Talanta 115 (2013) 755–760

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

An integrated approach to the simultaneous selection of variables, mathematical pre-processing and calibration samples in partial least-squares multivariate calibration

Franco Allegrini, Alejandro C. Olivieri *

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario S2002LRK, Argentina

ARTICLE INFO

Article history:
Received 2 May 2013
Received in revised form
24 June 2013
Accepted 25 June 2013
Available online 1 July 2013

Keywords:
Multivariate calibration
Variable selection
Pre-processing selection
Sample selection
Outlier detection
Partial least-squares

ABSTRACT

A new optimization strategy for multivariate partial-least-squares (PLS) regression analysis is described. It was achieved by integrating three efficient strategies to improve PLS calibration models: (1) variable selection based on ant colony optimization, (2) mathematical pre-processing selection by a genetic algorithm, and (3) sample selection through a distance-based procedure. Outlier detection has also been included as part of the model optimization. All the above procedures have been combined into a single algorithm, whose aim is to find the best PLS calibration model within a Monte Carlo-type philosophy. Simulated and experimental examples are employed to illustrate the success of the proposed approach.

1. Introduction

In multivariate spectroscopic calibration, variable selection intends to rationally choose, from the whole available spectrum, wavelengths where signals have maximum information regarding the analyte of interest, discarding at the same time those carrying irrelevant information (noise, saturation regions) or those heavily overlapped with other sample components which are not of analytical interest [1,2]. Although the concern is primarily directed toward spectral information, variable selection can also be applied to any multivariate technique where some sensors can in principle be more selective as to the analyte or property of interest, while others may give negligible signals. Improved PLS analytical performance has been reported upon variable selection, which supports the continuing interest in this chemometric activity [3,4].

Mathematical pre-processing techniques exist for removing variations in spectra from run to run, which are unrelated to analyte concentration changes [5,6]. The removal of these unwanted effects, e.g., dispersion in near infrared (NIR) spectra of solid or semi-solid materials, leads to more parsimonious partial least-squares (PLS) models requiring less latent variables than those based on raw data, and very often produce better statistical indicators.

Sample selection is another important activity in PLS regression analysis of complex samples (industrially manufactured or naturally occurring), and is intended to provide representativeness to the set of samples used for model building [7]. This means that their spectra should span most of the expected variability of future samples in spectral space.

Outlier detection has been extensively discussed in the literature, and several diagnostics have been proposed [8]. From a formal point of view, an outlier is a value which is not representative for the rest of the data [9]. In the context of PLS calibration, the main objective is to identify samples with features which make them significantly different from the remaining ones.

All the above activities are mutually connected. Spectral pre-processing modifies by definition the characteristics of the spectral space, and may lead to the selection of different samples for training, and also to different selected wavelengths. Changing the spectral regions, in turn, has a strong influence in the pre-processing method required to model the data in specific regions. Sample selection, on the other hand, is important during model optimization: if truly representative samples are included in the monitoring set instead of in the training set, the choice of model parameters may be misguided. Outliers (samples with wrong nominal concentrations or reference properties) could also be potentially harmful and should be removed. The selection process

* Corresponding author. Tel./fax: +54 341 4372704.
E-mail addresses: olivieri@iquir-conicet.gov.ar, aolivier@fbioyf.unr.edu.ar (A.C. Olivieri).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.06.051

could in principle be carried out on a trial and error basis until convergence, although it would be far more convenient to have a simultaneous variable, pre-processing, sample and outlier selection methodology. A step towards this integration has been previously done by combining pre-processing and variable selection with a single genetic algorithm (GA) [10]. A further integrated approach has been taken in the present report by combining all the above activities into a single algorithm, but using specific procedures for each task.

For variable selection, we propose ant colony optimization (ACO) [11,12] instead of GA. The former algorithm resembles the behavior of ant colonies in the search of the best path to food sources. It has been recently implemented with success in the field of variable selection, showing better performances than other approaches such as genetic algorithms [13–15] and particle swarm optimization [16]. This improved performance was due to two complementary reasons: (1) the effectiveness of the ant colony in their cooperative search for better solutions, and (2) the coupling of ACO with a Monte Carlo approach which provided increased reliability to the regression model.

The choice of a suitable pre-processing or combination of pre-processing methods could be extremely time consuming if performed on a trial and error basis. Thus this activity is proposed to be implemented by a suitable GA [17,18]. Each position ('gene') in a chromosome is either a '1' or a '0', indicating a selected pre-processing method or an ignored one, respectively. As in a previously described ACO algorithm, a Monte Carlo philosophy is applied [12]. If a certain pre-processing method is selected more times than those rejected over the Monte Carlo cycles, and consistently leads to lower average prediction errors, it is considered to be useful for the particular data set under study, and is thus included in the final PLS model.

Sample selection during model optimization is possible using several methods, such as those based on exchange [19], successive projections [20] or sample distances [21,22]. All of them appear to be very effective for providing a reasonably representative sample set. Two distance-based methods were implemented in our integrated strategy: Kennard–Stone [21] and joint *X*–*Y* distances [22].

Finally, in order to detect outlying samples, the usual criterion has been the comparison of a statistical *F* ratio with critical *F* values, both for training and monitoring samples [23]. The experimental *F* value may be based on either concentration or spectral residuals, and is computed as the ratio of squared error for a particular sample and the average squared error for the remaining samples. In this report, concentration residuals have been employed for outlier detection, because: (1) nominal concentrations are known for training and monitoring samples and (2) the objective of the algorithm is to produce a model whose main advantage is its improved prediction ability.

We illustrate the improvement in figures of merit which can be obtained by applying the proposed integrated approach with both simulated and experimental data sets. The approach has been implemented as a MATLAB graphical user interface (GUI) named ACOGASS (ant colony optimization+genetic algorithm+sample selection), which is freely available (see below).

2. Data

2.1. Simulated data

A synthetic data set was built by mimicking the spectra of three components and a sample-dependent non-linear background signal, with component 1 being the analyte of interest. All constituents are present in 70 training samples, 30 monitoring samples and 100 test samples, at randomly chosen concentrations ranging from 0 to 1 unit for constituents 1 and 2, and from 5 to 10 units for component 3 (in the latter case to ensure high relative concentrations of this latter component). Fig. 1A shows the pure component spectra, all at concentrations of 1 unit, as well as a typical background signal, as defined in a full spectral range of 100 sensors. From these noiseless profiles, training, monitoring and test spectra were built. Specifically, each training, monitoring and test spectrum **x** was created using the following expression:

$$\mathbf{x} = y_1\mathbf{s}_1 + y_2\mathbf{s}_2 + y_3\mathbf{s}_3 + \mathbf{b} \quad (1)$$

where $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are the pure component spectra at unit concentration, y_1, y_2 and y_3 are the component concentrations in a specific sample and **b** is the background signal. Gaussian noise with a standard deviation of 0.01 units was added to all concentrations, before inserting them in Eq. (1). A vector of signal noise (standard deviation = 0.05 units) was then added to each **x** vector after applying Eq. (1). Signals higher than 5 units were cut at this latter value, and noise was added to them with 1 unit of standard deviation (this mimics the saturation of the detector at high absorbances in a real experiment). Fig. 1B shows the resulting matrix of training signals. Notice the variations and non-linear nature of the added background signal, which makes it necessary, in general, to apply mathematical pre-processing for removing its effect.

2.2. Experimental BRIX data

This experimental data set was previously described [12], and consists of NIR spectra measured for 105 sugar cane juice samples

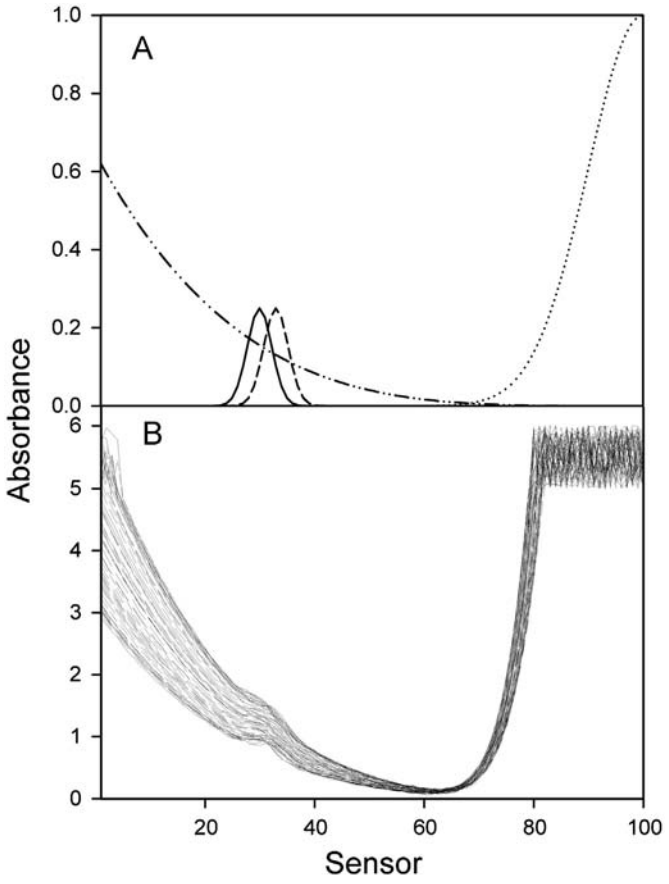

Fig. 1. (A) Plot of pure constituent spectra (analyte 1, solid line, constituent 2, dashed line, constituent 3, dotted line) and the background signal (dashed-dotted line), used to build the simulated data set. (B) Plot of the 70 simulated training spectra. Monitoring and test spectra are similar.

with a NIRSystems6500 spectrometer in the wavelength range 400–2498 nm each 2 nm (1050 data points). For each sample, reference Brix values were measured with a Leica AR600 refractometer, falling in the range 11.76–23.15.

2.3. Experimental CORN data

This is a freely available data set [24], consisting of NIR spectra of 80 samples of corn in the wavelength range is 1100–2498 nm at 2 nm intervals (700 channels). Several reference parameters were measured for this set, among which we selected the starch content, with values ranging from 62.83 to 66.47.

3. Software

The integrated algorithm has been incorporated into the ACOGASS graphical user interface which runs under MATLAB version 7.4.0 (R2007a) or higher [25]. Please refer to the document named 'ACOGASS_manual.pdf', which is provided with the software. The MATLAB codes, the manual and the simulated example data discussed in this report can be freely downloaded from www.iquir-conicet.gov.ar/descargas/acogass.zip. The manual is provided as Supplementary material for the present report.

4. Results and discussion

4.1. Setting algorithm parameters

In PLS calibration, it is usual to have two data sets: a calibration set, employed to build the regression model, and a test set to check the prediction ability of the PLS model after all calibration parameters have been optimized. For model optimization, on the other hand, the calibration set is further divided into a training set and a monitoring set. The purpose of the monitoring set is to guide choices during model optimization. In all three sets (training, monitoring and test), reference values (analyte concentrations or sample properties) should be known. When performing sample selection, the training and monitoring sets are merged into a single one, and then divided into new training and monitoring sets at each computation cycle, according to the results of the sample selection method. Two strategies are implemented for the latter activity: (1) the Kennard–Stone algorithm based on either PLS scores or principal component analysis (PCA) scores [21], and (2) selection based on joint *X*–*Y* distances, as described in Ref. [22]. On the other hand, if no monitoring set is provided, the whole calibration set is initially divided at random to create one.

Outliers are flagged if the F_i ratio for the ith. sample exceeds a critical value [23]. For calibration samples, F_i is given by:

$$F_i = \frac{(I-1)(y_{\mathrm{pred},i}-y_{\mathrm{nom},i})^2}{\sum_{i'\neq i}(y_{\mathrm{pred},i'}-y_{\mathrm{nom},i'})^2} \quad (2)$$

where $y_{\mathrm{nom},i}$ is the nominal concentration for sample i, $y_{\mathrm{pred},i}$ is the corresponding value as estimated by the regression model and I is the number of calibration samples. In the case of monitoring samples, the following ratio is computed [23]:

$$F_i = \frac{I(y_{\mathrm{pred},i}-y_{\mathrm{nom},i})^2}{\sum_{i'=1}^{I}(y_{\mathrm{pred},i'}-y_{\mathrm{nom},i'})^2} \quad (3)$$

where i' corresponds to the calibration samples and i to the monitoring samples.

As regards the selection of mathematical pre-processing methods, the algorithm uses a suitable GA to choose one or more pre-treatments among the following: (1) multiplicative scattering correction (MSC) [5], (2) standard normal variate (SNV) [6], (3) detrend, (4) first-derivative and (5) second-derivative (in the last two cases the derivatives were computed using the Savitzky–Golay approach [26]). These four methodologies are commonly applied in NIR/PLS applications [2]. The implementation of the GA requires one to set the number of the so-called chromosomes and the number of generations (see below). Notice that mean-centering is applied to all data sets as a default pre-processing method, as is regularly done in most NIR/PLS applications.

Finally, the most important activity is probably the selection of relevant variables (wavelengths in NIR/PLS studies). This is proposed to be done by ant colony optimization, given the success of this latter technique in related applications [12]. The implementation of ACO requires to set the number of ants, which are the variable-selecting artificial agents, and the number of evolving epochs during which the ants seek for the best combination of variables. Incidentally, in the proposed approach the number of ACO epochs is identical to the number of GA generations. Suitable default values for all ACO and GA parameters are suggested in the ACOGASS software manual (see Supplementary material).

One should be cautious concerning the sensor window (the number of individual sensors included in each of the selectable sensor blocks or variables). The selected window should reflect the typical width of a spectral band. For example, if a typical band has a width of 50 nm, and the spectrum is read in steps of 2 nm, then a reasonable value for sensor window is 25 (band width/step). During algorithm execution, the number of selected variables is allowed to vary within a certain range (i.e., between a minimum and a maximum, both input by the user).

It should be noticed that the parameter guiding the search for pre-processing methods and variables made by GA and ACO is the root mean square error of prediction in the monitoring set of samples (RMSEPmon). Therefore, a final parameter of crucial importance in this regard is the number of PLS factors for model building in each algorithmic step. An initial value to be input in ACOGASS may be estimated by leave-one-out cross-validation on

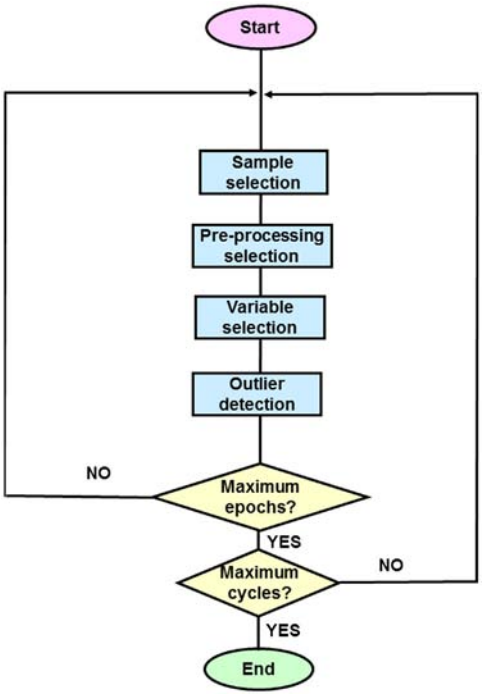

Fig. 2. Flow-sheet for the ACOGASS algorithm implementing sample, pre-processing and variable selection, and outlier detection within a Monte Carlo type strategy.

the raw data, i.e. full-spectral data with no pre-processing [23]. During algorithm execution, however, the number of latent variables is tuned at each step by examining the changes in RMSEPmon as a function of the number of PLS factors, and selecting the number for which no further significant changes in RMSEPmon occur. Leave-one-out cross validation is not employed because it significantly increases the computation time.

The flow sheet shown in Fig. 2 adequately summarizes the above discussed algorithmic steps. As can be seen, all the above activities are repeated for a certain number of times, allowing to obtain reliable results through a Monte Carlo type approach [12]. As is usual, a histogram is built reflecting the relative selection frequency for each variable. Those above a certain tolerance are finally chosen for PLS model building using the selected training sample set and mathematical pre-processing. The optimum model can then be applied, if desired, to the test sample set for checking its predictive ability.

A final note is in place regarding the activities described in the present report. It is likely that an experienced NIR/PLS worker will remove uninformative wavelength ranges upon visual inspection of the spectra (e.g., saturated or high-noise spectral regions), and will also most probably apply some form of mathematical pre-processing to the spectra if the material under analysis is solid or semi-solid. These intuitive forms of variable selection and pre-processing may improve the prediction performance of the PLS models. However, our intention is the development of a fully automated methodology, which could be incorporated into NIR/PLS instrument software in the future, and operated by rather unskilled personnel.

4.2. Simulated data

In this data set, three constituents occur, one of them being the analyte of interest, with an additional background signal. One of the constituents generates an intense signal causing saturation at sensors 80–100, while a non-linear, sample-dependent background signal occurs at sensors 1–50 (Fig. 1B). We expect the present ACOGASS approach to lead to reasonably low values of the RMSEP (both for monitoring and test), by selecting the apparently useful spectral region at sensors 25–40, applying a suitable pre-processing method to alleviate the effect of the variable non-linear background, and optimizing the number of PLS latent variables at two or at most three.

The ACOGASS algorithm was then run on this data set using the parameters shown in Table 1. Notice that each variable comprises two individual sensors (Table 1), which is about half the band width of individual analyte peaks (Fig. 1A). We initially set the number of latent variables at four (Table 1), since there are four spectrally active phenomena in this data set.

Table 1
Specific ACOGASS parameters.

Parameter	Simulated	BRIX	CORN
Number of ants	20	20	20
Blind proportion[a]	0.3	0.3	0.3
Minimum number of variables	4	4	4
Maximum number of variables	8	8	8
Number of chromosomes	20	20	20
Mutation frequency[a]	0.1	0.1	0.1
Cycles	10	10	10
Epochs	50	50	50
Sensor window	2	20	20
Tolerance	0.3	0.3	0.3
Latent variables[b]	4	12	17

[a] The blind proportion and mutation frequency are parameters introducing randomness in the search for minimum monitoring error (see Supplementary material).

[b] Estimated from leave-one-out cross-validation using no pre-processing in the complete spectral range.

Table 2
Figures of merit obtained by ACOGASS in the different data sets.

	Simulated	BRIX	CORN
Full spectrum			
RMSEPtest	0.28	0.75	0.23
REP%	53	4.2	0.36
R^2	0.1114	0.9238	0.9385
No. of latent variables	4	12	17
Pre-processing	None	None	None
After ACOGASS selection			
RMSEPtest	0.03	0.25	0.11
REP%	5.7	1.4	0.17
R^2	0.9900	0.9896	0.9902
No. of latent variables	2	9	14
Pre-processing	Detrend	None	MSC
Comparison of RMSEPtest			
p value[a]	5×10^{-4}	5×10^{-4}	3×10^{-3}

[a] Probabilities associated to the randomization test for comparing RMSEPs (see Ref. [27]).

According to the results presented in Table 2 for the figures of merit computed for the test sample set, which is different than that used for training and monitoring, it is apparent that the ACOGASS approach has found the correct answer. A large prediction error is obtained with no-preprocessing and full spectral data (Table 2). On the other hand, ACOGASS selected detrending as the best pre-processing method, which is reasonable given that this pre-treatment is able to effectively remove non-linear variable background signals, and an optimum number of latent variables of two, as expected. A reasonably low RMSEtest of 0.03 after ACOGASS selection is estimated. Comparison of both RMSEP values (before and after selection) was made using the randomization test suggested by van der Voet [27]. The result indicates that the RMSEP found by ACOGASS is significantly smaller than the one with no selection, since the probability value obtained (p) is much smaller than the critical level of 0.05 (Table 2). Additional indicators are the relative error of prediction REP%=5.7%, computed with respect to the average training value, and a correlation coefficient R^2=0.9900 (Table 2).

In comparison with the results obtained using the full spectra (Table 2), the improvement in predictive ability on variable and pre-processing selection is therefore very significant (Fig. 3).

4.3. BRIX data

The main spectral features of the BRIX data set involve a high absorbance signal due to water (around 1950 nm), regions with significant signals at 1450 and 2500 nm, as well as regions which are mainly dominated by noise below 1300 nm (Fig. 4A). The available set of 105 samples was randomly divided into training, monitoring and test, having 59, 23 and 23 samples respectively. Cross-validation using the full spectrum requires 12 PLS latent variables, which was subsequently employed as the maximum number of factors within ACOGASS (Table 1). Since the sensor window is 20, the minimum number of selectable sensors is 40 nm, because the recording step is 2 nm. This is reasonable in view of the spectral width at half height (Fig. 4A). The remaining ACOGASS parameters are shown in Table 1.

As can be seen in Table 2, the obtained figures of merit show a considerable improvement after selecting the spectral regions shown in Fig. 4A. The RMSEP significantly decreases in comparison to the value without applying a selection process, from 0.75 to 0.25

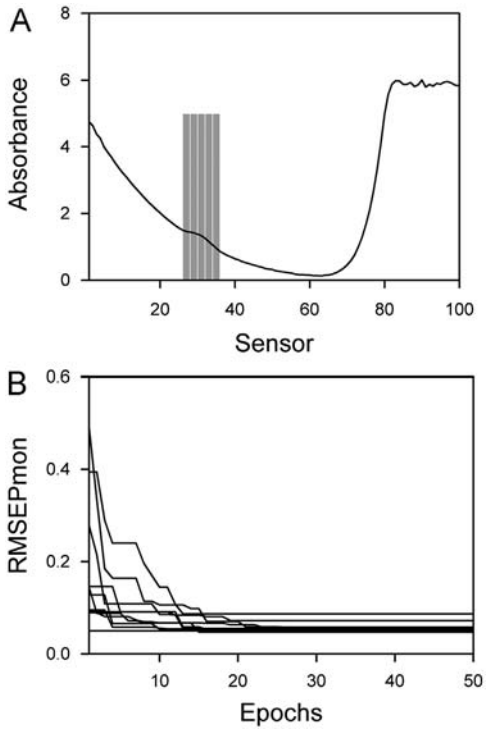

Fig. 3. (A) Gray bars showing the selected variables (sensor blocks) in the simulated data set. The black solid line is the average training spectrum. (B) Evolution of the monitoring error (RMSEPmon) as a function of epochs in the simulated data set.

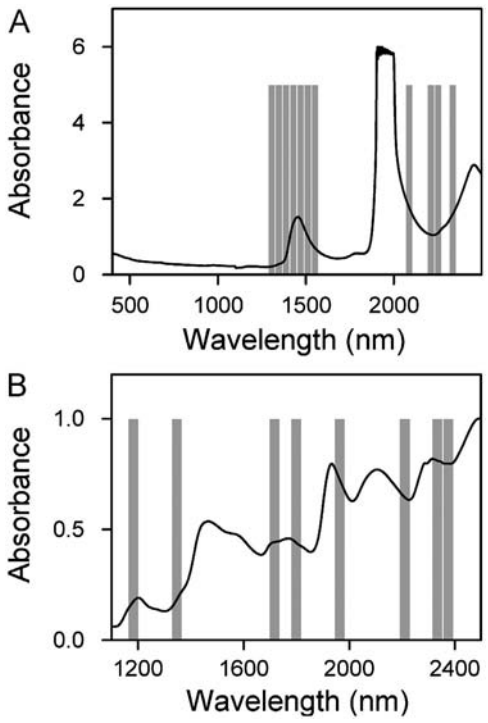

Fig. 4. (A) Selected variables (sensor blocks) in the BRIX data set shown as gray bars. The black solid line is the average training spectrum. (B) Same as (A) for the CORN data set.

Brix units, corresponding to a decrease in REP% from 4.2% to 1.4%. The improvement is confirmed to be significant by applying the randomization test for comparing RMSEPs (i.e., $p \ll 0.05$, see Table 2).

It may be noticed that the number of optimum ACOGASS latent factors is lower than when the full spectral model is applied, as expected from the reduction of spectral regions employed for training and the removal of spectral features which are unrelated with the Brix reference values. Furthermore, although many combinations of pre-processing methods have been tested in ACOGASS, no one was selected. This is in agreement with the features of these samples, which are liquid, so in principle there should be no scattering phenomena causing baseline deviations.

Notice that by visual inspection of the BRIX spectra and removal of the high-absorbance spectral region due to water absorption, PLS processing of the mean-centered resulting data (using 10 latent variables) leads to an RMSEP of 0.45 units for the test set. This value is lower than that for the raw data, although sup-optimal regarding the ACOGASS results (Table 2). We may stress again, however, that intuitive variable selection based on visual inspection of the spectra conspires against the aim of a fully automated process.

4.4. CORN data

This data set is available on the internet, and is intended for calibration of starch and other relevant parameters in corn seeds. The 80-sample set was divided into training (40 samples), monitoring (20 samples) and test (20 samples) at random. As regards the determination of the starch content, cross-validation indicated 17 PLS factors in the full spectral range. This number significantly decreased after variable selection, with a corresponding improvement in figures of merit (Table 2). Fig. 4B shows the regions selected by ACOGASS using the parameters shown in Table 1. As for the case of BRIX data, the reduction in RMSEP was found to be significant ($p \ll 0.05$ in Table 2), from 0.23 to 0.11, corresponding to REP% values of 0.60 and 0.17, respectively.

Notice that MSC was selected for mathematical pre-processing this data set, which is reasonable because in the case of solid samples such as grinded corn, a strong dispersion of the radiation leading to scattering effects is expected.

If full spectral CORN data are processed by applying the common scattering correction method (MSC), a 14-latent variable PLS model leads to an RMSEP of 0.21 units for the test set. This implies some improvement over the value quoted in Table 2, although sup-optimal in comparison with ACOGASS.

5. Conclusions

A new strategy is described for the combined implementation of three of the main optimization methods in partial least-squares calibration: variable, pre-processing and sample selection. It is based on a Monte Carlo procedure including ant colony optimization for variable selection, genetic algorithms for pre-processing selection and two usual sample selection methods. The algorithm has been tested using several sets of samples and the results were satisfactory. All these characteristics imply an innovative strategy based on the use of combined methods in order to obtain a fully optimized partial least-squares calibration.

Acknowledgment

Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project no. PIP 1950), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project no. PICT-2010-0084) are gratefully acknowledged for financial support. F.A. thanks CONICET for a doctoral fellowship.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.talanta.2013.06.051.

References

[1] R.K.H Galvão, M.C.U. Araújo, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, 2009p. 233.
[2] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-infrared Spectroscopy: Principles, Instruments, ApplicationsWiley-VCH, Weinheim, Germany, 2002.
[3] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Anal. Chim. Acta 348 (1997) 71–86.
[4] J.-P. Gauchi, P. Chagnon, Chemom. Intelligent Lab. Syst. 58 (2001) 171–193.
[5] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (1985) 491–500.
[6] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[7] A. Lorber, B.R. Kowalski, J. Chemom. 2 (1988) 67–79.
[8] D.L. Massart, B.G.M Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part AElsevier, Amsterdam, 1997p. 202.
[9] D.L. Massart, B.G.M Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997, p. 109.
[10] O. Devos, L. Duponchel, Chemom. Intelligent Lab. Syst. 107 (2011) 50–58.
[11] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, J. Chemom. 20 (2006) 146–157.
[12] F. Allegrini, A.C. Olivieri, Anal. Chim. Acta 699 (2011) 18–25.
[13] H.C. Goicoechea, A.C. Olivieri, J. Chem. Inf. Comput. Sci. 42 (2002) 1146–1153.
[14] C.E. Boschetti, A.C. Olivieri, J. NIR Spectrosc. 12 (2004) 85–91.
[15] H.C. Goicoechea, A.C. Olivieri, J. Chemom. 17 (2003) 338–345.
[16] N. Sorol, E. Arancibia, S.A. Bortolato, A.C. Olivieri, Chemom. Intelligent Lab. Syst. 102 (2010) 100–109.
[17] R. Leardi, M.B. Seasholtz, R.J. Pell, Anal. Chim. Acta 461 (2002) 189–200.
[18] R. Leardi, A. Lupiáñez González, Chemom. Intelligent Lab. Syst. 41 (1998) 195–207.
[19] J. Ferré, F.X. Rius, Anal. Chem. 68 (1996) 1565–1571.
[20] H.A. Dantas Filho, R.Kawakami Harrop Galvão, M.C. Ugulino Araújo, E.C. Silva, T.C.B. Saldanha, G.E. José, C. Pasquini, I.M. Raimundo Jr., J.J.Rodrigues Rohwedder, Chemom. Intelligent Lab. Syst. 72 (2004) 83–91.
[21] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[22] R.K.H. Galvão, M.C.U. Araujo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Talanta 67 (2005) 736–740.
[23] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193–1202.
[24] ⟨http://www.eigenvector.com/data/Corn/⟩.
[25] MATLAB. The Mathworks Inc., Natick, Massachusetts, USA.
[26] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627–1639.
[27] H. van der Voet, Chemom. Intelligent Lab. Syst. 25 (1994) 313–323.