen different sample sets were used in this study (Table 1). These sets include seven sets of gasoline data (density, benzene content, and fractional composition/boiling points),[4,6] two sets of ethanol–gasoline fuel data (density and ethanol content),[6] one set of diesel fuel data (total sulfur content), three sets of petroleum macromolecules data (weight percentage of asphaltenes, resins, and paraffins in toluene),[5] and one set of petroleum resins data (resins content in benzene).[7] In all cases, NIR spectra (Table 1) were used to build calibration models to predict the desired system property. See ref. 4–7 for a detailed description of the datasets used. Table 1 summarizes the main parameters of interest for all 14 datasets. See ref. 4–7 for a discussion of the reference data collection for each particular case.

2.2. NIR apparatus and experimental parameters

All NIR spectra (except those for diesel) were acquired with an NIR FT Spectrometer InfraLUM FT-10 (LUMEX, Russia) fitted with a special sampler for liquids. See Table 2 in the previous publication by Balabin *et al.*[4] for detailed spectrometer parameters. The spectra were acquired at room temperature (20–23 °C). Background spectra were recorded before and after each measurement to compensate for the absence of thermostating. The averaged background spectrum was subtracted from the sample spectrum before all pre-processing procedures. This resulted in an analytical signal with satisfactory accuracy and precision. The instrument calibration for wavelength and transmittance was performed using four pure hydrocarbons: toluene (C_7H_8), *n*-hexane (C_6H_{14}), benzene (C_6H_6), and isooctane (iso-C_8H_{18}). This calibration was repeated at least once per day to ensure stability of the experimental setup and data accuracy and reproducibility.

NIR spectra of diesel fuel were collected using a MPA Multi Purpose FT-NIR Analyzer (Bruker) at room temperature. The MPA NIR spectrometer was calibrated with benzene and cyclohexane (*c*-C_6H_{12}) at least twice per day to minimize the influence of variable laboratory conditions. The spectral range between 11 000 and 4000 cm^{-1} (909–2500 nm) was scanned with an 8 cm^{-1} resolution. Sixty-four scans were averaged for each spectrum. A background spectrum was measured every 45 min. A cylindrical glass cell with an 8 mm optical path was used throughout this study. Approximately 1 mL of diesel sample was required for each NIR measurement, much less than the 200 mL needed for distillation analysis to determine the fractional composition.[64] The NIR spectrum collection was repeated five

Table 1 General description of all fourteen (14) NIR datasets: systems, properties, and spectral ranges

Petroleum system	Property	Unit	Number of samples	Property range		Reference method accuracy[e]	Spectral range[f]/cm^{-1}	
				Min.	Max.		Max.	Min.
Gasoline[a]	Density at 20 °C	kg m^{-3}	95	640	800	0.5	14 000	8000
	Initial boiling point (IB)	°C	95	35	59	1–5	14 000	8000
	End boiling point 10% v/v (T10)	°C	95	58	117	1–5	14 000	8000
	End boiling point 50% v/v (T50)	°C	95	93	128	1–5	14 000	8000
	End boiling point 90% v/v (T90)	°C	95	121	175	1–5	14 000	8000
	Final boiling point (FB)	°C	95	178	205	1–5	14 000	8000
	Benzene content[b]	% w/w	57	0	10	0.10–0.25	13 500	8500
Biofuel: ethanol–gasoline[b]	Density at 20 °C	kg m^{-3}	117	672	785	0.5	13 500	8500
	Ethanol content[b]	% w/w	75	0	15	0.05[b]	13 500	8500
Diesel fuel	Total sulfur content	ppm	125	303	5100	2–20	11 000	4000
Petroleum macromolecules[c]	Asphaltene content	% w/w	120 (80)	0	10	0.01[c]	14 000	8000
	Resin content	% w/w	120 (80)	0	30	0.01[c]	14 000	8000
	Paraffin content	% w/w	120 (80)	0	10	0.01[c]	14 000	8000
Petroleum resins in benzene[d]	Resin content	mg L^{-1}	105 (54)	0	6000	1.1[d]	13 000	9000

[a] Ref. 4. [b] Ref. 6. [c] Ref. 5. [d] Ref. 7. [e] Ref. 74. [f] The range of [14 000; 8000] cm^{-1} refers to [714; 1250] nm.

Table 2 The results of linear (PLS) and quasi-nonlinear (Poly-PLS) methods application to near infrared spectroscopy and reference data of petroleum systems: partial least squares (PLS) and polynomial partial least squares (Poly-PLS) regression models

			PLS		Poly-PLS		
Petroleum system	Property	Unit	LV[g]	RMSEP	LV[g]	n[e,g]	RMSEP
Gasoline[a]	Density at 20 °C	kg m^{-3}	10	2.8	9	3	2.4
	Initial boiling point (IB)	°C	10	2.0	15	5	1.6
	End boiling point 10% v/v (T10)	°C	9	2.2	9	4	1.8
	End boiling point 50% v/v (T50)	°C	12	2.4	14	3	1.9
	End boiling point 90% v/v (T90)	°C	18	2.8	14	5	2.2
	Final boiling point (FB)	°C	19	2.8	18	3	2.1
	Benzene content[b]	% w/w	5	0.87	5	2	0.85
Biofuel: ethanol–gasoline[b]	Density at 20 °C	kg m^{-3}	11	2.70	9	3	2.40
	Ethanol content[b]	% w/w	5	0.22	3	2	0.22
Diesel fuel	Total sulfur content	ppm	6	344	6	3	341
Petroleum macromolecules[c]	Asphaltene content	% w/w	5	0.41 (0.43)[f]	5	2	0.25
	Resin content	% w/w	3	0.79 (0.79)[f]	5	2	0.71
	Paraffin content	% w/w	6	0.35 (0.39)[f]	6	2	0.35
Petroleum resins in benzene[d]	Resin content	mg L^{-1}	3	2.1 (2.1)[f]	2	2	2.1

[a] Ref. 4. [b] Ref. 6. [c] Ref. 5. [d] Ref. 7. [e] Also known as 'D' in Ref. 4. [f] The second number (in parentheses) refers to smaller sample set, see Table 1. [g] The optimal values were determined by the RMSECV minimization.

times with cell rotation inside the spectrometer between repetitions to minimize the interference from the cell or glass defects. Measurement of one sample took less than five minutes. The averaged and background-corrected spectra were used for subsequent data pre-processing.

See ref. 4–7 for experimental spectra examples and their discussion.

2.3. Model efficiency estimation

To characterize the prediction ability and efficiency of the created regression model, the root mean squared error of prediction (RMSEP) was calculated for each case. Validation set was constructed as one fifth of all samples from every sample set (19 out of 95 gasoline samples; 24 out of 120 diesel fuels; *etc.*). It was checked that the validation set consisted of samples from the entire property range.

The mean average percentage error (MAPE) was also calculated to estimate the relative accuracy of each calibration model. This is especially important for properties with a large range, such as sulfur in diesel fuel. See ref. 4–6 for the exact formulas and extra discussion.

Five-fold or ten-fold cross-validation was used to optimize the model's parameters based on the root mean squared error of cross-validation (RMSECV). It was checked that the cross-validation set consisted of samples from the entire property range. Other variants of the cross-validation procedure, *e.g.*, 7-fold version, leave-one-out cross-validation (LOOCV), were checked and found to produce almost identical results.

In all cases a negligible difference between RMSECV and RMSEP of PLS, Poly-PLS, and ANN methods was found as discussed in ref. 4–7. The use of either of them does not change the conclusions drawn here. This conclusion is not general—there are many cases, even among petroleum systems, where the RMSECV and RMSEP results can be quite different. For SVM-based methods prediction error was calculated.

Note that one needs to use the same dataset division for unbiased comparison with previously published results.[4–7]

There, of course, are some reservations about using cross-validation methods for optimizing regression models based on support vector machines. It is arguable that SV-type models cannot be compared directly as PLS-type models. There are a number of reasons for this. First of all, some samples (not SVs) do not contribute to the models, so removing them will make no difference for the final prediction of, *e.g.*, SVR. This is a complicated issue: removing too many samples may mean that there are different SVs, but removing a single non-SV sample usually means no change in the final model. Second, some parameters such as the error penalty term (C or γ) have a "quantized" effect on the model, that is a range of C values will result in an identical model. Neither of these issues are problems encountered when optimizing the PLS model.

2.4. NIR spectra pre-processing and outlier detection

Different types of spectra pre-processing (pre-treatment) methods were used, including normalization, magnitude normalization, multiplicative scatter correction (MSC), linearization, differentiation, double differentiation, autoscaling, and range scaling in different intervals. The best pre-processing technique was found for each calibration method and each petroleum system property. See ref. 4–7 for a detailed discussion of each particular system.

See ref. 4–7 for a detailed discussion of the outlier detection scheme for each particular petroleum system. In general, all results are reported for outlier-free sample sets. Note that for traditional statistical methods (such as PLS), it is sometimes indeed important to perform outlier detection prior to modeling, as outliers can have a huge influence on least squares approaches. However, for SVR this is not always necessary, because its behavior with respect to outliers can be controlled by the error penalty term. So, SVR can actually handle datasets with extreme outliers whereas some other approaches will fall down. Here we do not discuss the robustness of the techniques with respect to outliers; that is why the errors are reported for outlier-free sample sets.

2.5. Spectra reduction and feature selection

In order to create an effective and robust regression model, the spectral data, which have up to 10^4 independent variables, should be reduced.[65] Two common data reduction techniques, spectra averaging and principal component analysis (PCA), were used to achieve this goal for LS-SVM, SVR, and ANN methods.[1–9,26,27,52,65] Note that PLS-based techniques have an intrinsic data reduction ability (the latent variables). The PCA results are reported because this technique was found to produce the best results and the lowest errors in all cases. Other methods of feature selection (wavelets, UVE-PLS, *etc.*[66]) are out of the scope of current study. Of course, the optimal feature selection methodology leads to an increase in the prediction ability and a decrease in the error of any model discussed.[66]

2.6. Methods optimization

To compare the different classification models, the best results from each model need to be obtained; otherwise, the comparison is useless. The results from each model depend on the model parameters. We have used a wide range of model parameters to achieve the best results. RMSECV minimization was used for optimization in all cases and for all models.

These parameters and the corresponding model are as follows:

PLS: number of latent variables (LV);

Poly-PLS: LV and degree of polynomial (n);

ANN/MLP: number of input neurons (IN; equal to number of principal components, PC), number of hidden neurons (HN), and transfer function of hidden layer: $f(x) = \{logsig\}$; $\{tansig/tanh\}$. Detailed procedures for ANN training can be found in ref. 4. See for example, Table 4 in ref. 4 for the ANN training procedure for gasoline data.

SVR: the error weight (C), maximal error value (ε), and kernel-related parameters. The same set of kernels (linear, polynomial, and radial basis function (RBF)) was used for SVR and LS-SVM model building. See Table 4 in ref. 24 for a detailed list of parameters. See Section 3 for the parameter definitions and other clarifications.

LS-SVM: the regularization parameter (γ), determining the trade-off between the fitting error minimization and the smoothness of the estimated function, and the kernel-related parameters (*e.g.*, σ or σ^2 for the RBF kernel, Table 2). See Section 3 for the parameter definitions and other clarifications.

RBF kernels (default) were found to produce the lowest prediction errors in all cases studied. But the SVM-based methods were found not to be very sensitive to kernel choice; in many cases, polynomial kernels were able to produce very close results to RBF ones (compare with ref. 1–3).

Note that Spline-PLS, being a very time consuming technique, has not shown any considerable superiority over the Poly-PLS method for petroleum system analysis.[4] This is why it was not used in the current study.

So, the regression methods were optimized based on cross-validation procedure and tested using fully independent test (validation) sets (see also above).

2.7. Software and computing

MATLAB 2008b was used as the standard software for multivariate methods realization. The following toolboxes were used: MATLAB Statistics Toolbox, MATLAB Support Vector Machine Toolbox, MATLAB Neural Network Toolbox, N-way Toolbox for MATLAB, and PLS_Toolbox Version 4.0. For the SVR calculations, a MATLAB toolbox developed and described by Gunn was used.[67] The LS-SVM regression model was built using the LS-SVMlab1.5 MATLAB toolbox.[68] Ref. 67 and 68 contain a detailed description of the algorithms and procedures. The standard programs of these toolboxes were modified and extended by BRM (see also ref. 4 and 6).

2.8. Sample sets: their quality and representativeness

The current study deals with five chemical systems of petroleum origin. They are: gasoline, a classical sample for analytical chemistry in general and chemometrics in particular;[4,6] ethanol–gasoline biofuel, an increasingly popular type of motor fuel, partly produced from renewable sources that may have a colloid (dispersed) structure;[69,70] diesel fuel, a product of petroleum refining with a higher boiling range than gasoline due to a more complicated mixture of hydrocarbons and heteroatomic compounds;[13] a solution of all three classes of petroleum macromolecules (asphaltenes, the molecules responsible for the colloid structure formation in crude oil;[71] resins; and paraffins) in an aromatic solvent (toluene; each macromolecule class is an extremely complicated mixture); and a petroleum resins solution in benzene, a sample set used to calibrate NIR setup for adsorption studies.[7] Details about some of these systems have been published during the last 4 years by Balabin and co-workers.[4–7,53,54]

The four systems presented here greatly range in composition, properties, and behavior. While low molecular weight substances having 6–12 carbon atoms with low intermolecular forces (*n*-hexane, heptane isomers, isooctane, *etc.*) form gasoline,[13] heavy (above 500 Da) molecules with high tendency to aggregation and phase separation, like resins and asphaltenes, are found in the last two systems.[71] The number of effective components ranges from one in petroleum resins to millions. Therefore, rather general conclusions about algorithm behavior can be made based on the system studied.

The fourteen properties of the four petroleum systems described above form fourteen sample sets that are very different in nature (Table 1). For gasoline, these are the density at 20 °C, fractional composition (including initial boiling point (IB), end boiling points 10%, 50%, and 90% v/v (T10, T50, and T90, respectively), and final boiling point (FB)) and finally benzene content. For ethanol–gasoline fuel, these sample sets are based on density at 20 °C and ethanol content—[EtOH]. For diesel fuel, the sample set is based on the total sulfur content ([Sulfur]). For petroleum macromolecules, the sets are asphaltene content ([A]), resins content ([R]), and paraffins content ([P]). Finally, for petroleum resins the relevant sample set is the resin concentration in benzene ([R]).

Note that the quality (accuracy, repeatability, and reproducibility) of reference data ranges greatly from one property to another (Table 1). It is important to estimate the effect of initial data quality on final prediction results. The same can be said about property ranges; some are rather limited (*e.g.*, T50), some are very broad (*e.g.*, [Sulfur] or [R]). In industrial applications it is usually impossible to model the quality (in either accuracy or

range) of datasets. Therefore, the machine learning algorithms that show very good, even brilliant, results on model systems do not always show the same results when applied to real-world problems.[46,52] In this work we have tried to use wide ranges of reference data quality to help make our conclusions as general as possible.

The spectroscopic information for most sample sets (Table 1) was recorded in the short-wave part of the NIR region (above 8000 cm^{-1}). This is the region with the second to fifth overtones of characteristic molecular vibrations observed by standard IR and Raman techniques.[14,26,51] The only exclusion is the diesel fuel sample set, whose spectrum lies in the 4000–11 000 cm^{-1} region. In this particular case, it was important to get information from the long-wave part of the NIR spectrum due to the necessity of predicting the sulfur concentration in diesel samples.[14]

The number of samples in the sample sets ranged from 57 to 125 (Table 1). Since the number of samples can influence the quality of the multivariate model prediction, we tried to ensure that sample set saturation was observed at least in the case of the simplest (PLS) method, similar to the basis set limit (BSL) or complete basis set (CBS) methods in quantum chemistry.[50,72,73] Table 2 shows some representative examples of varying the number of training examples.

3. A short description of SVM regression methods: SVR *vs.* LS-SVM

Support vector machines were initially been developed by Vapnik[52,55] as a binary classification tool. SVMs are based on some "beautifully simple ideas"[56] and provide a clear intuition of what learning from examples is all about. Intuitively, an SVM model is a representation of the training sample set as vectors in space mapped so that the samples from the separate categories are divided by a clear gap that is as wide as possible. New samples from cross-validation or a test set are then mapped into that same space. Based on which side of the gap between classes they fall, they are predicted to belong to one category or another. SVMs show high performance in practical applications when solving sophisticated classification problems.[24,55,56]

The principles of SVM can easily be extended to regression tasks. For detailed in-depth theoretical background on SVMs for both classification and regression, the reader is referred to the ref. 1–3, 52 and 55. No equations will be used in the following text; see ref. 1–3 for all necessary equations and formalism.

Similar to the approach of ordinary least squares (OLS) and PLS, SVR also finds a linear relation between the regressors (input variables, X) and the dependent variables (y).[1] The cost function (the function that is minimized to obtain the best regression model) consists of a two-norm penalty on the regression coefficients, an error term multiplied by the error weight, C, and a set of constraints. Using this cost function, the goal is to simultaneously minimize both the coefficients' size and the prediction errors (function smoothness and accuracy). The first point is important because large coefficients might hamper generalization due to their tendency to cause excessive variance.[1]

In SVR, the prediction errors are penalized *linearly* with the exception of a deviation of below a certain value, ε, according to Vapnik's ε-insensitive loss function. Only predictions deviating more than ε ($|y - y_{pred}| > \varepsilon$, where y_{pred} is the SVR model prediction) are taken into account. The objects with prediction errors larger than ε are called "support vectors" and only these vectors determine the final prediction of the SVR model. Due to the fact that only the inner product is used in all calculations, it is possible to use kernel functions, or kernels, that enable nonlinear regression in a very efficient way. The values of ε and the parameter C have to be defined by the user; both are problem- and data-dependent.[1,55]

The ideology of the LS-SVM method is very close to that of SVR, but in this case the more usual sum of the *squares* of the errors is minimized, and no ε-based selection is made between samples. This is a general feature of least-squares (LS) methods.[3] This can make the final model more accurate and less computationally expensive; see ref. 3 for extra details. Parameter γ, the analog of parameter C in the SVR model, controls the smoothness of the fit.

So, if one forgets about kernel-specific parameters, the error weight (C) plus maximal error value (ε) and regularization parameter (γ) were optimized for SVR and LS-SVM methods, respectively.

As described above, SVM-based regression techniques solve many of the intrinsic ANN problems, such as its stochastic nature, the necessity to repeat network training many times, and the non-uniqueness of the final ANN solution. This makes SVR and LS-SVM interesting and promising alternatives to ANN. Note that the most important advantage, namely the possibility of building a nonlinear model, is still valid in the SVM regression case. Here we will try to understand the extent to which SVM-based techniques can substitute ANN-based approached techniques in real-world (industrial) applications. Are SVR and LS-SVM models accurate enough to really be regarded as alternatives to neural networks?

4. Results and discussion

4.1. PLS-based techniques: linear PLS and Poly-PLS

Table 2 shows the results of application of PLS-based techniques to NIR spectra of different petroleum systems.[4] Comparison of the property range and the reference method accuracy shows that very different results were obtained. In some cases, such as density and fractional composition of gasoline and ethanol–gasoline fuel, or resins in benzene, rather good accuracy of the PLS/Poly-PLS prediction was achieved. In other cases, such as petroleum macromolecules and total sulfur content in diesel fuel, only mediocre results were observed. It seems to be that the model quality is greatly dependent on the structure of the object under study. Compare: in the resins content in two different systems (Table 2), one is much more complicated than other (Table 1).

The structure of PLS-based models, namely the number of latent variables and the degree on polynomial, is inline with previous results for petroleum systems. The general trend is that the more complicated the quality (that is, the greater the nonlinearity), the greater the number of latent variables needed to extract all necessary information and to take into account the deviation from linear spectrum–property dependence (Table 2).

Note that in all cases, the Poly-PLS approach shows a RMSEP that is not worse than that of the linear PLS analog.[4,6,45] In other

words, for all petroleum systems under study, some kind of nonlinearity was observed and modeled with differing success by the Poly-PLS model.[4] The only property for which the Poly-PLS approach was really effective was the asphaltene content, in which the RMSEP was decreased by almost 40%. Almost no effect was observed for benzene, ethanol, and sulfur contents, where the RMSEP was decreased by only 2 ± 1%. Therefore, Poly-PLS approach is not the best model for increasing the accuracy of the calibration model, even though some effect (~10%) can be observed in a number of cases.[43,44]

4.2. ANN approach

The results of the ANN approach to the petroleum NIR data are summarized in Table 3. The accuracy of the ANN method is always much better than other methods, with the exception of resins concentration determination in benzene solution.

An average prediction error decrease relative to PLS of 41 ± 15% (±σ) was observed. The largest error decrease was observed for the asphaltene concentration (−63%), with resins and paraffin contents also showing large, and almost identical, decreases. This fact can be explained by the extremely high tendency of petroleum macromolecules to form dimers, oligomers, clusters, and aggregates.[12,29,71] Even phase separation, or asphaltene onset, can easily be observed in many petroleum systems. This is the process that is responsible for many troubles in the petroleum industry, from crude oil production to refining and transportation.[7,12,13,29] Since all of the described processes are concentration-dependent, a high degree of nonlinearity in spectrum–concentration dependence is expected. This leads to the need of nonlinear treatment of systems containing petroleum macromolecules (especially asphaltenes). ANN is the technique of choice in this case.

The absence of such a pronounced effect of ANN application for pure resins solution in benzene (−10% only) can be explained as follows. First, the system is simple and ANN is just not needed. Second, the PLS approach is itself highly accurate, close to the accuracy of the reference method (Table 1), and neither ANN nor other multivariate method can do better than the reference data allow (see below).[26,27]

In general, one can state that the ANN approach is extremely efficient for analysis of NIR spectra of petroleum systems, regardless of boiling range or composition. Very different properties and quality coefficients of industrially important products can be accurately predicted by neural networks.[4,46]

4.3. SVM-based techniques: SVR and LS-SVM

Table 4 summarizes the results of SVM-based approaches to petroleum data in all 14 datasets. The results of SVR and LS-SVM methods are presented for comparison.

One can see that, in general, both SVR and LS-SVM models show results not worse than those of ANN models. In cases of [Sulfur] prediction and petroleum macromolecules analysis, the SVM-based regression models have lower prediction error (−15 ± 1%) than ANN models. Good results are also shown by the SVR model for benzene concentration prediction (−9%). For T90, [EtOH] and petroleum resins in benzene, SVM regression models have higher RMSEP than ANN models (by 7%, 11%, and 7%, respectively). Note that in the last case all the methods show approximately the same results (±8%), so these data are not that representative (Table 4). The cause for this could be the relative system simplicity. In the five other cases, the results of the SVM approach are very close to those of neural networks (±2%).

The difference between SVR and LS-SVM results is small: −3 ± 7% with an advantage of LS-SVM regression model. A relatively significant difference (>10%) is observed for [Benzene], [A], [R] in toluene, and [R] in benzene. In the last three cases, the RMSEP of LS-SVM model is lower.

Based on data from Table 4, one can claim that both SVM-based methods are very effective for building calibration models (compare with Table 2). Both methods are recommended for analysis of petroleum products and biofuels (compare with Fig. 2 in ref. 3). Mostly due to computational aspects, the LS-SVM

Table 3 The results of artificial neural networks (ANNs) application to near infrared spectroscopy and reference data of petroleum systems: multi-layer perceptron—MLP or ANN-MLP

			ANN (MLP)[e]		
Petroleum system	Property	Unit	IN (PC)[f,g]	HN[g]	RMSEP
Gasoline[a]	Density at 20 °C	kg m^{-3}	10	7	2.0
	Initial boiling point (IB)	°C	16	8	1.3
	End boiling point 10% v/v (T10)	°C	19	6	1.4
	End boiling point 50% v/v (T50)	°C	15	9	1.6
	End boiling point 90% v/v (T90)	°C	14	9	1.7
	Final boiling point (FB)	°C	18	7	1.7
	Benzene content[b]	% w/w	12	5	0.58
Biofuel: ethanol–gasoline[b]	Density at 20 °C	kg m^{-3}	9	7	1.90
	Ethanol content[b]	% w/w	8	5	0.13
Diesel fuel	Total sulfur content	ppm	7	5	155
Petroleum macromolecules[c]	Asphaltene content	% w/w	5	3	0.15
	Resin content	% w/w	5	4	0.30
	Paraffin content	% w/w	5	3	0.13
Petroleum resins in benzene[d]	Resin content	mg L^{-1}	3	2	1.9

[a] Ref. 4. [b] Ref. 6. [c] Ref. 5. [d] Ref. 7. [e] ANN architecture is the following: IN − NH − 1; so, in the case of diesel fuel it will be "7 − 5 − 1". [f] The (optimal) number of input neurons (IN) is equal to the (optimal) number of principal components (PC) used for principal component analysis (PCA) of near infrared spectra.[4] Compare with LV in Table 2. [g] The optimal values were determined by the RMSECV minimization.

Table 4 The results of support vector machine regression (SVR and LS-SVM) application to near infrared spectroscopy and reference data of petroleum systems: support vector regression (SVR) and least-squares support vector machines (LS-SVMs)

			SVR		LS-SVM	
Petroleum system	Property	Unit	PC[e]	RMSEP	PC[e]	RMSEP
Gasoline[a]	Density at 20 °C	kg m^{-3}	5	2.0	6	2.0
	Initial boiling point (IB)	°C	8	1.4	8	1.3
	End boiling point 10% v/v (T10)	°C	8	1.4	8	1.4
	End boiling point 50% v/v (T50)	°C	7	1.5	7	1.6
	End boiling point 90% v/v (T90)	°C	10	1.8	9	1.8
	Final boiling point (FB)	°C	10	1.8	10	1.7
	Benzene content[b]	% w/w	5	0.53	6	0.58
Biofuel: ethanol–gasoline[b]	Density at 20 °C	kg m^{-3}	7	1.91	6	1.92
	Ethanol content[b]	% w/w	5	0.14	6	0.16
Diesel fuel	Total sulfur content	ppm	7	136	7	131
Petroleum macromolecules[c]	Asphaltene content	% w/w	4	0.15	4	0.13
	Resin content	% w/w	6	0.29	4	0.26
	Paraffin content	% w/w	5	0.12	5	0.12
Petroleum resins in benzene[d]	Resin content	mg L^{-1}	3	2.3	3	2.0

[a] Ref. 4. [b] Ref. 6. [c] Ref. 5. [d] Ref. 7. [e] The optimal values were determined by the RMSECV minimization.

regression model is preferred. This conclusion supports the early analysis of Buydens and co-workers[3] based on NIR spectra that were affected by temperature-induced spectral variation. Additional support for LS-SVM usage is the evidence that this model leads to robust models for spectral variations due to nonlinear interferences.[3]

4.4. General remarks. Trends and peculiarities

Fig. 1 shows the scatter plot of two relative error differences: $(RMSEP_2 - RMSEP_1)/RMSEP_1 \times 100\%$, where {1} and {2} refer to {PLS} and {ANN} differences and {ANN} and {LS-SVM} differences, respectively. Fig. 1 clearly shows that a correlation between model behaviors exists because the points form two distinct classes. The first, larger class with 10 points is characterized by a relative error decrease due to neural networks usage of 10–40%; in this case the use of LS-SVM regression does not lead to any significant error decrease compared to the ANN model. The second, smaller class of 4 points has an x-value below −50% and a y-value below −10%. So, if the PLS error is greatly decreased by ANNs by more than half, one can expect that the SVM-based regression model will be more effective than the ANN approach. Since the difference between the PLS and ANN models can be interpreted as a measure of object or property nonlinearity, the SVM-based approach is preferable for highly nonlinear objects.

This observation can be explained by the fact that the ANN method tends to overfit highly nonlinear objects. This behavior can significantly lower the generalization ability of the network. The same is not observed for the LS-SVM calibration model.

Note that the point with the smallest absolute x-value on Fig. 1 is the resins in benzene sample (see also the Discussion above).

The maximum accuracy achieved by each technique is the main, but not the only characteristic of model applicability to real-world (industrial) tasks. For example, one of many benefits of the SVM approach is its deterministic nature. It leads to the fact that the range of prediction errors for different training/test subsets separation for the SVM-based techniques is much smaller than for top-20 ANNs: [130−133] *vs.* [147−281] ppm for diesel fuel analysis, [0.15–0.17] and [0.13–0.30] % w/w for [EtOH] in biofuel, *etc.* for LS-SVM and ANN methods, respectively. In other words, one needs to repeat the ANN training many times to get a really accurate result.

5. Conclusions

The results of application of linear (PLS), quasi-nonlinear (Poly-PLS), and nonlinear (ANN, SVR, and LS-SVM) regression methods on NIR spectroscopy data are shown in Fig. 2. One can conclude the following:

(1) Fourteen different sample sets were studied by linear (PLS), quasi nonlinear (Poly-PLS), and three nonlinear (ANN, SVR, and LS-SVM) multivariate methods. NIR spectroscopy data were used in all cases.

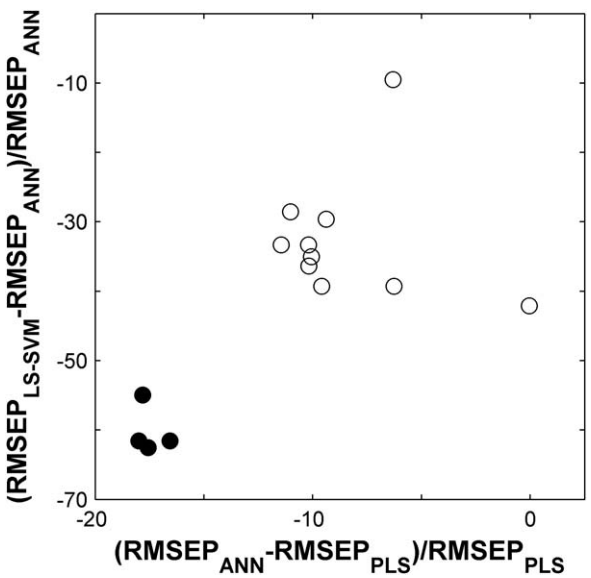

Fig. 1 Correlation between the decrease in relative error (%) using ANN and SVM (LS-SVM) regression methods: (x-axis) $100\% \times (RMSEP_{ANN} - RMSEP_{PLS})/RMSEP_{PLS}$; ($y$-axis) $100\% \times (RMSEP_{LS\text{-}SVM} - RMSEP_{ANN})/RMSEP_{ANN}$. Note the use of fourteen (14) different datasets.

(2) The accuracy of the SVM-based calibration models, SVR and LS-SVM, is comparable with the accuracy of the ANN-based approach.

(3) There is a correlation between the relative accuracies of the ANN- and SVM-based approaches.

(4) For highly nonlinear objects like petroleum macromolecules, SVM-based regression models are preferable to neural networks.

(5) Regression methodologies, based on the support vector machine ideology, are recommended for practical implementation. The regression models based on SVMs are sufficiently accurate and robust to be used for gasoline, biofuel, or diesel fuel analysis.

We hope that the role of SVM-based regression in chemometrics and multivariate data analysis is clearer after this study and that the possibilities of SVM-based approaches and obstacles to their application have become more evident to both analytical and industrial communities.

We believe that our results will help future chemometric investigations and investigations in the sphere of vibrational (IR, NIR, and Raman) spectroscopy of multicomponent systems.[1–3,56–63,75–82] The results presented herein can help achieve rapid and accurate analysis or classification of biofuels, products of petroleum refining, and petrochemicals. The use of NIR spectroscopy in other fields of analytical chemistry, such as pharmaceutical quality control, food quality control, and active pharmaceutical ingredient/pharmacon (pharmakon) analysis of tablets, can be enhanced by the application of modern methods of multivariate data analysis, including support vector machines and artificial neural networks as well as other machine learning techniques.

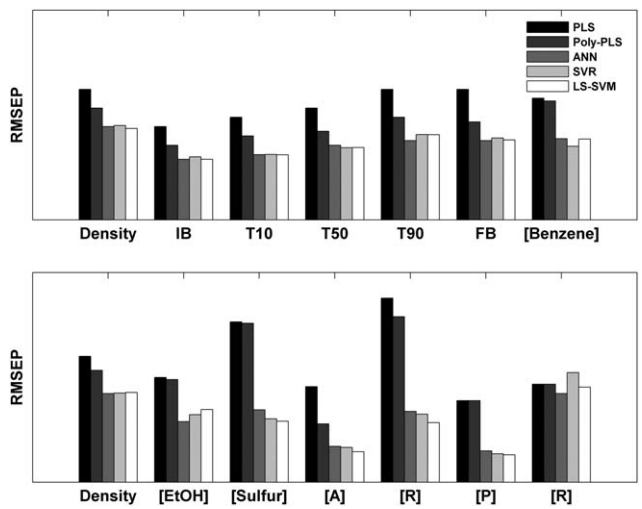

Fig. 2 Results of petroleum systems analysis by different multivariate techniques: LS-SVM *vs.* ANN and SVR *vs.* LS-SVM. Sample sets and properties: (top, from left to right) density—gasoline density at 20 °C, IB—initial boiling point, T10—end boiling point 10% v/v, T50–end boiling point 50% v/v, T90—end boiling point 90% v/v, FB—final boiling point, [Benzene]–benzene content in gasoline; (bottom, from left to right) density—ethanol–gasoline fuel density at 20 °C, [EtOH]—ethanol content, [Sulfur]—total sulfur content in diesel fuel, [A]—asphaltene content in petroleum macromolecule solution, [R]—resins content in petroleum macromolecule solution, [P]—paraffins content in petroleum macromolecule solution, [R]—petroleum resin concentration in benzene.[7] Calibration models: PLS—partial least squares regression (projection to latent structures), Poly-PLS—polynomial partial least squares regression, ANNs–artificial neural networks (multilayer perceptron), SVR—support vector regression, LS-SVM—least-squares support vector machine regression. The root mean squared errors of prediction (RMSEP) are presented. The errors are normalized for comparison among different systems.

References

1 U. Thissen, M. Pepers, B. Ustun, W. J. Melssen and L. M. C. Buydens, *Chemom. Intell. Lab. Syst.*, 2004, **73**, 169–179.
2 F. Chauchard, R. Cogdill, S. Roussel, J. M. Roger and V. Bellon-Maurel, *Chemom. Intell. Lab. Syst.*, 2004, **71**, 141–150.
3 U. Thissen, B. Ustun, W. J. Melssen and L. M. C. Buydens, *Anal. Chem.*, 2004, **76**, 3099–3105.
4 R. M. Balabin, R. Z. Safieva and E. I. Lomakina, *Chemom. Intell. Lab. Syst.*, 2007, **88**, 183–189.
5 R. M. Balabin and R. Z. Safieva, *J. Near Infrared Spectrosc.*, 2007, **15**, 343–346.
6 R. M. Balabin, R. Z. Safieva and E. I. Lomakina, *Chemom. Intell. Lab. Syst.*, 2008, **93**, 58–63.
7 R. M. Balabin and R. Z. Syunyaev, *J. Colloid Interface Sci.*, 2008, **318**, 167–171.
8 K. Brudzewski, A. Kesik, K. Kołodziejczyk, U. Zborowska and J. Ulaczyk, *Fuel*, 2006, **85**, 553–558.
9 R. P. Cogdill and P. Dardenne, *J. Near Infrared Spectrosc.*, 2004, **12**, 93–100.
10 B. Osborne and T. Fearn, *Near Infrared Spectroscopy in Food Analysis*, Wiley, New York, 1986.
11 E. W. Ciurczak and J. K. Drennen, *Pharmaceutical and Medicinal Applications of Near-infrared Spectroscopy*, CRC Press, 1st edn, 2002.
12 R. Z. Syunyaev, R. M. Balabin, I. S. Akhatov and J. O. Safieva, *Energy Fuels*, 2009, **23**, 1230–1238.
13 J. G. Speight, *The Chemistry and Technology of Petroleum*, CRC Press, 3rd edn, 1999.
14 J. M. Hollas, *Modern Spectroscopy*, WileyBlackwell, 4th edn, 2003.
15 J. Morosa, S. Garrigues and M. de la Guardi, *TrAC, Trends Anal. Chem.*, 2010, **29**, 578–591.
16 F. Sun, W. Ma, L. Xu, Y. Zhu, L. Liu, C. Peng, L. Wang, H. Kuang and C. Xu, *TrAC, Trends Anal. Chem.*, 2010, **29**, 1239.
17 A. J. Hobro and B. Lendl, *TrAC, Trends Anal. Chem.*, 2009, **28**, 1235–1242.
18 Y.-N. Shao, Y. He, Y.-D. Bao, *Spectroscopy and Spectral Analysis*, 2008, vol. 28, pp. 602–605.
19 M. Ala-Korpela, Y. Hiltunen and J. D. Bell, *NMR Biomed.*, 2005, **8**, 235–244.
20 I. Stanimirova, B. Üstün, T. Cajka, K. Riddelova, J. Hajslova, L. M. C. Buydens and B. Walczak, *Food Chem.*, 2010, **118**, 171–176.
21 D. Bullinger, H. Fröhlich, F. Klaus, H. Neubauer, A. Frickenschmidt, C. Henneges, A. Zell, S. Laufer, C. H. Gleiter, H. Liebich and B. Kammerer, *Anal. Chim. Acta*, 2008, **618**, 29–34.
22 Q. Xiong, Y. Zhang and M. Li, *Anal. Chim. Acta*, 2007, **593**, 199–206.
23 J. Zhang, M. Gao, J. Tang, P. Yang, Y. Liu and X. Zhang, *Anal. Chim. Acta*, 2006, **566**, 147–156.
24 R. M. Balabin, R. Z. Safieva and E. I. Lomakina, *Anal. Chim. Acta*, 2010, **671**, 27.
25 E. Pringsheim, E. Terpetschnig and O. S. Wolfbeis, *Anal. Chim. Acta*, 1997, **357**, 247–252.
26 T. Næs, T. Isaksson, T. Fearn and T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, UK, 2002.
27 B. F. J. Manly, *Multivariate Statistical Methods: A Primer*, Chapman and Hall/CRC, 3rd edn, 2004.
28 R. M. Balabin and E. I. Lomakina, *J. Chem. Phys.*, 2009, **131**, 074104.
29 R. Z. Syunyaev and R. M. Balabin, *J. Dispersion Sci. Technol.*, 2007, **28**, 419–427.
30 R. M. Balabin, *J. Dispersion Sci. Technol.*, 2008, **29**, 457–464.
31 V. N. Alves, R. Mosquetta, N. M. M. Coelho, J. N. Bianchin, K. C. Di Pietro Roux, E. Martendal and E. Carasek, *Talanta*, 2010, **80**, 1133–1138.

32 P. Geladi, *Spectrochim. Acta, Part B*, 2003, **58**, 767–782.
33 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
34 R. Z. Syunyaev and R. M. Balabin, *J. Dispersion Sci. Technol.*, 2008, **29**, 1505–1511.
35 I. G. Kaplan, *Intermolecular Interactions: Physical Picture, Computational Methods and Model Potentials*, Wiley, 1st edn, 2006.
36 N. S. Hush and J. R. Reimers, *Chem. Rev.*, 2000, **100**, 775–786.
37 C. M. Drain, A. Varotto and I. Radivojevic, *Chem. Rev.*, 2009, **109**, 1630–1658.
38 L. Brunsveld, B. J. B. Folmer, E. W. Meijer and R. P. Sijbesma, *Chem. Rev.*, 2001, **101**, 4071–4098.
39 V. A. Parsegian, *Van der Waals Forces: a Handbook for Biologists, Chemists, Engineers, and Physicists*, Cambridge University Press, 2005.
40 R. E. Johnson and R. H. Dettre, *J. Phys. Chem.*, 1964, **68**, 1744–1750.
41 S. Wang and L. Jiang, *Adv. Mater.*, 2007, **19**, 3423–3424.
42 F. H. de Kermadec, J. F. Durand, R. Sabatier, *Food Quality and Preference*, 1997, vol. 8, pp. 395–402.
43 I. E. Frank, *Chemom. Intell. Lab. Syst.*, 1990, **8**, 109–119.
44 S. Wold, N. Kettaneh-Wold and B. Skagerberg, *Chemom. Intell. Lab. Syst.*, 1989, **7**, 53–65.
45 S. Wold, *Chemom. Intell. Lab. Syst.*, 1992, **14**, 71–84.
46 S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 2nd edn, 1998.
47 V. Kůrkova, *Neural Networks 5*, 1992, pp. 501–506.
48 G. Andrejkova and M. Mikulova, *Neural Netw. World*, 1998, **8**, 501–510.
49 S. S. So and M. Karplus, *J. Med. Chem.*, 1996, **39**, 1521–1530.
50 R. M. Balabin, *J. Chem. Phys.*, 2009, **131**, 154307.
51 J. J. Voegel, U. von Krosigk and S. A. Benner, *J. Org. Chem.*, 1993, **58**, 7542–7547.
52 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.
53 R. M. Balabin and R. Z. Safieva, *Fuel*, 2008, **87**, 1096–1101.
54 R. M. Balabin and R. Z. Safieva, *Fuel*, 2008, **87**, 2745–2752.
55 V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
56 S. R. Amendolia, G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala and G. M. Mura, *Chemom. Intell. Lab. Syst.*, 2003, **69**, 13–20.
57 A. Borin, M. F. Ferrão, C. Mello, D. A. Maretto and R. J. Poppi, *Anal. Chim. Acta*, 2006, **579**, 25–32.
58 F. Liua, Y. He and L. Wang, *Anal. Chim. Acta*, 2008, **610**, 196–204.
59 F. Liua, Y. He and L. Wang, *Anal. Chim. Acta*, 2008, **615**, 10–17.
60 D. Wua, Y. He and S. Feng, *Anal. Chim. Acta*, 2008, **610**, 232–242.
61 F. Liu, F. Zhang, Z. Jin, Y. He, H. Fang, Q. Ye and W. Zhou, *Anal. Chim. Acta*, 2008, **629**, 56–65.
62 M. F. Ferrão, S. C. Godoy, A. E. Gerbase, C. Mello, J. C. Furtado, C. L. Petzhold and R. J. Poppi, *Anal. Chim. Acta*, 2007, **595**, 114–119.
63 F. Liu, Y. Jiang and Y. He, *Anal. Chim. Acta*, 2009, **635**, 45–52.
64 ASTM D86-09e1, *Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure*, 2009, DOI: 10.1520/D0086-09E01.
65 X. Michalet, et al., *Science*, 2005, **307**, 538.
66 D. Wu, Y. He, P. Nie, F. Cao and Y. Bao, *Anal. Chim. Acta*, 2010, **659**, 229–237.
67 S. R. Gunn, *Support Vector Machines for Classification and Regression, Technical Report, Image Speech and Intelligent Systems Research Group*, University of Southampton, UK, 1997, http://www.isis.ecs.soton.ac.uk/isystems/kernel/.
68 J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.
69 R. M. Balabin, R. Z. Syunyaev and S. A. Karpov, *Energy Fuels*, 2007, **21**, 2460–2465.
70 R. M. Balabin, R. Z. Syunyaev and S. A. Karpov, *Fuel*, 2007, **86**, 323–327.
71 O. C. Mullins, E. Y. Sheu, A. Hammami and A. G. Marshall, *Asphaltenes, Heavy Oils, and Petroleomics*, Springer, 1st edn, 2006.
72 P. Luoa and Y. Gu, *Fuel*, 2007, **86**, 1069–1078.
73 D. G. Truhlar, *Chem. Phys. Lett.*, 1998, **294**, 45–48.
74 R. A. K. Nadkarni, *Guide to ASTM Test Methods for the Analysis of Petroleum Products and Lubricants*, ASTM International, 2007.
75 R. G. Brereton and G. R. Lloyd, *Analyst*, 2010, **135**, 230–267.
76 M. Sattlecker, C. Bessant, J. Smith and N. Stone, *Analyst*, 2010, **135**, 895–901.
77 S. Zomer, S. J. Dixon, Y. Xu, S. P. Jensen, H. Wang, C. V. Lanyon, A. G. O'Donnell, A. S. Clare, L. M. Gosling, D. J. Penne and R. G. Brereton, *Analyst*, 2009, **134**, 114–123.
78 E. Widjaja, G. H. Lim and A. An, *Analyst*, 2008, **133**, 493–498.
79 R. M. Balabin, R. Z. Safieva and E. I. Lomakina, Near-infrared (NIR) spectroscopyfor motor oil classification: From discriminant analysis to support vector machines, *Microchem. J.*, 2011, DOI: 10.1016/j.microc.2010.12.007.
80 R. M. Balabin, E. I. Lomakina and R. Z. Safieva, Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy, *Fuel*, 2011, DOI: 10.1016/j.fuel.2010.11.038.
81 R. M. Balabin and R. Z. Safieva, Biodiesel classification by base stock type (vegetable oil) using near infrared (NIR) spectroscopy data, *Anal. Chim. Acta*, 2011, DOI: 10.1016/j.aca.2011.01.041.
82 M. Ventura, A. Sanchez-Niubo, F. Ruiz, N. Agell, R. Ventura, C. Angulo, A. Domingo-Salvany, J. Segura and R. de la Torre, *Analyst*, 2008, **133**, 105–111.

