Revised version

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

Franco Allegrini and 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

* Corresponding author. E-mail: olivieri@iquir-conicet.gov.ar. Telephone: +54-341-4372704.

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.

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

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 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). Figure 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 equation (1). A vector of signal noise (standard deviation = 0.05 units) was then added to each **x** vector after applying equation (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). Figure 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 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 *i*th. sample exceeds a critical value [23]. For calibration samples, F_i is given by:

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

where $y_{\text{nom},i}$ is the nominal concentration for sample i, $y_{\text{pred},i}$ 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]:

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

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 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.

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.

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 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., $p << 0.05$, 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). Figure 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 ($p << 0.05$ 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.

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.

References

[1] R.K.H Galvão, M.C.U. Araújo, Variable selection, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, 2009, p. 233.

[2] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-infrared spectroscopy: Principles, instruments, applications, Wiley-VCH, Weinheim, Germany, 2002.

[3] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry, Anal. Chim. Acta 348 (1997) 71-86.

[4] J.-P. Gauchi, P. Chagnon, Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, Chemom. Intell. Lab. Syst. 58 (2001) 171-193.

[5] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491-500.

[6] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772-777.

[7] A. Lorber, B.R. Kowalski, The effect of interferences and calibration design on accuracy: Implications for sensor and sample selection, J. Chemometrics 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 A, Elsevier, Amsterdam, 1997, p. 202.

[9] Ref. 8, p. 109.

[10] O. Devos, L. Duponchel, Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression, Chemom. Intell. Lab. Syst. 107 (2011) 50-58.

[11] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimization: A powerful tool for wavelength selection, J. Chemometrics 20 (2006) 146-157.

[12] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy / partial least-squares analysis, Anal. Chim. Acta 699 (2011) 18-25.

[13] H.C. Goicoechea, A.C. Olivieri, Wavelength selection for multivariate calibration using a genetic algorithm: a novel initialization strategy, J. Chem. Inf. Comp. Sci. 42 (2002) 1146-1153.

[14] C.E. Boschetti, A.C. Olivieri, A new genetic algorithm applied to the near-infrared analysis of gasolines, J. NIR Spectrosc. 12 (2004) 85-91.

[15] H.C. Goicoechea, A.C. Olivieri, A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy, J. Chemometrics 17 (2003) 338-345.

[16] N. Sorol, E. Arancibia, S.A. Bortolato, A.C. Olivieri, Visible/near infrared - partial least-squares analysis of Brix in sugar cane juice. A test field for variable selection methods, Chemom. Intell. Lab. Syst. 102 (2010) 100-109.

[17] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Anal. Chim. Acta 461 (2002) 189-200.

[18] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195-207.

[19] J. Ferré, F.X. Rius, Selection of the best calibration sample subset for multivariate regression, 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, A strategy for selecting calibration samples for multivariate modeling, Chemom. Intell. Lab. Syst. 72 (2004) 83-91.

[21] R.W. Kennard, L.A. Stone, Computer aided design of experiments, 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, A method for calibration and validation subset partitioning, Talanta 67 (2005) 736-740.

[23] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses, 1. Relation to other quantitative calibration methods and the extraction of qualitative information, 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, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627-1639.

[27] H. van der Voet, Comparing the predictive accuracy of models using a simple randomisation test, Chemom. Intell. Lab. Syst. 25 (1994) 313-323.

Figure captions

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.

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

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.

Revised version

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

Franco Allegrini and 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

* Corresponding author. E-mail: olivieri@iquir-conicet.gov.ar. Telephone: +54-341-4372704.

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.

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

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 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). Figure 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 equation (1). A vector of signal noise (standard deviation = 0.05 units) was then added to each **x** vector after applying equation (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). Figure 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 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 *i*th. sample exceeds a critical value [23]. For calibration samples, F_i is given by:

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

where $y_{\text{nom},i}$ is the nominal concentration for sample i, $y_{\text{pred},i}$ 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]:

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

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 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.

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.

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 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., $p << 0.05$, 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). Figure 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 ($p << 0.05$ 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.

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.

References

[1] R.K.H Galvão, M.C.U. Araújo, Variable selection, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, 2009, p. 233.

[2] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-infrared spectroscopy: Principles, instruments, applications, Wiley-VCH, Weinheim, Germany, 2002.

[3] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry, Anal. Chim. Acta 348 (1997) 71-86.

[4] J.-P. Gauchi, P. Chagnon, Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, Chemom. Intell. Lab. Syst. 58 (2001) 171-193.

[5] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491-500.

[6] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772-777.

[7] A. Lorber, B.R. Kowalski, The effect of interferences and calibration design on accuracy: Implications for sensor and sample selection, J. Chemometrics 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 A, Elsevier, Amsterdam, 1997, p. 202.

[9] Ref. 8, p. 109.

[10] O. Devos, L. Duponchel, Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression, Chemom. Intell. Lab. Syst. 107 (2011) 50-58.

[11] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimization: A powerful tool for wavelength selection, J. Chemometrics 20 (2006) 146-157.

[12] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy / partial least-squares analysis, Anal. Chim. Acta 699 (2011) 18-25.

[13] H.C. Goicoechea, A.C. Olivieri, Wavelength selection for multivariate calibration using a genetic algorithm: a novel initialization strategy, J. Chem. Inf. Comp. Sci. 42 (2002) 1146-1153.

[14] C.E. Boschetti, A.C. Olivieri, A new genetic algorithm applied to the near-infrared analysis of gasolines, J. NIR Spectrosc. 12 (2004) 85-91.

[15] H.C. Goicoechea, A.C. Olivieri, A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy, J. Chemometrics 17 (2003) 338-345.

[16] N. Sorol, E. Arancibia, S.A. Bortolato, A.C. Olivieri, Visible/near infrared - partial least-squares analysis of Brix in sugar cane juice. A test field for variable selection methods, Chemom. Intell. Lab. Syst. 102 (2010) 100-109.

[17] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Anal. Chim. Acta 461 (2002) 189-200.

[18] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195-207.

[19] J. Ferré, F.X. Rius, Selection of the best calibration sample subset for multivariate regression, 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, A strategy for selecting calibration samples for multivariate modeling, Chemom. Intell. Lab. Syst. 72 (2004) 83-91.

[21] R.W. Kennard, L.A. Stone, Computer aided design of experiments, 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, A method for calibration and validation subset partitioning, Talanta 67 (2005) 736-740.

[23] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses, 1. Relation to other quantitative calibration methods and the extraction of qualitative information, 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, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627-1639.

[27] H. van der Voet, Comparing the predictive accuracy of models using a simple randomisation test, Chemom. Intell. Lab. Syst. 25 (1994) 313-323.

Figure captions

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.

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

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.

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]).

Figure 1

Click here to download high resolution image

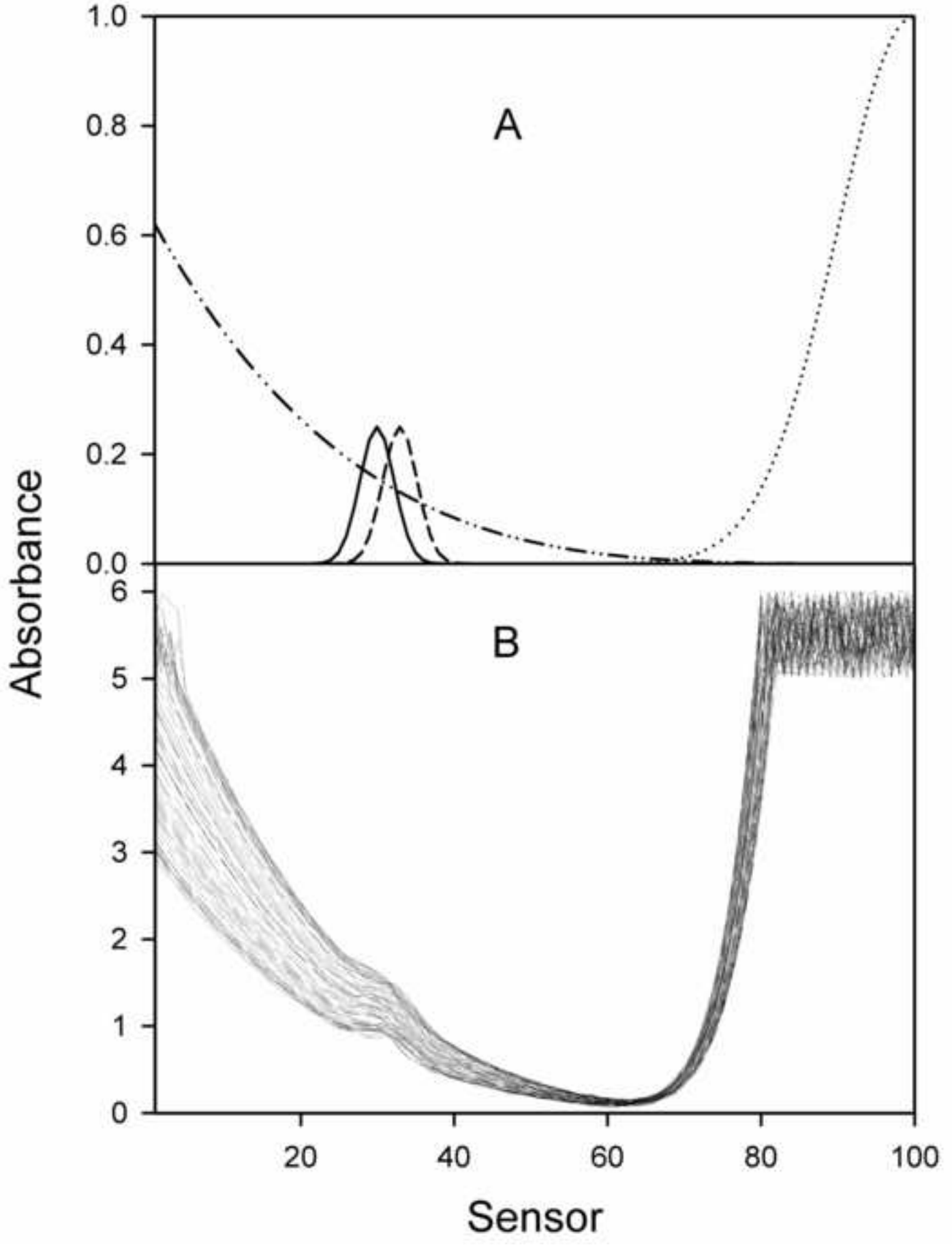


Figure 2

Click here to download high resolution image

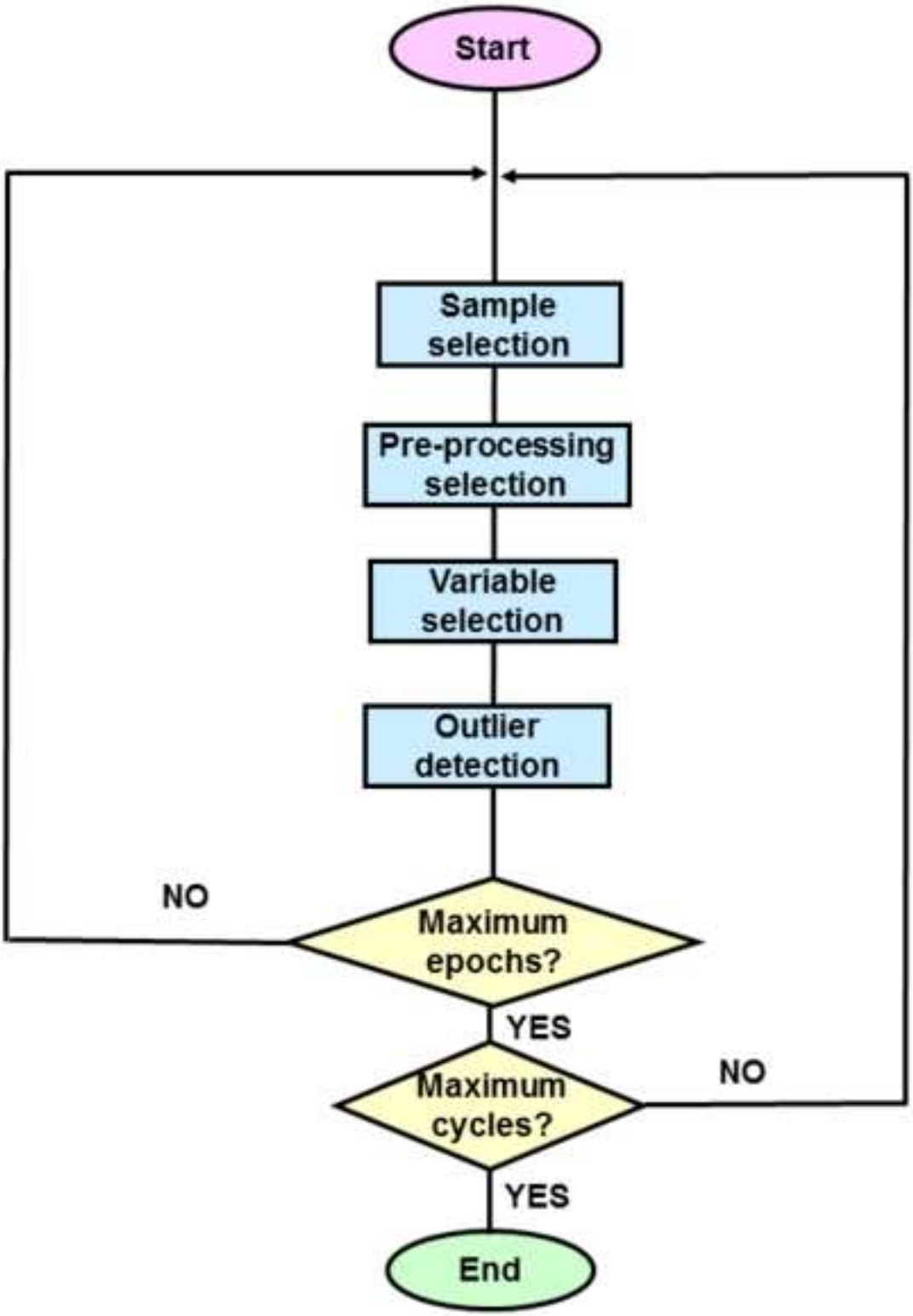


Figure 3

Click here to download high resolution image

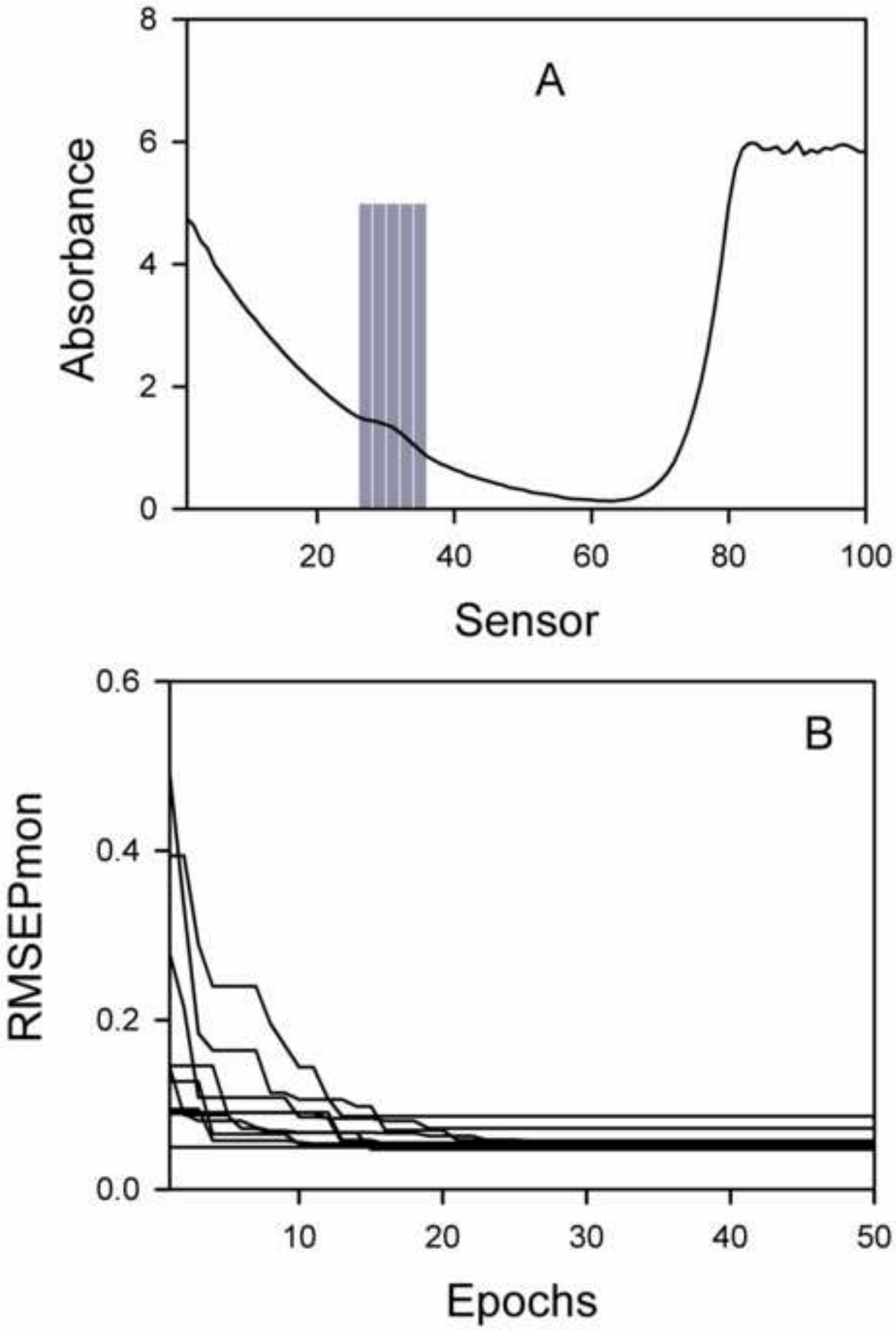
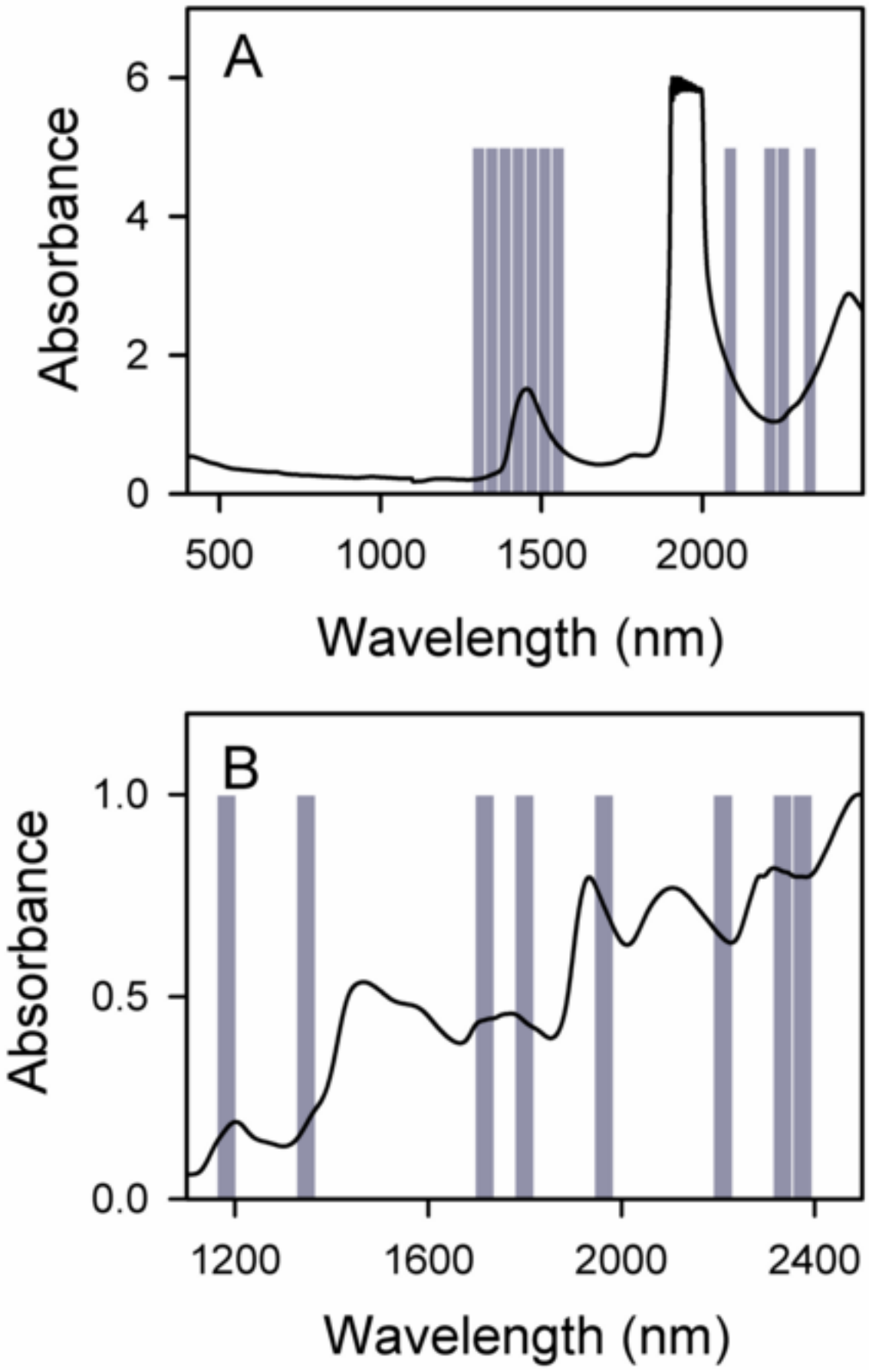


Figure 4

Click here to download high resolution image



INSTRUCTIONS MANUAL

ACOGASS Graphical User Interface

Franco Allegrini, Alejandro C. Olivieri
Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario and Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (S2002LRK), Argentina

INDEX

INTRODUCTION AND GENERAL DESCRIPTION OF THE ALGORITHM

Variable selection intends to rationally choose, from the whole available spectrum, sensors (i.e., 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. 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.

Mathematical pre-processing techniques exist for removing variations in spectra from run to run, which are unrelated to analyte concentration changes. 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 NIR-PLS regression analysis of complex samples (industrially manufactured or naturally occurring), and is intended to provide representativeness to the set of calibration samples. This means that their spectra 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. From a formal point of view, an outlier is a sample which is not representative for the rest of the data In the context of NIR calibrations, the main objective is to identify those samples with features which makes them significantly different from the remaining ones. The most common way to do this is by analyzing either concentration or spectral residuals. In both cases an *F* ratio is calculated as the ratio of squared errors for the sample of interest and the sum of squared errors for the rest of the set. This practical *F* ratio is then statistically compared with critical *F* values.

All the above activities are mutually connected. Pre-processing of spectra modifies by definition the characteristics of the spectral space, and may lead to the selection of different samples for calibration, and also to different selected wavelengths. Changing the spectral regions, in turn, will have an influence in the preprocessing method required to model those specific regions. The process 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.

The ACOGASS algorithm, the basis of the graphical interface to be described, follows the above mentioned process: stochastic algorithms together with classical sample selection methods are implemented in order to find the best compromise between representative calibration samples, preprocessing methods and significant spectral variables for the specific data set to be analyzed.

To achieve the proposed aim, the available set of samples should be divided into initial sets for calibration, monitoring, and validation. The algorithm is then trained leaving out the latter set, which is only used to validate the model at the end of execution. This convergence is reached by successive iterations in which the following steps are performed (see Figure 1):

1) Selection of calibration and monitoring samples by a sample division method (Kennard-Stone PLS o PCA, SPXY). This is optional; you may have your own monitoring set and may not wish to change it.

2) Selection of the optimal preprocessing method using GA (genetic algorithm). This is also optional; you may fix preprocessing methods and do not change them during execution.

3) Variable selection using ACO (ant colony optimization) algorithm.

4) Prediction using the models obtained after steps 1), 2) and 3). The monitoring RMSEPmon is calculated and outliers are removed from the model.

5) Evaluation of the models generated using RMSEPmon as objective function. If it decreases in a certain generation, the preprocessing method, the selected samples and variables are stored and replace the last optimal result.

6) Using the best model obtained in 4), steps 1), 2) and 3) are repeated as many times as generations have been included in each cycle.

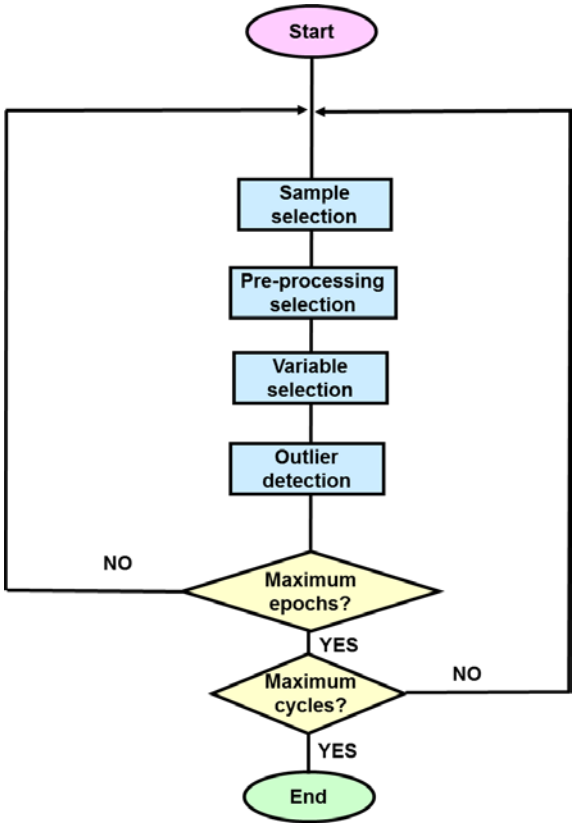


Figure 1. ACOGASS flow sheet.

MAIN WINDOW

To open the GUI, once the proper folder has been incorporated into the MATLAB path or selected as the actual folder, write "acogass" in the command window and press "ENTER". After this, the initial window will appear (see Figure 2). This window is divided into four panels, which are described below.

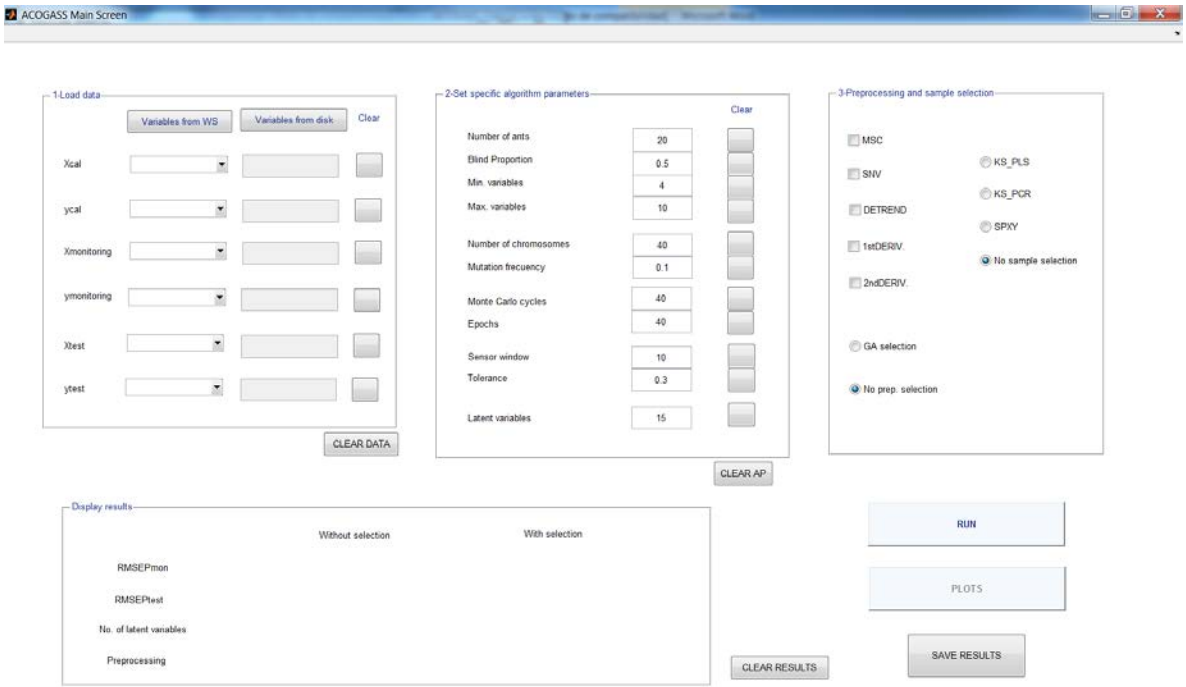


Figure 2. Main ACOGASS screen.

Panel 1: Load data.

In this panel, the data matrices as well as the vectors with reference values are introduced. There are two options to do this:

1) Select variables which have been previously loaded in the MATLAB workspace. In this case, use the button "Variables from WS" (Figure 3). The user may choose the name of the appropriate files from the drop-down menus.

2) Introduce the name of the files, in case they are in text format. To do this, click over the "Variables Manually" button (Figure 3). Notice that Xmonitoring, ymonitoring, Xtest and ytest may be optionally absent. However, Xcal and ycal are mandatory.

The correct sizes of these matrices and vectors are: for signal matrices (Xcal, Xmonit and Xtest), (sensors × samples), while for concentration (or property) vectors (ycal, ymonit and ytest), (samples × 1).

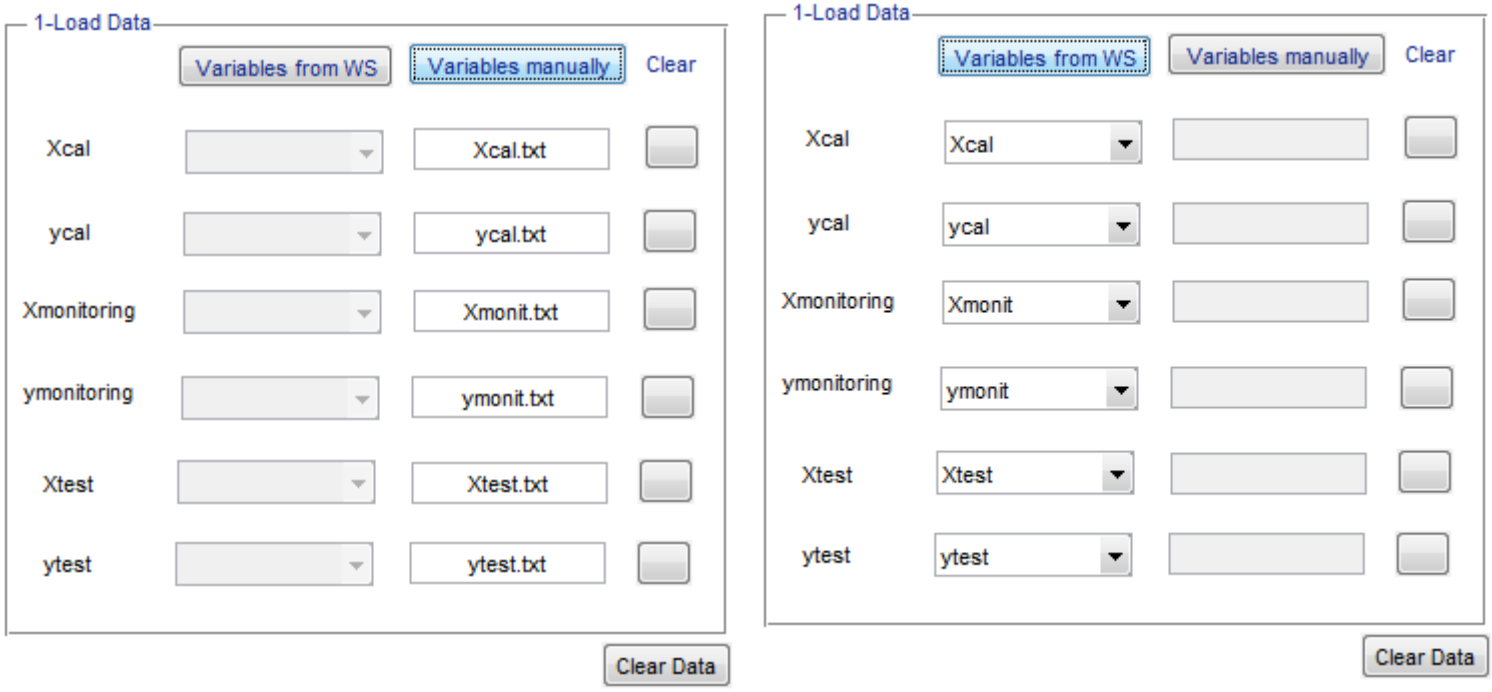


Figure 3. Load data panel. The variables can be loaded manually or from the MATLAB workspace.

Panel 2: Set specific ACO and GA parameters to select variables and optimal preprocessing methods respectively.

Brief description of the ACO algorithm

Ant colony optimization is a stochastic searching method inspired on the behavior of ants to find their food. During this process, they walk around the nest and when they find a food source, they start moving between both points, leaving a certain amount of pheromone in the trail, which works as an attraction signal for the rest of the population. This generates an evaporable trail which will be stronger as long as the path chosen is shorter and faster (optimal path), and it will disappear if this path is long and slow (less favorable path). Extending these concepts to the computational field, ants will work as searching agents in a space given by the total spectral sensors.

In successive epochs, each ant selects a certain number of variables that can vary withiin a range fixed by the user. After the variables are selected, the RMSEPmon is calculated for the monitoring set. According to the result obtained by each ant, a pheromone vector will be marked with more or less content in each of the selected variables. During the following epochs, those positions which show higher amounts of pheromone (meaning lower values of RMSEPmon), will have a larger chance of being selected than those having a lower amount of pheromone (coming from higher values of RMSEPmon).

The process of: 1) selecting a certain number of sensors, 2) evaluating the RMSEPmon for each ant, and 3) marking the corresponding position according to the result obtained in 2), is

repeated a certain number of epochs and Monte Carlo cycles. Finally, with the variables chosen in each cycle, a normalized histogram is generated. Positions with values over a certain tolerance limit will be selected and those with values below this limit will be discarded.

Brief description of genetic algorithms

Genetic algorithms constitute the most widely used variable selection strategy to optimize PLS calibrations. They are inspired in cross-over and mutation mechanisms performed by chromosomes, and as ACO, their nature is stochastic.

In the particular case of ACOGASS, each chromosome consists of a combination of particular preprocessing methods.

Parameters to be set

Considering the previous description, the parameters which can be configured by the user are:

Number of ants: number of artificial agents who select variables and construct different PLS models to be evaluated according their RMSEPmon.

Blind proportion: percentage of total ants that, from one generation to the next, will not use the pheromone vector as a guide to select variables. The larger this parameter is, the more disperse the search will be, but at the same time the probability of reaching a local minimum will also be lower.

Minimum number of variables to be selected (*Min. variables*): minimum number of variables (blocks of sensors) that each ant can select in a generation.

Maximum number of variables to be selected (*Max. variables*): maximum number of variables (blocks of sensors) that each ant can select in a generation. Notice that the product of *Max. variables* × *Spectral window* should not exceed the total number of sensors in the whole spectrum.

Number of chromosomes: number of chromosomes used to choose the optimal preprocessing methods. Considering the structure of the genetic algorithm employed, this number must be divisible by 4.

Mutation frequency: frequency of appearance of point changes in the chromosomes involved in the selection of the preprocessing method.

Cycles: number of times that the complete computation is repeated.

Epochs: number of times that the three steps of the ACO algorithm is repeated.

Sensor window: number of sensors included in each sensor block or variable. The selection is not performed on a sensor by sensor basis, but considering a group of contiguous sensors. The number of sensors is selected in such a way that their width is not larger than the width of the thinnest spectral band.

Tolerance: value that must be exceeded in the histogram to consider a variable as significant. Variables with values below the tolerance limit will not be included in the final model.

Maximum number of components *(Latent variables)*: maximum number of PLS factors used by the algorithm in the PLS predictions.

In all cases cited previously, a default value is set, to allow inexperienced users to use the algorithm.

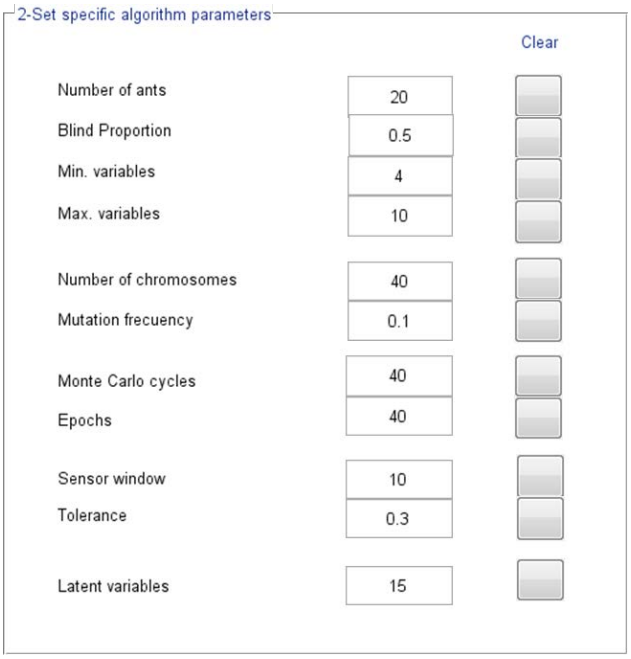


Figure 4. Panel with specific parameters to be introduced both in GA and ACO.

Panel 3: Sample selection method and preprocessing options.

In this panel two groups of options can be found:

1) Preprocessing options. The following alternatives are available:

- Use GA to select preprocessing (*GA selection*).

In this case it is not convenient to select a preprocessing method with anticipation, as the algorithm will evaluate the different options. It is important to notice that when using this configuration, the time of calculation significantly increases.

- Do not use GA to select preprocessing (*No prep. selection*).

The algorithm will only use ACO to select variables without including the preprocessing search step. In this situation, it is convenient to take advantage of the option boxes to choose the desired combination of preprocessing methods before starting the variable selection.

2) Sample selection options. The alternatives are:

KS_PLS (Kennard-Stone with PLS scores): in each generation, the samples of calibration and monitoring are gathered and then divided according to the selected variables and the combination of preprocessing methods. This is done by the traditional Kennard-Stone method, but based on PLS scores instead of PCA scores.

KS_PCR (Kennard Stone with PCR scores): the same as KS_PLS but using PCA scores (traditional Kennard Stone method).

SPXY (Sample Set partitioning based on the joint X-Y distance): this method works in a similar way to KS_PLS, but instead of using latent variables (scores) to classify samples, it utilizes the instrumental data matrix and the reference values, and calculates distances considering both parameters at the same time.

No sample selection: in this case the step consisting on joining the samples and dividing them every iteration is not applied, and the algorithm only focuses in selecting variables and preprocessing methods based on the calibration and monitoring matrices unchanged. If no monitoring set of samples is provided, the program divides the calibration set at random (2/3 for training and 1/3 for monitoring).

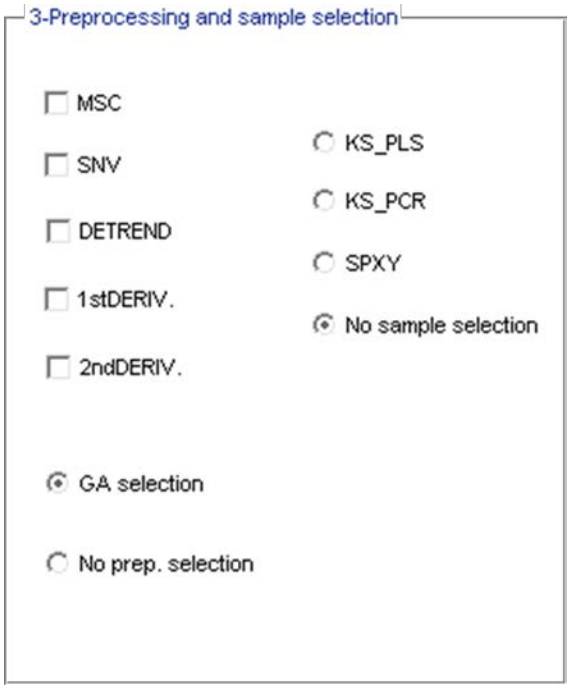
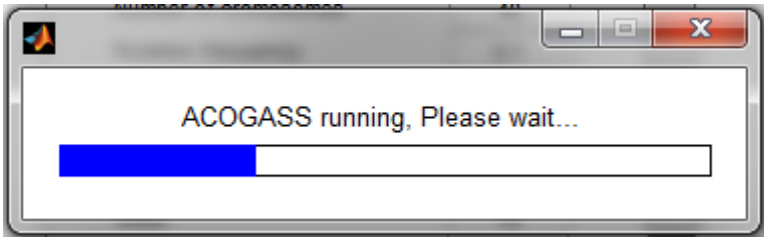


Figure 5. Sample selection options and preprocessing methods.

Panel 4: Results.

This panel will be empty until the algorithm has finished processing the data, action which begins by pressing the button “RUN”.

While the program is running, the progress is indicated by a *progress bar*:



Once the searching process finishes, the following results will appear:

- *RMSEPmon*: prediction error calculated over monitoring samples.

- *RMSEPtest*: prediction error calculated over test samples, using leave-one-out cross validation to determine the number of factors.

- *Number of latent variables*: number of explanatory latent variables that have been selected by cross-validation.

- P*reprocessing*: the result consists on a chain of ones and zeros, in which 0 indicates that the preprocessing corresponding to that position has not been selected, otherwise a 1 appears. The positions show the following relation with preprocessing:

1- MSC (Multiplicative scattering correction)

2- SNV (Standard normal variate)

3- Detrend

4- First derivative

5- Second derivative.

As an example, if the result obtained after running the algorithm is: 1 0 1 1 0, this indicates that the preprocessing methods selected are: MSC, Detrend, and 1st. Derivative.

Display results

	No Variable Selection	Variable Selection
RMSEPmon	0.28158	0.022948
RMSEPtest	0.29577	0.026382
No. of latent variables	1	2
Preprocessing	0 0 0 0 0	0 0 1 0 0

Figure 6. Results panel. It is presented in the form of a comparative table between the main figures of merit obtained after applying the ACOGASS algorithm, and without using it (i.e., using only raw spectra).

The MATLAB workspace shows a more complete report:

****** MONITORING SET RESULTS ******

After selection

Pre-processing: 0 0 0 0 0
RMSEmon: 0.021324
REPmon%: 4.3845
R2mon: 0.99826
Latent variables: 4
Monitoring outliers: 0
Calibration outliers: 0

No selection

RMSEmon: 0.30767
REPmon%: 63.2617
R2mon: 0.44099
Latent variables: 4
Monitoring outliers: 0
Calibration outliers: 0

****** TEST SET RESULTS ******
After selection

Pre-processing: 0 0 0 0 0
RMSECV: 0.022164
RMSEtest: 0.026954
REPtest%: 5.542
R2test: 0.99618
Latent variables: 4
Test outliers: 1

No selection

RMSECV: 0.27083
Biascv: -0.00099614
RMSEtest: 0.29836
REPtest%: 61.3474
Biastest: 0.03035
R2test: 0.079191
Latent variables: 1
Test outliers: 0

CLEAR BUTTONS

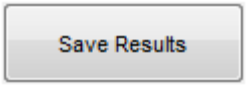
These buttons allow the user to:

- Clear the complete content of each panel individually (*Clear Data, Clear AP and Clear Results*).

- In panels 1 and 2, clear one by one the internal options of each panel (*Clear*).

SAVE RESULTS

The most important parameters resulting from applying the algorithm can be saved by pressing the *Save Results* button:



This will save the following variables in a "results.mat" file:

- Selected variables (variables_selected), in a vector with 1s and 0s, on a sensor-by-sensor basis (1 = included sensor; 2 = discarded sensor).

- RMSEPtest (*RMSEP_1, RMSEP_2*).

- RMSECV *(RMSECV_1, RMSECV_2).*

- Number of components (*number_of_components_1, number_of_components_2*).

The numbers at the end of each figure of merit indicates the following:

1) figure of merit obtained after applying the selection algorithm.

2) figure of merit obtained without selection.

PLOTS WINDOW

This window is called using the PLOTS button, once the run has finished.

The following plots are presented:

- Bar plot showing in blue the spectral zones which are over the tolerance limit (this means selected variables), and in red the ones which are below this limit (unselected variables),

- Change of RMSEPmon with generations and cycles.

- PRESS plot as a function of the number of components for leave-one-out cross validation, (performed over the complete spectral data and over the selected data).

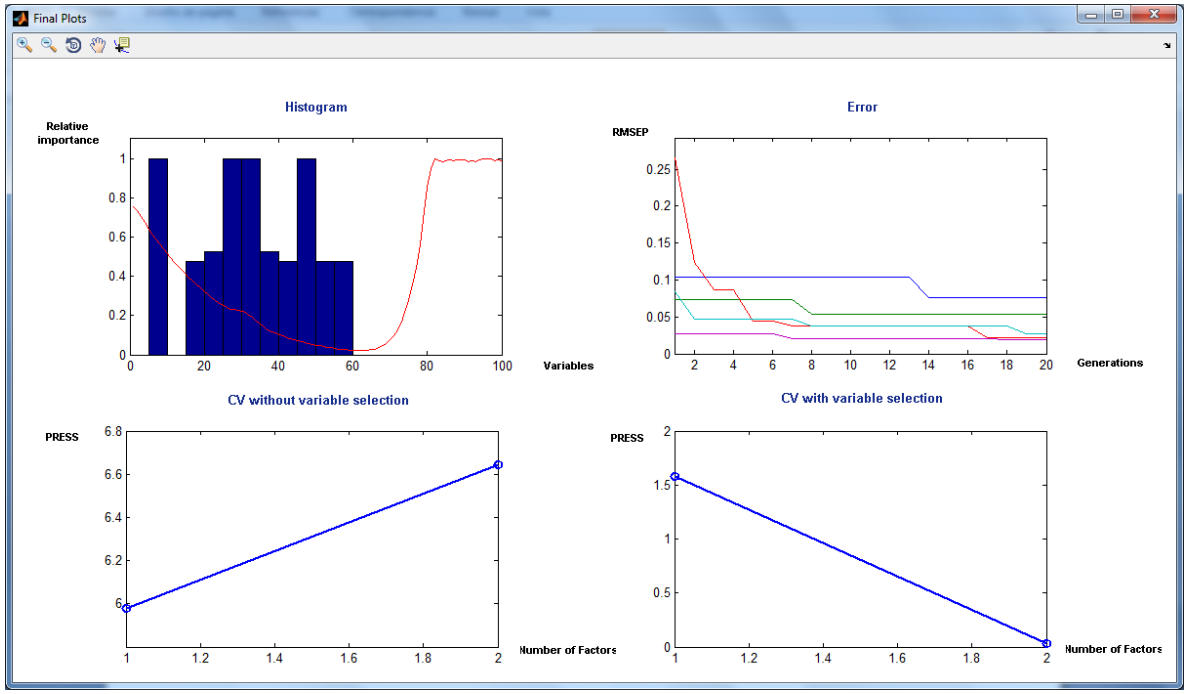


Figure 7. Plot window obtained for the example data.

EXAMPLE DATA

An example data set is provided with the program. It consists of six ASCII files: (1) calibration signals Xcal.txt, (2) calibration concentrations ycal.txt, (3) monitoring signals Xmonit.txt, (4) monitoring concentrations ymonit.txt, (5) test signals Xtest.txt, and (6) test concentrations ytest.txt.

These data can be processed with the parameters of Figure 4, with the results shown in Figure 7. The data contain a saturated region above sensor 80, a useful region in the range of sensors 20-40, and a variable, non-linear background signal. The results (Figure 7) shows that the useful region is selected, and the saturated region is avoided, while Figure 6 shows that suitable pre-processing is required to achieve a reasonably small RMSEPtest, in comparison with full spectral results.

***List of Three Potential Reviewers**

We suggest the following possible reviewers for this paper:

1) Prof. Ricardo Leardi, University of Genova, Italy, e-mail: riclea@dictfa.unige.it
2) Prof. Hai-Long Wu, Hunan University, China, e-mail: hlwu@hnu.cn
3) Prof. Ronei Poppi, University of Campinas, Brazil, e-mail: ronei@iqm.unicamp.br

Graphical Abstract