Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Analytica Chimica Acta 544 (2005) 292–305

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization

B. Üstün[a], W.J. Melssen[a], M. Oudenhuijzen[b], L.M.C. Buydens[a,*]

[a] *Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*
[b] *Department of Material Characterization and Analytical Technology, General Electric Advance Materials B.V., Plasticslaan 1, P.O. Box 117, 4600 AC Bergen op Zoom, The Netherlands*

Received 6 September 2004; received in revised form 10 December 2004; accepted 10 December 2004
Available online 13 January 2005

Abstract

Traditionally, the partial least squares (PLS) regression technique is most commonly used for quantitative analysis of near-infrared spectroscopic data. However, the use of support vector regression (SVR), a recently introduced alternative regression technique, for quantitative spectral analysis has increased over the past few years especially due to its high generalization performance and its ability to model non-linear relationships as well. Unfortunately, the practical use of SVR is limited because of its set of parameters to be defined by the user. For this reason, it was necessary to find an automated reliable, accurate and robust optimization approach to select the optimal SVR parameter settings. This paper presents a SVR parameter optimization approach based on genetic algorithms and simplex optimization, which satisfies all of the above-mentioned points. Furthermore, a comparison is made between the performance of SVR and PLS on various (noisy) data sets. From these results, it can be concluded that SVR is less sensitive to spectral noise, and hence, more robust with respect to spectral variations due to experimental circumstances. Generally, in the context of performance and robustness, the results demonstrate that SVR is a good well-performing alternative for the analysis and modelling of NIR data than the commonly applied PLS technique.

Keywords: Support vector regression (SVR); Partial least squares (PLS); Near-infrared (NIR) spectroscopy; Genetic algorithms (GA); Simplex optimization

1. Introduction

Near-infrared spectroscopy (NIR) is a potentially effective technique which has been widely applied in the pharmaceutical, polymer and food industry in recent years. Generally, NIR spectroscopy is used in combination with partial least squares (PLS) regression technique for quantitative analysis [1,2]. The ratio to use PLS is given by the characteristics of PLS, namely its simplicity, a high computational speed, for most problems a satisfying performance and, last but not least, its power of allowing the interpretation of the scores and loadings. However, PLS has some deficiencies like, modelling of data sets containing strong non-linear relationships [3]. Recently, a regression version of support vector machines (SVMs) has emerged as an alternative and powerful technique to solve this problem. In the sequel, this version is referred to as support vector regression (SVR). SVR, which is a statistical learning theory based machine learning formalism, developed by Vapnik [4], is gaining popularity due to its many attractive features and promising generalization performance. Some prominent features of SVR are: (i) the ability to model non-linear relationships, (ii) the ability to select only the necessary objects (spectra) to solve the regression function, which results in a sparse solution, (iii) the regression function is related to a quadratic problem (QP) which solution is global and in general unique. Apart form these features, SVR has also a drawback that limits the use of SVR on academic and industrial platforms: there are a number of free parameters that need to be defined by the user. Since the

[*] Corresponding author. Tel.: +31 24 36 53192; fax: +31 24 36 52653.
E-mail address: lbuydens@sci.kun.nl (L.M.C. Buydens).

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.12.024

generalization performance of the SVR models depends on a proper setting of these parameters, the main issue is to find the optimal settings for a given data set. Whereas the effects of these parameters are defined in different sources on SVR, there are no general guidelines to select these [4–8]. The problem of optimal parameter selection is further complicated by the fact that the SVR model complexity (and hence, its generalization performance) depends on all of these parameters together (interaction of parameters). This means that a separate optimization of each parameter is not sufficient enough to find the optimal regression model. For this reason, usually a very time-consuming grid search optimization method is invoked to find the optimal SVR parameter settings [5,9].

In this paper, an accurate, robust and fast approach based on genetic algorithms (GAs) and the simplex search technique is presented that identifies the optimal parameter settings, this instead of performing a time-consuming grid search. The optimization is conducted for SVRs equipped with a radial basis function (RBF, i.e., a Gaussian shaped function) and a variety of polynomial kernels. The motivation for selecting GAs in combination with simplex is that GAs performs a global coarse-grained search and thus is most likely to arrive at or nearby the global optimum solution. Taken this solution as new starting point, the simplex can be used to find possibly the exact optimal solution, due to its local fine-grained search character. To our opinion, this approach will stimulate the use of SVR on various research areas. Furthermore, this paper will focus on the importance of finding robust optimized parameter settings which lead to a good performance of the SVR model. Additionally, the robustness of SVR and PLS are compared for noisy data. Three NIR spectral data sets were selected to evaluate the intended SVR parameter optimization approach. The first two data sets, extensively described in literature, are mainly used as a benchmark to evaluate the validity and power of the combined GA/simplex SVR parameter optimization strategy. A straightforward and fair benchmark is possible because PLS is extensively applied to both data sets in literature [2,3]. Furthermore, the SVR results obtained for the first data set, but based on a grid search parameter optimization, performed by Thissen et al. [3], are also available.

The third data set, a real-life industrial data set provided by General Electric Advanced Materials B.V., The Netherlands [10], is beside evaluation of the optimization strategy also used to study the performance of SVR on this data set.

2. Theory

2.1. Support vector regression

Here a brief description of SVR is given. For a more detailed description the reader is referred to Vapnik [4,7], Schölkopf and Smola [5] and Cristianini and Shawe-Taylor [6].

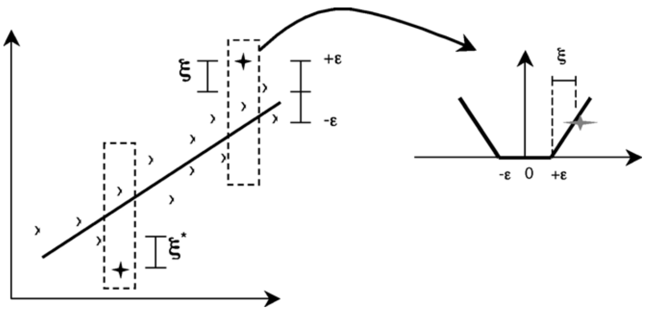


Fig. 1. In SVR, a tube with radius ε is fitted to the data. The trade-off between model complexity (flatness) and points lying outside the tube (slack variables ξ) is determined by minimizing Eq. (4). The points outside the ε zone are support vectors (black stars). On the right, the ε-insensitive loss function is shown in which the slope is determined by C (the star represents a support vector).

Like most linear regression models e.g. PLS, the SVR algorithm developed by Vapnik [4,7] relies on estimating a linear regression function:

$$f(\mathbf{x}) = w^{\mathrm{T}}x + b \quad (w, x \in R^d (d\text{-dimensional input space})) \tag{1}$$

where w and b are the slope and offset of the regression line. In case of SVR, the regression function is calculated by minimizing:

$$\frac{1}{2}w^{\mathrm{T}}w + \frac{1}{n}\sum_{i=1}^{n} c(f(\mathbf{x}_i), y_i) \tag{2}$$

where $(1/2)||w||^2$ is the term characterizing the model complexity (smoothness of $f(\mathbf{x})$) and $c(f(\mathbf{x}_i), y_i)$ the loss function determining how the distance between $f(\mathbf{x}_i)$ and the target values y_i should be penalized. In this so-called primal formulation, several different loss functions [5,8] are available, but in this paper we adopted the commonly used ε-insensitive loss function which was introduced by Vapnik [4]. This ε-insensitive loss function is defined by:

$$c(f(\mathbf{x}_i), y_i) = \begin{cases} 0, & |y - f(\mathbf{x})| \le \varepsilon \\ \text{if } |y - f(\mathbf{x})| - \varepsilon, & \text{otherwise} \end{cases} \tag{3}$$

In fact, this particular constraint defines a tube with radius ε around the hypothetical regression function (see Fig. 1) in such way that if a data point is positioned in this tube the loss function equals 0, while if a data point lies outside the tube, the loss is proportional to the magnitude of the Euclidean difference between the data point and the radius ε of the tube. In this particular case, the minimization of Eq. (2) is equivalent to solving the following constrained optimization problem:

$$\text{minimize} \quad \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4}$$

$$\text{subject to} \begin{cases} y_i - w^{\mathrm{T}}x_i - b \le \varepsilon + \xi_i \\ w^{\mathrm{T}}x_i + b - y_i \le \varepsilon + \xi_i^* \\ \varepsilon, \xi_i, \xi_i^* \ge 0 \end{cases} \tag{5}$$

where the constant $C > 0$ determines the trade-off between the model complexity of $f(\mathbf{x})$ and the amount up to which deviations larger than ε are tolerated. The slack variables ξ_i, ξ_i^* are introduced for the situation that the target value exceeds, this with respect to the origin of the original data space, more than ε above (ξ_i) and more than ε below the target (ξ_i^*), see Fig. 1. The points lying outside the ε tube are named support vectors (SVs), because these establish ('support') the fundaments of the estimated regression function. This implies that all other data points are in fact not important for inclusion into the model and can be removed after the SVR model has been constructed. Hence, usually (much) less training objects do constitute the regression model; therefore, such a solution is referred to as 'sparse'.

The constrained optimization problem given by Eqs. (4) and (5) can be reformulated into dual problem formalism (Eq. (6)) by using Lagrange multipliers. In this paper we adopted the strategy outlined by Vapnik [4], which leads to the solution:

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b \tag{6}$$

where α_i and α_i^* (with $0 \le \alpha_i, \alpha_i^* \ge C$) are the Lagrange multipliers and $\mathbf{K}(\mathbf{x}_i, \mathbf{x})$ represent the so called kernel function [4–6,11]. Intuitively, the primal formulation is suitable to solve problems where many objects (samples) are available, this with respect to the number of variables at hand. The dual Lagrangian formalism, on the other hand, eliminates the curse of dimensionality, and hence, is even suitable to find solutions for ill-posed problems. In the context of Eq. (6), data points with nonzero α_i and α_i^* value are SVs. It has been shown that a suitable kernel function makes it possible to map a non-linear input space to a high-dimensional feature space where linear regression can be performed [4]. Several kernel functions have been proposed in literature, but the particular choice of a kernel to map the non-linear input space into a linear feature space depends highly on the nature of the data representing the problem at hand. In this paper the focus is put on two widely used kernel functions, namely, radial basis function (RBF) and the polynomial function, which are defined in Eqs. (7) and (8), respectively:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right) \tag{7}$$

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + 1)^d \tag{8}$$

In case of the RBF kernel the parameter σ represents the kernel width whereas d in Eq. (8) denotes the degree of the polynomial kernel. It should be noted that in the latter case a $d = 1$ equals the standard vector inner product including a constant bias term. The kernel parameter and also the earlier mentioned parameters C and ε need to be selected properly by the user, because the generalization performance of the SVR model heavily depends on the right setting of these three parameters. Hence, a reliable and robust parameter selection optimization strategy is a pre-requisite to obtain a well-performing and robust SVR regression model.

2.2. Genetic algorithms

The genetic algorithm (GA) is a search algorithm based on the principles of natural evolution. A GA is stochastic search technique, formerly introduced by Holland [12], which can be used to find the global optimal solution in a complex multidimensional search space [13]. The algorithm starts with a set of random solutions called *population*. An individual solution is represented in encoded form and is called a *chromosome*. Each chromosome comprises of a sequence of individuals structures called *genes* which represent the actual parameters to be optimized. The solutions from one population are used to generate the next population. This is motivated by the idea (supported by the theory of building blocks [12]) that the new population, on average, will be better than the population of predecessors. In order to create a new population, GA uses genetic operators (i.e., crossover and mutation) and a selection process. Genetic operators are used to generate the new solutions (children population or offspring) from the current set of solutions (parent population). Selection reflects the principle of 'survival of the fittest' and is the driving mechanism of keeping and deleting some solutions from the parent population to generate an offspring with the same number of chromosomes. During this selection process, the solutions are selected according to their values of objective function (usually referred to as fitness). A high fitness value of a chromosome corresponds to a higher chance to be selected for the next generation. The GA will repeat this process until a termination condition is satisfied. The best solution is returned to represent the optimum solution.

The procedure of GA can be summarized in the following steps:

1. Choose a randomly generated population.
2. Calculate the fitness of each chromosome in the population.
3. Create the offspring by the genetic operators: selection, crossover and mutation.
4. Check the termination condition. If the new population does not satisfy the termination condition, repeat steps 2 up to 4 for the generated offspring as a new starting population.

Conventionally, the chromosomes in a GA are normally binary coded (strings of bits taking the values 0 or 1), which in fact lead to integer valued solutions [12–14]. For large problems, that is, where many parameters are involved or where a high resolution in the parameter domain is required, binary encoding results in very large strings (number of bits per variable) which slow down the evolutionary process. On

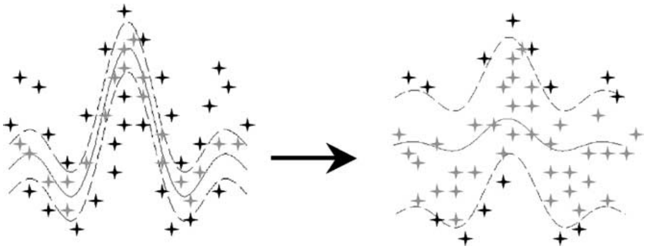


Fig. 2. A small ε value (left figure) allows more points to be outside the ε-tube (indicated by the dashed lines) and results in more SVs (depicted as black stars). A large ε value (right figure) results in less SVs and probably in a smoother regression function.

the other hand, if the length of the string is not long enough, it might be possible for the GA to get near to the region of the global optimum but never will arrive at it (due to the digitalization of the continuous value spectrum). Thus, since our optimization problem involves real-valued parameters, it is better to manipulate these directly in the original real-number space. For this reason we have applied real-value encoding [15–18].

2.3. Simplex

The simplex algorithm formulated by Nelder and Mead [19] is a robust direct local search method which, in n-dimensions, starts with $n+1$ points defining an initial simplex (polygon) and proceeds in a series of steps to find a local optimum (minimum or maximum) in the parameter space. The driving force underlying the simplex optimization algorithm is the set of rules which determine how the simplex polygon has to be rotated in order to move towards the local optimum. The actual dimension (n) of the simplex is equal to the number of control parameters. The shape of a simplex in one-, two- and three-dimensional parameter space is a line, a triangle or a tetrahedron, so the number of vertices of the simplex is equal to $n+1$. A geometric interpretation for more than three parameters is difficult, but a general description of the simplex method for n-parameters is formulated in [19,20].

3. Experimental

3.1. GA/simplex SVR parameter optimization

As mentioned in Section 2.1, three parameters, i.e., ε, C and one kernel parameter (σ in case of the RBF kernel and d for the polynomial kernel) need to be optimized for the SVR. The parameter ε regulates the radius of the ε tube around the regression function and thus the number of support vectors that finally will be selected to construct the regression function (leading to a sparse solution). A too large ε value results in less support vectors (more data points will be fit in the ε tube) and, consequently, in a more smooth (less complex) regression function (see Fig. 2). Unfortunately, in that case, the resulting regression model is not always applicable (large prediction errors on unseen data might be the result). It is known that the value of ε is related to the amplitude of the noise present in the training set [9]. Since the exact contribution of the noise to the real information in a data set is usually unknown, ε was varied in a range between 0 and 0.2. This range covered the full range of possible noise contributions well, for the particular data pre-processing methods applied throughout this paper.

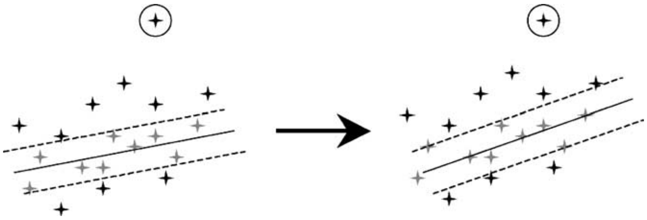


Fig. 3. In case of a small C value (left), the most remote SVs (dashed line) have almost the same contribution. For this reason, the regression function is more robust for possibly outliers (SV within the circle). Increasing of the C value result in a regression line, which is more influenced by the possible outlier (right). The distance between the outlier (SV within the circle) and the ε-tube will decrease in comparison to case if a small C value is chosen (left).

The parameter C determines the trade-off between the smoothness of the regression function and the amount up to which deviations larger than ε are tolerated. Furthermore, the robustness of the regression model depends on the choice of the C value, because the highest α_i and α_i^* values are by definition (according to the Lagrange optimization procedure) equal to C. This means that the choice of the C value influences the significance of the individual data points in the training set. For example, on one hand, a high C value results in support vectors with a high difference between the α_i and α_i^* values. In that case the support vectors with the highest α_i and α_i^* values are dominating the constructed regression function. On the other hand, a small C value can result in support vectors having small differences or even similar α_i and α_i^* values. In that case the data points selected as support vectors with similar α_i and α_i^* contribute equally to the regression function. Hence, a proper choice of C in combination with ε might result in a well performing and robust regression model, which is also insensitive to the presence of possible outliers (Fig. 3). A good choice of both parameters also prevents overtraining. In order to verify this, during the construction and selection of all SVR models, permanently an internal cross-validation procedure was applied. According to the theory of SVR, C takes values between 0 and ∞. In practice, we varied C between 1 and 1×10^8.

The kernel function represents the mapping instrument that is necessary to transform the non-linear input space to a high-dimensional feature space where linear regression is possible [4–6,17]. The mapping depends on the intrinsic topological structure of the data, implying that the kernel type and parameters need be optimized to approximate the ideal mapping [17]. In this paper we will focus on two commonly

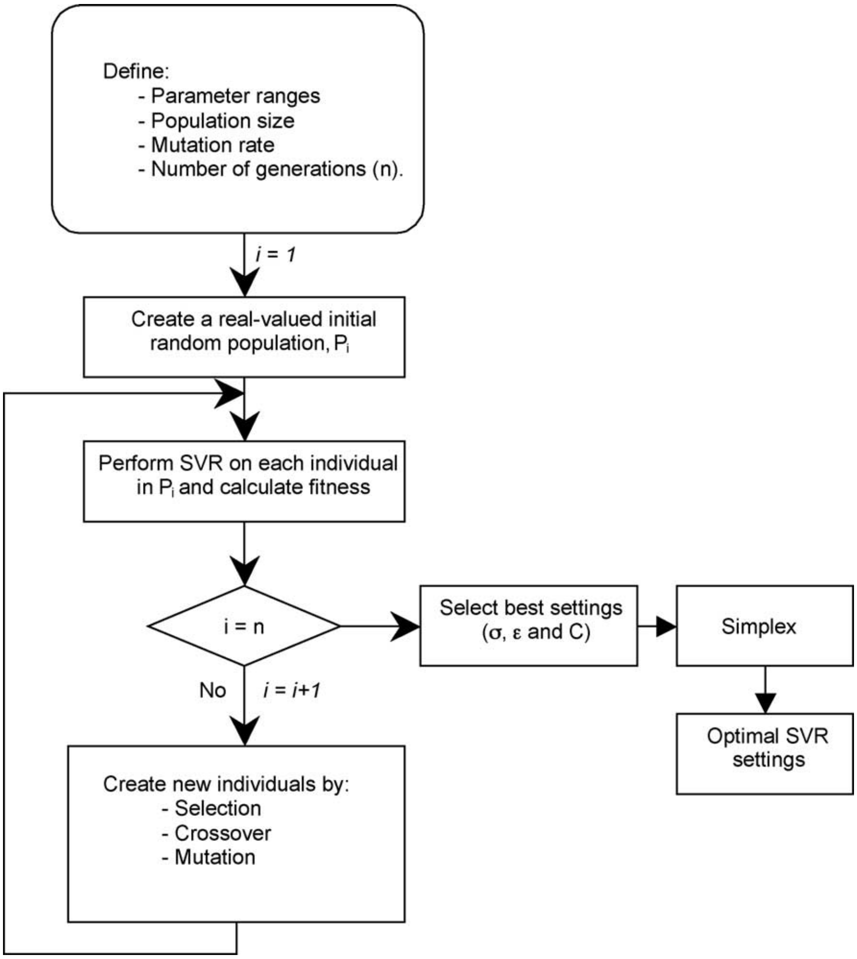

Define:
- Parameter ranges
- Population size
- Mutation rate
- Number of generations (n).
i = 1
Create a real-valued initial random population, P_i
Perform SVR on each individual in P_i and calculate fitness
i = n
Select best settings (σ, ε and C)
Simplex
Optimal SVR settings
No
i = i+1
Create new individuals by:
- Selection
- Crossover
- Mutation

Fig. 4. Flow diagram of the combined GA/simplex SVR parameter optimization procedure.

used kernel functions, namely, the RBF and the polynomial kernel. In case of the RBF kernel we have searched initially the space between 0.01 and 2.0 to find the optimal kernel width σ (high generalization performance). It should be noted that a RBF kernel with a large σ behaves similar like a polynomial or, at the lower bound, a linear function [21,22]. For this reason, if during the optimization process the maximum σ value for the RBF kernel was found, a polynomial kernel was used instead. The polynomial degree d was searched in the range between 1 and 6. Note that the polynomial kernel with $d=1$ corresponds to the linear case, implying that the model is strictly based on the space spanned by the inner-products of the objects in the training set.

As mentioned above, ε, C and the kernel-dependent parameter exhibit a (strong) interaction. As a consequence, these parameters cannot be optimized separately. Here, the parameters are optimized by using GA (real-value coded) in combination with a simplex optimization. Fig. 4 depicts the flow diagram of the followed procedure. The two main reasons that we have performed a GA/simplex optimization instead of a grid search for finding the best SVR parameters are accuracy and computational speed. The accuracy of the grid search optimization depends on the parameter range in combination with the chosen interval size (resolution in the parameter space). To increase the accuracy of the optimal solution one might need to increase the parameter range and/or decrease the step size substantially. However, this will result in a cumbersome time-consuming optimization process. The GA/simplex optimization depends exclusively on the pre-defined parameter range and, moreover, is faster in finding an optimal set of parameters as compared to a fine-grained grid search in the same search space.

Due to its stochastic character, the procedure outlined in Fig. 4 was repeated at least five times to select the optimal parameter settings. There exist various types of selection, crossover and mutation operators [13–16]. Table 1 gives an overview of those ones and their particular settings, which we found the most efficient, as was determined by a preliminary exploratory study on some NIR data sets. The fitness of each individual chromosome in the GA population was expressed as the root mean square error of cross-validation (RMSECV), calculated for the difference between the predicted value, $\hat{y}$, and the corresponding target value, y (Eq. (9)). Given the size of the training sets, a leave-15-out cross-validation [23] was

Table 1
An overview of the SVR and GA parameter settings

SVR settings	
ε	0–0.2
C	$1–1 \times 10^8$
σ	0.01–2.0
d	1–6
GA settings	
Population size	100
Number of generations	15
Selection type	Stochastic universal sampling[a]
Crossover type	Discrete recombination[a]
Mutation type	Real-value mutation[a]
Mutation rate	1/3

[a] A detailed description of these operators is given in [16].

applied during the training phase of the SVR model

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (9)$$

3.2. Data

3.2.1. Data set 1

The first data set contains NIR spectra of 19 ternary mixtures of ethanol, water and *iso*-propanol measured at five ambient temperatures (30, 40, 50, 60 and 70 °C, respectively) in a spectral range of 850–1049 nm at a resolution of 1 nm intervals. Fig. 5 shows the ternary mixture design and, in white and grey coding, the training and test set samples. Originally, the spectra were measured by Wülfert et al. [2] to investigate the influence of temperature-induced spectral variations on the predictive ability of partial least squares (PLS). In this investigation local and global regression models were set up by Wülfert et al. [2]. Additionally, SVR regression was applied by Thissen et al. [3] to this data set, according to the same approach as described in the original paper to compare the performance of SVR to PLS. In this paper, exclusively global SVR models were made by applying the GA/simplex optimization approach as described in Section 3.1. For a fair comparison with the results of previous publications, exactly the same data pre-treatment has been performed as described in [2,3].

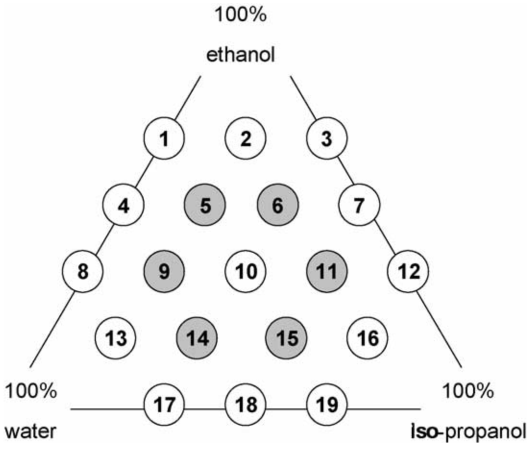


Fig. 5. Mixture design for ethanol, water and 2-propanol. The white circles were used as training set and the grey as test set, this for each temperature value. For more details, the reader is referred to the text.

3.2.2. Data set 2

The second data set contains 80 NIR spectra of corn samples measured on three different NIR spectrometers (indicated by the acronyms: m5, mp5 and mp6) for the prediction of moisture, oil, protein and starch content, respectively. The spectra were measured from 1100 to 2498 nm at a spectral resolution of 2 nm. The data were taken from the Cargill study and can be obtained at http://software.eigenvector.com/data/corn/index.html. In this paper we predicted the moisture, oil, protein and starch content of the samples originating from the set of spectra measured on the m5 NIR spectrometer. Because PLS is already applied on these spectra by Fuedale et al. [1], we used the outcome of this study as a benchmark for our SVR results. Again, as was the case for data set 1, the same data pre-treatment and training and test set selection has been performed as described in [1].

3.2.3. Data set 3

The third data set contains NIR spectra of methanol distillation samples, on-line measured at a process stream of a polymerization plant of General Electric Advanced Materials located in Bergen op Zoom, The Netherlands [10]. Methanol is used as the main solvent in the polymerization process and in order to minimize cost and environmental load, the methanol is recycled and purified by distillation. The main remaining impurity after distillation in the methanol is H_2O, and from the viewpoint of energy savings it is interesting to keep the percentage of water as close to the pre-defined specification limits as possible. To keep the distillation energy low and for a direct interaction with the process stream it is interesting to use an on-line measurement in the distillation process. One of the most suitable techniques for such a process stream control (at moderate ambient temperatures, monitoring the amount of water and methanol) is near-infrared spectroscopy in combination with multivariate calibration. In this paper, we applied SVR to this problem, because of its earlier mentioned advantages. For completeness also PLS was applied to enable a comparison of its performance to that of the SVR regression model.

In order to perform this work, 131 NIR spectra were collected of methanol distillation samples, measured in a temperature range of 20–40 °C. Roughly 30% of the samples was collected from the plant (46 plant samples). The remaining part was measured under controllable laboratory circumstances (86 laboratory samples). For these laboratory samples, an experimental design was used which encompassed the expected variations in the plant situation. The percentage water in the samples is determined with a gas chromatographic instrument (HP 5890A) equipped with a HP1 column

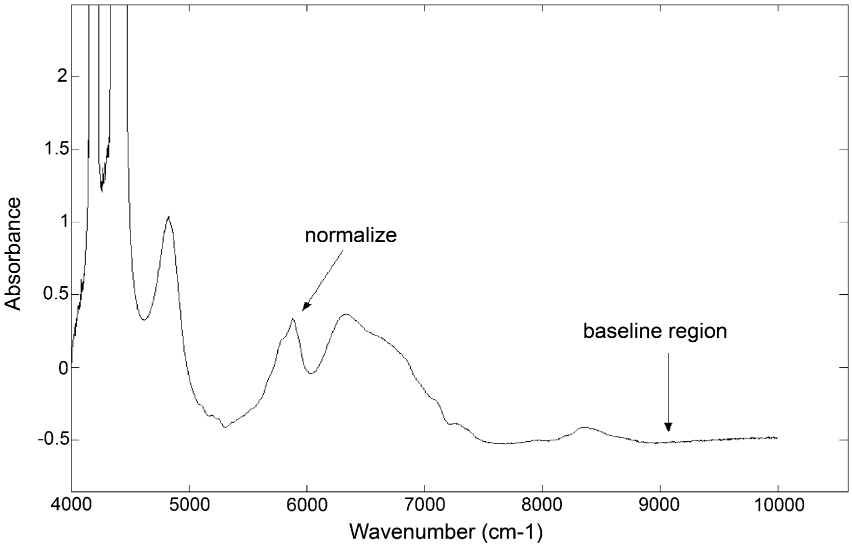


Fig. 6. Near-infrared spectrum before preprocessing. The baseline was subtracted for the entire spectrum by fitting a first-order function through the region 9200–9899 cm^{-1}. After baseline subtraction the spectra were corrected for small differences in pathlength by normalization on the height of the isolated peak around 5900 cm^{-1}.

and a FID detector. For each analysis, 1 μl of the sample was injected using an autosampler (HP G15113A) at an injection temperature of 200 °C in a Helium flow of 19 ml/min. The chromatogram was recorded during 10 min and both the methanol and water peaks were integrated automatically (using preset integration limits) in the HP ChemStation software. The intensity ratios were calibrated using the laboratory samples with known water percentages. The NIR spectra were collected on a Thermo-Nicolet Antaris FT-NIR spectrometer in a range of 4000–10,000 cm^{-1} with a resolution of 4 cm^{-1}. This spectrometer was equipped with a Thermo-Nicolet Sab-IR Fiber Optic probe with a transflectance unit on top, set at a fixed path length of 3.5 mm. Inherent to the design of such a flexible probe unit, the pathlength could vary between 0.1 and 0.2 mm, so the spectra were needed to be normalized to correct for such path length differences. A typical, unprocessed spectrum is shown in Fig. 6. The spectra were pre-processed by (1) a baseline subtraction by fitting a 1st order line in the region of 9800–9200 cm^{-1}, (2) a normalization of the spectra on the maximum intensity around 5900 cm^{-1}, and (3) removing the region where the absorption value was higher than 2 (4000–4500 cm^{-1}).

3.3. Software

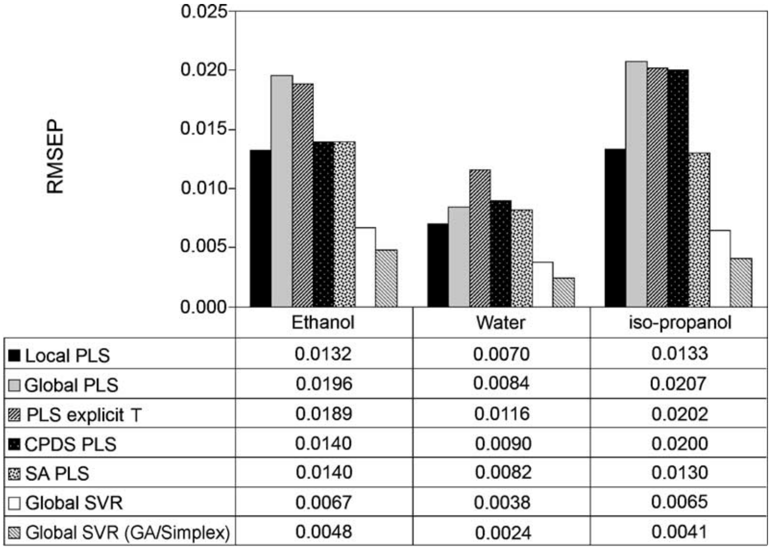
All calculations were carried out by using Matlab (V6.5, The Mathworks, Inc.) and the SVR toolbox developed by Gunn [8]. The latter toolbox can be obtained from http://www.isis.ecs.soton.ac.uk/isystems/kernel/. The SVR parameters are optimized by using the GA toolbox designed by Chipperfield et al. [16] and the Nelder–Mead simplex method taken from the standard Matlab optimization toolbox. The calculations are performed on a 3.0 GHz Intel Pentium IV PC with 1 GB RAM under windows XP.

4. Results and discussion

4.1. Data set 1

Within the framework of the investigation of the influence of temperature-induced spectral variations on the predictive ability of PLS, two types of models were built by Wülfert et al. [2], namely, local and global regression models. In case of local models, a PLS model is built for each particular temperature to predict the mixture mole fractions at that specific temperature. So, a total of five local models is the result. Opposed to this, a global PLS model contains information regarding all temperatures and, hence, is suitable to predict the mixture mole fractions at the full temperature range of 30–70 °C. Obviously, the advantage of such global model is that it is not necessary to know at which temperature an unknown sample is measured. Results of local and global PLS models, published by Wülfert et al. [2], are summarized in Table 2.

From Table 2 it can be conclude that the global model performs worse than each individual local model. In subsequent publications, several methods were applied on this data set to incorporate the temperature effect in a more explicit way in the PLS model to improve the prediction performance. Some of these methods are mentioned here below briefly. Different linear modelling techniques were performed by Wülfert et al. [24] by including the temperature directly as an extra variable into the PLS model. Alternatively, continuous piecewise direct standardization (CPDS) has been applied to the NIR spectra for correction of the (non-linear) temperature effect [25]. Next to this, Sweringa et al. [26] have applied a robust variable selection method by utilizing simulated annealing (SA). This, to decrease the temperature effect by removing undesired (i.e., temperature affected) spectral variations. In



	Ethanol	Water	iso-propanol
■ Local PLS	0.0132	0.0070	0.0133
□ Global PLS	0.0196	0.0084	0.0207
▨ PLS explicit T	0.0189	0.0116	0.0202
■ CPDS PLS	0.0140	0.0090	0.0200
▩ SA PLS	0.0140	0.0082	0.0130
□ Global SVR	0.0067	0.0038	0.0065
▧ Global SVR (GA/Simplex)	0.0048	0.0024	0.0041

Fig. 7. Comparison of prediction errors of the different modelling methods (see text) to predict the mole fractions of ethanol, water and *iso*-propanol in a ternary mixture.

contrast to previous approaches, which are based on the linear PLS technique, Thissen et al. [3] have performed SVR on the same data set, because of the ability of SVR to model non-linear relationships in the data. Fig. 7 depicts a comparison of the results of (the best performing models of) Wülfert et al. [2] and the other methods described above. It can be observed that the SVR model outperforms all PLS approaches. This probably attributes to the SVR results and parameter settings used by Thissen et al. [3] which are provided in Table 3 . These results and settings are used as benchmark for our GA/simplex SVR parameter optimization procedure as has been described in detail in Section 3.1.

The optimal SVR parameter settings and results using the GA/simplex optimization procedure are presented in Table 4 . Examination of this table leads to the conclusion the GA/simplex optimization approach considerably improves the results of Thissen et al. [3], by on average 34% (see Fig. 7). Our main goal was to develop an accurate and robust optimization method. From the results of Table 4 in comparison with that of Table 3, it can be conclude that our optimization approach is more accurately, i.e., results in substantial lower prediction errors. Probably, a more accurate solution is obtained because a GA can optimize more exactly and in a global sense. This indicates the necessity of determining the proper SVR parameter setting to achieve a well-behaving (expressed in terms of prediction errors) regression model. To test the selected SVR parameter settings on its robustness, we varied σ and C by ±10% and ±25% in combination with $\varepsilon + 0.001$ (five-level factorial experimental design) and calculated the RMSEP values. The outcome of this robustness test is only presented for ethanol. Water and

Table 2
Prediction results of the local and global models obtained by Wülfert

Model	Temperature (°C)		RMSEP		
	Sample	Model	Ethanol	Water	*iso*-Propanol
Local	30	30	0.0177	0.0092	0.0124
	40	40	0.0106	0.0067	0.0093
	50	50	0.0166	0.0111	0.0218
	60	60	0.0980	0.0043	0.0083
	70	70	0.0112	0.0038	0.0147
Mean			0.0132	0.0070	0.0133
Global	30	30–70	0.0138	0.0125	0.0113
	40	30–70	0.0132	0.0055	0.0164
	50	30–70	0.0377	0.0079	0.0405
	60	30–70	0.0159	0.0084	0.0174
	70	30–70	0.0175	0.0076	0.0175
Mean			0.0196	0.0084	0.0207

Table 3
SVR parameter settings and prediction results published by Thissen et al. [3]

	Optimal SVR settings[a]		
	Ethanol	Water	*iso*-Propanol
σ	0.5	0.5	0.5
ε	0.005	0.005	0.005
C	500	500	500

Model	Temperature (°C)		RMSEP		
	Sample	Model	Ethanol	Water	*iso*-Propanol
Global	30	30–70	0.0040	0.0029	0.0069
	40	30–70	0.0058	0.0024	0.0058
	50	30–70	0.0087	0.0064	0.0053
	60	30–70	0.0055	0.0028	0.0055
	70	30–70	0.0094	0.0043	0.0092
Mean			0.0067	0.0038	0.0065

[a] Note that the optimal SVR settings are obtained by an exhaustive grid search with a leave-one-out cross-validation on the training set. The SVR was equipped with a RBF kernel.

Table 4
SVR parameter settings and prediction results obtained by GA/simplex method

	Optimal SVR settings[a]		
	Ethanol	Water	*iso*-Propanol
σ	0.4296	0.3146	0.3976
ε	0	0	0
C	101744	232	490329

Model	Temperature (°C)		RMSEP		
	Sample	Model	Ethanol	Water	*iso*-Propanol
Global	30	30–70	0.0029	0.0030	0.0025
	40	30–70	0.0039	0.0019	0.0035
	50	30–70	0.0047	0.0028	0.0029
	60	30–70	0.0037	0.0014	0.0039
	70	30–70	0.0090	0.0028	0.0076
Mean			0.0048	0.0024	0.0041

[a] The optimal SVR settings are obtained by using the RBF kernel.

iso-propanol exhibited a similar behaviour. Compared to the RMSEP values obtained with the optimal settings, the maximal relative difference for the variation of ±10% is only 4% (0.0048 ± 0.0002) and 16% (0.0048 ± 0.0008) for a variation of ±25%. These observations led us to the conclusion that the proposed GA/simplex SVR optimization procedure yields parameter settings which are accurate as well as robust.

4.2. Data set 2

Recently, Fuedale et al. [1] have investigated the effects of various orthogonal signal correction (OSC) algorithms on the modelling power of PLS. Initially, OSC was developed by Wold et al. [27]. The basic idea underlying OSC is to remove systematic variations from the spectra (X-matrix) which are not related (thus orthogonal) to the information contained in the response matrix Y. Most common sources of (undesired) systematic variations in spectra are background noise and baseline drift. In order to investigate the influence of OSC, Fuedale et al. [1] applied PLS to non-processed spectra, OSC-corrected spectra according to the algorithms described by Wise and Gallagher [28] and Fearn [29], respectively, and on piecewise orthogonal signal correction (POSC) corrected spectra as well [1]. Especially, the PLS results obtained without any spectral correction will be used in this paper as a benchmark to evaluate the performance of our SVR approach. SVR will be applied on the non-processed spectra.

To evaluate the power of the GA/simplex SVR parameter optimization approach, we applied SVR on data set 2. Both data pre-treatment and the division into a training set and a test set were conducted according to Ref. [1]. The optimal SVR parameter settings and the prediction results for moisture, oil, protein and starch content are summarized in Table 5 .

Fig. 8 depicts the graphical comparison of the PLS results on non-processed spectra, OSC-filtered spectra attained by the Wise and Fearn algorithms, POSC-filtered spectra, and our SVR results for non-processed spectra. As can be seen, for most cases POSC pre-processing of the spectra improves the PLS results. However, SVR performs better than PLS without any spectral correction for moisture, protein, oil and starch content. Strikingly, SVR performs also better than PLS after POSC for the prediction of moisture, protein and starch content. It can be envisaged that the modelling capability of SVR is stronger than that of PLS combined with or without a spectral correction procedure. An exciting investigation would be to determine the performance of SVR after OSC/POSC spectral correction. However, this extended comparison is beyond the scope of this paper. But the current results suggest strongly that SVR does not require one or more data pre-processing steps, like OSC. Build on the raw data, SVR already yields a satisfactory model. In summary, also for data set 2 it appeared that a thoroughly optimized SVR model serves as an attractive and better performing alternative for the widely applied PLS technique. Furthermore, the SVR parameter robustness test performed on data set 1 is also applied on data set 2. However, now only ε and C are varied by ±10% and ±25% or $\varepsilon + 0.001$. Because, the polynomial degree (d) must change by integer values and it is on beforehand known that this will have large effect (e.g. $d = 1$ and $d = 2$ will give complete different results with the same ε and C values), d was kept constant. Compared to the RMSEP values obtained with the optimal settings, the highest relative difference is obtained for the moisture content. The maximal relative difference for the variation of ±10% is 1% (0.0100 ± 0.0001) and for a variation of ±25% is only 4% (0.0100 ± 0.0004). Together with the results of data set 1, it can conclude that the proposed SVR parameter optimization approach is able to find accurate and robust parameter settings.

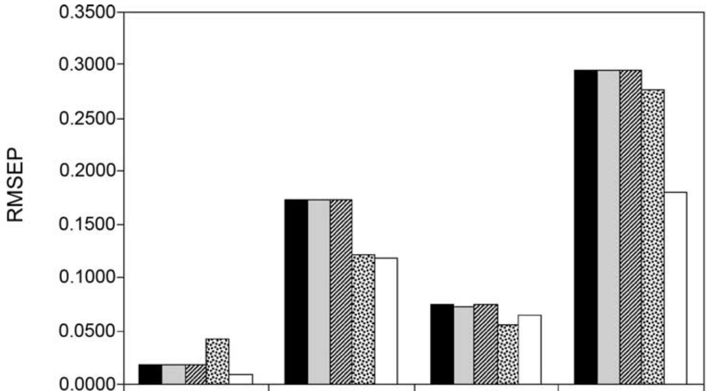
4.3. Data set 3

To predict the percentage of water impurity in the methanol distillation samples from NIR spectra SVR was applied in two ways. First, the SVR machinery was trained by invoking all spectra (plant and laboratory samples together) and calculated the prediction error by a leave-10-out cross-validation (method A). Second, the laboratory samples were selected as the training set, whereas the plant samples were used as an external validation set to determine the generalization abilities of the SVR regression model (method B). The latter has been done to check whether a SVR regression model exclusively based on controllable laboratory samples can be applied to predict real-world samples, on-line measured in the plant. The distribution of the plant and laboratory samples is shown in Fig. 9. Due to the applied experimental design, clearly, the spectra from the plant samples are situated between the extreme spectra of the laboratory sample set. Thus, theoretically, implicitly assuming that a non-transient relation is present between the spectra and the corresponding concentrations, it might be expected that the

Table 5
SVR parameter settings and prediction results

	Moisture	Protein	Oil	Starch
SVR settings				
d[a]	1	1	1	1
ε	1.08×10^{-5}	9.97×10^{-6}	0	6.37×10^{-2}
C	2.81×10^{3}	1.74×10^{4}	4.55×10^{3}	1.07×10^{6}
RMSEP	0.0100	0.1190	0.0654	0.1806

[a] d = degree of polynomial kernel.



	moisture	protein	oil	starch
■ PLS	0.0190	0.1732	0.0742	0.2946
□ PLS Wise OSC	0.0189	0.1733	0.0741	0.2946
▨ PLS Fearn OSC	0.0190	0.1732	0.0742	0.2946
▩ PLS POSC	0.0433	0.1211	0.0555	0.2759
□ SVR	0.0100	0.1190	0.0654	0.1806

Fig. 8. Comparison of the prediction errors of various PLS approaches and SVR.

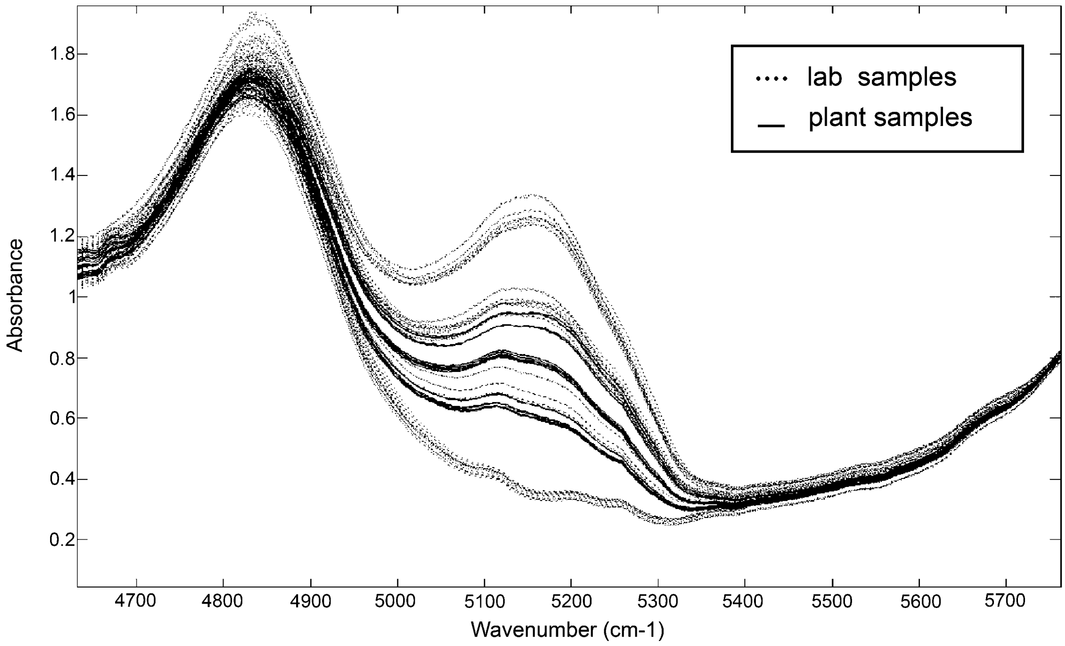
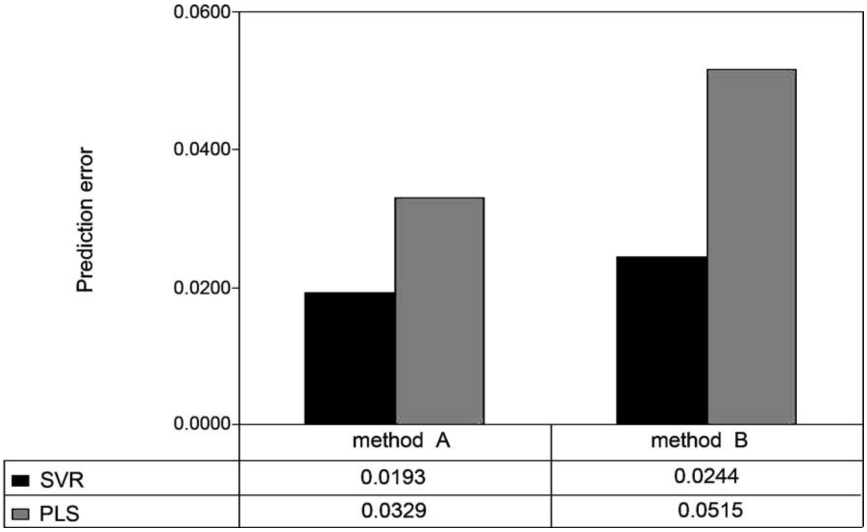


Fig. 9. Distribution of NIR spectra measured in the laboratory (dotted lines) and the plant (solid lines).



	method A	method B
■ SVR	0.0193	0.0244
▨ PLS	0.0329	0.0515

Fig. 10. Comparison of prediction errors obtained by SVR and PLS. In case of method A, the prediction error is determined by a leave-10-out cross-validation procedure (RMSECV). In case of method B, the plant samples are used as an external test set to determine the prediction error (RMSEP).

water impurity concentrations of the samples measured at the plant can be forecasted by using the regression model constructed for the set of laboratory samples. For completeness, we applied PLS for methods A and B in the same way as was done for SVR. Again, the performance of SVR is compared to PLS in terms of accuracy of prediction and model robustness. Fig. 10 depicts a visual comparison of the obtained performances of SVR and PLS. The optimal SVR parameters were for method A: $d = 2$ (polynomial kernel), $\varepsilon = 0.0044$, $C = 1$ and for method B: $d = 2$, $\varepsilon = 0.0024$, $C = 1$, respectively. In case of PLS, the lowest prediction error was obtained by taking nine latent variables into account for both validation methods.

Clearly, SVR outperforms PLS for both validation methods (A and B). The methanol samples are measured at a temperature range varying from 20 up to 40 °C. The possible temperature affected spectra (non-linear behaviour, cf. the discussion for data set 1) are most probably the main cause that SVR performs better than PLS. More interesting is that SVR is better capable, as compared to PLS, in predicting the concentrations of the plant samples by using a regression model consisting of exclusively laboratory samples. First, this serves as an indication that SVR is robust for small differences in spectra measured under laboratory and plant conditions. The advantage of the latter is that it is not necessary to collect more samples from the plant to make a new regression model, which is robust under real-world plant conditions (which saves, a.o. time, money, and human effort). Furthermore, as can be observed in Fig. 9, the laboratory spectra reach over a much larger absorption range than the plant samples do. So, in the ultimate situation, it is possible to predict accurately also the concentrations of plant samples possessing higher or lower water percentages as normally would have been expected.

Another interesting subject to investigate is: how robust is SVR in comparison to PLS for spectral deviations (e.g., in presence of noise caused by ambient influences)? To answer this question, artificially normally distributed noise (Gaussian white noise) has been added to the spectra to simulate spectral variations as typically can be expected in real industrial circumstances. To generate spectra with noise, first, the highest spectral variation was determined. This was done by calculating the spectral variance between sample spectra of comparable water percentages of laboratory and plant samples (Fig. 11). The highest spectral variations in intensity were observed between 4500 and 7500 cm^{-1}. Next, the spectral deviation (SD) in this range for each group of spectra is calculated by using Eq. (11), were σ^2 represents the variance per wavelength and N denotes the number of spectral variables. Subsequently, the highest spectral deviation (SD = 0.0164) is selected to add normally distributed noise to the external test set spectra (plant samples) at a level of 1, 2, 3, 6 and 10 times the calculated highest SD

$$\mathrm{SD} = \sqrt{\frac{\sum \sigma^2}{N}} \qquad (11)$$

To compare the robustness of SVR and PLS for noisy data, the concentrations of the altered plant sample spectra were predicted by the regression model build for the original laboratory sample spectra. The results are summarized in Figs. 12 and 13. The prediction of original and noisy spectra using SVR regression is a factor two better as compared to PLS. Although both regression techniques are sensitive to noise (indicated by the ascending profiles), the prediction error of PLS grows faster with the noise level than SVR. This result supports our previous observation that SVR is more robust and less sensitive

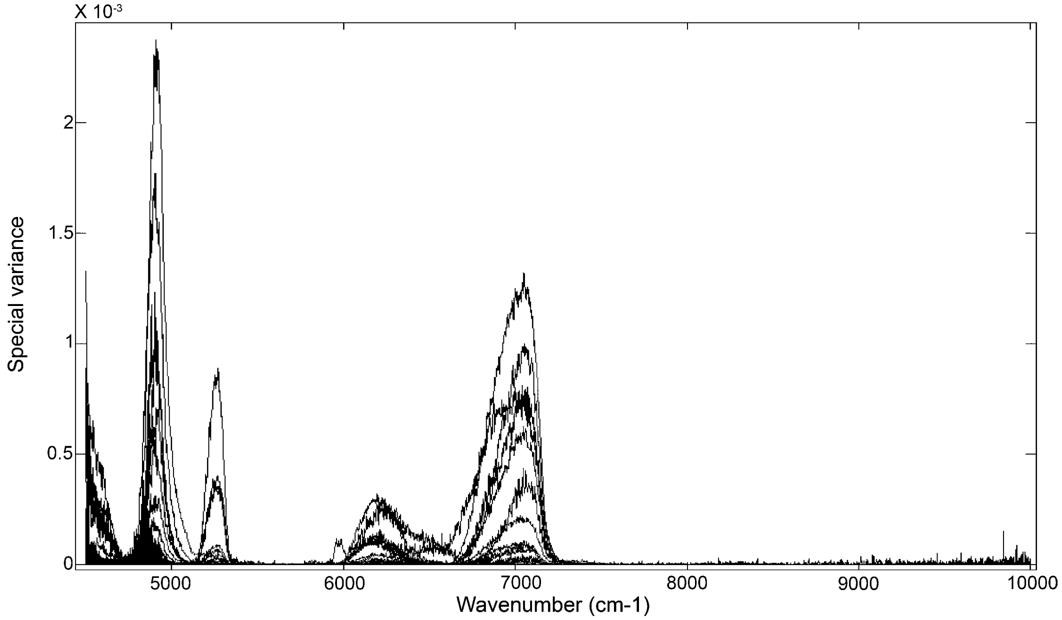
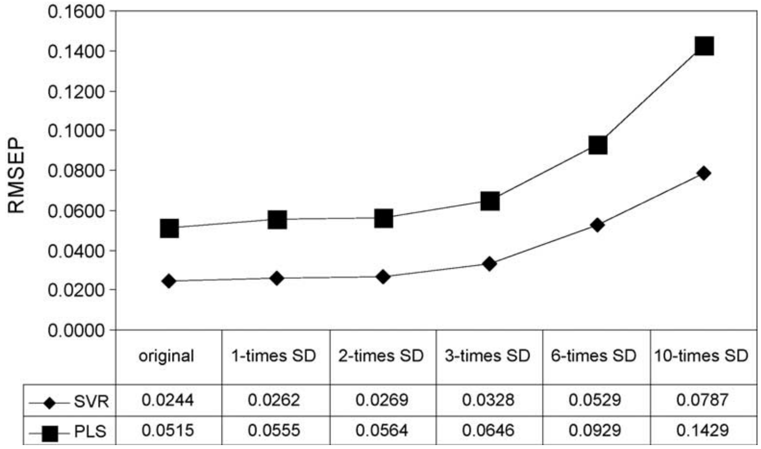
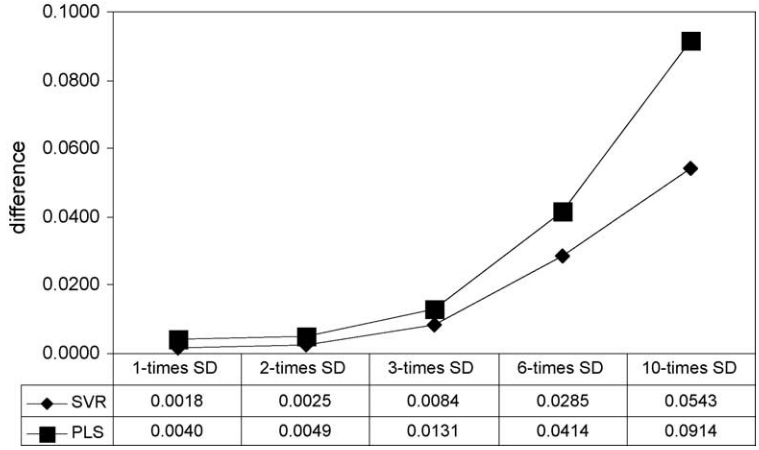


Fig. 11. Spectral variance of sample spectra of comparable water percentages.



	original	1-times SD	2-times SD	3-times SD	6-times SD	10-times SD
◆ SVR	0.0244	0.0262	0.0269	0.0328	0.0529	0.0787
■ PLS	0.0515	0.0555	0.0564	0.0646	0.0929	0.1429

Fig. 12. Prediction results of SVR (diamonds) and PLS (squares) at different noise levels.



	1-times SD	2-times SD	3-times SD	6-times SD	10-times SD
◆ SVR	0.0018	0.0025	0.0084	0.0285	0.0543
■ PLS	0.0040	0.0049	0.0131	0.0414	0.0914

Fig. 13. Absolute differences between the prediction results of SVR (diamonds) and PLS (squares) at different noise levels as compared to the predictions results for the original noise free data set.

to spectral variations as compared to PLS. This is probably due to the regulating effect of the ε-insensitive loss function mechanism build-in in the SVR algorithm. The ε-insensitive loss function is able to cope with the presence of noise in the data, thereby reducing or even eliminating the effect of it when the regression model is constructed.

5. Conclusion

In this paper a fully automated parameter optimization procedure is presented for estimating the optimal SVR parameter setting for a particular data set. It has been demonstrated that finding the optimal parameters is a pre-requisite, in case a SVR model is adopted for solving a given regression problem. Two well-described data sets and a new real-world chemical application in the field of NIR spectroscopy were considered as test cases.

The proposed cascade of genetic algorithms and simplex optimization yielded in all cases a set of parameters which resulted in accurate and robust SVR regression models which exhibited high levels of performance and generalization ability. Especially the latter (no over-fitting) was guaranteed by invoking an internal cross-validation procedure during the optimization process of the SVR model parameters.

The GA/simplex optimization procedure was developed as a relatively fast alternative for the time consuming (exhaustive) grid search approach, which usually is applied in optimizing the settings of a particular SVR model. An important advantage of the GA/simplex optimization approach compared to grid search is its accuracy. The accuracy of grid search depends on the selected parameter step size (parameter resolution), while the GA/simplex optimization does not depend on the step size. First, it can be concluded, in the context of the used data sets, that a well-optimized SVR model outperforms the commonly applied partial least squares regression technique. Benchmarks based on the RMSE measure demonstrated that considerable drops in prediction errors could be achieved. Even, a comparison with previously constructed SVR models taken from literature, which were optimized by means of a standard grid search, showed that the fine-grained search character of the GA/simplex procedure yielded deviating parameter settings which, in turn, led to an improved performance of the SVR regression model, as presented in this paper.

Although in many cases a rough search space (i.e., the error surface which had to be explored) was manifest, the GA/simplex procedure yielded SVR parameter settings which were very robust, as was indicated by the relatively small variations in model performance in presence of considerable perturbations of the three optimal SVR parameters (data sets 1 and 2). In conjunction, also the SVR model itself was robust. By adding high levels of normally distributed noise to the spectra in the test set (data set 3), the output of the regression model deviated just slightly as compared to the results for the noise-free test set. Moreover, it was demonstrated that PLS was much more affected in its performance if the same additive noise was imposed to the spectra in the test set.

One drawback of a SVR model, i.e., the lack of physicochemical interpretation possibilities as compared to the more transparent PLS regression technique, definitely needs to be tackled. This will be the subject of forthcoming investigations. Opposed to this, there are still various advantages. SVR is able to solve ill-posed problems, it does not require (apart from standard procedures like baseline correction and normalization) complex data pre-processing (like orthogonal signal correction) or feature selection (inherently due to its dual Lagrange formalism) and, last but not least, it is less sensitive to the presence of outliers and noise in the data.

Summarizing, the described GA/simplex optimization procedure in combination with SVR regression serves as a powerful alternative for PLS in the domain of NIR spectroscopy. An accurate model performance coupled to a robust behaviour appears to be the key properties of such optimized SVR regression models.

References

[1] R.N. Fuedale, H. Tan, D. Brown, Piecewise orthogonal signal correction, Chemom. Intell. Lab. Syst. 63 (2002) 129–138.

[2] F. Wülfert, W.Th. Kok, A.K. Smilde, Influence of temperature on vibrational spectra and consequences for predictive ability of multivariate models, Anal. Chem. 70 (1998) 1761–1767.

[3] U. Thissen, M. Peppers, B. Üstün, W.J. Melssen, L.M.C. Buydens, Comparing support vector machines to PLS for spectral regression applications, Chemom. Intell. Lab. Syst. (2004).

[4] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, USA, 1995.

[5] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, 2002.

[6] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.

[7] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, USA, 1998.

[8] S.R. Gunn, Support Vector Machines for Classification and Regression, University of Southampton, Southampton, UK, 1997.

[9] A.J. Smola, B. Schölkopf, A Tutorial on Support Vector Regression, University of London, UK, 1998.

[10] M. Oudenhuijzen, General Electric Advanced Materials B.V., Bergen op Zoom, The Netherlands, 2004. Personal communication.

[11] A.I. Belousov, S.A. Verzakov, J. Van Frese, A flexible classification approach with optimal generalisation performance: support vector machines, Chemom. Intell. Lab. Syst. 64 (2002) 15–25.

[12] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.

[13] D.E. Goldberg, Genetic Algorithms in Search, Optimisation and Machine Learning, Addison-Wesley, New York, USA, 1989.

[14] R. Wehrens, L.M.C. Buydens, Evolutionary optimisation: a tutorial, Trends Anal. Chem. 17 (1998) 193–203.

[15] Z. Michalewicz, Genetic Algorithms + Data Structures = Evaluation Programs, Springer, Berlin, 1996.

[16] A. Chipperfield, P. Flemming, H. Pohlheim, C. Fonseca, Genetic Algorithm Toolbox for Use with MATLAB, University of Sheffield, Sheffield, UK, 1994.

[17] M. Bessaou, P. Siarry, A genetic algorithm with real-value coding to optimize multimodal continuous functions, Struct. Multidisc. Optim. 23 (2001) 63–74.

[18] M. Su, H. Chang, Application of neural networks incorporated with real-valued genetic algorithms in knowledge acquisition, Fuzzy Sets Syst. 112 (2000) 85–97.

[19] J.A. Nelder, R. Mead, A simplex method for function minimization, Comput. J. 7 (1964) 308–313.

[20] R. Fletcher, Optimization, Academic Press Inc., London, 1969.

[21] W. Wang, Z. Xu, W. Lu, X. Zhang, Determination of the spread parameter in the Gaussian kernel for classification and regression, Neurocomputing 55 (2003) 643–663.

[22] B. Walczack, D.L. Massart, Local modelling with radial basis function networks, Chemom. Intell. Lab. Syst. 51 (2000) 219–238.

[23] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, 1998.

[24] F. Wülfert, W.Th. Kok, O.E. De Noord, A.K. Smilde, Linear techniques to correct for temperature-induced spectral variation in multivariate calibration, Chemom. Intell. Lab. Syst. 51 (2000) 189–200.

[25] F. Wülfert, W.Th. Kok, O.E. De Noord, A.K. Smilde, Correction of temperature-induced spectral variation by continuous piecewise direct standardization, Anal. Chem. 72 (2000) 1639–1644.

[26] H. Sweringa, F. Wülfert, O.E. De Noord, A.P. De Weijer, A.K. Smilde, L.M.C. Buydens, Development of robust calibration models in near infra-red spectrometric applications, Anal. Chim. Acta 411 (2000) 121–135.

[27] S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst. 44 (1998).

[28] B.M. Wise, N.B. Gallagher. http://www.eigenvector.com/MATLAB/OSC.html.

[29] T. Fearn, On orthogonal signal correction, Chemom. Intell. Lab. Syst. 50 (2000) 42–52.