A Chemometric Method for the Viability Analysis of Spinach Seeds by Near-Infrared Reflectance Spectroscopy with Variable Selection using Successive Projections Algorithm

Journal:	*Journal of Near Infrared Spectroscopy*
Manuscript ID	JNS-21-0075.R2
Manuscript Type:	Original Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	LAKSHMANAN, MADAN; Council of Scientific and Industrial Research Madras Complex, CEERI Boelt, Birte; Aarhus University, Department of Agroecology - Crop Health Gislum, René; Aarhus University, Agroecology
Keywords:	Seed quality assessment, Near Infrared Reflectance Spectroscopy, Variable selection for data classification, Successive projections algorithm, Chemometrics
Abstract:	This paper proposes a chemometric method for evaluating the viability of spinach seeds using Near Infrared Reflectance (NIR) spectroscopy and Successive Projections Algorithms (SPA). An essential step of the procedure is to apply the SPA to optimize the choice of variables for multivariate classification. Variable selection using SPA has been described as an optimization problem in which a cost function is minimized. Selecting the correct variables makes the chemometric models more complete, precise, accurate, and less complex. The obtained NIR spectra were processed using the Savitszky Golay and Multiplicative Scatter Correction techniques. After that, the best wavelength subset was selected using SPA. Different classification techniques are then applied to the dimension-reduced data to determine the seeds' viability. The results show that the proposed method is less complex than existing canonical variance methods (1.7 % miscalculation error in the proposed way) and is also easier to implement.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

projections_qr.m
Spinach_LDA.m
CostFunction.m

A Chemometric Method for the Viability Analysis of Spinach Seeds by Near-Infrared Reflectance Spectroscopy with Variable Selection using Successive Projections Algorithm

Madan Kumar Lakshmanan[1], Birte Boelt[2] and René Gislum[2]

[1] CSIR - Central Electronics Engineering Research Institute, CSIR Madras Complex, Pin – 600 113, India.

[2] Department of Agroecology - Crop Health, Aarhus University, Forsøgsvej 1, 4200 Slagelse, Denmark.

E-mail: mklakshmanan@ceeri.res.in

***Abstract*— This paper proposes a chemometric method for evaluating the viability of spinach seeds using Near Infrared Reflectance (NIR) spectroscopy and Successive Projections Algorithms (SPA). An essential step of the procedure is to apply the SPA to optimize the choice of variables for multivariate classification. Variable selection using SPA has been described as an optimization problem in which a cost function is minimized. Selecting the correct variables makes the chemometric models more complete, precise, accurate, and less complex. The obtained NIR spectra were processed using the Savitszky Golay and Multiplicative Scatter Correction techniques. After that, the best wavelength subset was selected using SPA. Different classification techniques are then applied to the dimension-reduced data to determine the seeds' viability. The results show that the proposed method is less complex than existing canonical variance methods (1.7 % miscalculation error in the proposed way) and is also easier to implement.**

***Keywords*- Seed quality assessment, Near Infrared Reflectance Spectroscopy, Chemometrics, Variable selection for data classification, Successive projections algorithm**

1. Introduction

The success of a crop largely depends upon the quality of the seeds sown. Ascertaining the quality and health of seed requires studying several factors that contribute to seed performance, including its physical quality, genetic character, seed lot characteristics (seed size, age/maturity of seed, moisture content, and impurities/injuries), and germination ability. Several organizations, such as the International Seed Testing Association (ISTA) [1], International Seed Federation [2], Association of Official Seed Certifying Agencies (AOSCA) [3], and Society of Commercial Seed Technologists [4], work globally to certify seeds and provide quality assurance to the farmers. Seed viability, vigor, and the content of Genetically Modified Organisms are the metrics commonly used by standardization agencies such as ISTA to determine the quality of a seed [1]. The number of seeds germinated from a given seed lot is defined as seed viability. On the other hand, seed vigor indicates the storage capacity of the seed and the prospects of it growing under different conditions [1-4].

Near Infrared Reflectance (NIR) Spectroscopy is a spectroscopic method that spans the region 780 to 2500 nm of the electromagnetic spectrum [5]. The overtone and combination tones of anharmonic vibrational oscillations of critical molecular bonds such as CH, NH, and OH are commonly found in water/moisture, oils/fats, and proteins/amino acids in the NIR spectroscopy region [5-6]. By carefully studying the NIR spectroscopy absorption bands with statistical and mathematical analysis, defined as Chemometrics, it is possible to conduct quantitative and qualitative studies of bio-substances, organic produce, polymers, and food ingredients for commercial applications [7-9].

A large body of literature exists on applying NIR spectroscopy and chemometrics to determine seed quality. In the work of Velasco et al. [10], vital parameters like the weight, oil content, and fatty acid composition of seed have been evaluated using NIR spectroscopy. The use of NIR spectroscopy in the study of seed quality for plant breeding has been explored by Font et al. [11]. The predictability of the viability of beechnuts using NIR spectroscopy has been investigated by Soltani et al. [12]. The

first study on applying the NIR spectroscopy and Chemometrics to analyze spinach seeds is presented by Olesen et al. [13]. Here, Olesen et al. have used a new classification tool named Extended Canonical Variates Analysis (ECVA) to distinguish viable seeds from non-viable ones. The ECVA provides an immediate solution to finding multivariate directions that separate groups even while simultaneously classifying them. The authors report that the technique is highly reliable, with a low misclassification percentage of about 1.7 %. The disadvantage of the ECVA method is that it is computationally intensive. Norgaard et al. [14] reported that it is also less sensitive to irrelevant variables, thereby mitigating an accurate understanding of the modeled relationship.

The NIR spectroscopy response typically consists of hundreds of variables, many of which may be redundant. The model predictions can be improved for better interpretation by removing irrelevant, noisy, and unreliable variables. Further, by optimizing the selection of variables, the discrimination analysis problem can be mathematically better conditioned. Such an approach could pave the way for reduced computational complexity, higher accuracy, and a greater understanding of the models developed.

The literature on the methods available for variable selection is diverse. Many authors have addressed the problem of variable selection for Partial least squares (PLS) estimators and classifiers, given their wide usage in chemometrics. As Tahir et al. [15] reported, such methods for PLS are classified broadly into the filter, wrapper, and embedded methods. Examples of the filter methods are Loading weights [16], Regression Coefficients [17], Jack-knife testing (JT) [18], variable importance in projection (VIP) [19], selectivity ratio (SR) [20] and significance multivariate correlation [21]. Wrapper methods include Monte-Carlo variable elimination (MVE) [22], sub-window permutation analysis [23], backward variable elimination (BVE) [24], and regularized elimination procedure(REP) [25]. Some of the commonly applied embedded methods are soft-threshold PLS (ST) [26], sparse-PLS (SPLS) [27], and distribution-based truncation for variable selection in PLS [28].

Apart from these methods, there are search-based methods such as Artificial Neural Network [29], Genetic Algorithm [30], and simulated annealing [31], which are applied to a broader class of chemometric problems. Other unclassified methods include principal component analysis (PCA) [32], multidimensional scaling [33], competitive adaptive reweight sampling methods (CARS) [34], Functional data analysis based PCA (FPCA) [35], and locality preserving projections (LPP) [36]. Recently, a data-driven method named supervised orthogonal locality preserving projection (SOLPP) [37], has been presented to reduce the feature dimension of NIR spectroscopy signals to assay and grade green tea.

The Successive Projections Algorithm (SPA) minimizes variable co-linearity through an iterative orthogonalization procedure where the wavelengths of interest are determined one at a time. The variables are chosen through an iterative process such that the Root Mean Square Error of Prediction (RMSEP) is the lowest. The SPA has been applied to a wide range of multivariate calibration and a few classification problems for signals collected from various instruments, including NIR spectrometer, UV-VIS spectrophotometer, and radars. The method was introduced by Paiva et al. [38], who showed its potential in multivariate calibration. The authors elucidate the procedure for Multiple Linear Regression (MLR), Principal Components Regression (PCR), and Partial Least Square Regression (PLSR) with open-source data consisting of NIR spectroscopy signals of corn data. In [39], Araujo et al. have applied SPA to analyze photo spectroscopy signals of analytes composed of cobalt, copper, manganese, nickel, and zinc ions [39]. Galvao et al. [40] have applied SPA-MLR to simultaneously determine manganese, molybdenum, chromium, nickel, and iron using a low-resolution plasma spectrometer diode array detector. Zhang et al. [41] have applied SPA to identify the correct pixels in remote sensing data representing the target.

Most of the SPA applications available in the literature are reported for quantitative problems. One of the few papers that deal with the classification problem depicts the work of Pontes [42]. In this

paper, the authors distinguish various edible oil samples from crops such as corn, canola, sunflower, and soya with the help of their UV-VIS spectrophotometry signals in combination with SPA-LDA. The right choice of wavelengths is obtained using an optimization problem minimizing a cost function. In [43], authors Chen et al. have combined dimension reduction using the Successive Projections Algorithm and classification with Partial Least Squares Discriminant Analysis (PLS-DA) to better distinguish normal and malignant colorectal cancer tissues from their NIR spectroscopy Signals.

Our work evaluates the viability of spinach seeds using the simpler (SPA). This study is novel in two ways: (a) in framing the optimization problem to select the best wavelengths for classification problems using SPA; and (b) by applying SPA-based multivariate classification to a new application: spinach seed viability analysis. SPA implementation for classification problems in Matlab is another objective. As of now, only calibration problems can be solved with SPA software. Matlab programs have been written in which a cost function is minimized when choosing the correct wavelength. The Matlab programs are included as a part of the 'Supplementary Materials.'

2. Materials and Methods

2.1 Description of the seed data set

The dataset used in the study comprises the NIR spectroscopy spectra from 282 spinach seeds. As per the guidelines described in [13], the samples have to be classified as belonging to either '0' – Non-germinated seeds or '3' – Germinated seeds.

NIR diffuse transmission spectra expressed in absorbance versus wavenumber were acquired using a Single Seed FT-NIR Analyser (Q-Interline A/S, QFAflex 600F; Tølløse, Denmark). Individual seeds were placed in a 30-sample carousel tray designed for round-shaped forms such as spinach seeds. To ensure optimal masking and uniform measurements, the dry seed was placed with the middle of the

seed covering the hole on the side of the carousel facing the incident light. A cover with 2.5mm apertures was placed on top of the carousel to avoid light leakage around the seed. Seeds were measured at resolution settings of 32 $cm^{-1,}$ and each spectrum was obtained using the mean of 64 successive scans at 15.4 cm^{-1} intervals between 12,000 cm^{-1} and 6000 cm^{-1} (833–1667 nm). Prior to scanning each 30-sample carousel, a reference (background) spectrum was taken using the built-in reference of the instrument. The reference is the first sample to be measured in the carousel. The schematic of the setup is shown in Figure 1.

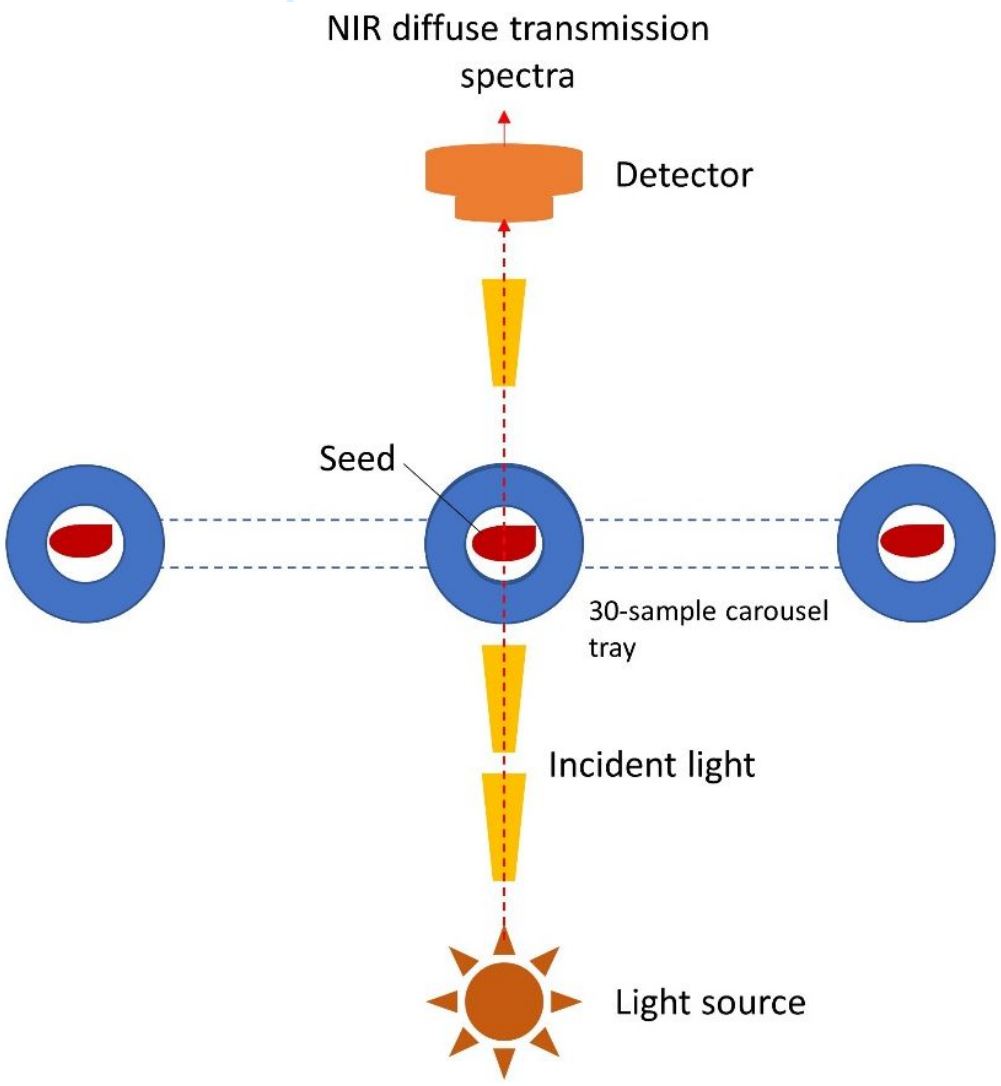


Figure 1. Schematic of single seed NIR diffuse transmission spectra acquisition.

2.2 Description of the reference method

We developed the chemometric models with the reference values obtained from germination tests conducted using the *between-paper* method as explained in [13]. The seeds are given the required resources like air, water, temperature, and light to germinate and grow into a seedling. The seeds are evaluated for germination after seven days, 14 days, and 21 days [13]. Seeds with a visible green neck and root hair after 21 days are given a score of 3 (germinated) [13]. The categorization of the seeds was done visually by an expert who could distinguish all seeds as per their respective category. It may be mentioned that no seed with a score of 2 was observed, possibly due to the high germination rate.

2.3 Multivariate Data Analysis

The steps performed to evaluate the NIR spectroscopy response of spinach seed are as follows. Principal component analysis (PCA) is employed to identify and remove observational errors as the first step. After this, the instrumental NIR spectroscopy response matrix is smoothed using the Savitzky Golay (SG) method [44]. In this work, Savitsky Golay smoothing was performed with a frame size of 21 and polynomial order of 2. After SG smoothing, the Multiplicative Scatter Correction (MSC) [45] technique is employed to correct for features related to scattering effects. This pre-processed NIR spectroscopy response is fed to the SPA to identify the right variables.

After selecting variables, the discrimination and categorization of the seeds as viable or non-viable is achieved using classification algorithms. In this work, we have considered the generative classifiers Naïve Bayes (NB) Classifier, Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA) [46- 47], and the non-parametric classifier Kernel Density Estimation as possible candidates [48]. The generative classifiers learn a model of the joint probability of the vector of inputs **x** (in this case, NIR spectroscopy signal intensities at the selected wavelengths) and the corresponding class C (viable/non-viable) from the training data [46]. A prediction of the most likely type, which is the

posterior probability $P(C|\mathbf{x})$, is then made using the Bayes rule [46]. In LDA and QDA, the class-conditional density models $P(\mathbf{x}|C)$, are assumed to be multivariate Gaussian [47].

On the other hand, in NB, the variables are considered to be independent, and the conditional density is treated as a product of univariate Gaussian distributions [47]. The Kernel Density estimator [48] is a non-parametric method of estimating the underlying probability density function from which the data instances were drawn. This method is highly effective as a classifier because of its ability to capture subtler aspects of the data [48].

The complete chemometric procedure is elucidated in Figure 2.

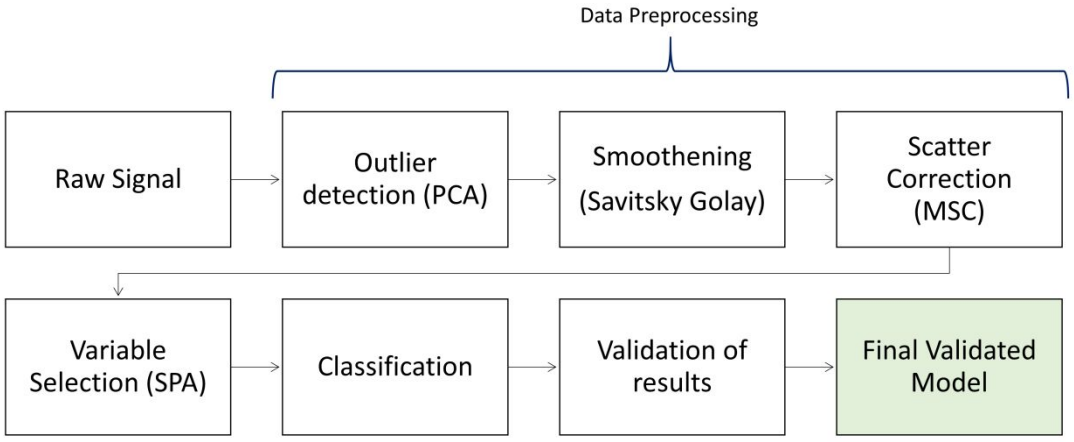


Figure 2. Flow diagram elucidating the chemometric procedure to evaluate the NIR spectroscopy response of the seed samples.

2.4 Successive Projections Algorithm (SPA) for classification

The consecutive projections algorithm (SPA) identifies a subset of variables that minimizes data collinearity. This way, it ensures that the information contained in each selected variable is minimally correlative with that contained in the other selected variables [38]. The variables are chosen one at a time through an iterative procedure of projecting the instrumental NIR spectroscopy response matrix onto a variable space [38]. The SPA algorithm operates in two phases. In the first phase, candidate

wavelength subsets are identified. In the second phase, the best wavelength subset which minimizes a cost function without compromising the accuracy is selected. It is worth mentioning that, in [49], Galvão et al. have demonstrated that in the selection of variables in SPA, the application of either a separate validation set or cross-validation is equally valid. This has the important consequence that even when there is a lack of adequate data sets for calibration or validation, as in the case of seed viability considered in this paper, the SPA may be effectively applied.

Phase 1- Identification of candidate wavelength subsets

The dataset consists of N samples, each described by K wavelengths (or variables) in the NIR spectroscopy region. The instrumental NIR spectroscopy response matrix **X** is then a dimension N x K matrix, as depicted in Figure 3. Out of the N samples, N_c samples are taken for calibration, and the remaining N_v samples are used for validation. The objective of the SPA is to identify L variables (where $L < K$), which maximizes the extraction of information without compromising accuracy.

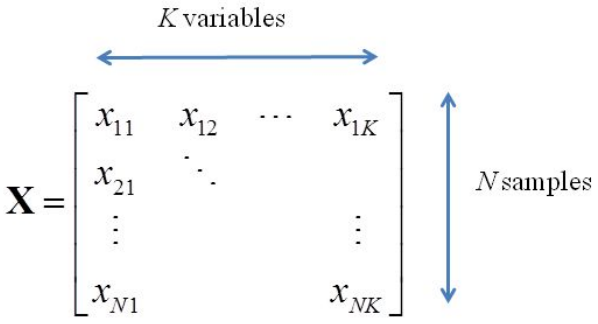
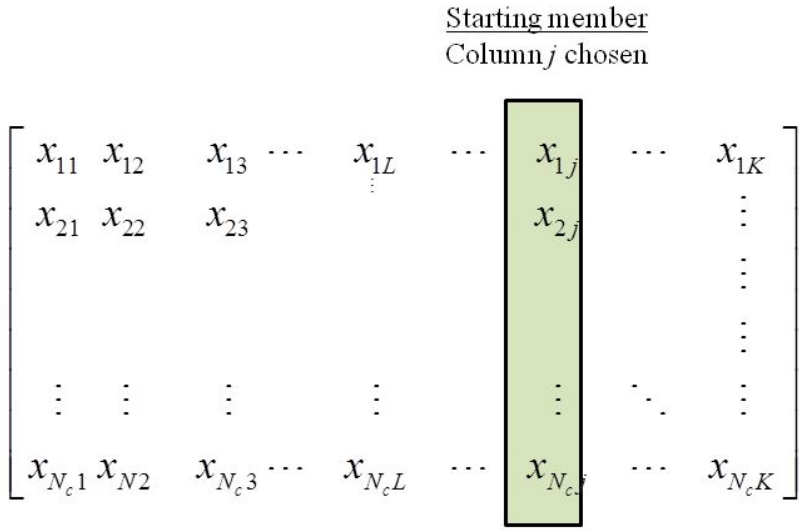


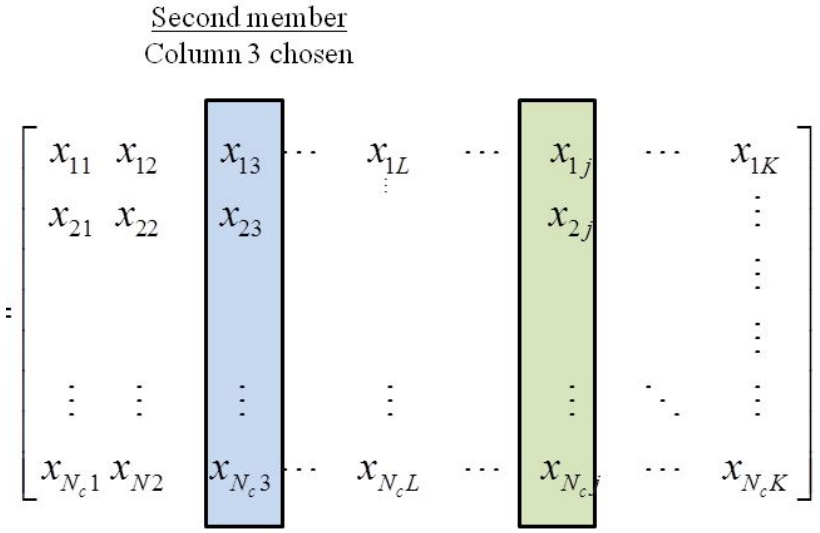
Figure 3. Matrix of the instrumental NIR spectroscopy response for N samples each described by K variables.

The SPA starts by randomly selecting a calibration instrumental response matrix Xc column, say x_j, as shown in Figure 4a. The subspace which is orthogonal to x_j is first computed. Then, those column vectors which have not been selected yet, i.e., $x_{k \neq j}$ are projected onto the subspace orthogonal to the

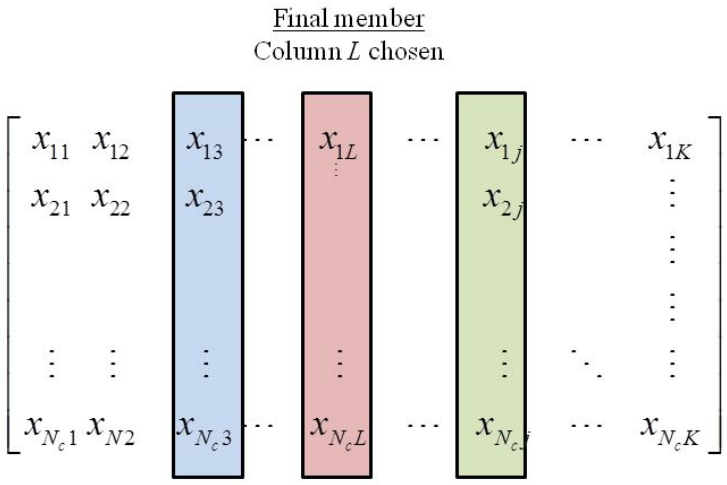
column vector x_j. The column vector with the largest length in the projected space is chosen as a candidate variable or wavelength (refer to Figure 5). In the example illustrated in Figure 4b, column 3 is the second vector of choice. The procedure is repeated until the desirable number of wavelengths L is identified (refer to Figure 4c). At the end of the process, K chains of candidate subsets, each containing L wavelengths, respectively, are identified.



(a)



(b)



(c)

Figure 4. Illustration of the selection of column vectors at different stages of the SPA. The algorithm starts with the vector x_j, then, x_3 and finally x_L at the end of the procedure.

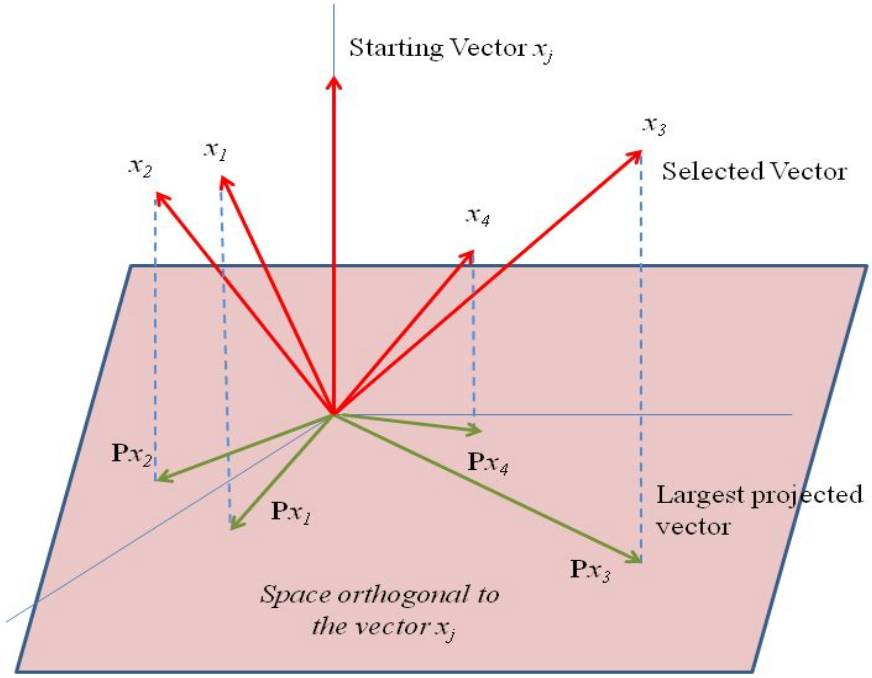


Figure 5: Illustration of selecting the second wavelength using the SPA after identifying the starting vector. In the figure x_j is the starting vector, and x_3 (which has the longest length in the projection space $\mathrm{P}\,x_3$ is the second variable specified. The operator P represents the projection operation.

Phase 2- Selection of an optimum subset of variables

Araujo et al. developed the SPA for regression analysis [38]. In the second phase, methods like Multiple Linear Regression (for regression problems) or metrics like the Mahalanobis distance (for classification/cluster analysis) are used to guide the wavelength selection to identify the best subset of wavelengths [38]. For classification problems, instead of MLR, the best subset of wavelengths is determined by solving an optimization problem that involves minimizing a cost function. The cost function G_{CF} for the optimization problem is defined as the average of the risk of misclassification g_n of each sample in the validation dataset [42].

The cost function G_{CF} for N_v validation samples is defined as [42]:

$$G_{CF} = \frac{1}{N_v}\sum_{n=1}^{N_v} g_n \tag{1}$$

The risk of misclassification g_n of the n^{th} validation sample $x^{v(C_i)}(n,k)$ belonging to the class c_i is defined as [42]:

$$g_n = \frac{D_M^2[x^{v(C_i)}(n,k), \mathrm{E}[x^{c(C_{j=i})}(m,k)]]}{\min\{D_M^2[x^{v(C_i)}(n,k), \mathrm{E}[x^{c(C_{j\neq i})}(m,k)]]\}} \tag{2}$$

In (2), $D_M^2(x,y)$ is the squared Mahalanobis distance defined as [42]:

$$D_M^2[x^{v(C_i)}(n,k), \mathrm{E}[x^{c(C_j)}(n,k)]] = \left(x^{v(C_i)}(n,k) - \mathrm{E}[x^{c(C_j)}(n,k)]\right)\Sigma^{-1}\left(x^{v(C_i)}(n,k) - \mathrm{E}[x^{c(C_j)}(n,k)]\right)^{\mathrm{T}} \tag{3}$$

With $x^{v(C_i)}(n,k)$ denoting the n^{th} validation sample belonging to the class c_i (represented by the superscript $v(C_i)$), $x^{c(C_i)}(m,k)$ representing all the calibration samples belonging to the class c_i (represented by the superscript $c(C_i)$), and $x^{c(C_{j\neq i})}(m,k)$ standing for all the calibration samples not belonging to the class c_i (represented by the superscript $c(C_{j\neq i})$). The operators $\mathrm{E}[\bullet]$ and $\min\{\bullet\}$ stand for the mean and minima operations, respectively. The operator Σ denotes the covariance or scatter

matrix of the pooled calibration instrumental response matrix $\mathbf{X}_c$. The SPA selects that subset of variables for which the cost function G_{CF} is minimum. A small value G_{CF} implies that the considered validation sample $x^{v(C_i)}(n,k)$ is proximal to its actual class c_i and away from all other classes $c_{j\neq i}$. The minimization of the cost function thus leads to identifying the correct variable subset that can better distinguish samples of different categories.

2.5 Validation of results

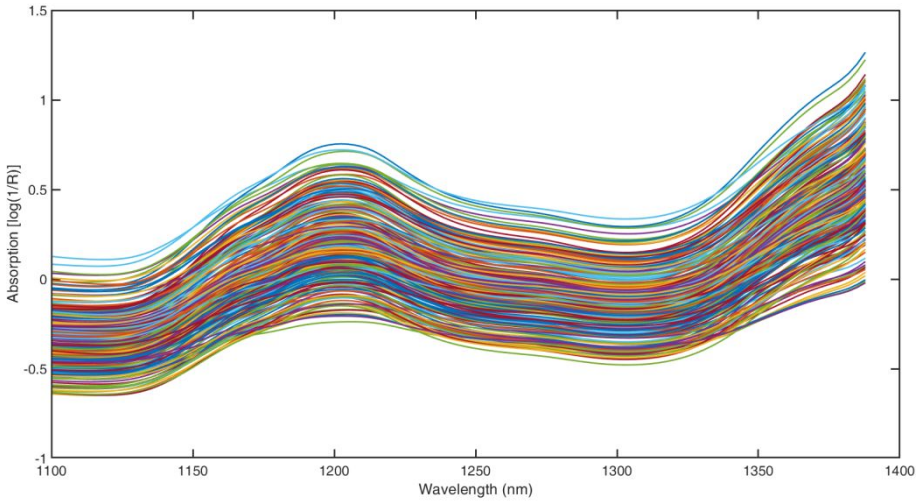
We used two performance metrics to verify the validity of the results. They are re-substitution error and general error.

1. Re-substitution error: The re-substitution error denotes the misclassification error (the proportion of misclassified observations) on the training set. Three samples were eliminated from the initial set of 282 samples because their Principal Components were found to be extreme in a PCA. The remaining collection of 279 samples was used for the analysis. Following an 80-20% dataset split, the data was divided into a training set of 223 samples and a test set of 56 samples. The training and test sets have the same class proportions as the original group. The prediction error of the training set is then calculated as the re-substitution error.
2. General error: The general error is calculated through a stratified 10-fold cross-validation process where the training set is divided randomly into ten disjointed subsets. Each subset has about the same size and class proportions as the training set. First, a subset is removed, and the classification model is trained using the other nine subsets. The classification of the extracted subset is predicted using the trained model. The same procedure is repeated for each of the ten subsets one at a time, and the cumulative error is obtained. In this work, each of the 10 test sets is portioned into a training set of 251 samples and a test set of 28 samples, following a 90-10% split of the dataset.

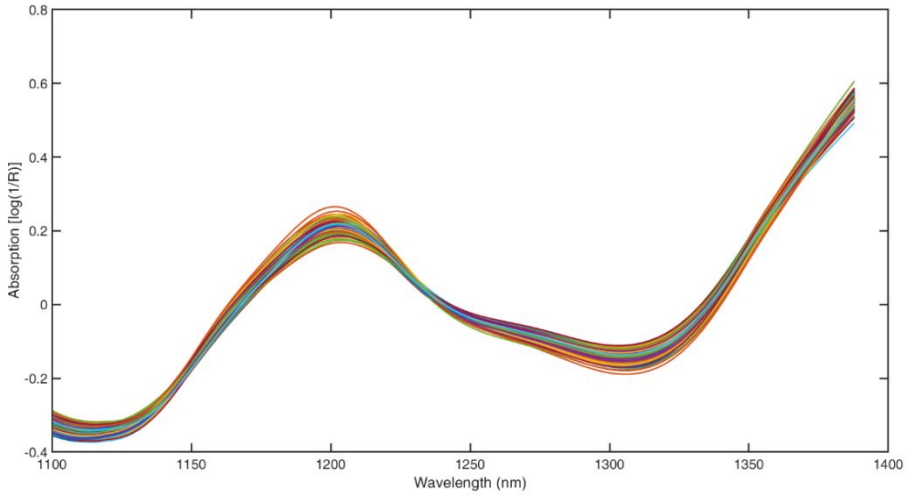
3. Results and Discussion

3.1.Raw and pre-processed NIR spectroscopy response

The spectra of the raw NIR spectroscopy response obtained are in the form of absorbance versus wavelength curves, as shown in Figure 6(a). The natural NIR spectroscopy response is smoothened with a Savitzky Golay method and then corrected using the MSC method for scattering effects. We can see from Figure 6(b) that the noise and scatter effects have been reduced.



(a)



(b)

Figure 6: Plot of NIR spectroscopy instrumental response of spinach seeds: (a) natural NIR spectroscopy response, (b) smoothed by Savitzky Golay (zero-order, frame size of 21 and polynomial order 2), and corrected for scattering effects using multiplicative scatter correction.

3.2. Dimension reduction through Successive Projections Algorithm

The cost function for various numbers of variables selected as per the SPA has been plotted in Figure 7. We may note from the figure that the value of the cost function is lowest when ten variables are chosen.

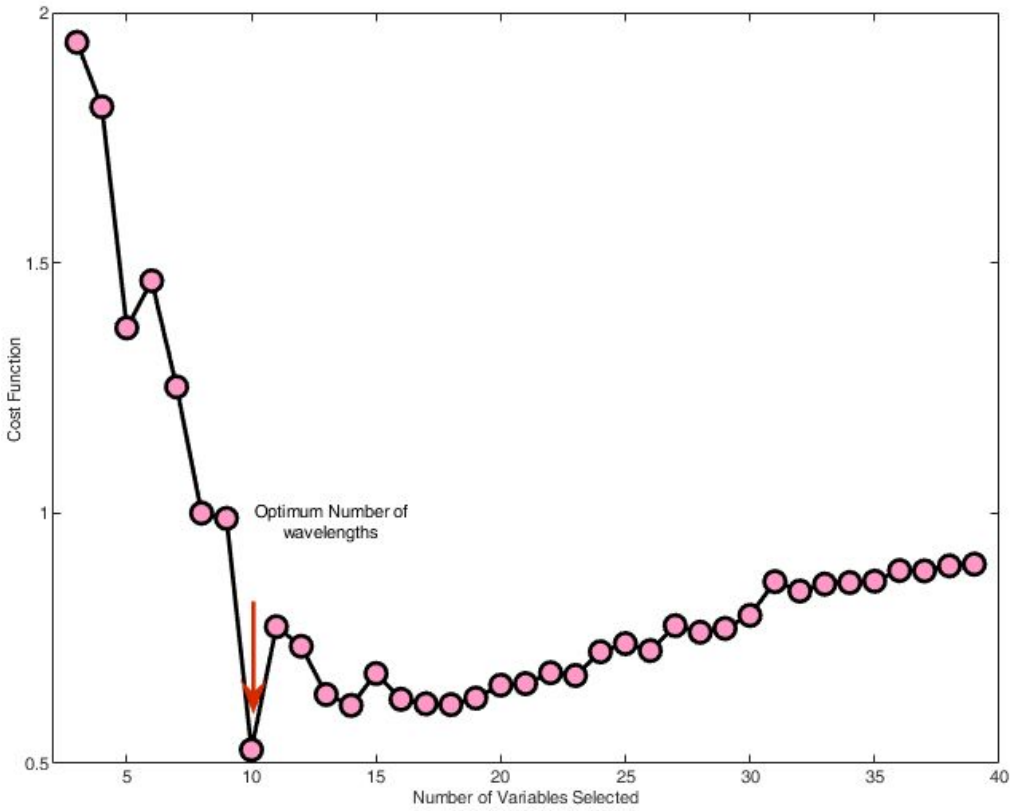


Figure 7. Root mean square error using a cost function over the selected variables.

Figure 8 is a 3D plot that depicts the variations in the cost function with respect to the selected starting variable and the number of variables. The plot shows that there are numerous local minima. The SPA algorithm does not stop at local minima but proceeds to identify the global minima.

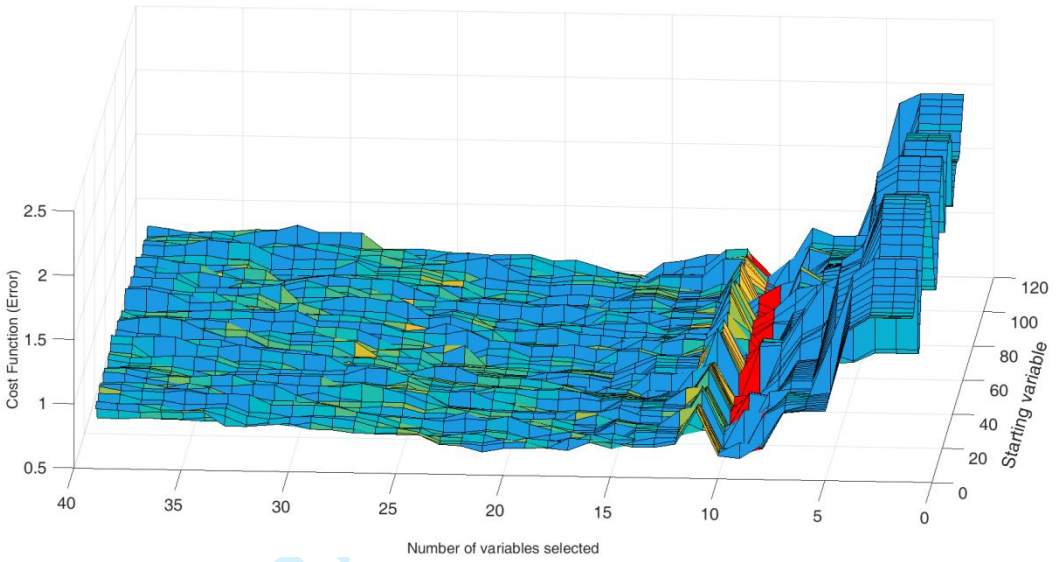


Figure 8. Visualization of Cost Function values versus the starting variable and the number of variables selected for the SPA algorithm. The plot shows that there are numerous local minima. The SPA algorithm is tuned to identify the global minima.

In Figure 9, the respective positions of the chosen variables are shown. One may observe that most of the selected variables are located near the peaks or at the necks of the peaks of the NIR spectroscopy response.

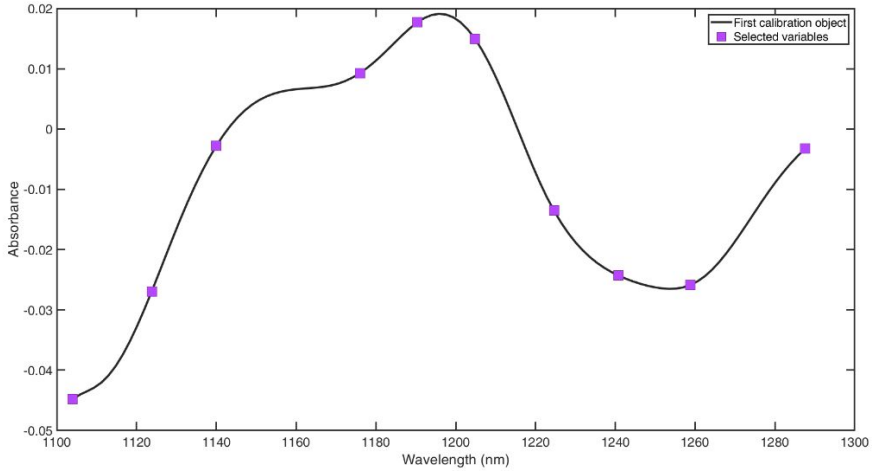


Figure 9. A plot of the NIR spectroscopy response of a single spinach seed along with the position of the ten variables selected by the SPA algorithm.

The selected SPA variable indices and their respective wavelengths are provided in Table 1.

Table 1. Variables selected by the SPA algorithm and their corresponding wavelengths

Index	41	1	87	57	68	103	77	12	21	49
Wavelength(nm)	1176	1104	1258	1204	1224	1287	1240	1123	1140	1190

3.3. Validation of results - categorization of spinach seeds using discriminant analysis

The results which validate the proposed chemometric method to classify the seeds are provided in Table 2. The errors for different classification methods range from 0.1% to 16%.

Table 2: Re-substitution and general error percentages for various discriminant analysis techniques

	LDA	**QDA**	**Naïve Bayes**	**Kernel Density**
Re-substitution Error	2.05%	5.45%	5.45%	0.11%
General Error	2.39%	5.25%	10.42%	15.56%

3.4. Comparison with ECVA-based approach

In comparison to the ECVA approach applied by Olesen et al. [13], the performance of the proposed method is marginally lower (2.1 % miscalculation error in the proposed way versus 1.7 % error in ECVA). However, the proposed method offers the advantage of simplicity. The ECVA algorithm is computationally complex, and in contrast, the SPA-based approach offers reduced complexity and increases execution speed. The most complex operation of the proposed method is variable selection using SPA. However, the SPA has to be used only once before the discriminant analysis is conducted, thus, significantly reducing the complexity and execution time.

4. Conclusion

A comprehensive analysis of the application of SPA to variable selection in classification problems has been conducted. In the present study, spinach seeds were assessed for viability using the method. SPA and data discrimination are used to select variables. Compared with existing methods based on canonical variance, the proposed method is comparable in terms of performance, reduced complexity, and ease of implementation. As a result of modelling only the selected variables instead of the whole spectrum, the developed models are more precise and accurate.

We are aware that no seeds had a score of 2 and we are consequently not able to predict this with our model. To predict this, we need seeds of this score, followed by retraining the model for multiclass classification involving three scores (0, 2, and 3). It is, however, not easy to obtain seeds with a score of 2 since the radicle has to range from 0.3 to 1.0 mm. It is most common for healthy seeds to reach a score of 3 quickly, while unhealthy seeds do not show radicles in most cases. As long as the overall aim is to predict whether seeds are healthy or not, not having seeds with a score of 2 is not a problem.

In the future, research on underlying chemical processes inside the seed can be considered based on an understanding of the critical variables that contribute to the models. A further research area would be to examine how data size impacts SPA's performance. An analysis of this kind can be conducted for both calibration and classification problems.

This publication includes MATLAB source code, which has previously been unavailable, to facilitate the application of SPA to classification problems.

Acknowledgment

The first author would like to express his heartfelt gratitude to Dr. P.C. Panchariya, the Director,

CSIR-CEERI, Pilani, Dr. C. Kumaravelu, Scientist-In-Charge, CSIR-CEERI, Chennai for facilitating this collaborative work and for their guidance, constant motivation, and support. He would also to thank Mr. V. Venkatraman, Senior Principal Scientist (retired), CSIR-CEERI, Chennai, who provided insight and expertise that greatly assisted the research.

References

1. International Seed Testing Association (ISTA), http://www.seedtest.org/en/home.html (accessed 4 June 2020).
2. International Seed Foundation (ISF) , http://www.worldseed.org/isf/home.html (accessed 5 June 2020)
3. Association of Official Seed Certifying Agencies (AOSCA), http://www.aosca.org (accessed 5 June 2020)
4. Society of Commercial Seed Technologists (SCST) http://www.seedtechnology.net/(accessed 5 June 2020)
5. Handbook of Near-Infrared Analysis, 3rd ed. Practical Spectroscopy, pp. 349–369, 2007.
6. Cen H, He Y. Theory and application of near-infrared reflectance spectroscopy in the determination of food quality, *Trends in Food Science & Technology*, Volume 18, Issue 2, 2007, pp.72-83.
7. Mark H, Workman J. Chemometrics in Spectroscopy, Academic Press, 1st ed, 2007.
8. Næs T, Isaksson T, Fearn T,Davies T. A User-Friendly Guide to Multivariate Calibration and Classification, *NIR Publications*, Chichester, 2002.
9. Varmuza K,Filzmoser P. Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, 2009.

10. Velasco L, Möllers C, Becker HC. Estimation of seed weight, oil content and fatty acid composition in intact single seeds of rapeseed (Brassica napus} L.) by near-infrared reflectance spectroscopy, Euphytica, 03-1999, Volume 106, Issue 1, pp 79-85.

11. Font R,Celestino MR, Bailón AH, The use of near-infrared spectroscopy (NIRS) in the study of seed quality components in plant breeding programs, *Industrial Crops and Products*, 01/2006; 24(3):307-313.

12. Soltani A, Lestander T, Tigabu M,Odén P.Prediction of the viability of oriental beechnuts, Fagus Orientalis, using near-infrared spectroscopy and partial least squares regression, *J Near Infrared Spectroscopy*, 01/2003; 11(1).

13. Olesen M, Shetty N, Gislum R,Boelt B. Classification of viable and non-viable spinach (Spinacia oleracea L.) seeds by single seed near-infrared spectroscopy and extended canonical variates analysis, *J Near Infrared Spectroscopy*, 19(3), 171–180 (2011). DOI: 10.1255/jnirs.928.

14. Norgaard L, Bro R,Westad F, Engelsen SB.A modification of Canonical Variates Analysis to handle highly collinear multivariate data, *J Chemometrics*, 2006, 20, 425-435.

15. Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, Solve Sæbø. A review of variable selection methods in Partial Least Squares Regression, *Chemometrics and Intelligent Laboratory Systems,* Volume 118, 2 August 2012, pp.62-69.

16. Martens M. Sensory and chemical quality criteria for white cabbage studied by multivariate data analysis. *Lebensmittel-Wissenschaft Technologie.* 1985;18(2):100-104.

17. Frenich AG, Jouan-Rimbaud D, Massart D, Kuttatharmmakul S, Galera MM, Vidal JLM. Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares. *Anal*. 1995;120(12):2787-2792.

18. Efron B, Tibshirani R. An *introduction to the bootstrap*, Vol. 57, New York, USA: Chapman & Hall/CRC; 1993.

19. Wold S, Johansson E, Cocchi M. PLS: partial least squares projections to latent structures. *3D QSAR Drug Design*. 1993;1:523-550.

20. Kvalheim OM, Karstang TV. Interpretation of latent-variable regression models. *Chemom Intell Lab Syst*. 1989;7(1-2):39-51.

21. Tran TN, Afanador NL, Buydens LM, Blanchet L. Interpretation of variable importance in partial least squares with significance multivariate correlation (SMC). *Chemom Intell Lab Syst*. 2014;138:153-160.

22. Cai W, Li Y, Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom Intell Lab Syst*. 2008;90(2):188-194.

23. Li HD, Zeng MM, Tan BB, Liang YZ, Xu QS, Cao DS. Recipe for revealing informative metabolites based on model population analysis. *Metabolomics*. 2010;6(3):1-9.

24. Frank IE. Intermediate least squares regression method. *Chemometr Intell Lab Syst*. 1987;1(3):233-242.

25. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. A partial least squares based algorithm for parsimonious variable selection. *Alg Mol Biol*. 2011;6:27.

26. Saebo S, Almoy T, Aaroe J, Aastveit AH. ST-PLS: a multi-dimensional nearest shrunken centroid type classifier via PLS. *J Chemometrics*. 2007;20:54-62.

27. Lê Cao KA, Rossouw D, Robert-Granie C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*. 2008;7(1):35.

28. Liland KH, Hoy M., Martens H, Saebo S. Distribution based truncation for variable selection in subspace methods for multivariate regression. *Chemom Intell Lab Syst*. 2013;122:103-111.

29. Boger Z. Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis. *Analytica Chimica Acta (2003)*, 490 (1-2) 31-40.

30. Swierenga H, de Groot PJ, de Weijer AP,Derksen MWJ,Buydens LMC, Improvement of PLS model transferability by robust wavelength selection (1998).*Chemometrics and Intelligent Laboratory Systems* (1998) 41:237–248.

31. Lucasius CB,Beckers MLM,Kateman G. Genetic algorithms in wavelength selection: a comparative study. *Analytica Chimica Acta* (1994), 286 (2) 135-153.

32. Kovalenko IV, Rippke GR,Hurburgh CR, Dimensionality Reduction of near Infrared Spectral Data Using Global and Local Implementations of Principal Component Analysis for Neural Network Calibrations, *J Near Infrared Spectroscopy*, Vol. 15, Issue 1, pp. 21-28, 2007.

33. Wish M,Carroll JD. Multidimensional scaling and its applications, *Handbook of Statistics*. Volume 2, 1982, pp.317-345.

34. Fan W,Shan Y,Li Y,Lang YZ. Application of Competitive Adaptive Reweighted Sampling Method to Determine Effective Wavelengths for Prediction of Total Acid of Vinegar, Food Analytical Methods 5(3), 2012.

35. Barati Z, Zakeri I,Pourrezaei K. Functional data analysis view of functional near-infrared spectroscopy data. *J Biomedical Optics*;18(11):117007, November 2013.

36. Feng G, Hu D, Zhou Z. A Direct Locality Preserving Projections (DLPP) Algorithm for Image Recognition, Neural Processing Letters 27(3):247-255, June 2008.

37. Liu P,Wen Y,Huang J, Xiong A, Wen J, Li H, Huang Y, Zhu X, Ai S, Wu R. A novel strategy of near-infrared spectroscopy dimensionality reduction for discrimination of grades, varieties, and origins of green tea, *Vibrational Spectroscopy*, Volume 105, November 2019.

38. Paiva HM, Soares SFC, Galvao RKH,Araujo MCU. A graphical user interface for variable selection employing the Successive Projections Algorithm, *Chemometrics and intelligent laboratory systems*, 2012, pp. 260-266.

39. Araujo MCU, Saldanha TCB,Galvao RKH,Araujo MCU. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, *Chemometrics and intelligent laboratory systems*, 57(2), 2001, pp. 65-73.

40. Galvao RKH , Pimentel MF, Araujo MCU,Yoneyama T,Visani V. Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry, *Anal Chem*, 443(1), 2001, pp.107-115.

41. Zhang J ,Rivard B ,Rogge DM. The Successive Projection Algorithm (SPA), an Algorithm with a Spatial Constraint for the Automatic Search of Endmembers in Hyperspectral Data, Sensors, 2008.

42. Pontes MJC, Galvao RKH,Araujo MCU,Moreira T,Neto ODP,Jose GE,Saldanha TCB(2005).The successive projections algorithm for spectral variable selection in classification problems*, Chemometrics and intelligent laboratory systems*, 78 (1-2). pp. 11-18.

43. Chen H, Lin Z,Mo L,Wu T,Tan C. Near-Infrared Spectroscopy as a Diagnostic Tool for Distinguishing between Normal and Malignant Colorectal Tissues, *BioMed Research International*, 2015, 472197.

44. Savitzky A, Golay MJE.Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Anal Chem*, 36, 1627-1639, 1964.

45. Isaksson T ,Næs T.The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy, *Applied Spectroscopy*, Vol. 42, Issue 7, pp. 1273-1284, 1988.

46. Ng AY,Jordan MI.On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, January 2001, pp.841–848.

47. James G,Witten D,Hastie T,Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, 2013.

48. Taylor C. Classification and kernel density estimation, Vistas in Astronomy, Volume 41, Issue 3. 1993.

49. Galvao RKH, Araujo MCU, Silva EC, Jose GE, Soares SFC, Paiva HM. Cross-validation for the selection of spectral variables using the successive projections algorithm, *J Braz. Chem*. Soc. 18 (8).2007.