

## Multivariate Calibration to Determine Phorbol Esters in Seeds of *Jatropha curcas* L. Using Near Infrared and Ultraviolet Spectroscopies

Jussara V. Roque,<sup>a</sup> Luiz A. S. Dias<sup>b</sup> and Reinaldo F. Teófilo<sup>\*a</sup>

<sup>a</sup>Departamento de Química and <sup>b</sup>Departamento de Fitotecnia, Universidade Federal de Viçosa, 36570-900 Viçosa-MG, Brazil

The building of partial least squares (PLS) regression models using near infrared (NIR) and ultraviolet (UV) spectroscopies to estimate the concentrations of phorbol esters (PEs) in *Jatropha curcas* L. is presented. The models were built using two algorithms for variable selection, ordered predictors selection (OPS) and genetic algorithm (GA). Chromatographic analyses were performed to determine the concentrations of PEs. Spectral data were obtained from seeds and oil extract. The results of PLS models were performed by analyzing statistical parameters of quality such as root mean square error of prediction (RMSEP) and correlation coefficient of external predictions ( $R_p$ ). The parameters obtained for NIR-PLS and UV-PLS models with OPS were respectively: RMSEP 0.48 and 0.22 mg g<sup>-1</sup> and  $R_p$  0.49 and 0.96. For GA were obtained, respectively: RMSEP 0.52 and 0.28 mg g<sup>-1</sup> and  $R_p$  0.12 and 0.95. The models built from seeds and oil extracts can be used respectively for screening and to accurately predict the PEs content. The OPS method provided simpler and more predictive models compared to those obtained by the selection of variables using the GA. Thus, the UV-PLS-OPS model can be used as an alternative method to quantification of PEs.

**Keywords:** phorbol esters, *Jatropha curcas* L., multivariate calibration, near infrared, ultraviolet spectroscopy

### Introduction

Currently, fossil fuels are the major energy source used by mankind. However, this has resulted in an increase in anthropogenic greenhouse gas emissions in the atmosphere that contributes to the increase in average temperature of the earth and complex changes in climate.<sup>1</sup> Fossil fuels, such as gasoline and natural gas, are limited and will be insufficient for the world's energy demands in the near future. Biofuels, such as biodiesel, are a strategy that combines energy security and sustainability.<sup>2,3</sup>

Biodiesel derived from vegetable seed oils is an alternative to diesel fuel as a renewable and environmentally clean source of energy. This fuel has been produced from a variety of vegetable sources, such as soybean,<sup>4</sup> sunflower,<sup>5</sup> palm oil<sup>6</sup> and physic nut.<sup>7,8</sup> In this context, *Jatropha curcas* L., also known as physic nut, stands out as one of the most promising non-edible oil seeds for the production of biofuel.<sup>9,10</sup>

*J. curcas* is a small tree or large shrub, which belongs to the *Euphorbiaceae* family and has a life expectancy

of up to 50 years.<sup>11</sup> Mexico is the center of origin and domestication of *J. curcas*,<sup>8,12</sup> but this plant spontaneously grows in different regions of Brazil and around the world, including Africa, India, Southeast Asia and China.<sup>13</sup> The *J. curcas* seed contains approximately 31% oil (ranging from 16 to 45%),<sup>14</sup> presenting a composition of fatty acids that provide biodiesel with excellent properties.<sup>15</sup> The press cake remaining after oil extraction is rich in proteins and has potential for use in animal feed, adding value to the productive chain of biodiesel.<sup>16,17</sup> However, the press cake contains toxic compounds, such as phorbol esters (PEs),<sup>18</sup> which limit its use for this purpose.<sup>19</sup> Fortunately, there are reports of the existence of germplasm free of these toxic compounds.<sup>20</sup>

PEs were first identified in 1984 by Adolf *et al.*<sup>21</sup> Six PEs were characterized from *J. curcas* and designated as *Jatropha* factors (C1, C2, epimers C3 and C4, C5 and C6). The determination of PEs in *J. curcas* is important to identify non-toxic germplasm or samples containing only traces of these compounds. This identification is essential for genetic breeding programs to select clones with lower contents of PEs, making possible the use of press cake as animal feed.<sup>22</sup>

\*e-mail: rteofilo@gmail.com

High-performance liquid chromatography with a diode array detector (HPLC-DAD) is broadly used to determine PEs. This technique presents good separation and low limits of detection.<sup>23,24</sup> Nevertheless, these analyses are time-consuming, laborious, expensive, environmentally harmful and destructive, which can, in some cases, preclude the evaluation and selection of genotypes for the development of cultivars. Alternatively, fast, environmentally friendly and non-destructive analyses are possible using spectroscopic techniques and multivariate calibration.<sup>25</sup>

Near infrared spectroscopy (NIR) has been widely used to measure the quality of agricultural and food products.<sup>26</sup> This non-invasive spectroscopic technique can rapidly provide information on physical and chemical samples. Baye *et al.*<sup>27</sup> used NIR to predict seed corn compositions. Choung *et al.*<sup>28</sup> used NIR for determining the protein and oil contents of soybean. Helgerud *et al.*<sup>29</sup> used NIR for online prediction of dry matter content in whole unpeeled potatoes.

Another spectroscopic technique, ultraviolet (UV), combined with multivariate calibration has become quite popular for quantification of interesting species. In the literature, several prediction models using this combination can be found, such as quantification of tannins,<sup>30</sup> determination of wine acidity<sup>31</sup> and simultaneous determination of caffeine and paracetamol in pharmaceuticals.<sup>32</sup> As UV spectroscopy is based on electronic transitions, in most cases, the technique has the ability to detect low concentrations of the analyte of interest. In contrast, NIR spectroscopy is based on vibrations and therefore has limitations at high limits of detection, larger than mg L<sup>-1</sup>.

There are several multivariate regression methods. Among them, we highlight regression by partial least squares (PLS), recognized as the most widely used method for building first order calibration models from chemical source data. This method does not require accurate knowledge of all the components present in the samples and it is able to perform prediction of an analyte of interest even in the presence of interference, as long as the interferents are also present in the model building.<sup>33</sup> PLS can be used in highly correlated noise and experimental data, allowing simultaneous modeling and prediction of more than one analyte, i.e., a matrix  $Y$ .<sup>34,35</sup> Firstly, the model is built and validated. After that, it can be used for predictions of unknown samples, without the need to carry out a reference analysis.

Thus, the aim of this work was to build multivariate calibration models for analysis of PEs using spectroscopic and chemometric methods. NIR and UV spectra were obtained from *J. curcas* seeds and oil extract. The

HPLC-DAD technique was used as the reference method to determine the PEs.

## Experimental

### Materials

The *J. curcas* samples used in this work belong to a germplasm active bank (GAB) and a progeny test (PT). A total of 77 samples of the GAB were used, being 74 taken from different regions of Brazil and 3 from Cambodia. A total of 121 samples were obtained from the PT. Both the GAB and PT were carried out in 2008 at the Universidade Federal de Viçosa (UFV). Based on this set (GAB plus PT, a total of 198 samples), 100 samples were selected using the principal component analysis (PCA). Each sample comprises a set of seeds with shells.

Phorbol-12-myristate-13-acetate (PMA) was obtained from Sigma-Aldrich ( $\geq 99\%$ ; TLC). The solvents, methanol (CH<sub>3</sub>OH) and acetonitrile (CH<sub>3</sub>CN), used in this study were of HPLC grade (J.T. Baker; 99.95%), while formic acid (HCOOH) was of analytical grade (Panreac;  $\geq 98.0\%$ ).

### *J. curcas* sample preparation

The oil was extracted from samples manually pressed with the aid of a mechanical press. Approximately 20 mL of oil was obtained, placed in amber glass bottles and stored under refrigeration ( $-20\text{ }^{\circ}\text{C}$ ).

PEs extraction was carried out by means of adjustments of the optimized methodology found in the literature.<sup>36</sup> PEs were mechanically extracted from *J. curcas* oil with methanol (1:2 m/v) using a magnetic stirrer at  $60\text{ }^{\circ}\text{C}$  for 5 min. The resulting mixture was gravity separated and the upper methanol layer was collected. The described process was repeated three times to quantitatively extract the PEs. Finally, the methanol extracts collected were evaporated and the residue was stored under refrigeration at  $-20\text{ }^{\circ}\text{C}$ .

### PEs analysis

PEs were determined in duplicate. Briefly, the stored residue was re-dissolved in 4 mL of methanol and filtrated (nylon filter with 45  $\mu\text{m}$  pore size and 25 mm diameter) to injection vials. Samples of 20  $\mu\text{L}$  were analyzed by HPLC.

The HPLC used for method development and validation was a Shimadzu® 20AT prominence separations module interfaced with a Shimadzu® SPD-M20A prominence DAD. Data acquisition and analysis were achieved with LabSolutions version 5.54 SP3 software. During method development and validation, a column reverse-phase C18

Kinetex® (Phenomenex, 100 × 4.6 mm, 2.6 μm) was used. A head column containing the same material was used to protect the column.

A central composite design (CCD) was performed to optimize the mobile phase composition in the chromatographic method to obtain the best peak separation and smallest analysis time. Thus, the acetonitrile percentage (CH<sub>3</sub>CN% 79-86%) and formic acid concentration (HCOOH% 0.45-0.75% v/v in water) were evaluated. The response variable, *y*, was obtained by the combination of chromatographic parameters to obtain the desired response (DR), as shown in equation 1.

$$DR = \frac{rtF}{np} \quad (1)$$

where *rtF* is the retention time of the last peak and *np* is the number of peaks that appear in retention time of PEs, not necessary separated. The quantification of the PEs is done by sum of the total area of all peaks that appear in the retention time of PEs. Thus, complete separation of the peaks is not necessary.

It is desirable to find the lowest *rtF* because this corresponds to a shorter run time and *np* = 5 is considered optimal since it indicates the separation of the five PEs present in *J. curcas* (the two *Jatropha* factors epimers appear in the same peak). Thus, the DR considered optimal, obtained by the equation, is the lowest possible. All calculations were performed using spreadsheets according to Teófilo and Ferreira.<sup>37</sup>

The mobile phase was further degassed and filtered through nylon membrane filters (0.45 μm). The column was eluted using an isocratic gradient at a flow rate of 1 mL min<sup>-1</sup>. The column temperature was maintained at 30 °C and the detection wavelength was 280 nm. Methanol was used as the diluent during the standard and test sample preparation.

PMA was used as an external standard and the area of the PEs peaks was summed and converted to PMA equivalent (mg mL<sup>-1</sup>) by taking its peak area and concentration.<sup>23,36</sup> Seven calibration solutions containing the PMA standard in the range 0.05 to 2.80 mg mL<sup>-1</sup> were used to build the inverse linear univariate regression.

In order to identify the chromatographic peaks, the spectra were compared to those in the literature.<sup>38,39</sup>

### Spectral analysis

NIR spectra were measured by a Fourier transform near infrared (FT-NIR) spectrometer (Thermo Scientific Antaris II) controlled with TQ Analysis software. The FT-NIR spectrometer was operated in an integrating sphere

diffuse reflectance module. Seed scans were the average result of 32 scans measured with 8 cm<sup>-1</sup> resolutions over the wavenumber range 10000-4000 cm<sup>-1</sup>. These spectra were obtained in reflectance mode as log (1/R), where R is the collected reflectance. Each intact seed was directly centered on the instrument window, without any sample preparation, and scanned on both sides. Each side was measured three times and the averaged spectrum of the two sides of each seed was used for the final analysis.

UV spectra were obtained by an Ocean Optics spectrometer (USB4000-UV-VIS) with the Spectra Suite Spectroscopy Operating software. Spectra were measured in duplicate using a quartz cuvette with a 10 mm optical path. Oil extract scans were the average result of 10 scans measured with 0.166 nm resolutions over the range 210-350 nm. These spectra were obtained in transmittance mode, where absorbance is collected. Oil extract was prepared from the dilution of the solution used in the PEs analysis in HPLC. The diluted solution was prepared in 2 mL of methanol with the addition of 600 μL of the stored oil extract.

### Multivariate calibration models

Spectra were exported from each instrument software and imported by Matlab 2016a (Math Works, Natick, USA). An inverse regression model (*y* = *Xb*) was built using the PLS.<sup>33,34,40</sup> The algorithms for the data import, the building and the validation models were written in our laboratory in a function .m to Matlab. All calculations were performed in Matlab.

The data matrices used as input for PLS regression are denoted by *X*<sub>NIR</sub> and *X*<sub>UV</sub>, where each row in *X*<sub>NIR</sub> and *X*<sub>UV</sub> corresponds to a given spectrum from the seeds in the NIR spectra and the oil extract in the UV spectra, respectively. The *y* vector used to build the regression is related to values experimentally obtained through the chemical analysis by the HPLC method. The variables *X*<sub>NIR</sub>, *X*<sub>UV</sub> and *y* have been pre-processed using centering in all calculations. Transformations were carried out on the rows of the matrix *X*<sub>NIR</sub> and *X*<sub>UV</sub> in order to find the best model for prediction. The transformations tested were first and second derivatives and the multiplicative scatter correction (MSC).

Regression models were evaluated on calibration and external validation steps by calculating the number of outliers excluded, root mean square error of calibration (RMSEC), correlation coefficient of calibration (*R*<sub>c</sub>), root mean square error of cross-validation (RMSECV), correlation coefficient of cross-validation (*R*<sub>c,v</sub>), root mean square error of prediction (RMSEP), correlation coefficient of external prediction (*R*<sub>p</sub>) and the bias.

The original data set was then split into calibration and validation sets using the Kennard-Stone<sup>41</sup> algorithm after outlier removal. According to ASTM E1655,<sup>42</sup> for regression models with five or less latent variables ( $h$ ), the validation set should have at least 20 samples. When  $h$  is higher than five, the set must have approximately four times the number of model latent variables.<sup>40</sup> Thus, a validation set was used for external prediction and the remaining samples were used in the calibration set. The RMSE was calculated according to equation 2 and the correlation coefficient ( $R$ ) is given by equation 3.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{I_m} (y_i - \hat{y}_i)^2}{I_m}} \quad (2)$$

$$R = \frac{\sum_{i=1}^{I_m} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I_m} (\hat{y}_i - \bar{\hat{y}})^2 (y_i - \bar{y})^2}} \quad (3)$$

where  $y$  is the experimental value,  $\hat{y}$  and  $\bar{\hat{y}}$  are the scalar and vector of the estimated values, respectively,  $\bar{y}$  is a scalar of mean values in  $y$  and  $I_m$  is the number of samples. The number of  $h$  in the model was determined by internal validation (cross validation, CV) applying the randomic removal of three samples. In cross validation, when  $I_m$  is the number of samples of calibration set (training), the error and correlation coefficients are named RMSECV and  $R_{cv}$ , respectively. When  $I_m$  is the number of predicting samples ( $P$ ), the error and correlation coefficients are named RMSEP and  $R_p$ , respectively.

Bias is defined as the difference between the value measured and a reference value, it is a measure of systematic errors in the model, calculated with just the validation samples. This parameter can be obtained by equation 5, at which  $y_i$  is the experimental value,  $\hat{y}_i$  is the estimated value,  $I_m$  is the number of predictions.

$$\text{bias} = \frac{\sum_{i=1}^{I_m} (y_i - \hat{y}_i)}{I_m} \quad (4)$$

The standard deviation of validation (SDV) calculated by equation 5 is used for evaluating the statistical significance of the bias using a  $t$ -test (equation 6), where  $nv$  is the degrees of freedom.

$$\text{SDV} = \sqrt{\frac{\sum_{i=1}^{I_m} [(y_i - \hat{y}_i) - \text{bias}]^2}{I_m - 1}} \quad (5)$$

$$t_{\text{calc}} = \frac{|\text{bias}| \sqrt{nv}}{\text{SDV}} \quad (6)$$

Additionally, methods of variable selection are applied to select regions that present relevant information and are better correlated with the concentration of PEs in *J. curcas* samples. In this work, two methods were used, ordered predictors selection (OPS)<sup>43</sup> and genetic algorithm (GA).<sup>44</sup>

The OPS method is based in obtaining an informative vector that contains information about the location of the best response variables for prediction. The original response variables ( $\mathbf{X}$  matrix columns) are differentiated according to the corresponding absolute values of the informative vector elements. The differentiated variables are sorting in descending order. Multivariate regression models are built and evaluated using cross validation. An initial subset of variables (window) is selected to build and evaluate the first model. Then, this matrix is expanded by the addition of a fixed number of variables (increment) and a new model is built and evaluated. New increments are added until all or some percentage of variables are considered. Quality parameters of the models are obtained for every evaluation and stored for future comparison. In the last step, the evaluated variable sets (initial window and its extensions) are compared using the quality parameters calculated during validations.<sup>43</sup> The calculations were performed using OPS\_Toolbox.<sup>45</sup>

In the OPS method, when using the information vector, special attention should be given to the number of latent variables,  $h$ , to be used to obtain this vector. Firstly, it is determined the number of latent variables,  $h = h\text{Mod}$ , to construct and validate the model. But maybe  $h\text{Mod}$  cannot generate a informative vector sufficiently informative for variable selection. To find the best informative vector, a study was then performed on the full data set by increasing the number of components in the model, starting from  $h = h\text{Mod}$  and carrying out the variable selection using the OPS algorithm. By varying the  $h$  value, different informative vectors are generated, from which the best component number ( $h = h\text{OPS}$ ) is selected. Therefore, two optimum numbers of components are employed in this work, one representing the component number for model building ( $h\text{Mod}$ ) and the other representing the component number employed to generate the best informative vector in OPS method ( $h\text{OPS}$ ).<sup>43</sup>

The GA is a method for solving optimization problems based on natural selection processes and genetics that mimic biological evolution. Initially, a start population is constituted for a series of different individuals and the amount of individuals is defined as the population with a defined size. These individuals are analyzed and crossed according to the introduced parameters into the algorithm. At the end of the process, a vector consisting of zeros and

ones indicates if variables (individuals) should be included (1) or not (0). The parameters previously optimized and introduced into the algorithm were a population of 54, a maximum generation of 300, a mutation rate of 0.008, a window width of 1, convergence of 80, 50 terms included at initiation, cross-over rule of 2, number of subsets to divide data into for cross-validation of 3, cross-validation iteration of 1 and replicate runs to perform of 3.

#### Figures of merit

Among the most important figures of merit, sensitivity, selectivity and limit of detection (LOD) are related to the concept of net analyte signal (NAS). This concept plays an important role in the calculation of figures of merit for characterizing a calibration model. Thus, NAS vector has been defined as the part of a mixture spectrum that is unique for the analyte of interest, i.e., it is orthogonal to the spectrum of interferences.<sup>46,47</sup>

The sensitivity (SEN) is the slope of the calibration curve. Multivariate models are built as inverse regression, therefore, the SEN, according to equation 7, is the inverse of the regression coefficients estimated by PLS.<sup>48</sup>

$$\text{SEN} = \frac{1}{\|\mathbf{b}\|} \quad (7)$$

where  $\|\mathbf{b}\|$  is the Euclidean norm of regression coefficients vector estimated by PLS regression.

The analytical sensitivity,  $\gamma$ , is the ratio between the sensitivity and standard deviation of reference signal ( $\hat{\sigma}_x$ ) estimated by the standard deviation of the NAS about the reference signal spectra,<sup>49,50</sup> as given in equation 8.

$$\gamma = \frac{\text{SEN}}{\|\hat{\sigma}_x\|} \quad (8)$$

The inverse of this parameter ( $\gamma^{-1}$ ) allows us to establish the lowest difference of concentration between samples, which can be distinguished by the method.

Selectivity (SEL) is the model's ability to determine particular analytes in mixtures or matrices without interference from other components or simply a measurement of the amount of the signal used in the determination. In multivariate calibration, the value of SEL measured is generally low. Nonetheless, the model might be able to estimate the concentration of the analyte in the presence of interferences. This parameter is calculated according to equation 9,<sup>46,51</sup> at which  $\text{nas}_{k,i}$  is the scalar value for sample  $i$  and  $x_{k,i}$  is the instrumental response vector for sample  $i$ .

$$\text{SEL} = \frac{\text{nas}_{k,i}}{\|x_{k,i}\|} \quad (9)$$

The LOD is the smallest amount of analyte that can be detected.<sup>52,53</sup> The LOD calculation in multivariate calibration is the most questioned and is not very well defined. Thus, several works have proposed calculations more precise for this parameter.<sup>54,55</sup> In this work, the method proposed by Ortiz *et al.*<sup>55</sup> was used, which evaluates the false positive and false negative hypothesis. This method is based on the extrapolation of a regression equation, obtained by the measured and predicted values, until a zero value of concentration. The LOD expression is calculated by equation 10.<sup>55,56</sup>

$$\text{LOD} = \frac{\Delta(\alpha, \beta) w_{y_0} \hat{\sigma}}{\hat{\mathbf{b}}} \quad (10)$$

where the term  $w_{y_0} \hat{\sigma}$  is an estimate of the error in the intercept and the factor  $\Delta(\alpha, \beta)$  depends on the probability  $\alpha$  (type I error) and  $\beta$  (type error II) and on the number of degrees of freedom.

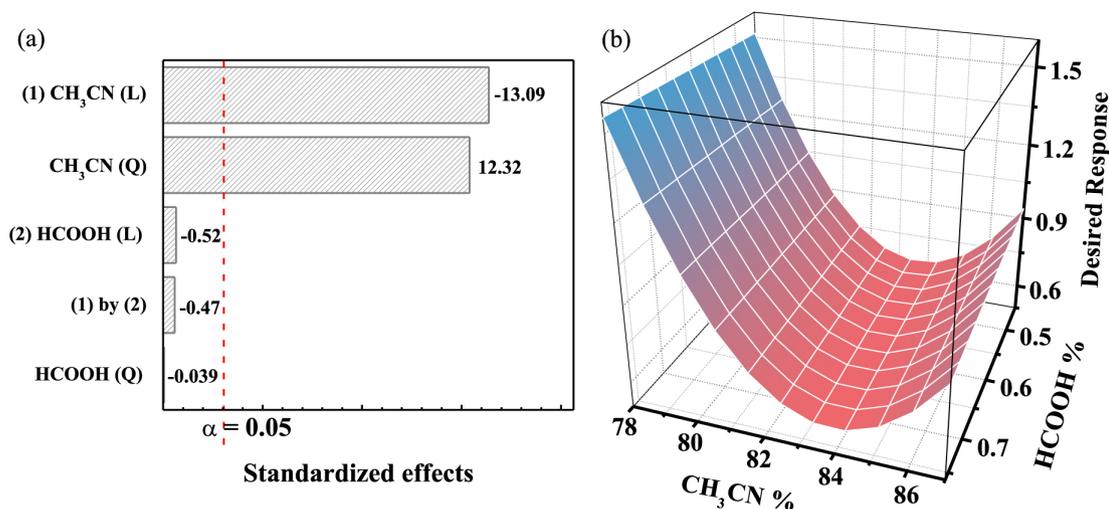
## Results and Discussion

### Central composite design

Statistical analysis by mean square residual error and the significance level ( $\alpha$ ) at 0.05 indicates that only  $\text{CH}_3\text{CN}\%$  is significant, as shown by the Pareto chart (Figure 1a). The linear regression coefficient for  $\text{CH}_3\text{CN}\%$  was negative, indicating that increasing the percentage decreases the last peak retention time to give the desired response, i.e., lower DR (equation 1). Moreover, the significant quadratic coefficient for  $\text{CH}_3\text{CN}\%$  indicated the DR decreases quadratically as the  $\text{CH}_3\text{CN}\%$  increases until 84%, while DR increases quadratically with a  $\text{CH}_3\text{CN}\%$  more than 84%, as shown in the response surface (Figure 1b). Both  $\text{CH}_3\text{CN}\%$  coefficients were significant ( $p = 1.20$  and  $1.70 \times 10^{-5}$ , linear and quadratic coefficients, respectively). However,  $\text{HCOOH}\%$  did not significantly influence the DR ( $p = 0.62$  and  $0.97$ , linear and quadratic coefficients, respectively) and the interaction between  $\text{CH}_3\text{CN}\%$  and  $\text{HCOOH}\%$  was also not significant ( $p = 0.65$ ).

Analysis of variance (ANOVA) shows that the regression model is significant ( $p = 3.62 \times 10^{-5}$ ) and with a non-significant lack-of-fit ( $p = 0.19$ ). These results suggest that the fitted response model can be applied to determine the optimum mobile phase composition.

Thus, it is observed that the lowest DR was found when the  $\text{CH}_3\text{CN}\%$  ranged from 82 to 86%. However,  $\text{HCOOH}\%$



**Figure 1.** (a) Pareto chart for the standardized main effects in the central composite design and (b) response surface estimated from the central composite design by plotting CH<sub>3</sub>CN% versus HCOOH% with desorption of the desired response.

was not significant, and it was obtained the lowest DR across of the studied range. Therefore, the optimal mobile phase composition chosen was 82.5% CH<sub>3</sub>CN and the HCOOH concentration was 0.60% (v/v). This mixture provided a good chromatographic separation of *J. curcas* samples in a run time of 4.5 min.

#### PEs quantification and identification

PEs peaks (five) appeared between 2.5 and 4.5 min (Figure 2a) and were monitored at 280 nm. The area of each peak was obtained and these results were expressed as equivalents of PMA, which was used as a marker and appeared between 5.2 and 5.8 min.

The inverse regression equation obtained for quantification of PEs was  $[PE] = 4.482 \times 10^{-7}SA - 0.01897$ , where SA is the sum of the chromatographic peaks areas and [PE] is the PEs concentration. The linearity, expressed by R, was equal to 0.9968. This value indicates high model linearity within the concentration range studied.

The precision was evaluated as a measure of reproducibility of the entire analytical method, it was expressed by the relative standard deviation (RSD) calculated by the ratio of the standard deviation and the average of the chromatographic peaks areas of five replicates. The RSD value of 11.16% obtained may be considered appropriate and acceptable because of the complexity of the samples.

The LOD and limit of quantification (LOQ) (i.e., the lowest concentration of an analyte in a sample that can be determined with acceptable precision and accuracy) were determined by calculating a signal to noise ratio of 3:1 (LOD) and 10:1 (LOQ) between the standard deviation value regarding the chromatographic signal of seven blank

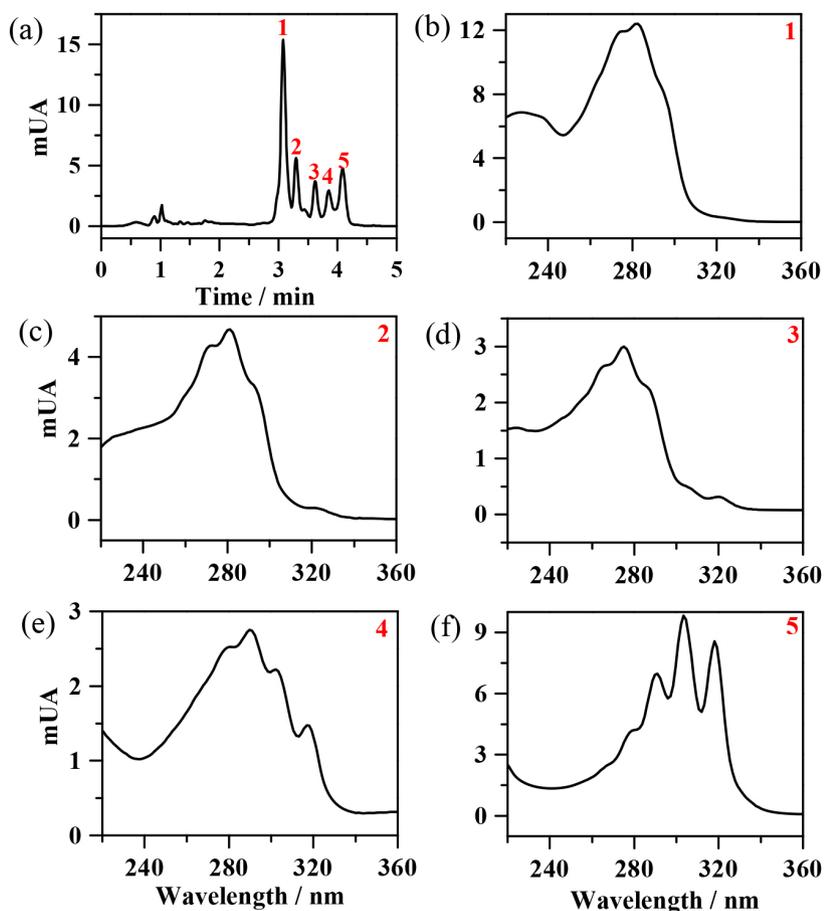
injections (CH<sub>3</sub>OH) and the slope of the calibration curve. LOD and LOQ found were 0.0051 and 0.0171 mg mL<sup>-1</sup>, respectively.

The seed PEs concentration in mg g<sup>-1</sup> was obtained according to oil content for each sample (access and progenies) and it ranged from 0.206 to 3.406 mg g<sup>-1</sup>. Several concentrations ranges have been reported for PEs depending on the region where the samples were obtained. In Malaysia, the concentration of  $3.0 \pm 0.16$  mg g<sup>-1</sup> was reported in 2001.<sup>57</sup> Makkar *et al.*<sup>58</sup> found concentrations between 0.87 and 3.32 mg g<sup>-1</sup> in countries of West and East Africa, North and Central America and Asia. Another study of Makkar *et al.*<sup>18</sup> evaluated four varieties of *J. curcas* originated in Cape Verde, Nicaragua, Nigeria and Mexico. Concentrations of 2.17, 2.30 and 2.70 mg g<sup>-1</sup> have been reported for samples from Nicaragua, Nigeria and Cape Verde, respectively. The Mexican variety was described as non-toxic with only 0.11 mg g<sup>-1</sup>. Recently, concentrations between 0.00 to 10.3 mg g<sup>-1</sup> were described containing samples from a wide variety of geographic regions around the world.<sup>25</sup> Ferrari *et al.*<sup>59</sup> evaluated *J. curcas* seeds from different regions of Brazil and the PEs concentrations found ranged from 1.41 to 8.97 mg g<sup>-1</sup>.

In order to confirm the presence of PEs, individual *J. curcas* PEs peaks, separated by the HPLC, were evaluated for absorption spectra. The absorption spectra of five peaks were similar to those reported in the literature.<sup>23,38,39</sup> The spectra of each of *Jatropha* factor (C1-C6) are shown in Figures 2b-2f.

#### Multivariate calibration models

In this section, we describe the model obtained using NIR spectra of the seeds, which presented results for



**Figure 2.** (a) Chromatogram of a *Jatropha* sample with five peaks relating to the PEs; (b)-(f) spectra of the chromatographic peaks of *Jatropha* factors: (b) C1 factor; (c) C2 factor; (d) C5 factor; (e) C6 factor and (f) C3 and C4 factors, which are epimers.

screening concentrations of PEs with quick decision-making for use in genetic breeding programs. The model built with oil extract UV spectra will also be presented, which showed satisfactory results and conditions to predict PEs concentrations with high accuracy. In addition, several models were built using the NIR and UV spectra of the kernel and oil of *J. curcas*. However, these models were not satisfactory and the results are not shown.

#### NIR spectra of *J. curcas* seed

Before building the model, transformations were carried out in order to find the best model for prediction. Thus, based on RMSECV values, it was observed that all transformations tested increased the error, and in this way, the spectra were just centered. Statistical parameters calculated for built models are shown in Table 1. The correlation coefficients ( $R_c$ ,  $R_{cv}$  and  $R_p$ ) are considerably higher for the NIR-PLS-OPS model. Thus, the model that used selected variables was considered more efficient and robust. NIR spectra of *J. curcas* seeds with the variables selected by OPS are shown in Figure 3a.

In Figure 3c, the measured *versus* predicted values from the NIR-PLS-OPS model show that although the model presents high prediction errors, there is a linear fit between these values.

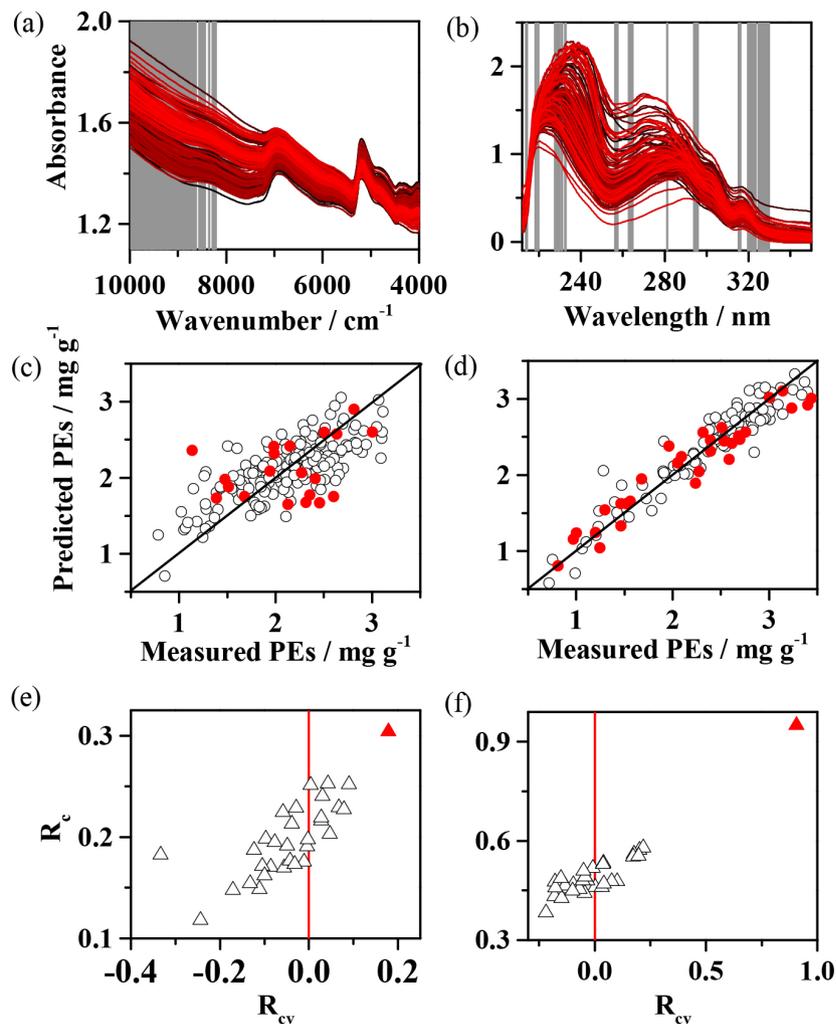
The NIR-PLS-OPS model was evaluated in order to verify if there was a correlation by chance. Hence, the vector  $y$  was randomized and models were built with this random  $y$ . Thus, the correlation parameters of these models were evaluated for both, random and measured  $y$  values. If these parameters of measured  $y$  distance themselves from the correlation values of the random  $y$ , it is an indication that the model did not occur by chance. Figure 3e shows that the NIR-PLS-OPS model (true model) is separated from the other models, so, the true model was not obtained by chance.

Montes *et al.*<sup>25</sup> presented multivariate calibration models to predict several *J. curcas* properties, including PEs concentrations. The authors showed that PEs could be determined by NIRS for qualitative classification purposes by excluding samples with low concentrations ( $< 0.1 \text{ mg g}^{-1}$ ). The NIR-PLS-OPS model obtained in our study can also be used for qualitative classification purposes. However, the model presented in this work was

**Table 1.** Statistical parameters of NIR-PLS and UV-PLS models with full variables and selected variables by OPS and GA methods, and figures of merit for UV-PLS models

	NIR-PLS model			UV-PLS model		
	Full	OPS	GA	Full	OPS	GA
h	4	5 (hOPS = 19)	4	8	8 (hOPS = 11)	8
nVars <sup>a</sup>	1556	130	85	656	85	85
RMSEC / (mg g <sup>-1</sup> )	0.4242	0.3131	0.4016	0.1914	0.1881	0.1913
R <sub>c</sub>	0.2957	0.7528	0.3518	0.9491	0.9508	0.9381
RMSECV / (mg g <sup>-1</sup> )	0.4408	0.4335	0.4121	0.3190	0.2747	0.2499
R <sub>cv</sub>	0.1776	0.4409	0.2861	0.8570	0.8947	0.8929
RMSEP / (mg g <sup>-1</sup> )	0.4441	0.4858	0.5271	0.2780	0.2273	0.2821
R <sub>p</sub>	0.0076	0.4976	0.1276	0.9326	0.9602	0.9526
Bias	0.0026	0.0001	0.0003	0.0041	-0.0061	0.0014
γ <sup>1</sup> / (mg g <sup>-1</sup> )	NC <sup>b</sup>	NC <sup>b</sup>	NC <sup>b</sup>	0.0404	0.0821	0.1007
SEL	NC <sup>b</sup>	NC <sup>b</sup>	NC <sup>b</sup>	0.1081	0.1664	0.1714
LOD / (mg g <sup>-1</sup> )	NC <sup>b</sup>	NC <sup>b</sup>	NC <sup>b</sup>	0.2895	0.2666	0.2607

<sup>a</sup>Number of variables; <sup>b</sup>not calculated. h: latent variables; RMSEC: root mean square error of calibration; R<sub>c</sub>: correlation coefficient of calibration; RMSECV: root mean square error of cross-validation; R<sub>cv</sub>: correlation coefficient of cross-validation; RMSEP: root mean square error of prediction; R<sub>p</sub>: correlation coefficient of external prediction; SEL: selectivity; LOD: limit of detection.



**Figure 3.** (a) NIR and (b) UV spectra of *Jatropha curcas* L. seeds ( $X_{\text{NIR}}$ ) and oil extract ( $X_{\text{UV}}$ ), respectively, with the variables selected by OPS algorithm (vertical lines); (c) NIR-PLS-OPS model and (d) UV-PLS-OPS model: measured *versus* predicted values of PE contents for the calibration (○) and prediction (●) sets; (e) NIR-PLS-OPS model (▲) and (f) UV-PLS-OPS model (▲) chance correlation plot.

obtained without any sample preparation, directly on seeds, while in the model obtained by the other authors,<sup>25</sup> the seed was ground to obtain the NIR spectra.

PEs concentrations found in this work ranged from 0.02-0.3% (m/m) and, according to Pasquini,<sup>60</sup> the NIR spectroscopy shows poor sensitivity, in general, the detection limit is about 0.1% (m/m).<sup>60</sup> Therefore, the low ability to predict PEs of *J. curcas* samples with accuracy can be explained by the low PEs concentrations in most samples. In addition, the obtained correlation might be explained by PEs association with other seed components, resulting in an indirect detection by NIR spectroscopy.

Thereby, the best model built was not able to predict PEs concentrations with high accuracy. However, it is possible to obtain a qualitative information about PEs contents (high or low concentrations) from NIR spectra obtained directly on *J. curcas* seeds.

#### UV spectra of *J. curcas* seed

Samples with concentrations lower than 0.72 mg g<sup>-1</sup> were excluded from the data set because these samples increased the model errors considerably. Thereby, before building the model, transformations were carried out in order to find the best model for prediction. Based on RMSECV values, it was observed that mean centering and second derivative of the matrix  $X_{UV}$  was the best pre-treatment.

Statistical parameters and figures of merit for models built using UV spectra are shown in Table 1. The RMSECV, RMSEC and RMSEP values and correlation coefficients ( $R_c$ ,  $R_{cv}$  and  $R_p$ ) are similar for all UV-PLS models.

Analytical sensitivity ( $\gamma^1$ ) defines the lowest concentration difference between samples that can be distinguished by the method. In this case, the UV-PLS-Full model is more sensitive. SEL values can range from 0 to 1, where 0 means that the analytical signal contains information of interferences, while a value of 1 indicates that it does not have interferences. Thus, the low selectivity for all UV-PLS models was expected because the model was built based in spectra obtained with interferences. In addition, LOD values are similar for the UV-PLS-OPS and UV-PLS-GA models. It is important to highlight that OPS was able to perform the calculations for this data set in minutes, while the GA took about an hour. On this basis, the OPS algorithm was considered more efficient than GA.

Thus, the UV spectra of *J. curcas* oil extract with the variables selected by OPS are shown in Figure 3b. In Figure 3d, the measured *versus* predicted values of PEs contents for the calibration and prediction show a linear fit between these values, indicating that the model is capable of accurately predicting PEs in *J. curcas* samples.

The UV-PLS-OPS model was evaluated in order to verify if there was a correlation by chance. The Figure 3f shows that the UV-PLS-OPS model (true model) is separated from the other models, so, the true model was not obtained by chance.

To the best of our knowledge, there are no data in the literature about use of UV spectroscopy combined with multivariate calibration to predict PEs concentrations. In this work, models using UV spectroscopy and PLS directly on seeds, without any sample preparation, were not obtained successfully. This occurred because the PEs contents found in *J. curcas* samples were very low. In addition, the obtained spectrum contains information from other components present in high concentrations, overlapping the contribution of PEs.

Thus, the multivariate calibration UV-PLS-OPS models obtained in this work were built after two extraction steps (oil extraction of seeds and PEs extraction of oil). The proposed method can be used as an alternative method to predict PEs concentrations with high accuracy.

## Conclusions

NIR spectroscopy was unable to make accurate predictions of PEs directly on seeds due to its limited sensitivity. The NIR-PLS-OPS model has potential for classification purposes. Thus, this model can only be used only for screening in genetic breeding programs of *J. curcas*. UV spectroscopy and multivariate calibration models were used for the first time in this work to predict PEs concentrations in *J. curcas*.

This study has proven that UV spectroscopy can be successfully used to predict PEs contents in *J. curcas* samples. Thereby, the UV-PLS-OPS model can be perfectly applied as an alternative method to determine these compounds. In addition, the PLS-OPS model, in both cases, using NIR or UV spectroscopy, proved to be simpler and more accurate than the GA.

Therefore, in order to achieve a quick and reliable response to include or exclude samples in the genetic breeding program, NIR seed spectra are enough to make this decision. Additionally, for more accurate information of PEs concentrations, oil extract UV spectra must be obtained to get an accurate value of concentration instead of just classifying the sample.

## Acknowledgments

The authors are grateful to CAPES for funding this research; FAPEMIG (Project: CEX - APQ-02254-15) and Rede Mineira de Química (RQ-MG) supported

by FAPEMIG (Project: REDE-113/10; Project: CEX-RED-00010-14).

## References

- Anderson, L. G.; *Renewable Sustainable Energy Rev.* **2015**, *47*, 162.
- Gui, M. M.; Lee, K. T.; Bhatia, S.; *Energy* **2008**, *33*, 1646.
- Dias, L. A. S.; Misso, R. F.; Ribeiro, R. M.; Freitas, R. G.; Dias, P. F. S.; *Bahia Anal. Dados* **2009**, *18*, 539.
- Demirbas, A.; *Energy Convers. Manage.* **2003**, *44*, 2093.
- Antolin, G.; Tinaut, F. V.; Briceno, Y.; Castano, V.; Perez, C.; Ramirez, A. I.; *Bioresour. Technol.* **2002**, *83*, 111.
- Crabbe, E.; Nolasco-Hipolito, C.; Kobayashi, G.; Sonomoto, K.; Ishizaki, A.; *Process Biochem.* **2001**, *37*, 65.
- Shah, S.; Sharma, S.; Gupta, M. N.; *Energy Fuels* **2004**, *18*, 154.
- Dias, L. A. S.; Missio, R. F.; Dias, D. C.; *Genet. Mol. Res.* **2012**, *11*, 2719.
- Contran, N.; Chessa, L.; Lubino, M.; Bellavite, D.; Roggero, P. P.; Enne, G.; *Ind. Crops Prod.* **2013**, *42*, 202.
- Dias, L. A. S.; *Crop Breed. Appl. Biotechnol.* **2011**, *11*, 16.
- Divakara, B. N.; Upadhyaya, H. D.; Wani, S. P.; Gowda, C. L. L.; *Appl. Energy* **2010**, *87*, 732.
- Pecina-Quintero, V.; Anaya-Lopez, J. L.; Zamarripa-Colmenero, A.; Nunez-Colin, C. A.; Montes-Garcia, N.; Solis-Bonilla, J. L.; Jimenez-Becerril, M. F.; *Biomass Bioenergy* **2014**, *60*, 147.
- Heller, J.; *Physic Nut, Jatropha curcas L.*; International Plant Genetic Resources Institute: Rome, 1996.
- Freitas, R. G.; Missio, R. F.; Matos, F. S.; Resende, M. D. V.; Dias, L. A. S.; *Genet. Mol. Res.* **2011**, *10*, 1490.
- Balat, M.; *Energy Convers. Manage.* **2011**, *52*, 1479.
- Openshaw, K.; *Biomass Bioenergy* **2000**, *19*, 1.
- Ahmed, W. A.; Salimon, J.; *Eur. J. Appl. Eng. Sci. Res.* **2009**, *31*, 429.
- Makkar, H. P. S.; Aderibigbe, A. O.; Becker, K.; *Food Chem.* **1998**, *62*, 207.
- Herrera, J. M.; Martinez, C. J.; Ayala, A. M.; Siciliano, L. G.; Escobedo, R. M.; Ortiz, G. D.; Cevallos, G. C.; Makkar, H.; Francis, G.; Becker, K.; *J. Food Qual.* **2012**, *35*, 152.
- Makkar, H. P. S.; Becker, K.; *Plant Foods Hum. Nutr.* **1999**, *53*, 183.
- Adolf, W.; Opferkuch, H. J.; Hecker, E.; *Phytochemistry* **1984**, *23*, 129.
- Freitas, R. G.; Dias, L. A. S.; Cardoso, P. M. R.; Evaristo, A. B.; Silva, M. F.; Araujo, N. M.; *Genet. Mol. Res.* **2016**, *15*, 1.
- Devappa, R. K.; Bingham, J. P.; Khanal, S. K.; *Ind. Crops Prod.* **2013**, *49*, 211.
- Wang, Z.; Tang, L.; Hu, H.; Guo, Y.; Peng, T.; Yan, F.; Chen, F.; *J. Agric. Food Chem.* **2012**, *60*, 9567.
- Montes, J. M.; Technow, F.; Bohlinger, B.; Becker, K.; *Ind. Crops Prod.* **2013**, *43*, 301.
- Cen, H.; He, Y.; *Trends Food Sci. Technol.* **2007**, *18*, 72.
- Baye, T. M.; Pearson, T. C.; Settles, A. M.; *J. Cereal Sci.* **2006**, *43*, 236.
- Choung, M.-G.; Baek, I.-Y.; Kang, S.-T.; Han, W.-Y.; Shin, D.-C.; Moon, H.-P.; Kang, K.-H.; *Korean J. Crop Sci.* **2001**, *46*, 106.
- Helgerud, T.; Wold, J. P.; Pedersen, M. B.; Liland, K. H.; Ballance, S.; Knutsen, S. H.; Rukke, E. O.; Afseth, N. K.; *Talanta* **2015**, *143*, 138.
- Aleixandre-Tudo, J. L.; Nieuwoudt, H.; Aleixandre, J. L.; Du Toit, W. J.; *J. Agric. Food Chem.* **2015**, *63*, 1088.
- Morais, C. D. S.; Leme, L. M.; Valderrama, P.; Março, P. H.; *Rev. Bras. Pesq. Aliment.* **2015**, *6*, 70.
- Maluf, D. F.; Nagata, N.; Farago, P. V.; Guillermo, P.; Zamora, P.; *Rev. Bras. Farm.* **2008**, *89*, 39.
- Brereton, R. G.; *Analyst* **2000**, *125*, 2125.
- Beebe, K. R.; Kowalski, B. R.; *Anal. Chem.* **1987**, *59*, 1007 A.
- Wold, S.; Trygg, J.; Berglund, A.; Antti, H.; *Chemom. Intell. Lab. Syst.* **2001**, *58*, 131.
- Devappa, R. K.; Makkar, H. P. S.; Becker, K.; *Biomass Bioenergy* **2010**, *34*, 1125.
- Teófilo, R. F.; Ferreira, M. M. C.; *Quim. Nova* **2006**, *29*, 338.
- Makkar, H. P. S.; Maes, J.; Greyt, W.; Becker, K.; *J. Am. Oil Chem. Soc.* **2009**, *86*, 173.
- Baldini, M.; Ferfua, C.; Bortolomeazzi, R.; Verardo, G.; Pascali, J.; Piasentier, E.; Franceschi, L.; *Ind. Crops Prod.* **2014**, *59*, 268.
- Ferreira, M. M. C.; *Quimiometria - Conceitos, Métodos e Aplicações*; Editora Unicamp: Campinas, SP, Brazil, 2015.
- Kennard, R. W.; Stone, L. A.; *Technometrics* **1969**, *11*, 137.
- ASTM E1655-05; *Standard Practices for Infrared Multivariate Quantitative Analysis*; ASTM International: West Conshohocken, PA, 2012.
- Teófilo, R. F.; Martins, J. P.; Ferreira, M. M. C.; *J. Chemom.* **2009**, *23*, 32.
- Leardi, R.; Lupiáñez González, A.; *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195.
- www.deq.ufv.br/chemometrics, accessed in December 2016.
- Lorber, A.; Faber, K.; Kowalski, B. R.; *Anal. Chem.* **1997**, *69*, 1620.
- Faber, N. M.; *Anal. Chem.* **1998**, *70*, 5108.
- Currie, L. A.; *Anal. Chim. Acta* **1999**, *391*, 105.
- Muñoz de La Pena, A.; Espinosa-Mansilla, A.; Acedo Valenzuela, M. I.; Goicoechea, H. C.; Olivieri, A. C.; *Anal. Chim. Acta* **2002**, *463*, 75.
- Rodríguez, L. C.; García-Campaña, A. M.; Linares, C. J.; Ceba, M. R.; *Anal. Lett.* **1993**, *26*, 1243.
- Ridder, T. D.; Brown, C. D.; Steeg, B. J. V. E. R.; *Appl. Spectrosc.* **2005**, *59*, 787.
- Currie, L. A.; *Anal. Chim. Acta* **1999**, *391*, 103.

53. Currie, L. A.; *Appl. Radiat. Isot.* **2004**, *61*, 145.
54. Boqué, R.; Larrechi, M. S.; Rius, F. X.; *Chemom. Intell. Lab. Syst.* **1999**, *45*, 397.
55. Ortiz, M. C.; Sarabia, L. A.; Herrero, A.; Sánchez, M. S.; Sanz, M. B.; Rueda, M. E.; Giménez, D.; Meléndez, M. E.; *Chemom. Intell. Lab. Syst.* **2003**, *69*, 21.
56. Sarabia, L.; Ortiz, M. C.; *TrAC, Trends Anal. Chem.* **1994**, *13*, 1.
57. Oskoueian, E.; Abdullah, N.; Ahmad, S.; Saad, W. Z.; Omar, A. R.; Ho, Y. W.; *Int. J. Mol. Sci.* **2011**, *12*, 5955.
58. Makkar, H. P. S.; Becker, K.; Sporer, F.; Wink, M.; *J. Agric. Food Chem.* **1997**, *45*, 3152.
59. Ferrari, R. A.; Casarini, M. B.; Marques, D. A.; Siqueira, W. J.; *Braz. J. Food Technol.* **2009**, *12*, 309.
60. Pasquini, C.; *J. Braz. Chem. Soc.* **2003**, *14*, 198.

Submitted: October 26, 2016

Published online: January 18, 2017