ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/222054519

PLS pruning: A new approach to variable selection for multivariate calibration based on Hessian matrix of errors

Article *in* Chemometrics and Intelligent Laboratory Systems · March 2005

DOI: 10.1016/j.chemolab.2004.09.007

CITATIONS
13

READS
107

3 authors, including:

Cesar Mello
IPclin
74 PUBLICATIONS **1,373** CITATIONS
SEE PROFILE

Ronei J Poppi
University of Campinas
343 PUBLICATIONS **7,116** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project
Raman mapping and chemometrics for the development of stable lipid-based pharmaceutical formulations View project

Project
Two Dimensional Correlation Spectroscopy View project

All content following this page was uploaded by Cesar Mello on 18 August 2017.

The user has requested enhancement of the downloaded file.

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 76 (2005) 73–78

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors

Silvio L.T. Lima[a], Cesar Mello[b], Ronei J. Poppi[a,*]

[a]*Institute of Chemistry, UNICAMP, P.O. Box 6154, 13083-970 Campinas, S.P, Brazil*
[b]*Institute of Chemistry, University of Franca, P.O. Box 32, 14404-600 Franca, S.P, Brazil*

Received 31 March 2004; received in revised form 27 September 2004; accepted 28 September 2004
Available online 6 November 2004

Abstract

In this article, a new approach called partial least squares (PLS) pruning is described for variable selection in PLS modeling. The aim of the method is the deletion of unimportant PLS coefficients of regression by using information from all second derivatives of the error function. The proposed approach was applied to Brix determination in sugar cane juice by near infrared spectroscopy. The results obtained were promising, leading to a meaningful variable reduction of 96% without loss of model prediction capability.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Partial least squares; Variable selection; Hessian matrix of errors

1. Introduction

Partial least squares (PLS) [1] is a linear multivariate calibration method based on factors that have been incorporated in several commercial softwares, being considered a standard multivariate calibration method. Also, multivariate calibration can be performed by using methods not based on factors and thus able to deal with non-linearities. Among these, we can find in the literature multivariate calibration methods based on neural networks such as multilayer perceptron [2], radial basis functions [3] and, more recently, support vector machines [4]. Both methodologies, whether based on factors or on neural networks, have wide acceptance and an enormous and increasing number of applications in the most different chemical areas.

In spite of the excellent results obtained with these methodologies, the selection and knowledge of important information from the variables still remains a field requiring considerable effort. Classically, the selection is accomplished from chemical knowledge of the samples or by simple correlation between the variables and concentration (or another property to be determined). However, several papers [5,6] have been shown that the utilization of mathematical algorithms to search for the best variables to be used in the model is more efficient and practical.

Several methods that are able to select variables in PLS calibrations have been proposed in the chemometrics literature in the last decade. Among the proposed methods for variable selection in PLS, we can point out iterative PLS (IPLS) [7], uninformative variable elimination by PLS (UVE-PLS) [8], interactive variable selection for PLS (IVS-PLS) [9] and the genetic algorithm [10]. These algorithms are based on different strategies, and they have been developed for different applications, but all are delineated to reduce data dimensions, avoiding colinearities and making the system more interpretable. The variables, once selected, must, ideally, capture in the analytical signals only the information tied to the interested compounds and exclude any other characteristic not related to the concentration or other measurable property of the sample.

* Corresponding author. Tel.: +55 19 37883126; fax: +55 19 37883023.
E-mail address: ronei@iqm.unicamp.br (R.J. Poppi).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.09.007

In this work, a novel method for variable selection in PLS models was developed and applied based on Hassib and Stork's work in neural networks named Optimal Brain Surgeon (OBS) [11]. It was thought that the basic idea of OBS, proposed specifically for neural networks, should be extended to multivariate calibration models based on factors, such as PLS. The advantage in applying OBS principles is to generate a local surface of error that allow to estimate how the error values vary when a specific variable is deleted. Then, it is possible to eliminate one correct noninformative variable at time, which produce a minimum increase or maximum decrease on error function. Thus, the proposed method, hereafter called PLS pruning, was applied for variable selection in near infrared spectroscopy for Brix determination in sugar cane juice.

2. PLS pruning

The basic idea of PLS pruning consists in, initially, building a multivariate calibration model using the PLS1 algorithm, where a model is developed for a data set (independent variables matrix—**X**) against just one dependent variable vector (**y**). The model developed can be expressed as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where **b** is the regression coefficients of the model and **e** the model error.

After optimization and conclusion of the PLS1 model, the next step is to select the most significant regression coefficients, using a method that is able to eliminate coefficients with lesser chemical or physical meaning, concomitant with error minimization of the model.

This simultaneous process can be mathematically done considering that the prediction value ($\hat{y}$) is governed by the following function:

$$\hat{y} = f(\mathbf{x},\mathbf{b}) \tag{2}$$

where **x** and **b** are the independent variables of the sample to be predicted and the coefficients vector of the model, respectively. Then it is possible to write an error function for the prediction:

$$E(\mathbf{b}) = \sum (y - f(\mathbf{x},\mathbf{b}))^2 \tag{3}$$

or

$$E(\mathbf{b}) = \sum (y - \hat{y})^2. \tag{4}$$

Here, the principal aim of the approach is reached, i.e., to minimize the error function. This minimization starts by the expansion of this function in a Taylor series up to the second order terms, around a possible single vector coefficient candidate $\mathbf{b}_0$, in order to guarantee that $E(\mathbf{b})$ has a minimum:

$$E(\mathbf{b}+\delta\mathbf{b}) = E(\mathbf{b}) + \nabla E(\mathbf{b})\delta\mathbf{b} + \frac{\mathbf{H}E(\mathbf{b})}{2!}(\delta\mathbf{b})^2 + \frac{\theta}{3!}(\delta\mathbf{b})^3 + \ldots \tag{5}$$

where $\delta\mathbf{b}$ is the increment added to the vector **b**, given by:

$$\delta\mathbf{b} = \mathbf{b}_0 - \mathbf{b} \tag{6}$$

$\nabla E(\mathbf{b})$ is the gradient of the error function, as follow:

$$\nabla E(\mathbf{b}) = \frac{\partial E(\mathbf{b})}{\partial \mathbf{b}} \tag{7}$$

$\mathbf{H}E(\mathbf{b})$ is the Hessian matrix:

$$\mathbf{H}E(\mathbf{b}) = \frac{\partial^2 E(\mathbf{b})}{\partial \mathbf{b}^2} \tag{8a}$$

or

$$\mathbf{H} = \begin{bmatrix} \nabla^T \frac{\partial E(\mathbf{b})}{\partial \mathbf{b}_1} \\ \vdots \\ \nabla^T \frac{\partial E(\mathbf{b})}{\partial \mathbf{b}_n} \end{bmatrix}. \tag{8b}$$

To evaluate $\delta\mathbf{b}$ in order to yield the lowest increase in the error function, two approximations must be applied:

(1) Around the error function minimum, a quadratic function should be considered, leading to disregarding the higher order terms of the Taylor series;
(2) Inasmuch as the elimination process starts only after the PLS model is optimized, it is possible to assume that the model could reach a local or global minimum of error surface, or in other words, that the gradient term should be equal to zero.

Considering that these suppositions are true, Eq. (5) can be written as:

$$E(\mathbf{b}+\delta\mathbf{b}) = E(\mathbf{b}) + \frac{\mathbf{H}E(\mathbf{b})}{2!}(\delta\mathbf{b})^2 \tag{9}$$

$\Delta E(\mathbf{b})$ can be written as:

$$\Delta E(\mathbf{b}) = E(\mathbf{b}+\delta\mathbf{b}) - E(\mathbf{b}) \tag{10}$$

and the final Equation is

$$\Delta E(\mathbf{b}) = \frac{1}{2}\delta\mathbf{b}^{\mathrm{t}}\mathbf{H}E(\mathbf{b})\delta\mathbf{b}. \tag{11}$$

3. Saliency calculation

During the pruning process, it is desirable that one of the coefficients (b_i) be set to zero, simultaneous with the minimization of the error function (ΔE) in Eq. (11). The elimination of this coefficient (b_i) is established as:

$$\delta b_i + b_i = 0 \tag{12a}$$

or more generically

$$\mathbf{1}_i^{\mathrm{t}}\delta\mathbf{b} + b_{\mathrm{i}} = 0 \tag{12b}$$

where $\mathbf{1}_i^{\mathrm{t}}$ is a vector with all elements equal to zero, except the i-th element that is equal to 1.

Coefficients elimination with no meaning for the model and the simultaneous minimization of the error function is a typical problem that can be solved applying Lagrange's multipliers method, which can be written in a generic way in the following form:

$$S(\mathbf{b}) = \Delta E(\mathbf{b}) + \lambda\left(\mathbf{1}_i^{\mathrm{t}}\delta\mathbf{b} + b_{\mathrm{i}}\right) \tag{13}$$

where λ is the Lagrange's multiplier.

Relating the $S(\mathbf{b})$ derivative to $\delta\mathbf{b}$, and letting the coefficient b_{i} be zero, the optimal $\delta\mathbf{b}$ is given by:

$$\delta\mathbf{b} = \frac{b_{\mathrm{i}}}{\left[\mathbf{H}^{-1}\right]_{ii}}\mathbf{H}^{-1}\mathbf{e}_{\mathrm{i}}. \tag{14}$$

The so-called Saliency, $S_{\mathrm{i}}(\mathbf{b})$, corresponds to Eq. (13) solution, optimized with relation to $\delta\mathbf{b}$, subjected to the i-th coefficient elimination constraint:

$$S_{\mathrm{i}}(\mathbf{b}) = \frac{1}{2}\frac{b_{\mathrm{i}}^2}{\left[\mathbf{H}^{1}\right]_{ii}} \tag{15}$$

where $\mathbf{H}^{-1}$ is the Hessian inverse matrix and $[\mathbf{H}^{-1}]_{ii}$ is the ii-th element of this inverse matrix.

Actually, the saliency S_{i} represents the increase in the error function due to elimination of b_{i}. Considering all procedures needed to eliminate the correct coefficients, the hardest is to compute the Hessian matrix inversion, inasmuch as it must not be singular. The inverse Hessian matrix calculation was made based on the work of Hassibi and Stork [11], who applied the lemma matrix inversion to solve this problem.

4. The pruning procedure

The pruning procedure intends to promote an ordered and selective PLS regression coefficients elimination based on specific criteria, trying to reach an improvement in the prediction capability of the model. This procedure finds the best set of coefficients through error function evaluation using the Hessian matrix. The simplified procedure consists in:

(1) start from an optimized PLS1 model, with "m" independent variables and therefore "m" regression coefficients and evaluate its error function using the validation set;
(2) eliminate just one regression coefficient (b_{i}) according to saliency analysis that estimates the function error variation caused by the exclusion of this parameter;
(3) create a new regression model by PLS1 using the "$m-1$" remaining variables in order to achieve a new regression coefficients vector (**b**). This model is used to determine a new value for the error function;
(4) execute steps 2 and 3 until only one regression coefficient remains to be eliminated.

The optimization of the number of latent variables used in each model was done iteratively, obeying the following criteria:

(1) Evaluate the PRESS (Prediction Error Sum of Squares) using a number of latent variable (LV) equal to 10% of all variables. Once this method reduce gradually the number of variables, this procedure is executed until achieve to, at least, 20 variables. If there are 19 variables or less, then the number of latent variables becomes the same of these remaining variables.
(2) Find LV with minimum PRESS. Search another lower LV with PRESS up to 10% higher this minimum. If it was found, then this one is let to be the LV used in the model. The purpose of this procedure is to avoid models over fitting.

By this procedure, it is possible to collect a set of "m" models. The first one with "m" regression coefficients; the second one with "$m-1$" regression coefficients, and so on, until the last few with 4, 3 and 2 regression coefficients, ending with the model that uses only one regression coefficient. For each model, there is an associated error value. The best model is going to be that which shows the minimum validation error with the fewest regression coefficients at the same time.

In order to verify the reliability of the best model, the prediction set was used, and its RMSEP was then compared to the RMSEP from the model with all variables.

5. Experimental

The spectra of the sugar cane juices were acquired in the near infrared region (NIR), using a FEMTO spectrophotometer with spectral resolution of 2 nm, in the transmittance mode, in a continuous flow cell with an optical path of 1 mm. The data set used for the application of the PLS-pruning was composed of 300 NIR spectra of sugar cane juice, whose BRIX were obtained by using a refratometer [12].

The spectra were filtered by noise minimization using a Fourier Transform Filter [13], before application of the modeling process, once it is able to reduce undesirable gross noise, not smoothing significantly the spectra.

6. Performance evaluation of the model

To evaluate the relative deviation of models from known values (experimentally evaluated), the root mean square

error prediction (RMSEP) was utilized and is defined as

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_p} (y_i - \hat{y}_i)^2}{n_p}} \quad (16)$$

where n_p is the number of samples used in the prediction set, $\hat{y}_i$ is the predicted value by the i-th model and y_i is the measured value by the accepted method.

Thus, the RMSEP was used to compare PLS-pruning to the well-established full spectrum method PLS1. Moreover, the F-test [14] at the 95% confidence level was applied to verify whether there were significant differences between models. These estimates were carried out by comparison of RMSEP of two different calibration methods according to Eq. (17):

$$F(p_i, p_j) = \left(\frac{RMSEP^2_{PLSpruning}}{RMSEP^2_{PLS1}} \right) \quad (17)$$

where p is the number of prediction samples.

7. Computer programs

The routines used to select the optimal coefficients by PLS-pruning were implemented using a program in Matlab by the authors, from a code developed by Norgaard from the Institute of Automation Technical University of Denmark [15]. The PLS1 calculations were performed using the PLS toolbox for use with Matlab by Eigenvector Research [16].

8. Results and discussion

The data used to carry out this application were 300 near infrared spectra of sugar cane juice measured in 626 wavelengths (1200–2500 nm with a resolution of 2 nm) collected at the Copersucar Technology Center, Brazil. The raw data that constitutes the complete data set (sugar cane juice) studied was originated from different regions and acquired on different days, ensuring a satisfactory heterogeneity and noncorrelations among samples.

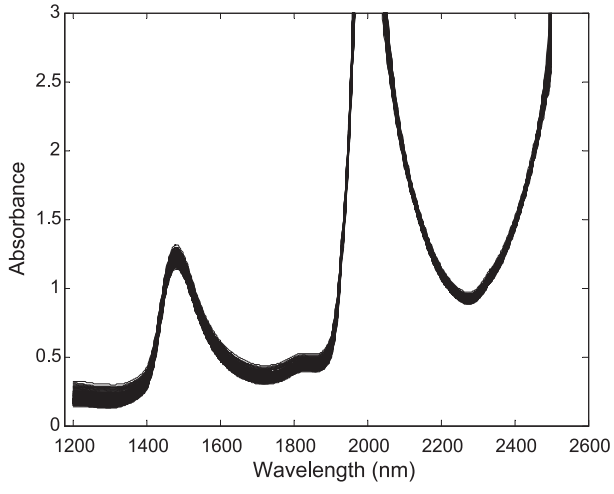

Fig. 1. Sugar cane juice spectra in NIR region.

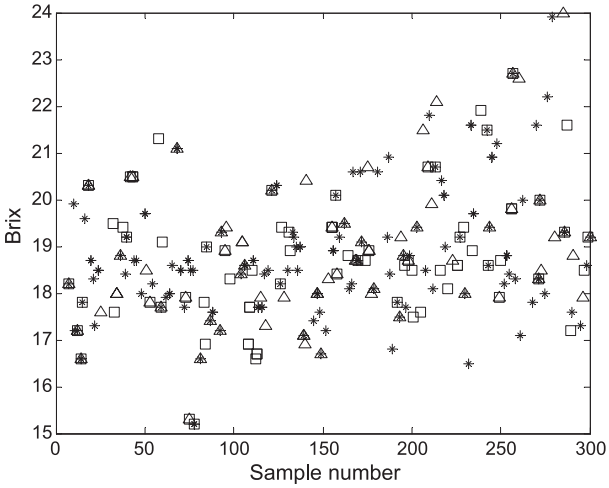

Fig. 2. Brix values of the samples used in (*) calibration; (△) validation and (□) prediction sets.

Fig. 1 shows the sugar cane spectra, and it indicates the existence of a spectrum region, near to 2000 nm, where the signal is saturated due to the water band. The information contained in this region is not useful to model development, and, moreover, it can make a satisfactory construction of any model type unfeasible. It was decided therefore to eliminate this region of the spectra.

After this elimination, the spectra still had 555 variables, indicating therefore that 71 variables were eliminated. Also, it was decided to decrease by half the number of variables due to processing time, inasmuch as its iterative character demands a computational time-consumption proportional to the number of variables in order to complete the pruning approach. Thus, only those absorbance values acquired from even numbered wavelengths were chosen, giving 278 variables. The goal of this procedure was to guarantee that even if important wavelengths were eliminated, its neighboring elements could play their role, inasmuch as the resolution (2 nm) allows this approach.

After the preprocessing, the 300 samples were distributed in three distinct sets: calibration; validation and prediction. The criteria employed to proceed to this distribution was as follows: the original set of 300 samples were reorganized randomly but obeying an approximately homogeneous distribution. After that, the first 150 samples were assigned to be the calibration set, the next 75 samples were used to compose the validation set and the last 75 samples were used to be the prediction set. Care was taken to always choose the calibration set with Brix values over the whole range, avoiding extrapolations. Fig. 2 illustrates this arrangement.

Observing Fig. 2, it can be seen that the validation and prevision samples have their Brix values surrounded by a sufficient number of calibration samples. The calibration and validation sets were used to elaborate the PLS1 model

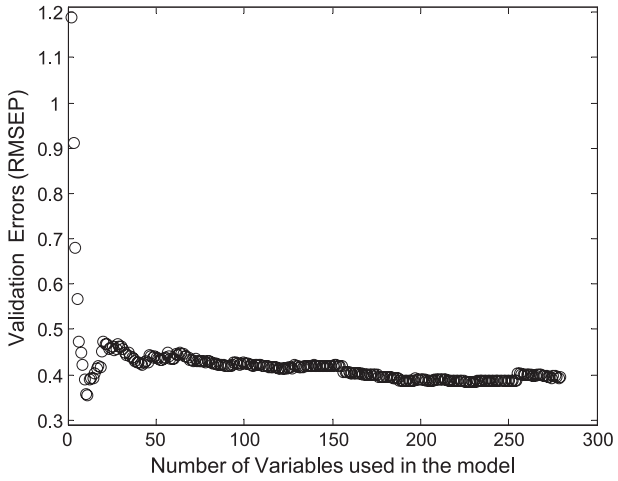

Fig. 3. Validation errors against variable reduction.

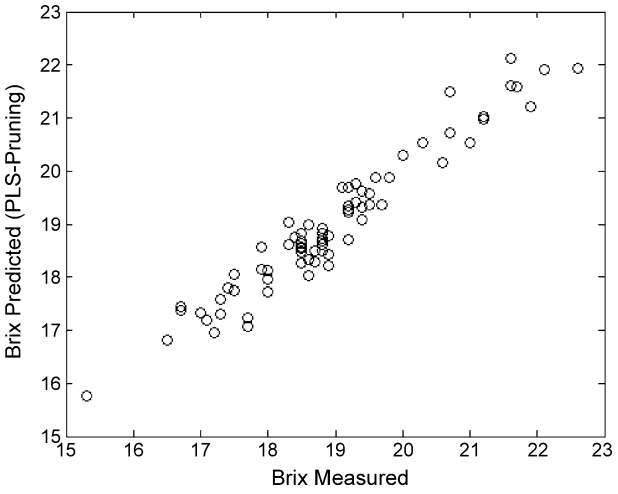

Fig. 5. BRIX predicted by PLS-Pruning against the measured values.

and to evaluate the pruning approach. The prediction set was used to verify the generalization capacity of the model because it was not directly present in its development.

9. Variables selection

The selection of the important variables for the calibration model was done by submitting the coefficients vector **b**, found by the PLS1 method, to pruning. Fig. 3 shows the general behavior of how the validation errors vary, as the number of variables used for a calibration model decreases.

The first model built during this processing started with all the variables, corresponding to the PLS1 method, having its own validation error. As the coefficient is pruned one by one, the coefficients remaining are employed to generate a simplified model with fewer variables in it.

The model choice is made by seeking that which reach is the smallest validation error. Having this reasoning in mind, it was possible to reduce the number of variables by about 96% of the initial number, from 278 variables to only 11, which seems to exhibit a smaller prediction error than PLS1 with all 278 variables.

Analyzing Fig. 3, a characteristic behavior can be seen. Inasmuch as the pruning procedure runs until only one coefficient remains, it could be observed that a greater reduction in the number of variables that compose the model (<11) implies a sudden increase in the value of validation errors, indicating that a very reduced number of variables are not able to model the relation between the input spectrum and the property measured from the samples (Brix). Another important fact is that the variables selected could be identified, and their attributions could be comprehended [17], as shown in Fig. 4.

Most of the selected wavelengths shown in Fig. 4 can be assigned for overtones and combination bands of –CH, $–CH_2$, $–CH_3$ and –OH bonds. These wavelengths are correlated with the sugars (glucose, sucrose and fructose) present in the cane juice samples.

There is a band, at about 2500 nm, assigned to the stretchings of C–H, C–C and C–O–C. There are, as well,

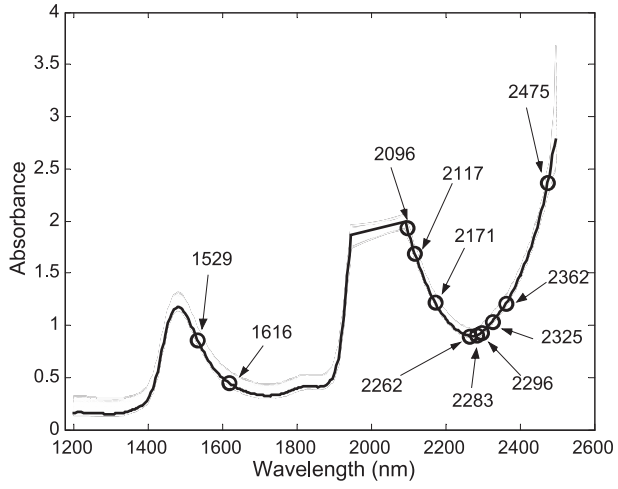

Fig. 4. Selected wavelengths by the PLS-pruning approach.

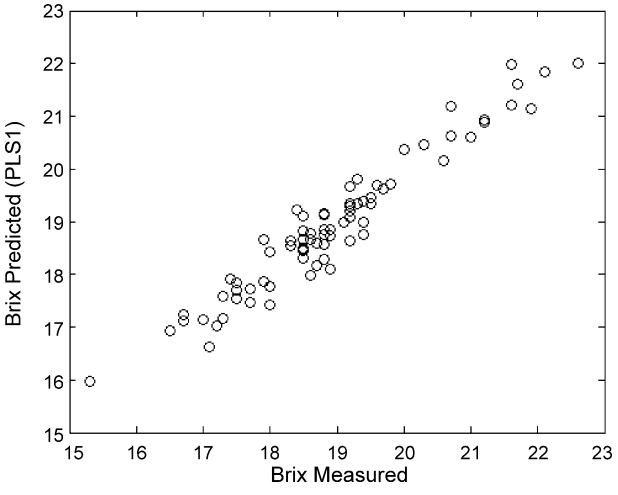

Fig. 6. BRIX predicted by PLS1, without variable selection, against the measured values.

several bands in the region of 2280–2330 nm, relative to the stretching of C–H and to deformation of $–CH_2–$. Other characteristic bands occur near 2100 nm as a result of a combination of stretching and deformation of O–H bonds. Finally, in the region near to 1450 nm, the appearance of the band due to the first overtone of the stretching O–H is observed.

10. Performance of the model for Brix determination

An initial comparison between the PLS-pruning and the PLS1 method was performed by plotting the graphics of the measured value of Brix against the predicted value for both models, as represented in Figs. 5 and 6.

Visually it is possible to verify that there is a high similarity between the graphics. The RMSEP for PLS1 using the whole spectra and the PLS-pruning were very close to each other: 0.38 and 0.40, respectively. A comparison of these values leads to conclude they are statistically equivalent—it was confirmed by *F*-test. Thus, models with the 11 selected coefficients will produce equivalent results to the model built with all 278 coefficients.

11. Conclusion

The results obtained for the sugar cane juice samples were promising, showing the potentiality of the PLS-pruning method that, in its turn, permits generation of very simple models with no degradation in prediction ability, inasmuch as they were statistically equivalent to the model built from the established PLS1 method, which is widely employed in multivariate calibration. Further investigations have to be made to confirm the potentiality of this procedure in relation to other accepted PLS variable selection methods.

References

[1] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.
[2] J. Zupan, J. Gasteiger, Neural Networks for Chemists: An Introduction, VCH, New York, 1993.
[3] M. Carlin, T. Kavli, B. Lillekjendlie, Chemom. Intell. Lab. Syst. 23 (1994) 163.
[4] C.J.C. Burges, Data Min. Knowl. Disc. 2 (1998) 121.
[5] P.J. Brown, J. Chemom. 6 (1992) 151.
[6] E.V. Thomas, Anal. Chem. 66 (1994) 795.
[7] S.D. Osborne, R.B. Jordan, R. Kunnemeyer, Analyst 122 (1997) 1531.
[8] V. Centner, D.L. Massart, O.E. de Nord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
[9] F. Lindgren, P. Geladi, S. Rannar, S. Wold, J. Chemom. 8 (1994) 349.
[10] R. Leardi, J. Chemom. 14 (2000) 643.
[11] B. Hassibi, D.G. Stork, in: S.J. Hanson, J.D. Cowan, C.L. Giles (Eds.), Advances in Neural Information Processing Systems, vol. 5, Morgan Kaufmann, San Mateo, 1993, p. 164.
[12] F.L. Hart, H.J. Fisher, Modern Food Analysis, Springer-Verlag, New York, 1971.
[13] R. Bracewell, The Fast Fourier Transform and its Application, McGraw-Hill, New York, 1965.
[14] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, 4th ed., Prentice Hall, London, 2000.
[15] N. Norgaard, Neural Network Based System Identification Toolbox, Tech. Report 95-E-773, Institute of Automation Technical University of Denmark, 1995.
[16] B.M. Wise, N.B. Gallagher, PLS Toolbox for use with MATLAB, ver. 2.11,202.
[17] D.A. Burns, E.W. Ciurczak (Eds.), Handbook of Near-Infrared Analysis, 2nd. ed., Marcel Dekker, New York, 2001.

View publication stats