# Kernel chemometrics: an introduction to support vector machines

**R.P. Cogdill[a] and P. Dardenne[b,*]**
[a]*U.S. Fulbright Laureate, Cemagref-ITAP, 361 rue J.-F. Breton, BP 5095, 34033 Montpellier, France, Cedex 1*

[b]*Centre de Recherches Agronomiques de Gembloux, Chaussée de Namur, B-5030 Gembloux, Belgium E-mail: dardenne@cragx.fgov.be*

## Introduction

Linear methods of least-squares regression[1] are used more often than any other for chemometric modelling of near-infrared (NIR) spectroscopic data. There are, however, some cases where a nonlinear model is clearly required to accurately fit the training data. Furthermore, experience has shown that even though a linear model may be adequate, the performance of some calibrations may be significantly improved with the use of a nonlinear model.

Support vector machines (SVM),[2-5] are a family of semi-parametric, nonlinear modelling techniques that are rapidly gaining practical application in a number of fields. Despite their theoretical and practical advantages, support vector machines have only recently been applied to chemometrics,[6-8] and only as a nonlinear classification scheme. The intent of this work is to present SVM regression in a way that is more familiar to the NIR spectroscopy community, and illustrate its performance in some practical applications. For a more detailed explanation of SVM theory, those interested should consult some of the many references available on kernel methods[4, 5] and support vector machines.

## Theory

At its core, SVM regression is quite similar to least squares regression. However, rather than seeking to minimize prediction error only, the SVM objective function has been augmented with terms to minimize the complexity (rms magnitude) of the coefficient vector, **b,** while searching for a solution with good predictive ability (1). The proportional influence of prediction error and model complexity on the objective function optimization is controlled by a regularization constant (γ). With these changes, the ordinary least squares objective function is replaced by the so-called primal-dual form:

$$\min[2^{-1}\Sigma(y - \hat{y})^2) + \gamma\Sigma(2^{-1}\mathbf{b}^T\mathbf{b})] \qquad (1)$$

Thus, as γ is increased, more emphasis during model training is placed on reducing the magnitude of the model coefficients. This is a concept familiar to both regularization training of artificial neural networks (ANN)[9] and linear ridge regression.[10] In this representation, **b** is the [*n* x 1] vector of model coefficients (assuming *n* variables, or wavelengths). At this point, SVM theory begins to deviate from traditional regression theory, by altering the loss function[4] (which will not be discussed here), and by optimizing in "sample-space", rather than "variable-space", by using kernel substitution[11,12].

Kernel substitution involves supplanting the [$m$ x $n$] matrix of spectra, **X**, with an [$m$ x $m$] kernel matrix, **K**, where each element describes the relationship between two calibration vectors. Generally, some form of Gaussian radial basis function (RBF) is used for the kernel function:

$$k_{ij} = e^{\left[\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\sigma^2}\right]} \qquad (2)$$

The choice of kernel function will determine the amount of nonlinearity that can be modelled during regression. For the RBF kernel, the degree of model nonlinearity can be adjusted by changing $\sigma^2$ (Figure 1). For the traditional SVM, optimization is performed in a space of Lagrangian multipliers, using quadratic programming.

Driven by the desire to make SVM regression as simple as possible (but no simpler), Suykens, *et. al.* proposed an alternate formulation of the SVM strategy called the least-squares support vector machine (LS-SVM),[4] which uses the traditional least-squares loss function. While the primal-dual form of the objective function is retained (1), the LS-SVM can be trained much more efficiently by solving a linear Karush-Kuhn-Tucker (KKT) system of the following form:

$$\left[\begin{array}{c|c} 0 & \mathbf{1_m}^{\mathrm{T}} \\ \hline \mathbf{1_m} & \mathbf{K} + \mathbf{I}/\gamma \end{array}\right] \left[\begin{array}{c} b_0 \\ \hline \mathbf{b} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{y} \end{array}\right] \qquad (3)$$

Where, **I** refers to an [$m$ x $m$] identity matrix, $\gamma$ is the regularization constant, $\mathbf{1_m}$ is a [$m$ x 1] vector of ones, **y** is the vector of reference values, **b** is the [$m$ x 1] vector of model coefficients (Lagrangian multipliers), $b_0$ is the model bias term, and **K** is the [$m$ x $m$] kernel matrix.

Following transformation into a positive definite form, the LS-SVM KKT system can subsequently be solved by any of the myriad of methods for solving sets of linear equations, such as conjugate gradient descent. To implement LS-SVM algorithm, after choosing suitable pre-processing, the user must only specify two parameters: $\gamma$ and $\sigma^2$. Predictions from new samples are derived by creating a kernel vector between each calibration sample and the test sample, followed by taking the inner product with the **b**-coefficient vector.

## Experimental

### Datasets

For the performance comparison, datasets of NIR spectra and reference values were compiled for four, diverse products: apples, meat, corn, and animal feed (Table 1). Three of the datasets (apples, meat, and corn) were used for regression analysis. Each consisted of spectra from a typical

NIR analyzer, and each had multiple analytes. The fourth dataset (animal feed) was a discriminate analysis problem with the objective of detecting meat and bone meal contamination in ruminant feed; the spectra for this dataset were collected in a novel manner using the imaging spectrometer at CRAGx (Gembloux, Belgium).

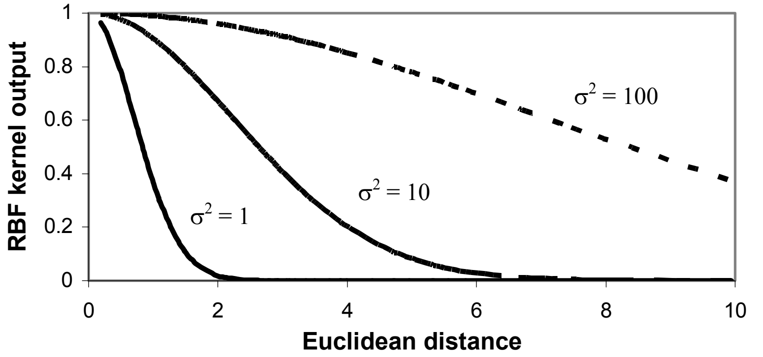

Figure 1. Graphical illustration of the relationship between Euclidean distance and RBF kernel output.

Each dataset consisted of calibration and test subsets, except the meat dataset, where a portion of the dataset was randomly selected for parameter optimization; then, after the optimal parameters were set for each algorithm, the entire dataset was broken into seven subsets according to the year in which the sample was drawn. Predictions were derived for each subset by calibrating on the remaining subsets, using the previously-determined parameter settings.

Table 1. Product and constituent datasets used for performance comparison.

| | | Samples (n) | | Instrument Description | | Wavelength Range | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Product | Cal. | Test | Make | Model | MIN (nm) | MAX (nm) | inc |
| Apples | Sucrose | 601 | 139 | NIR Systems | 6500 | 796 | 2416 | 2 |
| | pH | 514 | 118 | | | 1104 | 2300 | 2 |
| | Acidity | 519 | 120 | | | 800 | 2498 | 2 |
| | Firmness | 506 | 114 | | | 800 | 1998 | 2 |
| Meat | Moisture | 691 | | NIR Systems | 5000 | 1300 | 2398 | 2 |
| | Protein | 658 | | | | | | |
| | Fat | 651 | | | | | | |
| | Collagen | 472 | | | | | | |
| Feed | | 3603 | 654 | Spectral Dimensions | MatrixNIR | 1200 | 1700 | 10 |
| Corn | Moisture | 891 | 429 | FOSS-Tecator | Infratec 1229, 1241 | 850 | 1048 | 2 |
| | Protein | 907 | 429 | | | | | |
| | Oil | 899 | 429 | | | | | |

## Software

Four regression methods were tested during the comparison: MPLS and LOCAL (ISI II v1.50, Infrasoft International, LLC, Port Matilda, PA, USA), ANN (FOSS NIRSystems, Silver Spring, MD, USA), and the LS-SVM[4] toolbox for MATLAB. The ANN and LS-SVM calculations were carried out in MATLAB 6.5 (The Mathworks, Inc., Natick, MA, USA).

The optimization of MPLS, LOCAL, and ANN calibrations were carried out by Dr. Pierre Dardenne. Selection of pre-processing, removal of training outliers, and parameter optimization were completed using standard methods for each algorithm and dataset. Depending on the capabilities of the individual algorithms, optimization was performed with cross-validation, or by splitting the calibration into a training and validation set. The optimization of LS-SVM was carried out independently by R. P. Cogdill, using the same datasets. The LS-SVM cross-validation was performed using MATLAB functions custom-written for this paper, since the cross-validation procedure supplied with the LS-SVMlab toolbox often suggested overly optimistic $\gamma$ and $\sigma^2$ settings, leading to over-fit of the training data during regression.

## Results and Discussion

The apples dataset was the first to be analyzed using LS-SVM. Though the spectra had been truncated, training with many hundred wavelengths and data points was prohibitively slow for feasible parameter optimization. It was found that training time increases with the square of the number of training samples, and linearly with the number of independent variables. With this in mind, the decision was made to replace the spectra with (an excess of) PLS factors when there were too many wavelengths to feasibly include the entire spectrum. PLS factors were used with the apples and meat datasets (Table 2); for the corn dataset, whose spectra consisted of only 100 variables, using the entire spectrum produced better prediction results.

The results of the regression tests (apples, meat, and corn) are shown in Table 2. While LS-SVM was superior to MPLS, LOCAL, and ANN, in all but two tests, it is more surprising that LS-SVM performed best even for calibrations that are generally considered to be linear, such as protein in corn. For the ruminant feed discriminate analysis problem, LS-SVM misclassified 6 out of 654 samples, while MPLS and ANN misclassified 9 and 15 samples, respectively; LOCAL was not found to be applicable to the problem. For the discriminate analysis problem, LS-SVM was trained using 12 PLS factors, with $\gamma$ and $\sigma^2$ levels of 6000 and 1000, respectively.

Using the corn protein dataset, a test was devised whereby progressively smaller subsets were randomly drawn from the original set of 920 samples. For each subset, PLS, LOCAL, ANN, and LS-SVM calibrations were derived (with update of the parameter settings) and tested using the same set of 429 independent test samples. The results are shown in Figure 2, relative to the PLS results (upper and lower confidence limits are included). The performance of LOCAL and ANN, relative to PLS, were much as expected; both algorithms required somewhere between 250 and 500 samples before their performance began to significantly improve on that of PLS. Surprisingly, regardless of the size of calibration dataset, LS-SVM always performed better than LOCAL and ANN, and in only one case was not significantly better than PLS.

Table 2. Performance comparison results and LS-SVM training parameters.

| | | SEP | | | | LS-SVM Parameters | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Product | MPLS | LOCAL | ANN | LS-SVM | Input | $\gamma$ | $\sigma^2$ |
| Apples | Sucrose | 0.37 | 0.34 | 0.33 | **0.32** | 19 lv | 3000 | 3000 |
| | pH | 0.12 | 0.11 | 0.13 | **0.10** | 20 lv | 8000 | 4000 |
| | Acidity | 1.47 | 1.47 | 1.36 | **1.28** | 20 lv | 4000 | 3000 |
| | Firmness | 1.00 | 1.03 | 1.02 | **0.87** | 20 lv | 6250 | 4500 |
| Meat | Moisture | 0.62 | 0.85 | **0.61** | 0.69 | 15 lv | 5000 | 1000 |
| | Protein | 0.88 | 1.00 | 1.05 | **0.81** | 20 lv | 3500 | 7000 |
| | Fat | 0.52 | 0.56 | 0.91 | **0.47** | 10 lv | 4000 | 1000 |
| | Collagen | 0.87 | 1.94 | 1.08 | **0.71** | 16 lv | 6500 | 4500 |
| Corn | Moisture | 0.73 | 0.75 | 0.59 | **0.57** | 100 λ | 5000 | 2000 |
| | Protein | 0.41 | 0.45 | 0.43 | **0.36** | 100 λ | 3000 | 3000 |
| | Oil | 0.41 | **0.40** | 0.47 | **0.40** | 100 λ | 3000 | 4000 |

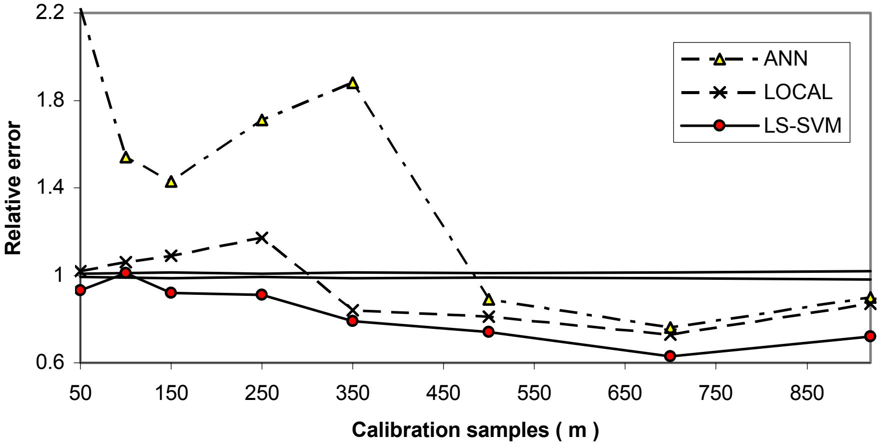

Figure 2. Calibration dataset size versus error (relative to PLS) for ANN, LOCAL, and SVM. The confidence interval for PLS is shown as solid lines immediately above and below one.

## Conclusions

Given the observed results of this work, some conclusions may be made regarding the application of least-squares support vector machines to chemometrics:

- LS-SVM was shown to be generally superior to MPLS, LOCAL, and ANN in predictive performance.
- A large sample database is not required for calibration development using LS-SVM regression.

- While the effect of various pre-processing methods on LS-SVM performance has not been shown; it can be concluded that latent variable compression is not always necessary during LS-SVM calibration.
- The proper selection of tuning parameters ($\gamma$ and $\sigma^2$) is critical to avoid over-fitting during LS-SVM training.

Just as for any traditional chemometric technique, proper use of kernel methods, including LS-SVM, requires some understanding and experience; with the power to model virtually any nonlinear function, the ever-present danger of over-fitting, and subsequent poor predictive performance, places an even greater demand on the skills of the chemometrician.

## References

1. H. Martens and T. Næs, *Multivariate Calibration*. Wiley, New York (1993).
2. V. Vapnik and A. Chervonenkis, *Automation and Remote Control*. **24**, 774 (1963).
3. V. Vapnik, in Nonlinear Modelling: *Advanced Black-box Techniques*, Ed by J.A.K. Suykens and J. Vandewalle. Kluwer Academic Publishers, Boston, p. 55 (1998).
4. 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).
5. www.kernel-machines.org, May 1, 2003
6. A.I. Belousov, S.A. Verzakov and J. von Frese, *J. Chemometrics*. **16**, 482 (2002).
7. A.I. Belousov, S.A. Verzakov and J. von Frese, *Chemom. Intell. Lab. Syst.* **64**(1), 15 (2002).
8. R.P. Cogdill and P. Dardenne, *J. Near Infrared Spectrosc*. **12**, 93 (2004).
9. F. Girosi, M. Jones and T. Poggio, *Neural Computation* **7**, 219 (1995).
10. H.R. Draper and H. Smith, *Applied Regression Analysis*, Second Edition, Wiley, New York (1981).
11. W. Wu, D.L. Massart and S. de Jong, *Chem. Intell. Lab. Syst.* **36**, 165 (1997).
12. B. Walczak and D.L. Massart, *Anal Chim. Acta.* **331**, 177 (1996).