ELSEVIER

Chemometrics and Intelligent Laboratory Systems 57 (2001) 65–73

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

The successive projections algorithm for variable selection in spectroscopic multicomponent analysis

Mário César Ugulino Araújo[*], Teresa Cristina Bezerra Saldanha,
Roberto Kawakami Harrop Galvão[1], Takashi Yoneyama[1], Henrique Caldas Chame,
Valeria Visani

Departamento de Química, Universidade Federal da Paraíba, CCEN, Caixa Postal 5093, CEP 58051-970-João Pessoa, PB, Brazil

Received 11 September 2000; accepted 20 January 2001

Abstract

The “Successive Projections Algorithm”, a forward selection method which uses simple operations in a vector space to minimize variable collinearity, is proposed as a novel variable selection strategy for multivariate calibration. The algorithm was applied to UV–VIS spectrophotometric data for simultaneous analysis of complexes of Co^{2+}, Cu^{2+}, Mn^{2+}, Ni^{2+} e Zn^{2+} with 4-(2-piridilazo)resorcinol in samples containing the analytes in the 0.02–0.5 mg l^{-1} concentration range. A convenient spectral window was first chosen by a procedure also proposed here and applying Successive Projections Algorithm to this range allowed an improvement of the predictive capabilities of Principal Component Regression, Partial Least Squares and Multiple Linear Regression models using only 20% of the number of wavelengths. Successive Projections Algorithm selection resulted in a root mean square error of prediction at the test set of 0.02 mg l^{-1}, while the best and worst realizations of a genetic algorithm used for comparison yielded 0.01 and 0.03 mg l^{-1}. However, genetic algorithm took 200 times longer than Successive Projections Algorithm, and this ratio tends to increase dramatically with the number of wavelengths employed. Finally, unlike genetic algorithm, Successive Projections Algorithm is a deterministic search technique whose results are reproducible and it is more robust with respect to the choice of the validation set. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Successive projections algorithm; Variable selection; Multicomponent analysis; UV–VIS spectrophotometry; Multivariate calibration

1. Introduction

Nowadays, the majority of spectrophotometric quantitative analysis of multicomponent systems are made by using multivariate calibration methods allowing simultaneous determination of several analytes. Among these, the most used are multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) [1]. The application of these methods to spectrophotometric multicomponent simultaneous analysis usually requires spectral variable selection for building well-fitted models [2]. MLR yields models, which are simpler and easier to interpret than PCR and PLS, since these calibration techniques perform regression on latent

[*] Corresponding author. Fax: +55-83-216-7437.
E-mail address: laqa@quimica.ufpb.br (M.C.U. Araújo).

[1] Electronics Engineering Division, Technological Institute of Aeronautics, São José dos Campos, SP, 12228-900, Brazil.

0169-7439/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(01)00119-8

variables, which do not have physical meaning. On the other hand, MLR is more dependent on a good choice of spectral variables.

Selecting from the full spectrum the wavelengths that result in the maximum accuracy is still a challenging task, mainly when spectra display strong overlapping and have imperceptible distinctive features, as is the case with UV–VIS spectrophotometry. For overcoming this, several approaches have been proposed to select optimal sets of variables for multivariate calibration, such as the “branch and bound” algorithm commonly applied in combinatorial optimization [3], the least condition number of the calibration matrix [4], generalized simulated annealing [5], incertitude modelling [6], genetic algorithms [7–10], artificial noise introduction in PLS modelling [11], analysis of weights resulting from MLR [10], hybrid linear analysis [12], hierarchical multiblock PLS models [13], cyclic subspace regression [14], artificial neural networks [15,16], wavelet transform [17,18], iterative predictor weighting PLS [19], discriminant partial least squares [20]. Among these different variable selection strategies, genetic algorithms (GA) are an interesting, flexible and widely used alternative. GA are guided random search techniques inspired on natural selection mechanisms, which explore the solution space in an efficient manner and are suitable for parallel processing implementations. However, due to their stochastic nature, results are realization-dependent and variable selections may not be reproducible.

In this work, the “Successive Projections Algorithm” (SPA) is proposed as a novel variable selection strategy for multivariate calibration. SPA is compared with a genetic algorithm using visible spectrophotometric data for simultaneous analysis of complexes of Co^{2+}, Cu^{2+}, Mn^{2+}, Ni^{2+} and Zn^{2+} with 4-(2-piridilazo)resorcinol (PAR) in mixtures containing the analytes in the concentration range of 0.02–0.5 mg l^{-1}.

2. Background and theory

2.1. Notation

Matrices and linear operators are represented by bold capital letters, column vectors by bold lowercase letters, and scalars by italic characters. Elements of a sequence are denoted by italic characters with an index between parenthesis. The superscript T means transposed. Let $k(n)$ be the wavelength selected at the nth iteration of SPA. Let $\mathbf{X}_{cal}$ be the matrix $(M_{cal} \times J)$ of instrumental response data (independent variables) for M_{cal} calibration mixtures and J wavelengths. Let $\mathbf{Y}_{cal}$ be the matrix $(M_{cal} \times A)$ of calibration concentrations (dependent variables) for A analytes and $\mathbf{Y}_{test}$ be the matrix $(M_{test} \times A)$ of concentrations for the M_{test} test mixtures. The hat symbol ($\hat{}$) indicates a predicted value.

2.2. Description of the successive projections algorithm

SPA is a forward selection method [1], that is, it starts with one wavelength, then incorporates a new one at each iteration, until a specified number N of wavelengths is reached. Its purpose is to select wavelengths whose information content is minimally redundant, in order to solve collinearity problems. SPA steps are described below, assuming that the first wavelength $k(0)$ and the number N are given.

Step 0: Before the first iteration $(n = 1)$, let $\boldsymbol{x}_j =$ jth column of $\mathbf{X}_{cal}$; $j = 1, \ldots, J$.

Step 1: Let S be the set of wavelengths which have not been selected yet. That is, $S = \{j \text{ such that } 1 \le j \le J \text{ and } j \notin \{k(0), \ldots, k(n-1)\}\}$.

Step 2: Calculate the projection of $\boldsymbol{x}_j$ on the subspace orthogonal to $\boldsymbol{x}_{k(n-1)}$ as

$$\mathbf{P}\mathbf{x}_j = \overbrace{\mathbf{x}_j -}^{\text{obtém vetor ortogonal à projeção de xj na direção de xk(n-1)}} \underbrace{\left(\mathbf{x}_j^{T}\mathbf{x}_{k(n-1)}\right)\mathbf{x}_{k(n-1)}\left(\mathbf{x}_{k(n-1)}^{T}\mathbf{x}_{k(n-1)}\right)^{-1}}_{\text{projeção de xj na direção de xk(n-1)}} \quad (1)$$

for all $j \in S$, where $\mathbf{P}$ is the projection operator.

Step 3: Let $k(n) = \arg(\max \|\mathbf{P}\boldsymbol{x}_j\|, j \in S)$.

Step 4: Let $\boldsymbol{x}_j = \mathbf{P}\boldsymbol{x}_j$, $j \in S$.

Step 5: Let $n = n + 1$. If $n < N$ go back to Step 1.

End: The resulting wavelengths are $\{k(n); n = 0, \ldots, N-1\}$.

The number of projection operations performed in the selection process can be shown to be $(N-1)(J - N/2)$. The above steps are exemplified in Fig. 1, which illustrates the first iteration of SPA.

Remark that, although SPA may resemble the Gram–Schmidt orthogonalization procedure [21], these algorithms have different purposes. Gram–

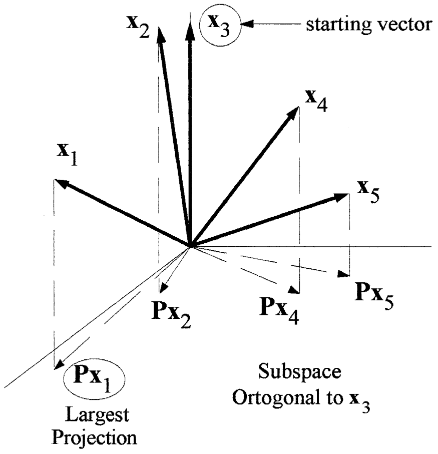

Fig. 1. Example of SPA with $J = 5$, $M_{cal} = 3$ and $k(0) = 3$. Result of first iteration: $k(1) = 1$.

Schmidt manipulates data in order to generate a new set of orthogonal vectors, which, in the general case, do not have physical meaning. SPA, on the contrary, does not modify the original data vectors, since projections are used only for selection purposes. Thus, the relation between spectral variables and data vectors is preserved.

If N and $k(0)$ are not known a priori, then the following criterion can be used to find optimum values N^* and $k^*(0)$.

não diz que é ótimo

```
Specify a set of validation mixtures.
Specify $N_{min}$ and $N_{max}$ (minimum and maximum values to be sought for $N^*$).
For $N = N_{min}$ to $N_{max}$ do
    For initial = 1 until $J$ do
        -Using Steps 0–5 above, select $N$ wavelengths starting from $k(0) =$ initial.
        -Build a MLR calibration model using the selected wavelengths.
        -Use the model to predict the concentrations of the validation set.
        -Calculate the root mean square error (RMSE) as
```

$$\text{RMSECV} = \sqrt{\frac{1}{AM_{cal}} \sum_{i=1}^{M_{cal}} \sum_{j=1}^{A} \left[\hat{Y}_{cal}(i,j) - Y_{cal}(i,j)\right]^2} \tag{2}$$

```
        if cross validation is the method used (internal validation), where the validation set is composed of the calibration mixtures themselves or as
```

$$\text{RMSEP} = \sqrt{\frac{1}{AM_{test}} \sum_{i=1}^{M_{test}} \sum_{j=1}^{A} \left[\hat{Y}_{test}(i,j) - Y_{test}(i,j)\right]^2} \tag{3}$$

```
        if a test set which was not employed in the calibration phase is used in validation (external validation).
        -Let $\rho(\text{initial}) =$ RMSE.
    Next initial
    Let $r(N) = \min[\rho(\text{initial})]$, initial $= 1, \dots, J$.
    Let $s(N) = \arg[\min \rho(\text{initial})]$, initial $= 1, \dots, J$.
Next $N$
Let $N^* = \arg(\min r(N),\ N = N_{min}, \dots, N_{max})$.
Let $k^*(0) = s(N^*)$.
```

número de amostras de calibração

Notice that: número de variáveis

- $N_{min} \geq A$, because using a number of wavelengths smaller than the number of analytes is not recommended in multivariate calibration [22];
- $N_{max} \leq M_{cal}$, because this is the maximum number of wavelengths that can be selected by SPA, since, after that, all projections become a point with null dimension;
- the expressions for RMSECV and RMSEP use an overall prediction error for all chemical components being modeled.

2.3. A comparison of SPA with guided random search techniques (GRSA)

GRSA, such as simulated annealing [5] and genetic algorithms [7–10], are generally employed when the gradient of the function to be optimized with respect to its parameters is unknown and exhaustive search is not feasible. These techniques try to explore the solution space in an efficient manner, usually incorporating some kind of random factor (for instance, mutations in GA) to avoid being captured by local minima.

Given enough computation time, a GRSA will eventually match and possibly surpass the SPA solution. However, its stochastic nature makes it difficult to estimate the number of iterations which would be required.

For the sake of simplicity, assume that a GRSA is expected to yield an acceptable solution after evaluating the objective function at aP points, for a fixed $a < 1$, where P is the number of possible solutions in the search space. If the problem consists of selecting N variables out of J, then

$$P = \binom{J}{N} = \frac{J(J-1)\cdots(J-N+1)}{N!} \tag{4}$$

and thus, the time spent by the GRSA, for a fixed N, will increase with J^N.

On the other hand, the time spent by SPA, if the starting point is optimized according to the previously described criterion, will have a component proportional to $J(N-1)(J-N/2)$ (J selections of N wavelengths) and another proportional to J (J calibrations). Thus, SPA computing time, for a fixed N, increases only with J^2. For instance, if a selection of $N = 10$ wavelengths is desired and the total number of wavelengths J increases 10 times, the time spent by SPA will increase 100 times, while the time expected for the GRSA to yield an acceptable solution will increase 10^{10} times.

Apart from computational workload, a qualitative comparison between SPA and GRSA should also be made. The performance of GRSA is highly dependent on the choice of the objective function, which could follow two approaches.

(A) Using only the matrix of instrumental responses. In this case, selection is aimed at minimizing some measure of collinearity, such as variance inflation factor or condition number [23,24].
(B) Using the matrices of instrumental responses and concentrations. In this case, selection quality can be assessed by evaluating prediction errors by a validation set.

Both alternatives have intrinsic limitations.

(A) The set of least-correlated variables is not always the one that yields the best models. For instance, particularly noisy variables, though weakly correlated, may not contain useful information for calibration purposes.
(B) If the set employed for validation is not representative, large errors may appear when the resulting models are tested in a different data set. Also, collinearity problems are considered only indirectly in the selection process.

SPA may be regarded as a compromise between approaches (A) and (B). By construction, the algorithm tries to minimize collinearity, while the problem of useful information content may be tackled by the use of a RMSEP criterion for choosing the starting point and the best number of wavelengths.

A GRSA could attempt to overcome this problem by combining criteria (A) and (B). However, this requires a judicious choice of a weight parameter, which would be an additional problem for the analyst.

3. Experimental

3.1. Apparatus

Zero-order and first-derivative absorption spectra were recorded with a model 8453 Hewlett-Packard diode array UV–VIS spectrophotometer, using 1.00 cm quartz cells and 1-s integration time. The instrument operates from 190 to 1100 nm, with 1-nm resolution. From the full spectrum, a range pre-selection was made based on the maximum differences between the spectra of the individual complexes Metal-PAR. For recording, the individual spectrum of each analyte, the concentration of each metal was fixed in 0.5 mg l^{-1}, a value that falls within the linear concentration range of the five metallic cations of interest.

3.2. Range pre-selection procedure

Pre-selection follows these steps: (a) the absorption spectra of the individual complexes are recorded; (b) the differences between each pair of spectra are calculated; (c) first order derivatives of these differences are then obtained; (d) finally, wavelengths for which the derivative does not exceed 10% of the absolute value of the derivative maximum in all pairs

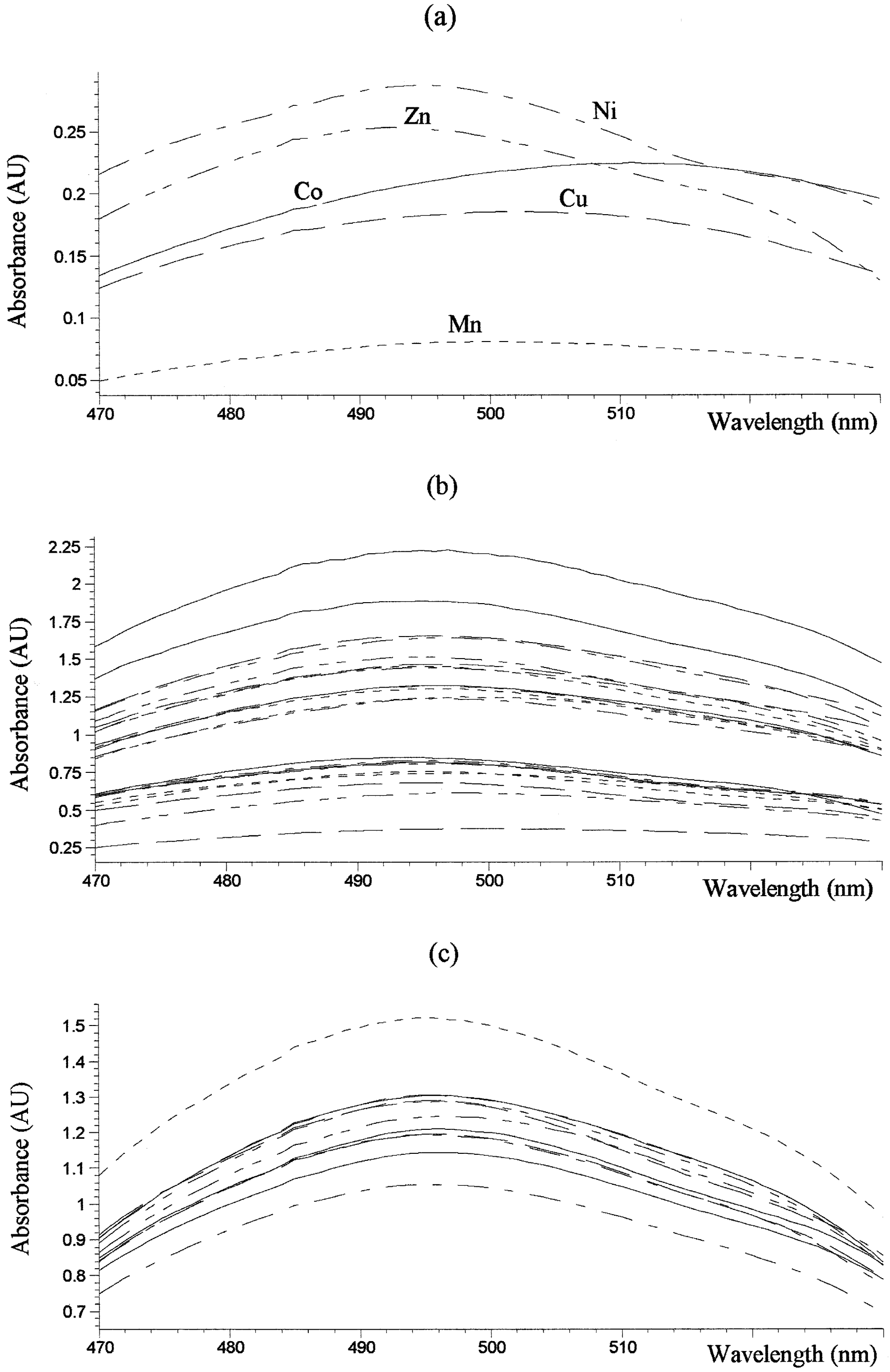

Fig. 2. VIS spectra of individual metal-PAR complexes with a concentration of 0.5 mg l^{-1} (a), of calibration mixtures (b) and of test mixtures (c).

are eliminated. This criterion was met by 61 wavelengths, from 470 to 530 nm. Fig. 2a shows the spectral profiles of the individual chelates in this range.

3.3. Reagents

The PAR stock solution (5.0×10^{-3} mol l^{-1}) was prepared from the sodium salt monohydrate-$C_{11}H_8N_3Na.H_2O$ (Merck) in an ammonium medium (NH_3 0.05 mol l^{-1}), where it is stable for about 1 month [25]. The solution was kept in a polyethylene flask covered with aluminum foil. The pH 9.0 buffer solution was prepared by 1:5 dilution of a 0.05 mol l^{-1} borax solution [4]. Analyte stock solutions (1.000 g l^{-1}) were prepared by diluting Tritrisol (Merck) or Fixanal (Riedel de Haën) ampoules as recommended by the manufacturer. The standard solutions of each complex and the calibration and test mixtures were prepared by diluting the analyte and PAR stock solutions. The reagents were added in the order metal solution–buffer solution–PAR, and then the volume was completed with the buffer solution. PAR concentration (in mol l^{-1}) in all mixtures was set equal to three times the sum of the metal concentrations in the most concentrated calibration mixture, that is, $[\mathrm{PAR}]/\Sigma[\mathrm{Metals}] \geq 3$. A solution containing the buffer and PAR in the same proportion as the mixtures was used as the blank. Analytical grade reagents and water purified by a Milli-Q (Millipore) system were used throughout.

3.4. Procedure

The linear working ranges were the following: 0.02–0.20 mg l^{-1} Co^{2+}, 0.03–0.30 mg l^{-1} Mn^{2+} and 0.05–0.50 Cu^{2+}, Ni^{2+} and Zn^{2+}. The compositions of the calibration mixtures were selected according to a 2^{5-1} fractional factorial design [26]. For model assessment, it was used a set of 12 test mixtures, with concentrations randomly picked within the calibration range. As seen in Fig. 2b and c, the spectra of calibration and test mixtures display high collinearity even after the pre-selection procedure.

3.5. Software

Version 6.1 of the Unscrambler chemometrics software (CAMO A/S) was used for specifying the factorial design and for PLS and PCR calculations [27]. SPA and GA selections, as well as MLR calibration, were performed with homemade programs written in MATLAB 4.2.c.1 [28]. The GA had the following features [29]: The number of wavelengths to be selected was provided by the user; solutions were coded in binary chromosomes with J genes (a "1" gene marks a selected wavelength); the probability of a given chromosome being selected for the mating pool is proportional to its fitness; population size is kept fixed (old generation is completely replaced by the new one); elitism was not employed; each generation is composed of 100 chromosomes; each realization has 100 generations; crossover consists of exchanging genes whose position is selected randomly, but with the restriction that the total number of ones in either chromosome must be kept constant (otherwise, the number of selected wavelengths would vary); crossover probability = 60%; mutation consists of flipping one "1" gene and one "0" gene randomly selected in a chromosome and mutation probability = 10%.

4. Results and discussion

Before proceeding with wavelength selection, a preliminary calculus using the pre-selected spectral range for multivariate calibration was performed. In this case, MLR is only feasible if QR decomposition [30] is performed on the calibration matrix, in order to stabilize the pseudo-inverse calculus (yielding a RMSEP of 0.07 mg l^{-1}). If MLR without QR decomposition is attempted, absurd errors are obtained (RMSEP of 10 mg l^{-1}).

Prior to the application of SPA, each column of $\mathbf{X}_{\mathrm{cal}}$ was mean-centered and auto-scaled by its standard deviation. In the present application, this procedure showed to be advantageous, since all variables had approximately the same signal-to-noise level. Also, by forcing all vectors $\boldsymbol{x}_j$ to have the same norm, the selection process will focus on the angles between them, which are directly related to collinearity problems. The choice of N and the starting point for SPA were based on the RMSEP (that is, the test samples are not employed in the calibration). From Fig. 3, the best number of wavelengths can be seen to be 14.

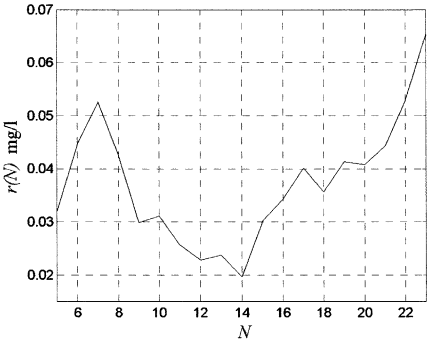

Fig. 3. Choice of the best number of wavelengths (N), according to the RMSEP criterion.

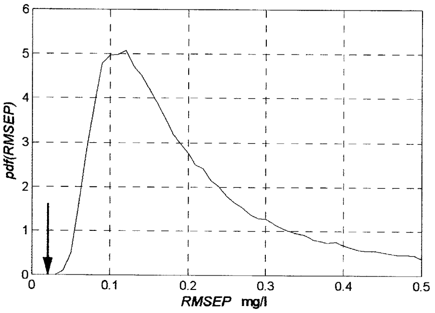

Fig. 4. Estimated *pdf* of the RMSEP yielded by random selections of 14 wavelengths. The arrow marks the RMSEP resulting from the SPA selection.

The best starting point for $N = 14$ was $k^*(0) = 22$ which corresponds to 491 nm. The selected wavelengths, in crescent order, are indicated in Table 1 (remark that $k(n)$ were not obtained in that order), resulting in RMSEP = 0.02 mg l^{-1}. Remark that the same RMSEP is obtained, whether QR decomposition is employed or not.

To have a raw notion of the quality of the SPA solution, RMSEP was evaluated with 100,000 random selections of 5–23 wavelengths (a number smaller than 5 is not recommended, as discussed before. Also, if the number is larger than 23—the number of calibration samples—numerical problems arise in MLR). Remark that an exhaustive search is unfeasible, since the number of possible selections is larger than 8×10^{16}. The result was used to generate an estimated probability density function *pdf* [31], which is depicted in Fig. 4. From this graphic, it should be noted that the probability of a random selection being better than the SPA solution is almost zero (area below the *pdf* curve between 0 and 0.02 mg l^{-1}). In fact, the best result among the randomly generated selections was RMSEP = 0.03 mg l^{-1}.

4.1. SPA × GA comparison

The GA was run 100 times, starting from different initial conditions, and using the inverse of RMSEP as fitness criterion. As mentioned before, the search space for GA was restricted to selections with a fixed number of wavelengths, which was set to 14, in order to allow a comparison with the SPA solution. Remark that, even with this restriction, the number of possible selections is still larger than 2×10^{13}. The best and worst GA realizations yielded RMSEP of 0.01 and 0.03 mg l^{-1}, respectively. The

Table 1
Average absolute errors ($\times 10^{-2}$ mg l^{-1}) obtained in the predictions of the test set concentrations

Regression	Pre-selection (61[a])					GA (best–worst values[b]) (14[a])[c]					SPA (14[a])[d]				
	Co	Cu	Mn	Ni	Zn	Co	Cu	Mn	Ni	Zn	Co	Cu	Mn	Ni	Zn
MLR	e	e	e	e	e	1	2–4	1–2	1	1	1	3	2	1	1
MLR + QR	4	11	4	2	2	1	2–4	1–2	1	1	1	3	2	1	1
PCR	1	5	3	2	1	1	2–5	1–2	2	1	1	3	2	1	1
PLS1	1	5	3	1	2	1	2–5	1–2	1	1	1	2	1	1	1

e = errors larger than 1 mg l^{-1} due to numerical instability.

[a] Number of employed wavelengths.

[b] If best and worst results do not coincide, both are indicated.

[c] Wavelengths (nm) selected by GA: 470, 476, 481, 484, 485, 486, 490, 492, 498, 509, 518, 519, 526 and 527.

[d] Wavelengths (nm) selected by SPA: 470, 475, 478, 485, 486, 488, 490, 491, 495, 512, 516, 523 and 530.

average RMSEP obtained was 0.02 mg l^{-1}, which matches the value yielded by SPA. The best set of selected wavelengths indicated in Table 1 resembles the SPA-selected set. Remark, however, that obtaining 100 GA realizations took about 3300 s, while SPA spent only 16 s in the same microcomputer. These times do not include the pre-selection of the spectral window, nor the search for the best number of wave lengths. However, if the computational effort required by the pre-selection procedure were to be taken into account, it would have to be added both to GA and to SPA. It could be argued that GA does not require pre-selection steps, nevertheless the use of full spectrum would increase the dimension of the search space, thus decreasing the probability of GA finding a good solution in a given number of realizations. The same line of reasoning applies if GA is not provided with the number of wavelengths to be selected.

Concentrations of the test set were predicted using MLR (with and without QR decomposition), PCR and PLS1 models calibrated with the selected wavelengths by SPA and GA. Results are displayed in Table 1. On the overall, results for Cu and Mn were worse because their spectra are very similar and have no distinctive features, when compared to the other analyte spectra (Fig. 2).

Wavelength selection improved the predictive quality for all calibration models. Average absolute errors obtained with SPA and GA-selected wavelengths are very similar, despite the fact that GA selection may take much longer than SPA selection. It is important to note that wavelength selection allows good results to be obtained with a MLR model, which is simpler to calibrate and interpret than PCR and PLS1 models. Moreover, QR decomposition is no longer needed, because collinearity has been satisfactorily removed.

4.2. *Internal validation*

Results above were obtained using the RMSEP criterion to select SPA starting point and to evaluate chromosome fitness in GA. For means of comparison, the RMSECV criterion was used in both algorithms to perform new 14-wavelength selections, again running GA 100 times. The selected wavelengths were employed to build a MLR calibration model, which was then used to predict the concentrations of the test set, in order to calculate RMSEP.

Both SPA and the best GA realization yielded RMSECV = 0.01 mg l^{-1} and RMSEP = 0.04 mg l^{-1}. However, the average RMSEP obtained by GA was 0.08 mg l^{-1} and 98 GA realizations yielded a RMSEP larger than SPA. The reason for the poor performance of GA with respect to RMSEP becomes clear if the condition number (CN) of matrix $\mathbf{X}_{\text{cal}}^{\text{T}}\mathbf{X}_{\text{cal}}$ at the selected wavelengths is inspected. The CN for SPA and the best GA realization were 4×10^8 and 1×10^9 respectively, that is, SPA was more successful in removing data collinearity. As mentioned before, if GA is only aimed at minimizing prediction errors, the results are very dependent on the validation set employed. In this sense SPA seems to be more robust than GA.

5. Conclusion

This paper described a novel forward selection algorithm (SPA) for variable selection in multivariate calibration. SPA employs simple operations in a vector space to obtain subsets of variables with small collinearity. It is shown to demand a smaller computational workload than guided random search techniques (GRSA), such as genetic algorithms (GA), mainly when the total number of variables gets large. Also, unlike SPA, the stochastic nature of GRSA makes them non-dependable for finite optimization time.

The restriction in the number of wavelengths to be selected (which cannot be larger than the number of calibration samples) is a limitation of SPA. However, this was not a major handicap in the present application. It can also be argued that, if many spectral variables are needed to discriminate the analytes, then a large number of samples will also be required to perform the calibration.

The use of SPA with data sets gathered by different multicomponent instrumental techniques is being investigated. Future research could also attempt to employ SPA to provide a initial solution to be further refined by GA. Preliminary results have shown that such procedure allows GA to approach the optimum selection in a smaller time and also help alleviate the uncertainty in its performance.

Acknowledgements

The authors would like to thank Prof. Roy Edward Bruns for his suggestions and to PADCT science funding program for partial financial support. The research fellowships granted by the Brazilian agency CNPq and CAPES are also gratefully acknowledged.

References

[1] H. Martens, T. Naes, Multivariate Calibration. Wiley, London, 1993.

[2] C.H. Spielgman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Cote, Anal. Chem. 70 (1998) 35–44.

[3] K. Sasaki, S. Kawata, S. Minami, Appl. Spectrosc. 40 (1986) 185–190.

[4] M. Otto, W. Wegscheider, Anal. Chem. 57 (1985) 63–69.

[5] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024–2030.

[6] L. Xu, I. Schechter, Anal. Chem. 68 (1996) 2392–2400.

[7] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135–153.

[8] R. Leardi, J. Chemom. 8 (1994) 65–79.

[9] D. JouanRimbaud, D.L. Massart, R. Leardi, O.E. deNoord, Anal. Chem. 67 (1995) 4295–4301.

[10] R. Leardi, R. Boggia, M. Terrile, J. Chemom. 6 (1992) 267–281.

[11] V. Centner, D.L. Massart, O.E. deNoord, S. Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851–3858.

[12] H.C. Goicoechea, A.C. Olivieri, Analyst 124 (1999) 725–731.

[13] S. Wold, N. Kettaneh, K. Tjessen, J. Chemom. 10 (1996) 463–482.

[14] G.A. Bakken, T.P. Houghton, J.H. Kalivas, Chemom. Intell. Lab. Syst. 45 (1999) 225–239.

[15] F. Despagne, D.L. Massart, Chemom. Intell. Lab. Syst. 40 (1998) 145–163.

[16] R. Todeschini, D. Galvani, J.L. Vilchez, M. del Olmo, N. Navas, Trends Anal. Chem. 18 (1999) 93–98.

[17] B.K. Alsberg, A.M. Woodward, M.K. Winson, J.J. Rowl, D.B. Kell, Anal. Chim. Acta 368 (1998) 29–44.

[18] D. JouanRimbaud, B. Walczack, R.J. Poppi, O.E. deNoord, D.L. Massart, Anal. Chem. 69 (1997) 4317–4323.

[19] M. Forina, C. Casolino, C.P. Millan, J. Chemom. 13 (1999) 165–184.

[20] B.K. Alsberg, D.B. Kell, R. Goodacre, Anal. Chem. 70 (1998) 4126–4133.

[21] E. Kreyszig, Introductory Functional Analysis with Applications. Wiley, New York, 1978.

[22] C. Jochum, P. Jochum, R.B. Kowalski, Anal. Chem. 53 (1981) 85–92.

[23] D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. Wiley, New York, 1980.

[24] G.H. Stewart, Stat. Sci. 2 (1987) 68–100.

[25] E. Gomez, J.M. Estela, V. Cerdá, M. Blanco, Fresenius' Z. Anal. Chem. 342 (1992) 318–321.

[26] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters. Wiley, New York, 1998.

[27] The Unscrambler User's Guide 6.1 in CAMO A/S, Trondhein, 1996.

[28] Matlab in The Mathworks, South Natick, MA, 1994.

[29] L. Lawson, R.J. Hanson, Solving Least-Squares Problems. Englewood Cliffs, Prentice-Hall, 1974.

[30] A. Papoulis, Probability, Random Variables and Stochastic Processes. 3rd edn., Hardcover, McGraw Hill, 1991.

[31] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst 25 (1994) 99–145.