SUPERVISED PATTERN RECOGNITION TECHNIQUES FOR CLASSIFICATION OF *EUCALYPTUS* SPECIES FROM LEAVES NIR SPECTRA

ROSARIO CASTILLO^A, DAVID CONTRERAS^{B*}, JUANITA FREER^{A,B*}, JOSE RUIZ^A, SOFÍA VALENZUELA^{A,C}.

^a Renewable Resources Laboratory, Biotechnology Center, University of Concepcion, Chile

^b Faculty of Chemical Sciences, University of Concepcion, Chile

^c Faculty of Forestry Sciences, University of Concepcion, Chile

*Corresponding author. Tel.: +56-41-2204601; Fax: +56-41-2245974

(Received: January 24, 2008 - Accepted: July 20, 2008)

ABSTRACT

Three supervised pattern recognition methods (*SPRM*) were evaluated to discriminate between *Eucalyptus globulus* and *Eucalyptus nitens* species applying near infrared (*NIR*) spectroscopy on leaves. The methods used were k-nearest neighbor (*KNN*), soft modeling class analogy (*SIMCA*) and discriminant partial least squares (*PLS-DA*). First and second derivatives were used as transform techniques and mean-center (*MC*) and autoscaling (*AS*) as preprocessing techniques. The training set was constituted by 288 samples and 20 samples were used as validation set. A significant difference between the assayed methods was not observed, however best results for separation of classes and prediction rate were obtained when first derivative and *MC* were used for all the recognition pattern methods. Use of leaves and *NIR* spectroscopy avoids the destructive usual wood analysis in forest industries and facilitates the fast classification of these species for forest applications.

Key words: Pattern recognition, *NIR*, *Eucalyptus*.

INTRODUCTION

The genus *Eucalyptus* (Myrtaceae), gathers around of 600 species, widely distributed in the Southern hemisphere, mainly in Australia and Tasmania¹. Within its 7 subgenus², *Symphomyrtus* includes the species with major commercial value, among them, *Eucalyptus nitens* and *Eucalyptus globulus*³. These species are widely used in forest industries, due to its fast growing, easy adaptability to new soils and environments and its high pulp yield^{3,4}. The wood of these species have different chemical and mechanical properties and prices. The fast differentiation of these species is not possible without the use of botanical or genomic methodologies over the plant.

Supervised pattern recognition methods (*SPRM*)⁵, have been used for classification of biological species⁶⁻⁸ and to determine the geographical and botanical origin of plant derivatives⁹. Near infrared (*NIR*) spectroscopy, has been used together with pattern recognition methods in order to solve problems about plants classification¹⁰⁻¹², based on morphological^{13,14}, chemical¹⁵ and genetics properties¹⁶. *NIR*, has also been used on wood samples for forestry applications, specially in prediction of its chemical composition^{7,13,17-20} mechanical properties,^{18,20} and species identification^{6,21}.

On *Eucalyptus* species, Michell et al.⁷ developed a method based in principal components analysis (*PCA*) and a soft independent modeling of class analogy (*SIMCA*) for classification of *E. globulus* and *E. nitens* species. Also, Schimleck et al.²² used *NIR* spectroscopy with partial least squares (*PLS*) for predicting *E. globulus* and *E. nitens* wood chemical composition and Bailléréis et al.²³ used *NIR* spectroscopy as a tool for rapid screening of some major wood characteristics in an *Eucalyptus* breeding program. It is highlighted that in all of those works, wood samples were used for *NIR* spectra, *PCA* does not allow to predict classes, and *SIMCA* models has not properly separated the classes. In our knowledge there are not classification models developed for a differentiation of *Eucalyptus* species using *NIR* reflectance spectroscopy and leaves as samples.

The aim of this study was evaluate the supervised pattern recognition methods (*SPRM*): k-nearest neighbor (*KNN*), *SIMCA* and discriminant partial least squares (*PLS-DA*) with different techniques of preprocessing and transforms in order to obtain a fast and accurate classification of *E. globulus* and *E. nitens*, using *NIR* reflectance spectra. The comparison between the three methods was developed evaluating parameters of each method on training and prediction sets and using statistical McNemar's Test¹⁵ with the objective to find the best model for separation of the species and the prediction of the class of unknown samples.

2. Methods description

2.1. Exploratory analysis

Two unsupervised pattern recognition methods, principal component

analysis (*PCA*) and hierarquical cluster analysis (*HCA*) were used for the exploratory analysis of the data.

*PCA*²⁴, provides a way to reduce dimensionality of the data finding lineal combinations of independent variables. *PCA* express a matrix *X* as a product of the other two matrices, the score matrix *T* and the transpose of the score loadings matrix *P* according to:

$$X = TP^T \quad (1)$$

The columns of *P* are the principal components (*PC*'s), the elements of the first column of loadings, indicate the contribution of the originals variables to the first principal component (*PC1*) and the matrix of scores *T* is the projection of the samples over the axes defined by the loadings²⁵. Associated with each factor is a *PC* which expresses the magnitude of variance captured by each factor.

HCA^{26,27}, calculates and compares distances between pairs of samples. Relatively small distances between samples implies that the samples are similar. Dissimilar samples will be separated by relatively large distances. Euclidian distance d_{ab} between two samples vectors, *a* and *b*, is determined by computing differences at each of the *m* variables⁵ according to:

$$d_{ab} = \left[\sum_{j=1}^m (a_j - b_j)^2 \right]^{1/2} \quad (2)$$

Inter-sample distances are transformed in a standard scale corresponding to a similarity index calculated by:

$$\text{similarity}_{ab} = 1 - \frac{d_{ab}}{d_{\max}} \quad (3)$$

where $d_{\max}$ is the largest distance in the data set.

2.2. Supervised Pattern Recognition Methods

SPRM use the objects of known class for which a certain number of variables have been measured, as a training set, and a later and independent set for the evaluation of the performance of the model through of a validation step^{5,28}. In this work we used the *KNN*, *SIMCA* and *PLS-DA* methods.

KNN method²⁹, attempts to categorize an unknown based on its proximity to samples placed in categories³⁰. The predicted class of an unknown sample depends on the class of its *k* nearest neighbors. The multivariate distance used in *KNN* is similar to the euclidean distance d_{ab} , calculated in *HCA*, where *a*_{*j*} and

b_j are the data vector for the two samples for variable j and m is the number of variables. Euclidian distance separates each pair of samples in the training set and stores them in a table of distances. For class prediction of any sample, the classes of its neighbors can be tallied and the sample is assigned to the class to which most of nearest neighbors belong. The optimal number of k nearest neighbors is determined by cross validation procedure³¹, where each object in the training set is taken out and considered as a validation sample. This process is performed for $k=1$ to $n-1$ and the number k nearest neighbors with the lowest error rate is chosen.

SIMCA method³², develops a PCA for each training set category by iteratives NIPALS algorithm³³. SIMCA determines the number of PC's or eigenvectors needed to describe the structure of the training class by cross validation⁵. PC models has the structure described by:

$$X_K = T_K P_K^T + E_K \quad (4)$$

Where X_K is the centered X -matrix for the class K , T_K the (un-normed) score matrix ($n \times p^*$), P_K the loading matrix ($m \times p^*$) and E_K the matrix of residuals ($n \times m$).

The residuals from one model can be computed from the scores on the nonretained eigenvectors for obtain the residual variance of each sample in the class q (S_p^2) and the total residual variance of class q (S_o^2), calculated by the following equations:

$$S_p = \sqrt{\frac{\sum_{j=1}^J (res_{pj}^q)^2}{J - Aq}} \quad (5)$$

$$S_o = \sqrt{\frac{\sum_{p=1}^{Iq} \sum_{j=1}^J (res_{pj}^q)^2}{(Iq - Aq - 1)(J - Aq)}} \quad (6)$$

Where, J is the number of variables, Aq the number of PC used by class q and Iq the number of samples in the class q .

The total residual variance of class is a measure of the samples compaction inside the class.

For a new object x_{new} the scores would be determined by the following equation.

$$t_{new}^T = (x_{new} - \bar{x}_K)^T L_K \quad (7)$$

Then, S_p^2 is calculated by each sample similarly to Eq. (5), and a statistic test F with determined level of probability is applied. If S_p of the new object is less to the critical standard deviation (S_{crit}) obtained by the residuals of a training class, the new object belongs to class K , otherwise it does not⁵. If the new object qualifies as a member of both classes, the class having the smallest sample residual is considered the best, and the other class is deemed next best. If the new object exceeds both critical values, it is assigned to neither class. The values of these residuals are often called distances and accumulated in a class distance object. The information in this object can be presented in a multiplot view where pairwise combinations of classes form the subplot axes²⁵. This plot currently denominated Coomans Plot³⁴ shows four quadrants divided by two lines indicating the S_{crit} values for each training set.

PLS-DA³⁵, uses regression for classification trough of partial least square (PLS), a regression method that find latent variables in the features spaces which have a maximum covariance within the predictor variables³⁶. The input feature of dependent variables is the matrix X , and the assignment to class is described in the Y matrix, which has columns with value of 0 or 1 for each object. Regression coefficient matrix B is calculated with the training set according by:

$$Y_{training} = X_{training} B \quad (8)$$

PLS algorithms give the scores matrix T and the weight loadings the matrix P to have Eq. (9)³⁷.

$$T = X_{training} P \quad (9)$$

Prediction of dependent variables on a new set of objects is estimated by PLS-DA calculated by:

$$Y_{validation} = X_{validation} B \quad (10)$$

Some applications of PLS-DA have been reported in^{15, 31, 38-40}.

2.3 Pre-processing and transform methods

MC permits center the data about the mean. This mean is computed for each variable according to:

$$\bar{x}_j = \frac{1}{n} \sum_i x_{ij} \quad (11)$$

The mean is then subtracted from each data value to produce a mean-centered matrix defined by Eq. (12), where i is the sample and j is the variable.

$$x_{ij(mc)} = x_{ij} - \bar{x}_j \quad (12)$$

AS, is a preprocessing technique that apply a mean-centering followed by variance scaling on the data. Autoscaling of the data is calculated according to the follow equation:

$$x_{ij(as)} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (13)$$

Where s_j is the standard deviation for the variable j .

First derivative and second derivative are based on a Savitzky – Golay polynomial filter⁴¹. This method applies a convolution to independent variables in a window containing a center data point and n points on either side. A weighted second-order polynomial is fit to these $2n + 1$ points and the center point is replaced by the fitted value.

EXPERIMENTAL

3.1 Samples

Leaves of *Eucalyptus* species were obtained from forestry industries of Bio-bio Region, Chile. Three hundred eight samples corresponding to 258 samples of *E. globulus* (8 maintained under field conditions and 258 under nursery conditions) and 50 samples of *E. nitens* (all maintained under nursery conditions), were divided by random selection in two sets as follows: 288 samples as training set for the construction of classification models and 20 samples as validation set. Validation set had 10 samples of *E. globulus* and 10 samples of *E. nitens*. Leaves were dried at 50°C by 24 hours, pulverized and sifted through 30 Mesh for NIR spectroscopic analysis.

3.2 Acquisition of NIR spectra

A NIR reflectance Perkin Elmer spectrometer (Identichack FT-NIR, Beaconsfield, England) was used for the collection of NIR reflectance spectra expressed as percentage of reflectance (%R). For the NIR spectra acquisition, 0.5 g of *Eucalyptus* leaves were dried and pulverized, then 50 scans in a range of wavenumber from 10000 to 4000 cm^{-1} (1000 – 2500 nm) with resolution of 8 cm^{-1} were performed for each sample. Duplicate spectra were measured

for each sample and an average of them was obtained using the Perkin Elmer software. The spectra were transformed to absorbance using $\log(1/R)$ for chemometric analysis (Fig. 1).

Fig. 1. NIR spectra of *Eucalyptus* species.

Fig. 2. PCA scores plot of different data preprocessing and transformations: a) *MC* and first derivative, b) *MC* and second derivative, c) *AU* and first derivative, d) *AU* and second derivative. Filled circles are *E. globulus* samples and empty circles are *E. nitens* samples.

3.3 Spectral data pre-treatments

MC and *AS* were the pre-processing techniques while first and second derivatives were used as transform techniques.

3.4 Chemometric Analysis

PCA, HCA, SIMCA, KNN and PLS-DA methods were performed using Pirouette 3.11 (Infometrix Inc). Recognition and prediction ability were evaluated by the members percentage correctly classified on the training and validation set (rate prediction), respectively.

a) Exploratory analysis

For PCA, the computation of PC's were developed using nonlinear iterative partial least squares (NIPALS) algorithm^{42,43} over the training set data. The optimal number of PC's was determined based on the cumulative

variance percent of PC's.

For the HCA, the Euclidian distances between pairs of training set samples was calculated and a dendrogram was built using incremental linkage method²⁵ for calculate the distance between clusters.

b) Pattern recognition methods

KNN, SIMCA and PLS-DA were applied over the training set previously transformed with first or second derivative and pre-processed with *MC* or *AS* techniques for obtain predictive models. Then, these models were used to classify samples of validation set. Rate prediction (percentage of samples correctly classified in the validation set) was calculated for all the methods.

Cooman's plots⁴⁴ were used for the discrimination of classes and S_o was calculated according Eq. (6) for evaluate the class compaction in each model.

3.5. Statistical comparison of methods

McNemar's test, a particular case of Fisher's test, is used when the same sample is assayed on two occasions⁵. This test is based on a χ^2 test with one degree of freedom. Details of this test in classification are described in Roggo et al. (2003)¹⁵. This test has been used for SPRM comparisons in application to qualitative analysis of sugar beet¹⁵. The χ^2 critical value with 5% level of significance and one degree of freedom, $\chi^2_{(1,0.95)}$ is 3.84. If χ^2 value is less than 3.84 the null hypothesis is true, and the two algorithms are not significantly different, but if χ^2 is over 3.84, the null hypothesis is false and the two algorithms are significantly different, with a probability of 95%.

RESULTS AND DISCUSSION

4.1 Exploratory Analysis

PCA of the NIR spectra shows separation of the samples in two non overlapping groups, corresponding to *E. nitens* and *E. globulus*. Scores plot (Fig. 2) shows the best separation when *MC* and first derivative were used, with 65% of the cumulative variance explained by the three first PC's. Use of second derivative shows less cumulative variance values and when it is accompanied with *AU*, the species cannot be properly separated. The data classification in species by PCA allows to use classification methods based on PC's; in this way, SIMCA and PLS-DA methods were performed.

When HCA was carried out, all the used pre-treatments, separated the samples in two main clusters pertaining to the *Eucalyptus* species. However, only first derivative shows less similarity index between the species and greater similarity index between the samples in each species. The two main clusters showed 0.214 and 0.267 as maximal similarity index between the species for the *AS*-first derivative and *MC*-first derivative pre-treatments, respectively. The species classification by HCA, allows to create a KNN model.

4.2 Classification methods

a) KNN

Optimal number of neighbors in KNN was 1 for all the algorithms used. The results showed that the first and second derivative were good transformation techniques for classification of the modeling set. Non significant differences were found between the transform and preprocessing techniques over the validation set, since that all the methods classified correctly 100% of the samples.

b) SIMCA

All the utilized algorithms classified the data of the training set in two separated classes. The Cooman's plot (Fig. 3), shows that *MC* - first derivative pre-treatment produces the best separation of the samples in two classes, placing all the samples of each class under their respective S_{crit} values. Both pre-processing techniques accompanied by second derivative, showed all samples into of overlap classes quadrant, over the S_{crit} values, indicating that the samples can belong to both classes, although no overlap between *E. globulus* and *E. nitens* samples was observed. Models using *MC* - first derivative showed also the best interclass distance and greater cumulative variance explained with the optimal number of PC's, calculated by cross validation (Table 1); however, models with second derivative showed less values of S_o^2 than models with first derivative, indicating higher compaction of the classes (Table 1).

Table 1. Parameters of SIMCA analysis in classification of *E. globulus* and *E. nitens*.

Pretreatment	Interclass distance	<i>E. globulus</i> class			<i>E. nitens</i> class		
		# of PC's	Cumulative variance	Total residual variance of class (So) ²	# of PC's	Cumulative variance	Total residual variance of class (So) ²
MC - First derivative	1.6	10	73	1.4×10^{-9}	7	70	1.4×10^{-9}
AS / First derivative	1.2	8	41	1.5×10^{-9}	10	53	1.3×10^{-9}
MC / Second derivative	0.22	5	14	6.4×10^{-10}	5	27	9.6×10^{-11}
AS / Second derivative	0.12	5	9.6	6.5×10^{-10}	6	25	6.1×10^{-10}

It is highlighted that non significant differences between rate prediction over validation set were found using McNemar's test with the different transform and pre-processing techniques; thus, *MC* – first derivative and *AS* – first derivative classified correctly 95% of the samples, *MC* – second derivative, 100% and *AU* – second derivative 90%. SIMCA models for classification of *eucalyptus* species previously reported, using wood as samples⁷, showed a Cooman's plot with non properly separated classes, it shows the advantages to use leaves and the application of pre-treatments to the spectra.

Fig. 3. Cooman's plot: a) *MC* and first derivative, b) *MC* and second derivative, c) *AU* and first derivative, d) *AU* and second derivative. Lines show the S_{crit} value of each class with 95% of probability.

c) PLS-DA

PLS-DA on the training set, with optimum number of PC's determined by the lowest values of the standard calibration error (SEC), predicted residual error sum of squares (PRESS), and the percentage of cumulative variance for each class, showed also that *MC* – first derivative is the best model for classification of the classes (Table 2), although all the PLS-DA models predicted correctly 100% of validation set samples.

Table 2. PLS-DA in classification of *E. globulus* and *E. nitens*.

Pre-treatment	Optimal number of PC's	Cumulative variance	SEC	PRESS
MC / First derivative	15	79	0.0082	0.018
MC / Second derivative	5	14	0.041	0.47
AS / First derivative	11	44	0.012	0.041
AS / Second derivative	4	8.0	0.038	0.41

4.3. Comparison of classification models

There are no significant differences in the prediction rate on validation set between algorithms used in independent studies into SIMCA, KNN and PLS-DA models, either between the three studied models; however, all the methods show best results for separation of the classes in the training set when first derivative and *MC* were used. Although the prediction rate in the validation set is good for all the used methods, the comparison between the pre-treatments of the data in the training set, shows the negative influence of the application of the second derivative over our data. This is in agreement with the knowledge of the noise increase when this transform technique is applied. SIMCA model is able to distinguish the best separation of the classes and the influence of the pre-treatments of our data in the training or prediction set, with a known significance level, and thus this method could be considered the best choice for posterior classification of new samples.

It is important to mention that this work shows the use of *SPRM*, not only for separation of the classes, as it has been previously reported with the use of wood^{7,21}, but also, shows the predictive ability of these methods in order to get a fast, reliable and accurate classification of new samples in the forest industry. The models grouped the samples of *E. globulus* maintained under field and nursery conditions in the same class. Since there are not significant differences in the prediction rates on validation set, results of McNemar test are not shown.

CONCLUSIONS

All the analyzed methods are able to classify samples of *Eucalyptus* species. There are not significant difference between the prediction ability of the methods utilized, however the SIMCA method is more adequate for differentiation of botanical species by NIR spectroscopy, because this method is not determinist

like KNN and PLS-DA, this means that SIMCA gives a probability to belong to a class, both classes or it is possible to recognize when the sample does not belong to the classes studied. Best pre-treatment of the data was the application of mean-centered and first derivative of the spectra.

The NIR spectroscopy of leaves together with *SPRM* models could avoid the need of botanical classification by genetic or morphologic methods and the use of wood as samples for classification in forest applications.

ACKNOWLEDGMENTS

The authors thank financial support of INNOVA Bio-bio N° 03-B1-210-L1 project and DIUC 206.021.024-1.0. R. Castillo thanks to DAAD for PhD. scholarship. Authors thank the collaboration of Prof. Dr. Matthias Otto.

REFERENCES

- G. H. Chippendale, Myrtaceae - Eucalyptus - Angophora, Australian Gout. Publishing Service, Canberra, 1988.
- L. D. Pryor and L. A. S. Johnson, A classification of the eucalyptus, Australian National University Press. Canberra, 1971.
- J. W. Turnbull, *New Forests*. **17**, 37, (1999).
- C. A. Raymond, *Ann. Forest. Sci.* **59**, 525, (2002).
- B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier, Amsterdam, 1998.
- N. Gierlinger, M. Schwanninger and R. Wimmer, *J. Near Infrared. Spec.* **12**, 113, (2004).
- A. J. Michell and L. R. Schimleck, *Appita J.* **51**, 127, (1998).
- Y. A. Woo, H.-J. Kim and J. Cho, *Microchem. J.* **63**, 61, (1999).
- R. Fernandez-Torres, J. L. Perez-Bernal, M. A. Bello-Lopez, M. Callejon-Mochon, J. C. Jimenez-Sanchez and A. Guiraum-Perez, *Talanta*. **65**, 686, (2005).
- M. J. Saiz-Abajo, J. M. Gonzalez-Saiz and C. Pizarro, *J. Near Infrared. Spec.* **12**, 207, (2004).
- M. Andre, *Anal. Chem.* **75**, 3460, (2003).
- Q. S. Chen, J. W. Zhao, H. D. Zhang, M. H. Liu and M. Fang, *J. Near Infrared. Spec.* **13**, 327, (2005).
- P. D. Jones, L. R. Schimleck, G. F. Peter, R. F. Daniels and A. Clark, *Wood Sci. Technol.* **39**, 529, (2005).
- M. Tigabu and P. C. Oden, *New Forests*. **25**, 163, (2003).
- Y. Roggo, L. Duponchel and J.-P. Huvenne, *Anal. Chim. Acta*. **477**, 187, (2003).
- L. Xie, Y. Ying, T. Ying, H. Yu and X. Fu, *Anal. Chim. Acta*. **584**, 379, (2007).
- A. J. Michell, *Appita J.* **47**, 29, (1994).
- J. B. Hauksson, G. Bergqvist, U. Bergsten, M. Sjostrom and U. Edlund, *Wood Sci. Technol.* **35**, 475, (2001).
- A. Terdwongworakul, V. Punsuwan, W. Thanapase and S. Tsuchikawa, *J. Wood Sci.* **51**, 167, (2005).
- S. S. Kelley, T. G. Rials, L. R. Groom and C. L. So, *Holzforschung*. **58**, 257, (2004).
- M. Brunner, R. Eugster, E. Trenka and L. BerganminStrotz, *Holzforschung*. **50**, 130, (1996).
- L. R. Schimleck, A. J. Michell and C. A. Raymond, *Appita J.* **53**, 318, (2000).
- H. Bailleres, F. Davrieux and F. H. Pichavant, *Ann. Forest. Sci.* **59**, 479, (2002).
- J. T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.
- I. T. i. A. C. Inc, *Pirouette Multivariate Data Analysis Software User Guide*, Bothell, WA, 2003.
- J. H. Ward, *J. Am. Stat. Assoc.* **58**, 236, (1963).
- R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, New York, 2001.
- P. K. Hopke, *Anal. Chim. Acta*. **500**, 365, (2003).
- M. P. Derde, L. Buydens, C. Guns, D. L. Massart and P. K. Hopke, *Anal. Chim. Acta*. **59**, 1868, (1987).
- B. R. Kowalski and C. F. Bender, *Anal. Chem.* **44**, 1405, (1972).
- B. K. Alsborg, R. Goodacre, J. J. Rowland and D. B. Kell, *Anal. Chim. Acta*. **348**, 389, (1997).
- S. Wold and M. Sjostrom, in: B. R. Kowalski (Ed.), *Chemometrics: Theory and Application*, ACS Symposium Series, Washington, DC, 1977, p. 243.
- M. Otto, *Chemometrics. Statistics and computer application in analytical chemistry*, Wiley-VCH, Weinheim, 2007.
- D. Coomans and I. Broeckaert, *Potential Pattern Recognition in Chemical and Medical Decision Making*, John Wiley, Letchworth, 1986.
- S. W. Lars Stähle, *J. Chemometr.* **1**, 185, (1987).
- D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, 1997.
- A. J. Burnham, R. Viveros and J. F. MacGregor, *J. Chemometr.* **10**, 31, (1996).
- S. Masoum, D. J.-R. Bouveresse, J. Vercauteren, M. Jalali-Heravi and D. N. Rutledge, *Anal. Chim. Acta*. **558**, 144, (2006).
- Z. Ramadan, D. Jacobs, M. Grigorov and S. Kochhar, *Talanta*. **68**, 1683, (2006).
- C. Y. Pierce, J. R. Barr, A. R. Woolfitt, H. Moura, E. I. Shaw, H. A. Thompson, R. F. Massung and F. M. Fernandez, *Anal. Chim. Acta*. **583**, 23, (2007).
- A. Savitzky and M. J. E. Golay, *Anal. Chem.* **36**, 1627, (1964).
- H. Wold, in: P. R. Krishnaiah (Ed.), *Multivariate Analysis*, Academic Press, New York, 1966, p. 391.
- T. I. Yoshikatsu Miyashita, Hiroyuki Katsumi, Shin-Ichi Sasaki, *J. Chemometr.* **4**, (1990).
- D. Coomans, I. Broeckaert, M. P. Derde, A. Tassin, D. L. Massart and S. Wold, *Comput. Biomed. Res.* **17**, 1, (1984).