Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes

F. Chauchard, R. Cogdill, S. Roussel, J.M. Roger, Véronique Bellon Maurel

Application of LS-SVM to non-linear phenomena in NIR spectroscopy : Development of a robust and portable sensor for acidity prediction in grapes

F. Chauchard(1), R. Cogdill(2) ,S. Roussel,
J.M. Roger(1) and V. Bellon-Maurel(1)

1- Information and Technologies for Agro-processes - Cemagref BP 5095, 34033 Montpellier Cedex 1, France
2 -Duquesne Unviversity, Graduate School of Pharmaceutical Sciences, 410 Mellon Hall, 600 Forbes Avenue, Pittsburgh, PA 15282, USA
3- Ondalys , Cemagref BP 5095, 34033 Montpellier, www.ondalys.com

Abstract

Nowadays, near infrared (NIR)technology is being transferred from the laboratory to the industrial world for on-line and portable applications. As a result, new issues are arising, such as the need for increased robustness, or the ability to compensate for non-linearities in the calibration or instrument. Semi-parametric modeling has been suggested as a means for adapting to these complications. In this article, Least-Squared Support Vector Machine (LS-SVM) regression, a semi-parametric modeling technique, is used to predict the acidity of three different grape varieties using NIR spectra. The performance and robustness of LS-SVM regression are compared to Partial Least Square Regression (PLSR) and Multivariate Linear Regression (MLR). LS-SVM regression produces more accurate prediction. However SNV pretreatment is required to improve the model robustness.

NIR Spectroscopy Robust calibration LS-SVM PLSR MLR Grapes tartaric and malic acidity.

1 Introduction

Near Infrared (NIR) spectroscopy provides non-destructive measurement of many chemical compounds in heterogeneous products [1] and has proved its efficiency for laboratory applications. To go further, many current studies are focused on NIR sensor design for on line [2] or portable operation [3]. However many constrains are placed upon the design of an efficient portable NIR sensor : the sensor must have a small size, be low cost, and deliver robust performance. Different approaches have been followed. Micro-spectrometers sometimes suffer from poor performance compared to conventional spectrometers but are perfectly suited for use with fiber optics[4]. Therefore some recent studies have illustrated the use of micro-spectrometers for portable NIR applications [5] [6]. Yet, these devices have been built only with the size-constraint in mind. Rather than relying on miniature holographic gratings, or bandpass filters another possibility is to use monochromatic light sources, e.g. LED or Laser diodes, selected for the measurement of specific chemical components, coupled with a silicon photodiode detector [7]. As far as portable spectrometer are concerned, this alternative

is ideal because of its combination of small size, low cost and good robustness. Beyond sensor design, a growing amount of research is being focused on methods of developing robust calibration models which are less disturbed by the challenges of portable applications. Partial Least Square Regression (PLSR) is the most commonly used method for prediction using numerous correlated variables (such as NIR spectra). However, when a small number of wavelengths is used for the NIR analysis (monochromatic light source for instance) Multiple Linear Regression (MLR) is adequate. Factors such as experimental conditions (e.g. temperature, external light), instrument variations (lamp aging, sensor sensitivity) and analyte characteristics (matrix change) induce non-linearities in the spectra. In these cases non-linear methods may provide a more optimal solution than classical PLSR and MLR [8]. Support Vector Machines (SVM) is one of these new and attractive methodology [9]. SVM have already been used in various fields such as diagnosis ovarian tumor malignancy prediction [10], image classification [11] [12] and spam categorization [13]. Only recently has SVM technology been applied to chemometric issues as a non-linear discrimination [14] [15] and quantitative predictions [16]. An alternate formulation of SVM strategy for regression problems is the Least-Square Support-Vector Machine (LS-SVM)[17].
This paper aims at studying the capabilities of LS-SVM to derive accurate and robust calibration models for the prediction of total acidity of fresh grapes from NIR data gathered using a portable sensor. The performance of LS-SVM regression will be compared to PLSR and MLR in terms of accuracy of prediction, and model robustness.

2 Theory

2.1 Notation

Bold, upper-case characters will be used for matrices, e.g $\mathbf{X}$; bold lower-case characters for column vectors, e.g $\mathbf{x}_i$ will denote the i^{th} column of $\mathbf{X}$; row vectors will be denoted by the transpose notation, e.g $\mathbf{x}_j^T$ will denote the j^{th} row of $\mathbf{X}$; non bold characters will be used for scalars, e.g matrix elements x_{ij} or indices i. $\mathbf{X}$ will represent a $[n \times p]$ matrix containing the p spectral responses of the n samples.

2.2 Regression methods

Since MLR [18] and PLSR [19] are both well known methods for multivariate linear regression, the theory of these methods will not be presented explicitly herein. Linear least squares models attempt to correlate the spectrum $\mathbf{x}_i^T$, and reference value y_i, of all the samples in $\mathbf{X}$. The predictions $\hat{\mathbf{y}}$ are computed by the following equation :

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{b}_0 \tag{1}$$

Where $\hat{\boldsymbol{\beta}}$ is a $[p \times 1]$ vector of regression coefficients ; and $\hat{b}_0$ is the model offset.

2.3 SVM Regression

The roots of SVM date back to the discrimination work of Vapnik and Lener [20]. However the general non-linear version of SVM is quite recent [21]. In 1998 Vapnik extended the theory to non-linear regression using the SVM framework [9].

SVM regression is based on a kernel substitution, where $\mathbf{X}$ $[n \times p]$ is replaced by a $[n \times n]$ kernel matrix $\mathbf{K}$ and $\hat{\boldsymbol{\beta}}$ $[p \times 1]$ is replaced by $\hat{\mathbf{b}}$ $[n \times 1]$ vector. In order to model non-linear processes, the Gaussian radial basis function (RBF) kernel has been chosen. $\mathbf{K}$ is then defined as :

$$\mathbf{K} = \begin{pmatrix} k_{1,1} & \dots & k_{1,n} \\ \vdots & \ddots & \vdots \\ k_{n,1} & \dots & k_{n,n} \end{pmatrix} \tag{2}$$

Where $k_{i,j}$ is defined by the RBF function :

$$k_{i,j} = e^{\frac{-\|\mathbf{x_i^T} - \mathbf{x_j^T}\|^2}{\sigma^2}} \tag{3}$$

According to equation 3, more similar samples will produce RBF output near 1, while less similar sample provide output near 0. Hence $k_{i,j}$ is conceptually a non-linear measure of similarities between two samples, thus $\mathbf{K}$ can be thought of as a sort of sample-sample correlation matrix. The kernel width parameter, σ, is related to the confidence in the data, or SNR ; adjustment of σ also influences the non-linear nature of the regression. As σ increases, the kernel becomes wider, forcing the model toward a less complex (more linear) solution.

The calculation of the regression vector $\hat{\mathbf{b}}$, here a $[n \times 1]$ vector, follows a different objective than the classical PLSR or MLR models. Rather than trying to minimize the prediction error only, the SVM objective function is augmented with the root mean square (rms) magnitude of the regression vector $\hat{\mathbf{b}}$, which represents the model complexity. The classical objective function for PLSR and MLR is :

$$min(\mathbf{e}) = min(\sum_{i=1}^{n}(\mathbf{y} - \hat{\mathbf{y}})^2)$$

It is replaced by a so-called primal-dual form :

$$min(\mathbf{e}) = min\Big(\frac{\sum_{i=1}^{n}\xi_i}{2} + \Gamma\frac{\sum(\hat{\mathbf{b}}^T\hat{\mathbf{b}})}{2}\Big) \tag{4}$$

Where Γ is a regularization parameter, such that increasing Γ places a greater significance on reducing the rms magnitude of the model coefficients. ξ_i is called the e-insensitive error of generalization. It replaces the classical least square criterion $(y_i - \hat{y}_i)^2$ and is defined using the significance threshold ϵ :

$$\xi_i = \begin{cases} 0, \text{ if } |y_i - \hat{y}_i| < \epsilon \\ |y_i - \hat{y}_i| - \epsilon, \text{ otherwise} \end{cases} \tag{5}$$

Consequently, any single residual error of magnitude less than ϵ is set to zero. This assumes that any error lower than ϵ is uncertain (insignificant), and to fit below ϵ would likely produce an over-fitted solution. On the other hand, when the residual is larger than ϵ, the absolute value of the error is summed rather than the squared error ; this tends to limit the influence of outliers during model training. The defined objective function modifies the approach of model training. All calibration samples with residual error lower than ϵ are given a $\hat{b}_i$ coefficient equal to zero, which means that this sample is redundant, and can be easily predicted by the other ones. Those samples whose b-coefficients are nonzero are referred to as support vectors. This process is conceptually analogous to thresholding the coefficients of a PLSR model for automatic variable selection. Also as a consequence of the inequality constraints imposed on the model, the $\hat{\mathbf{b}}$ solution cannot be obtained directly by solving a linear system. To do so, the model is optimized in the space of Lagrangian multipliers by using quadratic programming. This optimization is slower than least square methods, but it is still a convex, deterministic process, and is guaranteed to converge to a single global minimum.

2.4 LS-SVM

LS-SVMs are an alternate formulation of SVM regression proposed by Suykens[17]. The objective function (equation n^o 4) is the same as for SVMs, but the e-insensitive loss function is replaced by the classical squared loss function. The benefits of automatic sparseness are lost (all b_i coefficients will be non zero) but the model can be trained much more efficiently after constructing the Lagrangian by solving the linear Karush-Kuhn-Tucker (KKT) system :

$$\begin{bmatrix} 0 & \mathbf{l_n^T} \\ \mathbf{l_n} & \mathbf{K} + \frac{\mathbf{I}}{\gamma} \end{bmatrix} \begin{bmatrix} \hat{b}_0 \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{6}$$

Where $\mathbf{I}$ refers to an $[n \times n]$ identity matrix, $\mathbf{l}_n$is a $[n \times 1]$ vector of ones, $\frac{1}{\gamma}$ correspond to Γ and $\mathbf{y}$ is the vector of reference values. The solution of equation 6 can be found using most standard methods of solving sets of linear equations, such as conjugate gradient descent. While SVM implementation requires the tuning of three parameters $(\gamma, \sigma, \epsilon)$, the implementation of LS-SVM requires the specification of only two parameters (γ, σ). A disadvantage of the both methods is that training time increases with the square of the number of training samples and linearly with the number of variables (dimension of spectra) which is the opposite of classical least squares methods.

3 Material and methods

3.1 Material

During data collection for this study, NIR scans were collected for 371 grape berries belonging to three different varieties : carignan(188), mourverdre(84) and ugniblanc(99). Each spectrum is the mean of 5 sub-scans, collected in transmission trough the berry. The total acidity (TA = malic + tartaric acid concentrations) was measured by the CTIFL[1] via HPLC assay. The NIR spectrometer utilized was a Zeiss MMS1 polychromatic diode array spectrometer, sensitive in the 300nm-1160nm range, with a 3.3nm sampling interval, which provided 256-point spectra. The light source consisted of five, 14V, 50mA, 3 Lumens micro lamps powered by a 12VDC supply.

3.2 Methods

For all of the regression methods tested in this study, the spectral range was restricted to the 680-1100nm short wave NIR window, which included the chlorophyll peak in the visible region (685nm). The 830-920nm spectral window, corresponding to the sugar absorbance peaks [22], was removed to avoid the concomitant influence of the major chemical constituent in grapes. For portable application test, only a few wavelengths were selected.

3.2.1 PLSR calibration

In order to optimize the PLSR model, wavelength selection was performed based on the BQ method (Backward Q^2_{cum}) [23]. This method was observed to be the most suitable feature selection procedure among 20 methods compared empirically [24]. The Q^2_{cum} fitness criterion is defined as :

$$Q^2_{cum} = 1 - \Pi^k_{j=1} \frac{PRESS_j}{RSS_{j-1}}$$

where k is the number of latent variable, calculated in leave-one-out cross-validation, in the PLSR model and

$$PRESS = \Sigma^n_{i=1}(y_i - \hat{y}_{-i})^2$$

$$RSS = \Sigma^n_{i=1}(y_i - \hat{y}_i)^2$$

$\hat{y}_{-i}$ is defined as the prediction of y_i when y_i is removed from the data before constructing the model. $\hat{y}_i$ is defined as the prediction of y_i when y_i is included in the calibration data. At each step, the variable with the smallest regression coefficient (in terms of absolute value) is eliminated and the Q^2_{cum} value is calculated. The variable subset selection showing the highest Q^2_{cum} is retained. The number of factors retained in the final PLSR model was chosen using a leave-one-out cross validation procedure.

3.2.2 MLR calibration

To simulate the situation of a portable NIR sensor, in which only few wavelengths are available, wavelength selection was performed using a forward-backward stepwise MLR procedure and cross-validation. The selection process

[1] Fruit and Vegetable Interprofessional Technic Center, St Remy de provence, FRANCE

was stopped when the number of wavelengths was equal to the number of latent variables of the PLSR.

3.2.3 LS-SVM calibration

The LS-SVM regression was trained using two method. First, to compare LS-SVM with PLSR, the LS-SVM model was derived using scores derived from the PLSR model factors. The number of factors included in the model was chosen using cross-validation. This model is referred to : LS-SVM^{lv}. Second, an LS-SVM model (LS-SVM^{sv}) was derived using the variables selected by the stepwise MLR (portable sensor framework). In each case, the tuning of γ and σ parameters was performed using cross validation.

3.2.4 Model performance evaluation

Prior to model development, the dataset was split into training and test sets using the 'Venetian blinds' method [25] according to total acidity (TA). The training set S^0 and the test set S^1 had the same TA distribution ($\mathbf{y}_0$ and $\mathbf{y}_1$), each containing 186 and 185 samples, respectively.
For each model the Standard Error of Calibration (SEC), the Standard Error of leave-one-out Cross Validation (SECV) and the coefficient of determination in cross validation (R^2_{cv}) were calculated using S^0. The Standard Error of Prediction (SEP), R^2_{test}, the Bias, the bias-corrected SEP (SEPC) were computed on the test set S^1.
In order to assess the relative robustness of each methods, seven noisy test sets were generated by modifying S^1 spectra following a procedure described in [26]. The simulated noises were : Gaussian Noise, Multiplicative Noise, Baseline Shift, Baseline Slope, Wavelength Shift, Stretch/Shrink, Bandwidth perturbation. For all these data sets, the SEP and SEPC were calculated. In order to evaluate the real LS-SVM interest for a portable sensor, the performances of Standard normal variate transformation (SNV)([27]) corrected model was also calculated.

3.2.5 Chemometrics software

All calculations were performed using MATLAB 6.0 (The MathWorks, Inc., Natick, USA), and the *PLS_toolbox* from Eigenvector Research, Inc. (Manson ,USA). The free LS-SVM toolbox (LS-SVM v1.4 [2], Suykens, Leuven, Belgium) was used with MATLAB to derive all of the LS-SVM models.

[2] www.esat.kuleuven.ac.be/sista/lssvmlab/

4 Results

4.1 Model tuning

During PLSR optimization, the BQ method selected 68 variables from the spectra; from these variables, eight factors were derived for the PLSR model, and 10 factors for LS-SVM^{lv} (figure 1) calibration. The optimal LS-SVM tuning parameters were found to be $\gamma = 10$ and $\sigma^2 = 50$. During optimization of the LS-SVM parameters, it was observed that the LS-SVM must be tuned very cautiously. Figure 2 shows the SEC and SECV values depending on γ and σ values. The tendency of SEC to be much lower than SECV for LS-SVM regression suggests a greater tendency to overf-fit.

To match the number of factors used during PLSR modeling, the stepwise MLR procedure was stopped after 8 variables were selected; the MLR and LS-SVM^{sv} calibrations were derived with these 8 wavelengths. As is shown in figure 3, the tuning of this model produces a different result : the surface tends to a minimum value as γ increases toward infinity. Increasing γ to such a magnitude, however would likely lead to an over-fitted calibration. Thus we chose γ value where SECV is almost constant (SECV variation under 0.5%), here, $\gamma = 4000$ and $\sigma^2 = 40$.

The large difference in γ values between the two LS-SVM models has to be explained. For LS-SVM^{lv}, the data compression based on latent variables acts as a pre-processing filter : the PLSR factors were calculated to maximize the correlation between spectra and acidity which provides a kernel matrix $\mathbf{K}$ depending mainly on the acidity concentration (the measurement noise is suppressed, while spectra-reference covariance is maximized). LS-SVM^{sv} is applied with no variable pre-processing, therefore each variable contains not only linear and/or non-linear acidity information, but also measurement noise as well as other chemical information. As a consequence, $\mathbf{K}$ contains sample/sample correlation near 0 and, the amplitude of the regression coefficients is increased (higher γ) to compensate this phenomena.

For both models, σ^2 has approximatively the same value which means that a consistent degree of non linearity is modeled. σ^2 is quite low, suggesting that the relationships between the spectra and the total acidity of grapes is highly non-linear.

4.2 Model accuracy

The performances of LS-SVM were found to be better than the classical linear methods (table 1). LS-SVM^{lv} produces the best SEP of $1.03 g.l^{-1}$, compared to $1.28 g.l^{-1}$ for PLSR. Similarly, LS-SVM^{sv} greatly improved the prediction accuracy compared to MLR. Table 1 draw a further comment when regarding SEC and SECV : for the two LS-SVM models the gap between these two criteria was shown wider than for PLSR and MLR, which suggests a higher likelihood of over-fit with LS-SVM. This enlights γ importance which permits to find the good balance between over-fitting and non-linear modeling.

The calibration plots (table 2) for the PLSR and MLR prediction demonstrate the presence of non-linearities. The high acidity samples are not well predicted. This could be due to changes in chemical interactions or in the fruit matrix (unripe fruits with dense structure). Both LS-SVM models took into account this non-linearity. LS-SVM^{lv} corrected the non linearities for the high acidity but also improved the low prediction performance level. LS-SVM^{sv} succeeded in predicting the high nonlinearities with only 8 selected wavelengths. However in this case, the dispersion is more scattered in the middle range of acidity. On the contrary, the prediction values in the low acidity were more compact.

4.3 Model robustness

Table 3 contains the results of the robustness tests for each of the four methods. The solid line at $1\,g.l^{-1}$ represents a satisfactory predictive model (Standard deviation(Y_1)/SEP $= 2.72$) while the other line at $1.5\,g.l^{-1}$ represents the performance of the worst model (MLR : Standard deviation(Y_1)/SEP $= 1.81$). PLSR appeared sensitive to Baseline slope, Wavelength Shift and Stretch/Shrink and to a smaller extebd to Baseline Shift. In these cases, the model was unusable. Nevertheless, since the SEPC remained under $1.5\,g.l^{-1}$, it may be possible to correct the prediction with a simple bias.

As far as SEP is concerned, LS-SVM^{lv} is more robust than PLSR be for all the noises, the sum of SEP^2 is much lower. However it is sensitive to four type of noise : Gaussian noise, Baseline slope, Wavelength Shift and Stretch/Shrink. However unlike the PLSR, the LS-SVM SEPC is above the $1.5\,g.l^{-1}$ circle and follows the SEP tendency. Hence, it is not possible to correct the model since the error is no longer a simple bias.

MLR was very sensitive to only Gaussian noise and was shown to be the most robust model, but it was also the most limited in performance since the model was barely usable. Due to this stability, there is practically no difference between the SEP an SEPC.

LS-SVM^{sv} is very sensitive to 4 noises : Gaussian noise, Multiplicative noise, Baseline Shift, Baseline slope (SEP $= 14\,g.l^{-1}$!), which exceeded the significance circles, precluding the correction of the model error via bias removal. LS-SVM^{lv} appeared more robust than LS-SVM^{sv} because it was derived using PLSR factors which acts as a data pre-processing filter ; hence, the method aggregates the advantages of PLSR and LS-SVM regression. The LS-SVM model behavior (SEPC which follows SEP trend) can be explained by the sensitivity of RBF functions to multiplicative and additive noise, since the RBF kernel is calculated as a function of the difference between each training spectrum.

The SNV preprocessing improved the robustness of all models (table 4), in particular for baseline-shift and multiplicative noise (by nature since $x_{SNV} = \frac{x-\bar{x}}{standard\,deviation}$) but also for baseline slope. The models based on single wavelengths (MLR, LS-SVM^{sv} with or without SNV) are very sensitive to gaussian noise, whereas the latent variable (PLSR factors) computation partially removes the noise in PLSR and LS-SVM^{lv}. Thus, except for the gaussian noise, $SNV +$ LS-SVM^{sv} is the most robust model.

5 Conclusion

In this paper, we compared classical linear regression techniques and LS-SVM regression for the prediction of total acidity in fresh grapes using NIR spectroscopy. The LS-SVM models were implemented using both PLSR factors (LS-SVM^{lv}), and selected variables from the raw spectrum as input (LS-SVM^{sv}). LS-SVM was shown to increase prediction performance by correcting the non-linearities which limited the performance of the classical linear methods. The most accurate calibration was the combination of LS-SVM, SNV preprocessing and PLSR latent variables.
LS-SVM^{lv} was also found to be more robust than PLSR or LS-SVM^{sv}. However, though it was the least accurate method, MLR was the most robust among the methods tested. Both latent variable compression and SNV preprocessing proved to be very important for LS-SVM performance. The combination of LS-SVM and SNV on selected wavelengths appeared to be as robust as MLR model. Based on these results, LS-SVM regression seems to be an interesting tool for chemometrics in the area of quantitative prediction, and a valuable solution for portable sensor applications. Further research efforts will focused on including LS-SVM regression during the preprocessing optimization and variable selection.

6 Acknowledgments

The authors gratefully acknowledge Prof. Dr. J. Suykens for the LS-SVM toolbox and helpful comments. They also wish to thank Dr. Crochon and Jean-Michel Deseigne (ITV) for the grape data and CTIFL for acidity analyse, sponsored by ACTA.

References

[1] B. Osborne, T. Fearn, Near infrared spectroscopy in food analysis, John Wiley and Sons, N.Y., 1986.

[2] P. Fayolle, J.-M. Roger, V. Steinmetz, L. Dusserre-Bresson, V. Bellon, An on-line near infrared system to sort fruits according to sugar content, in: Sensoral 98 Colloque international sur les capteurs de la qualité des produits agro-alimentaires, ENSAM-Cemagref-INRA, Montpellier, France, 1998, pp. 533–541.

[3] T.Hyvarinen, E. Herrala, J. Malinen, P. Niemla, Nir analysers can be miniature, rugged and handheld, in: K. Hildrum, T. Isaksson, T. Naes, A. Tandberg (Eds.), Near Infrared Spectroscopy, Bridging the Gap Between Data Analysis and NIR Applications, Ellis Horwood, london, 1992.

[4] A. Davies, H. Heise, P. Lampen, R. Kurte, L. Kupper, Wavelength selection and probe design for the customisation of micro-spectrometers, Spectroscopy Europe 13/3 (2001) 22–26.

[5] T. Temma, K. Hanamatsu, F. Shinoki, Development of a portable near infrared sugar-measuring instrument, J. Near Infrared Spectrosc. 10 (2002) 77–83.

[6] S. Saranwong, J. Sornsrivichai, S. Kawano, Performance of a portable near infrared instrument for brix value determination of intact mango fruit, J. Near Infrared Spectrosc 11 (2003) 175–181.

[7] J. Malinen, M. Kansakoski, R. Rikola, C. G. Eddison, Led-based nir spectrometer module for hand-held and process analyser applications, Sensors and Actuators B: Chemical 51 (1-3) (1998) 220–226.

[8] B. Walczak, D. L. Massart, The radial basis functions – partial least squares approach as a flexible non-linear regression technique, Analytica Chimica Acta 331 (3) (1996) 177–185.

[9] J. A. K. Suykens, J. Vandewalle (Eds.), Nonlinear Modeling : advanced black-box techniques, Kluwer Academic Publishers, Boston, 1998.

[10] C. Lu, T. Van Gestel, J. A. K. Suykens, S. Van Huffel, I. Vergote, D. Timmerman, Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines, Artificial Intelligence in Medicine 28 (3) (2003) 281–306.

[11] O. Chapelle, P. Haffner, N. Vapnik Vladimir, Support vector machines for histogram-based image classification, IEEE Transactions on Neural Networks 10 (5) (1999) 1055–1064.

[12] G. Guo, S. Z. Li, K. L. Chan, Support vector machines for face recognition, Image and Vision Computing 19 (9-10) (2001) 631–638.

[13] H. Drucker, D. Wu, N. Vapnik Vladimir, Support vector machines for spam categorization, IEEE Transactions on Neural Networks 10 (5) (1999) 1048–1054.

[14] A. Belousov, S. Verzakov, J. von Frese, Applicational aspect of support vector machine, J. of Chemometrics 16 (2002) 482–489.

[15] R. Goodacre, Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules, Vibrational Spectroscopy 32 (1) (2003) 33–45.

[16] R. P. Cogdill, P. Dardenne, Least-squares support vector machines for chemometrics: An introduction and evaluation, J. Near Infrared Spectrosc. In press - (-) (2003) –.

[17] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific Publishing Co., Pte, Ltd., Singapore, 2002.

[18] H. Martens, T. Naes, Multivariate Calibration, John Wiley and Sons, New York, 1989.

[19] S. Wold, M. Sjostrom, L. Eriksson, Pls-regression: A basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2) (2001) 109–130.

[20] V. Vapnik, A. Lerner, Pattern recognition using generalized portrait method, Automation and Remote Control 24 (1963) 774–780.

[21] V. Vapnik, The nature of Statistical Learning Theory, Springer Verlag, New Yorkn, 1995.

[22] N. S. Sanchez, S. Lurol, J. Roger, V. Bellon-Maurel, Robustness of models based on nir spectra for sugar content prediction in apples, Journal of Near Infrared Spectroscopy 11 (2) (2003) 97–107.

[23] A. Lazraq, R. Cleroux, J.-P. Gauchi, Selecting both latent and explanatory variables in the pls1 regression model, Chemometrics and Intelligent Laboratory Systems 66 (2) (2003) 117–126.

[24] J. Gauchi, P. Chagnon, Comparison of selection methods of explanatory variables in pls regression with application to manifacturing process data, Chemometrics and Intelligent Laboratory Systems 58 (2001) 171–193.

[25] R. D. Snee, Validation of regression models : methods and examples., Technometrics 19 (1977) 415–428.

[26] S. Roussel, C. R. Hurburgh, D. B. Funk, Noise robustness comparison of multivariate calibration models based on near infrared spectroscopy measurements, in: PITTCON Pittsburgh conference on analytical chemistry and applied spectroscopy, New Orleans, LA, USA, March 2002.

[27] R. Barnes, M. Dhanoa, J. Lister Susan, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Applied Spectroscopy 43 (5) (1989) 772–777.

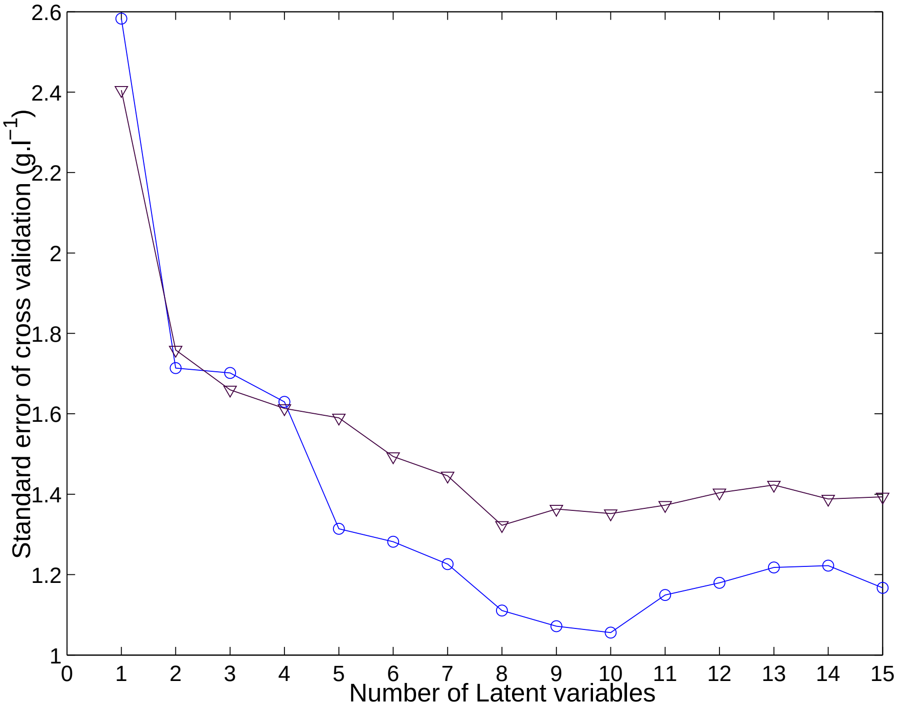


Figure 1: SECV for LS-SVM^{lv} (-○-) and PLSR (-∇-) models depending on number of latent variable. For LS-SVM, optimum γ and σ^2 were calculated at each step.

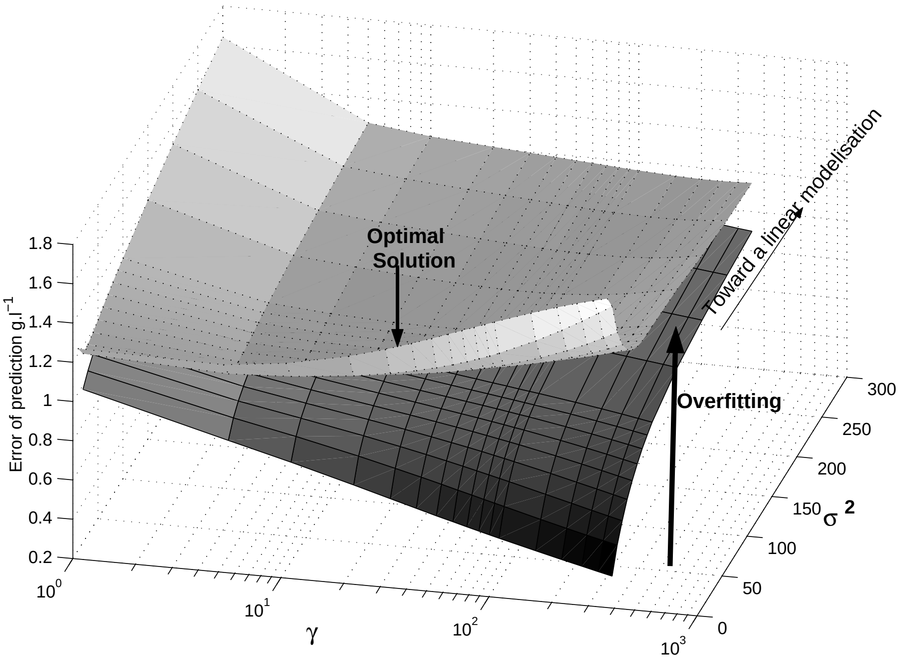


Figure 2: Tuning of γ and σ^2 for LS-SVM^{lv}. The above surface (dotted lines) represents the SECV and the beneath one represents the SEC. Note that γ is on logarithm scale.

	PLSR	LS-SVM^{lv}	MLR	LS-SVM^{sv}
R^2_{cv}	0.76	0.83	0.69	0.77
SECV ($g.l^{-1}$)	1.32	1.11	1.50	1.30
SEC ($g.l^{-1}$)	1.23	0.89	1.42	0.95
R^2_{test}	0.77	0.86	0.68	0.78
SEP ($g.l^{-1}$)	1.28	1.03	1.53	1.30
SEPC ($g.l^{-1}$)	1.28	1.03	1.53	1.30
Bias ($g.l^{-1}$)	-0.02	0.01	-0.11	0.03

Table 1: Performances of the four models for acidity prediction in fresh grape. Result in leave-one-out cross-validation and in test.

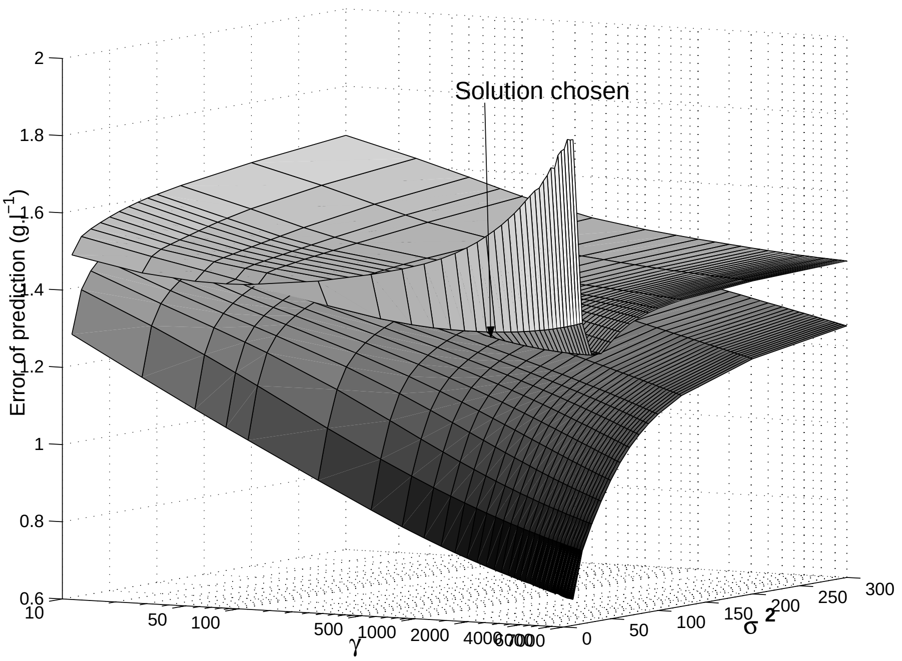
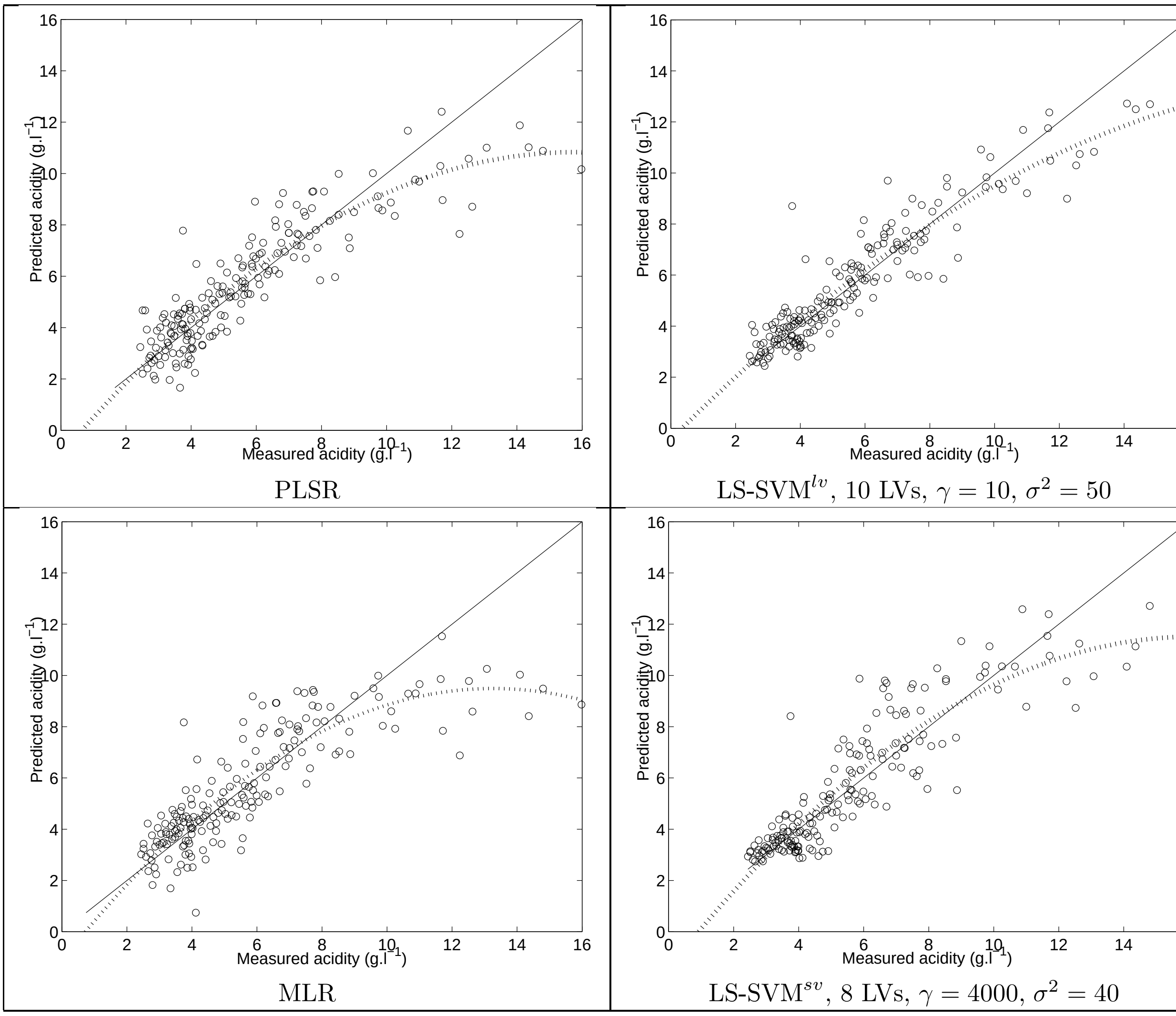


Figure 3: Tuning of γ and σ^2 for LS-SVM^{sv}. The above surface (dotted lines) represents the SECV and the beneath one represents the SEC. Note that γ is on logarithm scale.



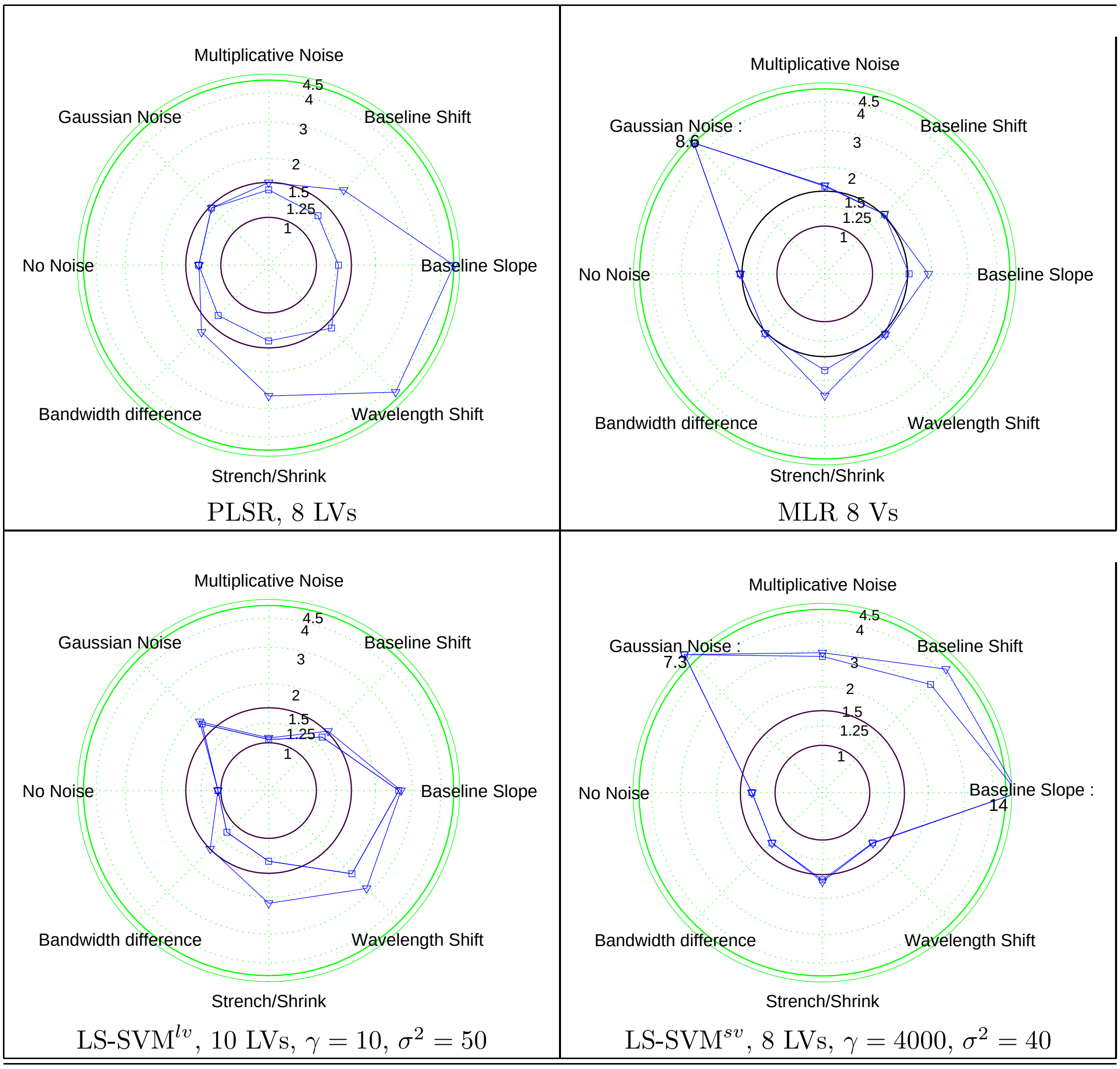
PLSR

LS-SVM^{lv}, 10 LVs, $\gamma = 10$, $\sigma^2 = 50$

MLR

LS-SVM^{sv}, 8 LVs, $\gamma = 4000$, $\sigma^2 = 40$

Table 2: Calibration plots : Measured acidity ($g.l^{-1}$) vs predicted acidity ($g.l^{-1}$) for the different models. -○- predicted values, ······ Quadratic tendency curve.



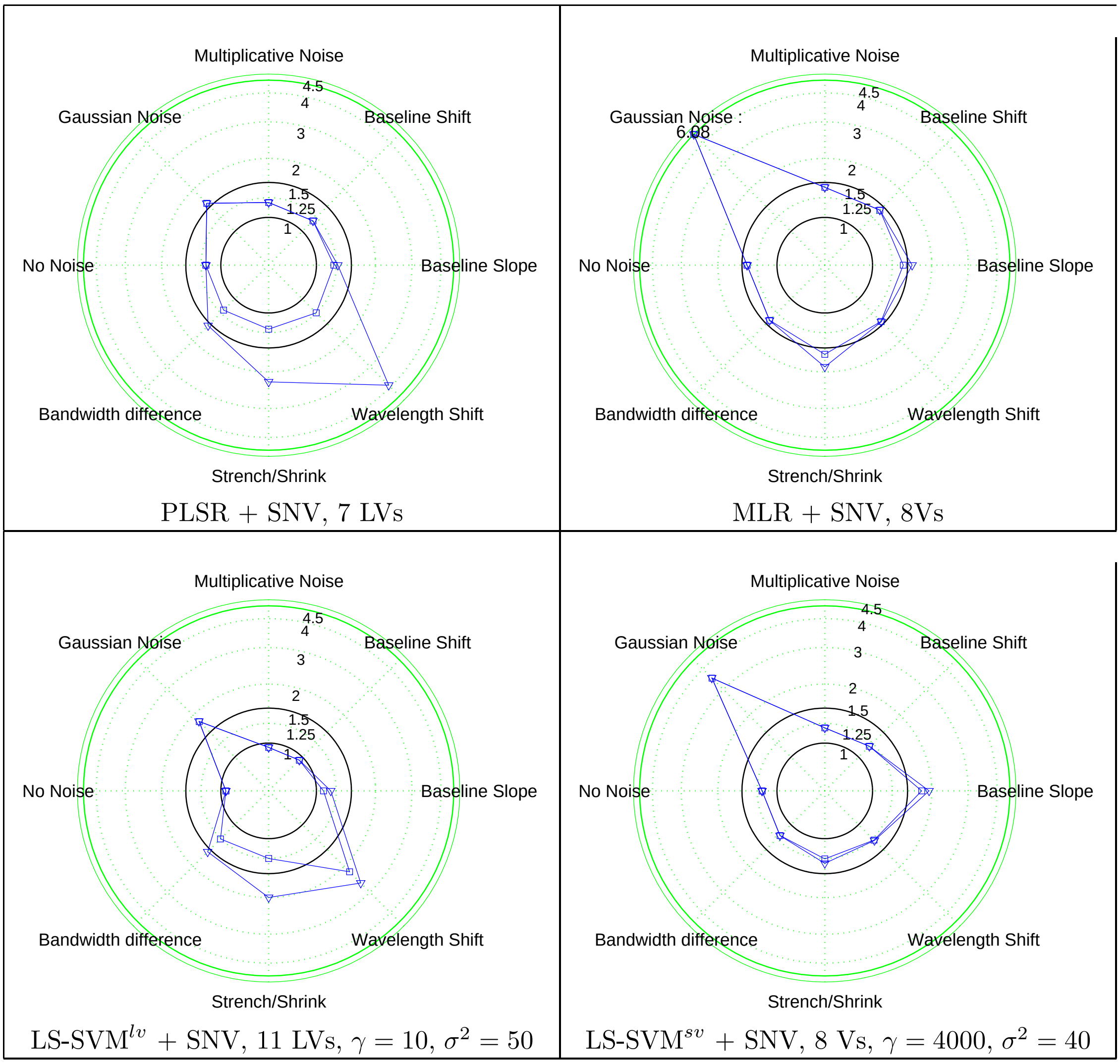
PLSR, 8 LVs

MLR 8 Vs

LS-SVM^{lv}, 10 LVs, $\gamma = 10$, $\sigma^2 = 50$

LS-SVM^{sv}, 8 LVs, $\gamma = 4000$, $\sigma^2 = 40$

Table 3: Prediction error($g.l^{-1}$) for the different models depending on the noise type -∇- SEP, and -□- bias-corrected SEP(SEPC). The solid line circle at $1g.l^{-1}$ shows the threshold for predictive model, and the one at $1.5g.l^{-1}$ indicates MLR performance (the worst SEP with no noise).



PLSR + SNV, 7 LVs

MLR + SNV, 8Vs

LS-SVM^{lv} + SNV, 11 LVs, $\gamma = 10$, $\sigma^2 = 50$

LS-SVM^{sv} + SNV, 8 Vs, $\gamma = 4000$, $\sigma^2 = 40$

Table 4: Prediction error$(g.l^{-1})$ for the different models depending on the noise type -∇- SEP, and -□- bias-corrected SEP(SEPC). The solid line circle at $1g.l^{-1}$ shows the threshold for predictive model, and the one at $1.5g.l^{-1}$ indicates MLR performance (the worst SEP with no noise).