

# LASSO MODELING AS AN ALTERNATIVE TO PCA BASED MULTIVARIATE MODELS TO SYSTEM WITH HEAVY SPARSITY: “BIODIESEL QUALITY BY NIR SPECTROSCOPY”

Cesar Mello<sup>1</sup>, Cassiano Escudeiro<sup>1</sup>, and Isao Noda<sup>2</sup>

<sup>1</sup>Integrated Institute of Clinical Research - IPclin, Jundiaí, State of São Paulo, Brazil.

<sup>2</sup>Department of Materials Science and Engineering, University of Delaware, Newark, DE, USA.

## ABSTRACT

*Principal component analysis (PCA) is a widespread and widely used in various areas of science such as bioinformatics, econometrics, and chemometrics among others. Once that PCA is based in the eigenvalues and the eigenvectors which are a very weak approach to high dimension systems with degrees of sparsity and in these situations the PCA is no longer a recommended procedure. Sparsity is very common in near infrared spectroscopy due to the large number of spectra required and the water absorption broad bands what makes these spectra very similar and with heavy sparsity in matrix dataset, demoting the precision and accuracy, in the multivariate modeling and within projections of data matrix in smaller dimensions. To overcoming these shortcomings the LASSO, a not PCA based method, model was applied to a NIR spectra dataset from Biodiesel and its performance was, statistically, compared with traditional multivariate modeling such as PCR and PLSR.*

## KEYWORDS

*lassomodel, sparsity, multivariate modeling, NIR spectroscopy, biodiesel from recyclable sources.*

## 1. INTRODUCTION

The multivariate modeling is a traditional chemometric tool and very useful to build models near infrared (NIR) spectroscopy data sets to ensure products quality and to evaluate, in a very fast way, the performance of industrial processes, in many areas, such as foods, fuels and polymers, for short. This spectroscopy technique provides fast, noninvasive and nondestructive analysis of samples, predicting a large number of samples properties, from NIR spectra [1]. The great majority of multivariate modeling use, at least in initial steps, the principal components analysis (PCA) [2]. The PCA based methods are very important especially in spectroscopy data sets, where the number of independent variables is greater than the number of acquired spectra; the equations system is overdetermined. However, PCA based methods as principal component regression (PCR) [3] and partial least squares regression (PLSR) [4] have at least two important shortcomings, when the matrix from NIR dataset tends to sparsity and the relationship among the sample properties and spectra, is lightly nonlinear. At this point in this text, it is important to make it clear that when we refer to matrix from NIR spectra set, we are referring to correlation matrix (C) calculated by the following steps:

### i. Data centering on average:

$$x_{i,j} = x_{i,j} - \bar{x}_j \quad (1)$$

Where  $\bar{x}_j$  is the average of the  $j$ -th column of data.

$$\bar{x}_j = \frac{1}{I} \sum_{i=1}^I x_{i,j} \quad (2)$$

Aonde  $I$  é o número de linhas,  $j$  o número de colunas e  $I$  o número total de colunas.

### *ii. Autoscale*

$$x_{i,j} = \frac{x_{i,j} - \bar{x}_j}{s_j} \quad (3)$$

Where  $S$  is the standard deviation from data matrix ( $X$ )

### *iii. Correlation matrix C*

$$C = \frac{\mathbf{x}_{\text{autoscaled}}^T \mathbf{x}_{\text{autoscaled}}}{I-1} \quad (4)$$

Although some authors claim that their PCR and PLSR models can model small nonlinearities and sparsity in very small degree, with the addition of extra principal components (PC's). However in practice, external disturbances such as light scattering, baseline fluctuations and noise, for short, will introduce nonlinearities and increasing to the sparsity in spectral data, deteriorating the prediction accuracy step and precision from PCR and PLSR modeling. However, in many cases it is possible to minimize these sources of nonlinearities in a very effective way. There are several mathematical techniques to do these tasks such multiplicative scattering correction (MSC)[5], noise minimization by Fast Fourier Transform (FFT) [6] or wavelets filter[7] Gram-Schmidt orthogonalization [8] and detrend baseline correction[9] just to name a few. Such methods of pre-processing spectral data set, allow the use of PCA based models without great problems. However, the spectroscopic technique NIR presents several shortcomings very difficult to overcoming. The intrinsic NIR broad absorption bands called overtones and highly undesirable effects due to water absorption broad bands making the NIR spectra a very broad bands spectrum, referred as overtone bands, of difficult Physical and/or Chemical interpretation and analysis. This broad band affects the NIR spectrum as whole making the NIR spectra from the same sample very similar, with heavy overlaps. All of these physical effects, together, such effects together, generate heavy sparsity in  $C$ . In such cases the PCA based multivariate modeling methods completely lose accuracy and precision, with orders of magnitude dependent on the system under evaluation. Moreover, in such cases it is necessary to acquire a large number of NIR spectra for prediction and validation procedures of these models.

Thus we have a reasonably complicated situation, that is, a large data matrix with heavy sparsity. Accordingly, in such cases multivariate modeling methods based on principal component analysis can fail. Therefore, in such cases it is recommended to analyze, previously, the asymptotic behavior of the Principals Components i.e. the direction of the main components (PC's) into large dimension systems as in matrices from NIR spectroscopy dataset. The asymptotic behavior analysis basically shows that if the first few eigenvalues of a population covariance matrix are large enough compared to the others, then the corresponding estimated PC directions are consistent or converge to the appropriate subspace and most other PC directions are strongly inconsistent, that is, such PC's do not have anything physical meaning. After such mathematical analysis one can decide on whether it is possible to use PCA or not. However it is much simpler to use  $C$  to calculate its condition number before applying to PCA. Thus, in a first approximation one can establish the very simple rule to the use of the PCA by calculating the condition number (CN) [10] from NIR matrix of dataset. However a little bit further analysis over CN from  $C$  it is necessary. The CN from a matrix is a measure of whether

the problem has "good condition" to be treated numerically. A problem with a small CN is called well-conditioned, while problems that have a high condition number are called poorly conditioned due to, among others effects, the sparsity. For example, the CN associated with a linear system is a number that estimates the accuracy that can be obtained for an approximate solution of the system. Note that this is before the effects of rounding errors are taken into account. Paired with this problem there are any number of algorithms that can be used to solve such kind of problem, that is, to calculate the solution. Some numeric algorithms have a property called backward stability. In general, a backward stable algorithm can be expected, to accurately solve, only, well-conditioned problems. However, CN is an intrinsic property of the C and we can use the following simple rule of thumb: if the CN is small them is possible, in a first approximation, to apply to principal component analysis (PCA) and to use the traditional and well established multivariate models like PCR and PLSR. However, this type of mathematical analysis is a little cumbersome and very unusual in daily use of NIR so it is much more practical and efficient to use the LASSO modeling as a choice method. Despite the possible difficulties mentioned above, the PCA is still an ordinary and very efficient mathematical method for dimensionality reducing from the spectral data set with the shortcomings aforementioned and the vast majority of NIR spectrometers have software to perform PCA based methods such as PCR and PLSR, among others. The widespread use of these nonparametric methods relies on its easy computational implementation, and such methods, in general, are algorithmic methods. In the 1980s, routine interactive computing was just beginning to raise its head, and exploratory data analysis was a new idea. Since then, we have witnessed a number of remarkable developments in local computing power and data storage. Besides that, efficient and interactive chemometrics packages have enabled sophisticated data analyses to carry out effortlessly and the vast majority of NIR spectrometers have software to perform PCA based methods such as PCR and PLSR, among others. However such kind of nonparametric methods provide approximates solutions and not necessarily robust ones, causing a lot of wrong solutions and misinterpretations even in quite simple systems. These nonparametric methods are not robust inmathematically sense due to its high sensibility to the nonlinearities, outliers and heavy sparsity in C. Hence, before using the nonparametric methods, it might be a good strategy to analyze the sparsity of C, before building PCA based models as extensively mentioned. In this work, a parametric method not largely usedin multivariate modeling mainly in, at least, three very important areas such as chemometrics, econometrics and financial sciences, called Least Absolute Shrinkage, Selection Operator (LASSO) was used and its performance was evaluated against models done with PCR and PLSR. The LASSO method, also reduce the dimensionality of a data set such as PCA does, but using a different mathematical approach as we will describe ahead [11, 12]. The LASSO method, also reduce the dimensionality of C such as PCA does, but using a different mathematical approach as we will describe ahead. PCR and PLSR as along with LASSO method were used to build multivariate calibration models to determinate the percentage in mass (w/w) of Biodiesel[13], from animal fat recyclable sources, contained in Diesel B20, which is a complex blend of soybean, animal fat Biodiesel and petroleum Diesel. Depending on the percentage (w/w) of animal fat Biodiesel and operational temperatures, it can cause several and different engine failures. Hence, the multivariate modeling very robust, less sensitive to outliers and mainly sparsity is a very attractive alternative to determine the content of animal fat Biodiesel in Diesel B20, by NIR spectroscopy, to improve the quality control of fuel to avoiding operational failures related to the Diesel quality, in engine vehicles. Nowadays, the use of multivariate modeling in modeling NIR spectroscopic data sets is widely employed, being considered a standard procedure for quantitative analysis in several analytical techniques and precisely because of this fact, very robust multivariate modeling methods with low sensitivity to sparsity of the NIR spectra matrix should be used.

## 1.1. The LASSO Modeling

The LASSO method is a hybrid of variable selection and shrinkage estimators. The procedure shrinks the coefficients of some of the variables not simply towards zero, but exactly to zero, giving an implicit form of variable selection. In the simple standard multiple regression we have the Eq.5:

$$y_i = \lambda_0 + \sum_{j=1}^p \lambda_j \cdot x_{ij} + \varepsilon \quad (5)$$

where  $y_1, y_2, \dots, y_n$  are measurements on a response variable  $y$ ;  $x_{ij}$ ,  $i = 1, 2, \dots, n$ , and  $j = 1, 2, \dots, p$ , are corresponding values of  $p$  predictor variables;  $\lambda, \lambda_0, \lambda_1, \dots, \lambda_p$  are the parameters in the regression of the Eq.5. The  $\varepsilon$  value is an error term. In least squares regression, the parameters are obtained minimizing the residual sum of squares:

$$\min. \sum_{i=1}^n (y_i - \lambda_0 - \sum_{j=1}^p \lambda_j x_{ij})^2 \quad (6)$$

The LASSO modeling imposes an additional restriction on the coefficients, namely:

$$\sum_{j=1}^n |\lambda_j| \leq t \quad (7)$$

Thus in LASSO modeling we have a typical problem of constrained minimization problem where  $\lambda$  is a kind “tuning parameter”, suitable choices of this constraint,  $t$ , has an interesting property, which forces some of the coefficients in the regression equation to become zero. The computation of the solution to Eq.6, with the restriction given by Eq.7 is a quadratic programming problem with linear inequality constraints. There are several efficient and very stable algorithms for solving this kind of inequality problem. Finally, it is possible to summarize the LASSO theory as a simple ordinary least square(OLS)[14] with restrictions leading to an exact solution, obtained by quadratic programming. The LASSO parametric method was applied in this work to build a direct modeling to a spectral data set with one hundred of NIR spectra, to determine the percentage of animal fat Biodiesel in Diesel B20. However, it can be used to C with two, three hundred or even more NIR spectra. Finally, the results were compared with two traditional nonparametric methods, the partial least square regression(PLSR) and principal component regression (PCR).

## 2. EXPERIMENTAL

The process of producing biodiesel from renewable natural sources is well established and widely known and references can be found anywhere as aforementioned. Therefore, it is not necessary to present in great detail such a process in this work. We emphasize once again that the most important is the evaluation of biodiesel produced through NIR spectroscopy. With this fact in mind the experimental part can be simplified. Therefore, experimental part can be simplified. A set of one hundred blends of Diesel B20 and animal fat Biodiesel were initially prepared, with different concentrations of animal fat Biodiesel in Diesel B20. The animal fat Biodiesel content in Diesel B20 ranged from 0.00% to 13.80 % (w/w), where the Diesel B20 is a blend of soybean Biodiesel and petroleum diesel, containing at most 50 ppm in mass of sulfur. Near infrared (NIR) spectra were acquired, after each addition of animal fat Biodiesel using a NIR spectrometer Perkin Elmer 100N spectrometer equipped with a liquid reflectance accessory in the range from 4,000 to 400  $\text{cm}^{-1}$  using a resolution of 4  $\text{cm}^{-1}$  and 64 scans per sample. All samples were acquired in triplicate and the average spectrum was used to build the calibration model. The set of one hundred NIR spectra was split into two data sets, using the Kenston[16] algorithm, one of them for calibration with 70 samples and the other one with 30 samples, for predictions and modeling validation. The C was pre-processed, using MSC and the

Applied Mathematics and Sciences: An International Journal (MathSJ), Vol. 7, No. 1, March 2020  
noise was minimized by using Fast Fourier Transformfilter. All computational procedures were performed with routines developed in GNU Octave 5.1.0 [15].

### 3. RESULTS AND DISCUSSION

Using the above procedure, in materials and methods section, a set of one hundred NIR spectra, were acquired in reflectance mode. To better visualization the usual transformation in the reflectance measurements ( $R$ ), was applied:

$$\log \frac{1}{R} \quad \text{Eq.8}$$

The set of spectra was plotted with  $\log 1/R$  in abscises and the wavenumbers in  $\text{cm}^{-1}$  in the ordinates, as shown the Figure 1.

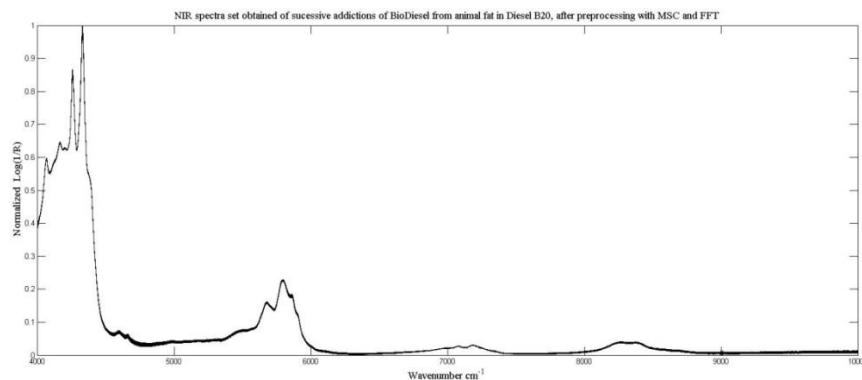


Figure 1. The NIR Spectra obtained for 99 additions of animal fat Biodiesel in Diesel B20, and one with pure Diesel B20, in a final set of 100 NIR spectra.

The coefficients  $\lambda$  were calculated using simple quadratic programming [16], as shown in the Figure 1. The Figure 2 points out the shrinkage of the coefficients ( $\lambda$ ) to zero, as described in the introduction of LASSO method.

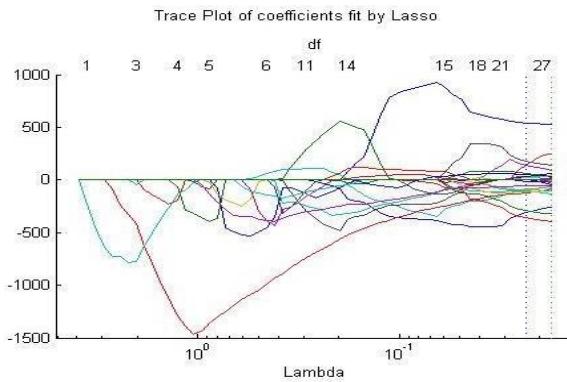


Figure 2. Graphical representation of the Lambda ( $\lambda$ ) set values fitted by LASSO against degrees of freedom (df).

In the Figure 2 is shown  $\lambda$  values obtained by cross-validation. The cross validation was done by using blocks [17], since it is impossible to use the leave one out (LOO)scheme for LASSO method.

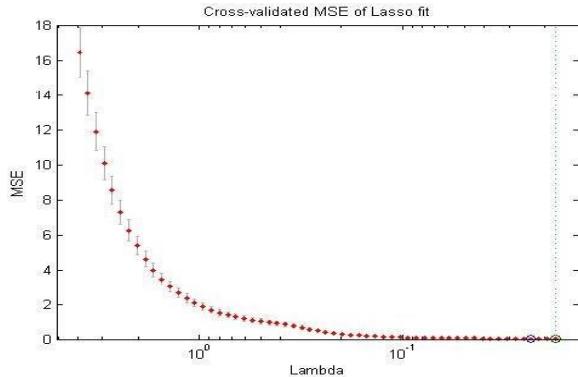


Figure 3. Plot of  $\lambda$  against mean square error (MSE), to choose the  $\lambda$  value with MSE.

The calibration curve obtained using the LASSO method with in the values of the abscissas being the actual value of % (w/w) animal fat Biodiesel and the values in the ordinates, the obtained values of % (w/w) animal fat Biodiesel from LASSO modeling. The merit figures are inside of the Figure 4.

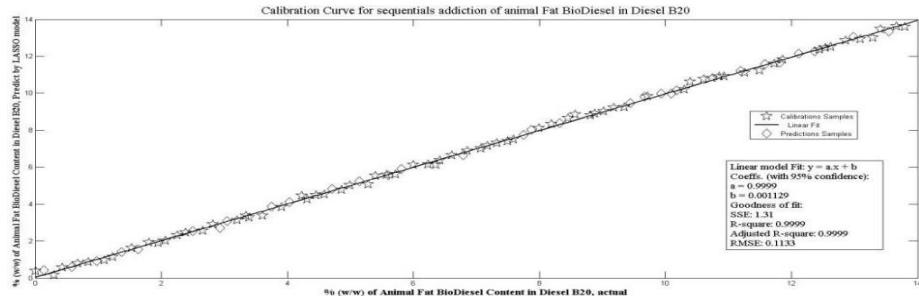


Figure 4. Plot of calibration curve obtained by Lasso method. The goodness of fit is show inside the figure;  $R^2 = 0.9999$ .

The plot of Q residuals vs. Hoteling's  $T^2$  statistic gives a compact view of both, residual and score outliers (as well as inliers). The axis shows that scores capture 90.25% (abscises) of the total variance and 9.75% (ordinates) of the variance remains in the residuals. The Figure 5 shows there are not so many scores outliers; some of them present a small value. As can be observed in the Figure 5 there, the LASSO method outperforms the PCA-based models, because the presence of some outliers does not degrade the performance as it does in PCAbased models.

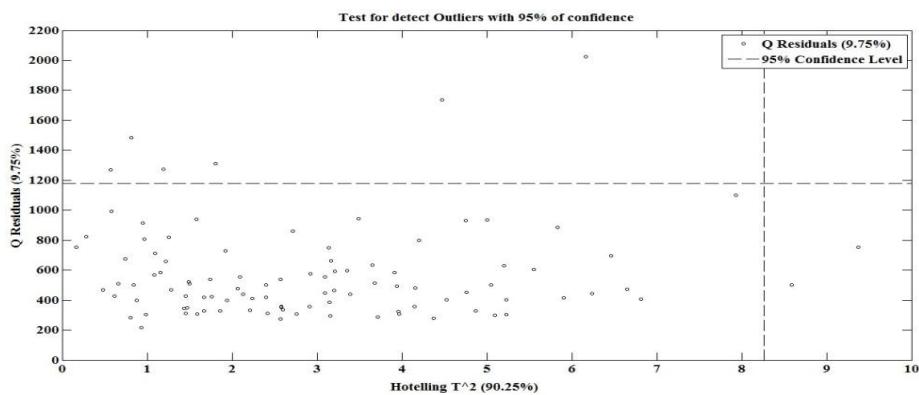


Figure 5. Plot of Hotelling  $T^2$  test to verify the presence of outliers.

A possible solution for PCA based models is to remove the outliers before, which is a cumbersome and in some sense, a subjective task. Moreover, it is essential to emphasize that a reduction in the size of the data set also reduces, at same time, the performance of the PCA based methods due to condition number, is always a trade-off. Nevertheless, if it is necessary to remove some outliers, it is possible to apply a heuristic method to remove them from the large data set. This heuristic method is quite simple; manually removing the outliers whose values are very discrepant in the data set. Such heuristic method goes against one of the main idea of chemometrics, which is to extract maximum information with a minimal experimental and subjective work. In cases in which the data set are small, as observed in many chemometrics applications, perhaps the best way is to keep all outliers and hope that the error does not increase so much. The LASSO method is a very robust alternative for these cases, that is, do not show high sensitivity to outliers, especially for small datasets. The optimal number of Principal Component (PC's) values to PCR were calculated by cross-validation using the LOO method. From the Figure 6 is easy to choose the optimal number of PC's, four to this data set. Using the parsimony principle and observing that after the value 4 the curve of RMSECV/RMSEC against PC's tend toward asymptotically to a constant value as shown in the Figure 6.

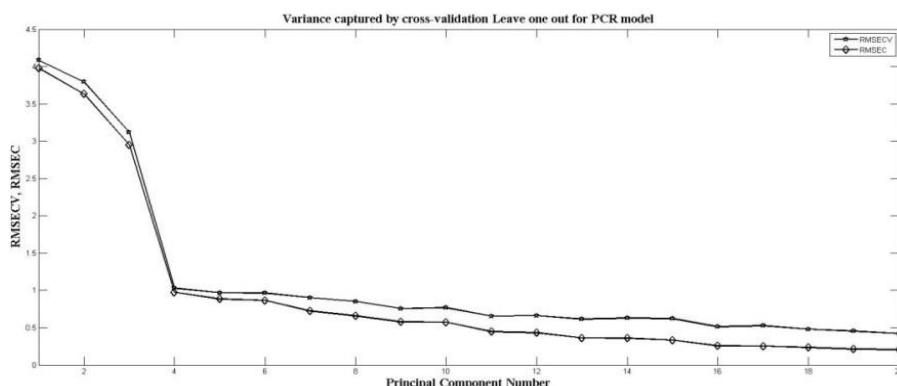


Figure 6. Plot of root mean square error for calibration (RMSEC) and root mean square error for cross validation samples (RMSECV), for PCR using model leave one out method.

The Figure 7 shows the calibration curve to PCR, in the values of the abscissas are the actual value of % (w/w) animal fat Biodiesel, and in the values of the ordinates are the obtained values of % (w/w) animal fat Biodiesel, with PCR model. The goodness of fit is inside the Figure 7.

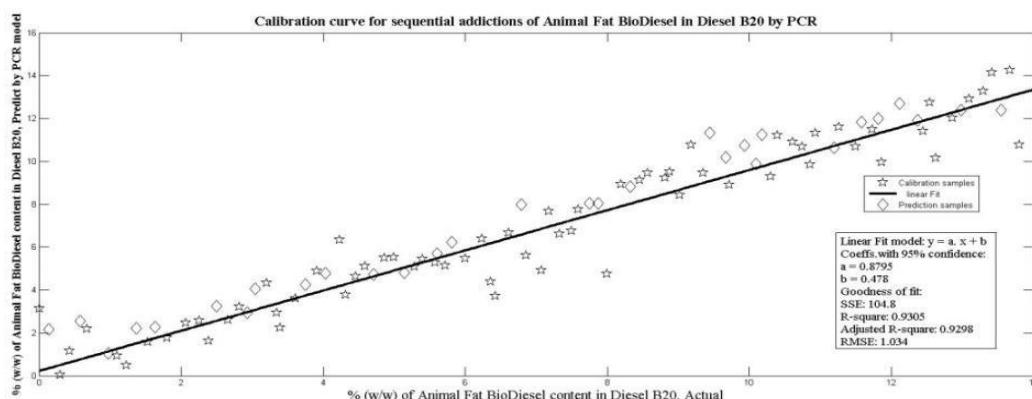


Figure 7. Plot of calibration curve obtained by PCR method, the goodness of fit are show inside of the figure;  $R^2 = 0.9298$ .

The optimal number of latent variables (LV's) value to PLSR was obtained by cross-validation using the LOO method. From the Figure 8 is easy to choose the optimal number of LV's, such that for this data set, these number is around four or five once that after these values of RMSEC/RMSECV the curve tend asymptotically to zero. Then, using the parsimony principle of was chosen four LV's to build the PLSR model.

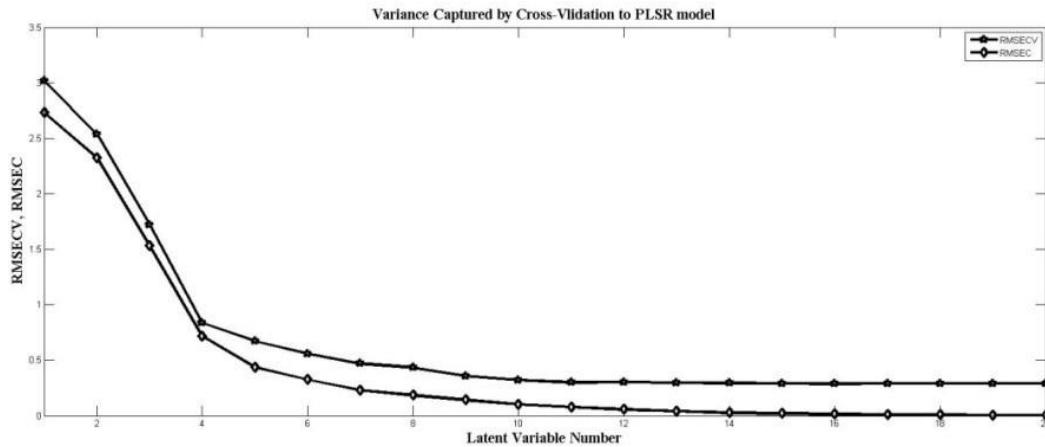


Figure 8. Plot of (RMSEC) and (RMSECV), for PLSR using model leave one method.

The calibration curve for PLSR shown in the Figure 9, in the abscissas values are the actual values and in the values of the ordinates are the obtained values of (%(w/w) animal fat Biodiesel, with PLSR model. The merit figures are shown inside of the Figure 9.

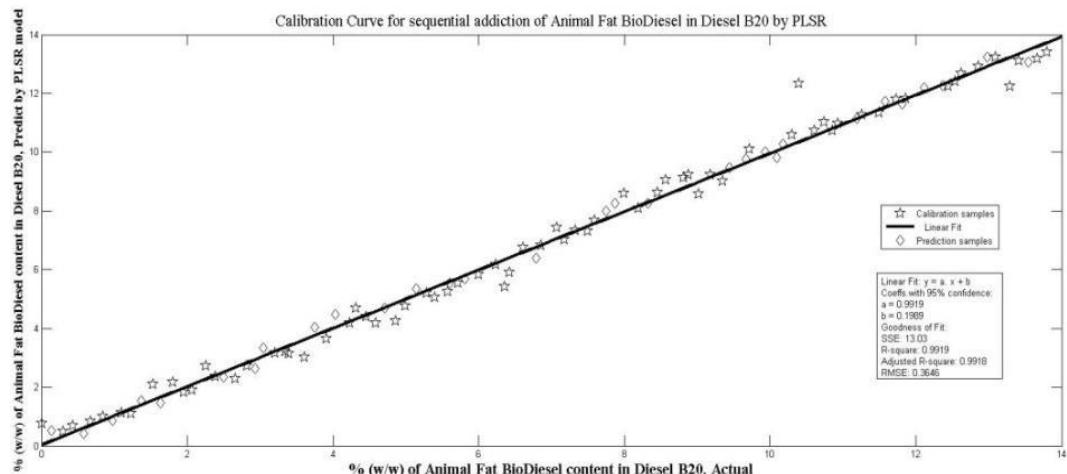


Figure 9. Plot of calibration curve obtained by PLSR method, the goodness of fit are shown in the figure;  $R^2 = 0.9918$ .

The next three graphics (Figures 10, 11 and 12), the models obtained will be further checked, evaluating the possible presence of overfitting or underfitting. Such assessment will be performed in a simple, but statistically efficient and correct way. This procedure will be carried out by fitting a normal probability distribution function for to fit the residues leave of each one model performed in this work. Naturally, a model with the best fit to the experimental data set should show invariably residues with a tendency to a perfect Gaussian distribution shape in ideal conditions. In Figure 10 it is possible to observe that the model obtained with PCR leaves an approximate Gaussian distribution without tail. The Gaussian fit of the residues indicates clearly that the PCR model does not present overfitting or underfitting.

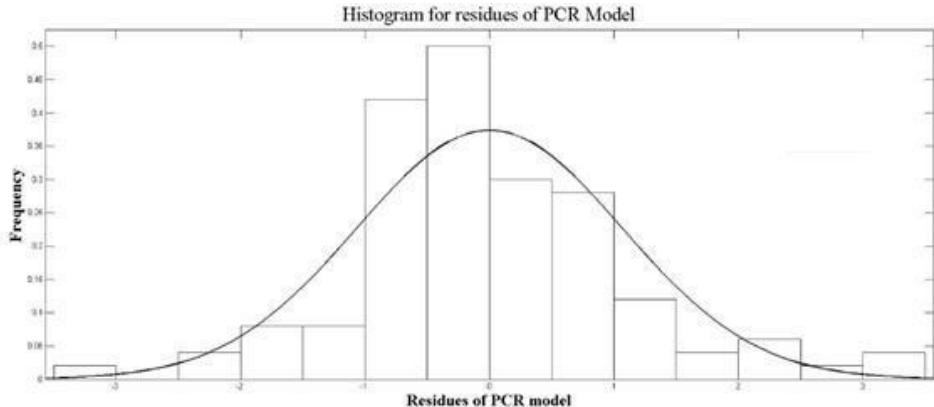


Figure 10. Gaussian fit of residues leave by PCR, the goodness of fit are shown in the Figure 10;  $\bar{x} = 427562 \cdot 10^{-14} = 4.7562$  and  $s = 1,1381$ .

In Figure 11 it is possible to observe that the model obtained with PLSR leaves a Gaussian distribution with a light tail, indicating that the PLSR model presents a small overfitting and/or underfitting.

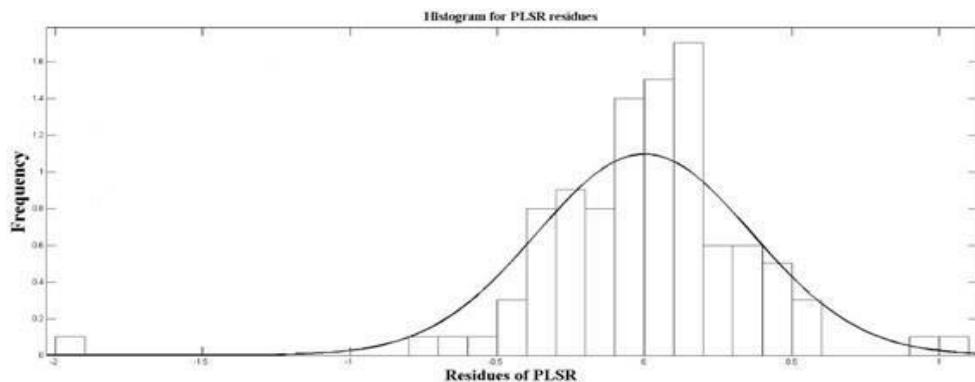


Figure 11. Gaussian fit of residues leave by PCR, the goodness of fit are shown in the Figure 10;  $\bar{x} = 1,6954 \cdot 10^{-14}$  and  $s = 1,1381$ .

A definitive discussion about a comparative approach of PCR against PLSR is not a simple task, once that there are so many possible interpretations to overfitting or underfitting. However is possible to point out, specifically for this dataset without being flippant, that the PCA applied for independent variables revealed a poor performance, possibly due to the size of data set. There is no consensus, in the literature, to explain the better performance of PCR than PLSR in some data sets, since the optimal number of LV's was chosen by cross validation, as shown in Figure 8. Perhaps, if the dataset was larger, the light tail would not happen. However, obtaining a large data set is not a good idea too, as aforementioned and this little overfitting presented by PLSR model, is not a big problem for practical purposes, once that, local models have been constructed for both kind of model and in this range of concentrations PLSR model is valid and useful too.

In Figure 13 it is possible to observe that the model obtained with LASSO leaves a Gaussian distribution with a very light tail as PLSR, indicating clearly that the LASSO model present a smaller overfitting and/or underfitting than PCR and PLSR Models in an overall way.

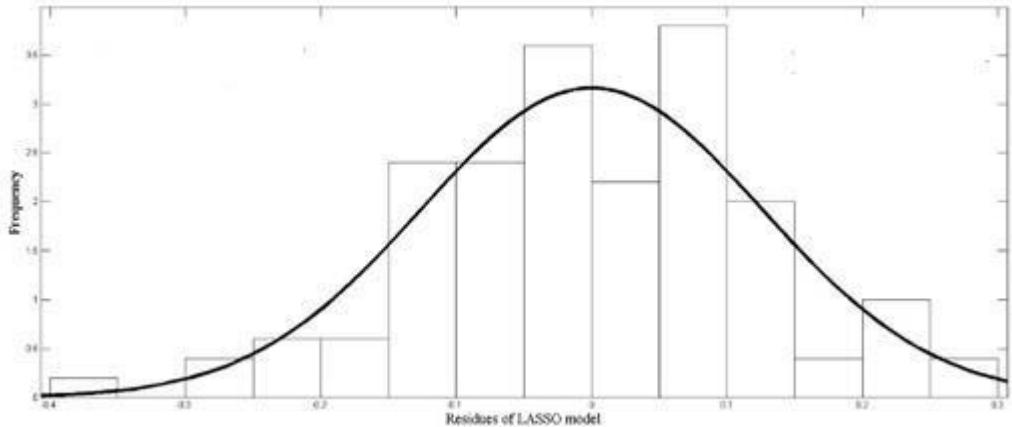


Figure 12. Gaussian fit of residues leave by PCR, the goodness of fit are shown in the Figure 10;  
 $\bar{x} = 1,2093 \cdot 10^{-14}$  and  $s = 0,0629$ .

The relative performances from different models, was evaluated in terms of root-mean-square-error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_{\text{model},i})^2}{n-1}} \quad (9)$$

Where  $y$  is the actual value and  $\hat{y}$  is the predicted value by the model and  $n$  is the number of the samples used, in the prediction set. A very simple analysis of RMSE values, shown in Table 1, indicates that the LASSO method was the most efficient multivariate modeling, followed by PLSR and PCR respectively.

Table 1. Results of RMSE for different types of models evaluated: LASSO, PLSR and PCR.

Model	
LASSO	0.1133
PLSR	0.3646
PCR	1.3040

A very simple analysis of RMSE values, shown in Table 1, indicates that the LASSO method was the most efficient calibration method, followed by PLSR and PCR respectively. Such differences observed in the models, are due to the nature of the methods and size of NIR spectroscopy data set used in this article. The parametric model carried out with the LASSO method has a very robust numeric solution while the nonparametric PLSR and PCR models are PCA based methods, which, in some data sets, provides a mathematically weak and unstable approximated solution. Perhaps, if the data used in this article was larger, the results obtained with the nonparametric model could be approximately equal to the parametric method LASSO. It is possible to affirm that the nonparametric methods used in this article are very sensible to outliers and the size of data set, that is, they are not robust methods for small data sets or ill conditioned data set. To check if such the assumptions previously made are correct an F-test [18], with 95% confidence level was used to compare RMSE for the different multivariate modeling used in this article, the F-test is given by Eq.10:

$$F(p_i, p_j) = \frac{RMSE_{\text{parametric model, LASSO}}^2}{RMSE_{\text{Non-parametric model, PCR and PLSR}}} \quad (10)$$

Where,  $p$  is the number of validation samples, in our case  $p = 30$ . The F-test *critical value* at the 95% confidence level is  $F_{critical} = 1.61$ . When the F-test is applied to the RMSE values obtained from different modeling methods, it is observed that the LASSO model is significantly better than PLSR ( $F = 2.1$ ) and PCR ( $F = 2.4$ ). Despite of the best performance, LASSO is a little bit more difficult to implement, computationally, its require more computational time, around 1 minute while the PLSR and PCR modeling only few seconds. However, this statement may not have a long life since the new low cost microcomputers and new numerical algorithms that are more efficient have grown exponentially, in addition to enhanced processing power, what is drastically reducing CPU use time.

#### 4. CONCLUSIONS

Based on the results obtained and in the discussion before, it is possible to conclude that the method LASSO showed better results than nonparametric methods, like PCR and PLSR, due to size of data set as aforementioned. Another very important outcome about PCA based inverse models is the necessity of repeat the NIR spectrometer calibration frequently. Once that with the pass of the time and the high use of NIR spectrometer the internal calibration is lost and the NIR spectra are no longer reliable. When we decided to use the LASSO method as multivariate modeling method for the quality determination of Biodiesel by NIR spectrometry, our intention was not only to evaluate the use NIR spectroscopy applied to the quality of Biodiesel but also verify if the LASSO method overcomes the performance of traditional inverse chemometrics as PCR and PLSR. Although we have presented the main advantages of the LASSO over PCA based methods in spectroscopic datasets where the CN is large, the LASSO method can present severe discrepancies if the distribution of regression errors shows a heavy tail for a finite sample. Sparse principal component analysis (SPCA) aims at estimating a PCA-like model where sparsity is induced on the model parameters; scores and / or loadings. However, this is still an algorithm-based method, the PCA, without an exact solution. The SPCA works reasonably well in certain types of spectroscopy, such as mass spectroscopy where the spectroscopy matrix dataset, can have its sparsity reduced experimentally, only in some cases, using spectrometers of high accuracy and precision, as well as chemical standards of very high purity, making the very high-priced method accessible only in big researcher centres, inaccessible to the vast majority of researchers worldwide. Therefore, the SPCA goes in the opposite direction of the PCA-based methods which were developed, precisely, to obtain calibration models in situations where high instrumental accuracy was not possible or important, as occurred in the NIR spectroscopy in the early 80s. However, the purpose of this article is not to criticize the use of the PCA or SPCA-based methods [19], but to present alternatives with exact solutions concerning purely algorithmic methods.

The better results obtained with LASSO method shown clearly that PCA based models had an important shortcoming to data sets with tendency to sparsity, which is quite usual in diary practical situations of NIR spectroscopy in reflectance mode. Besides that, the LASSO is a very simple alternative to PCA based models where the matrix describing the spectroscopic data set tends to become still sparser. Finally, it is possible to notice that the parametric methods works better in presence of outliers than nonparametric methods, such PCR and PLSR.

#### 5. ACKNOWLEDGEMENTS

The Authors thanks to FAPESP, CNPq, University of Delaware and Institute for Integrated Clinical Research – IPclin.

## REFERENCES

- [1] Siesler, H. W. , Y. Ozaki and Kawata, S. , (2001) “Near-Infrared Spectroscopy: Principles, Instruments”, Applications, Wiley-VCH Verlag Gmbh.
- [2] Jolliffe, I.T., 2nd edition, (2002), “Principal Component Analysis”, Spring.
- [3] R, De Maesschalck, F, Estienne , Verdú-Andrés, Candol. J, Centner V. A. ,Despagne, F., D. Jouan-Rimbaud, Walczak. B. ,Massart , D.L., de Jong, S, de Noord, O.E , Puel, C, Vandeginste, B.M.G.
- [4] Brereton, R. , Jansen, J. , Lopes, J., Marine, F., Pomerantse, A., Rodionova O., Roger, M. J., Walczak, B. and Tauler. R.( 2018), Anal. and Bioanal. Chem., Vol. 410 (26), pp 540-543.
- [5] Naes, T., Isaksson, T and Kowalski, B., (1990), Anal. Chem. Vol. 62(7), pp.664-673.
- [6] Mello. C.; Ozório, E.; Kubota, L. T.; (2000), Química Nova, 2000, Vol. 25, pp690-698.
- [7] RIsmail, A.; SAsfour. S. (1999), Jour. Biomech., Vol. 32, pp317-322
- [8] Å.Björck., (1994), Linear Algebra and its Applications, Vol. 197-198, pp. 297-316.
- [9] Tanabe. J.; Miller. D.; Tregellas. J.; Freedman. R. and Mayer. F.G. (2002); NeuroImage, Vol. 15, pp-902-907.
- [10] Charpentier. S.; Fouchet. K and Zarouf. S. (2019), Analysis and Mathematical Physics, Vol.9 (3), pp-971-990.
- [11] Friedman. J; Hastie. T. and Tibshirani. R.,(2008) Biosataistics,Vol. 9, pp432-441
- [12] Horowitz. J. L. (2019), Annual Review Economics, Vol. 11, pp193-224.
- [13] Knothe. G (Editor), Krah. Jl (Editor), Van Gerpen. J (Editor), (2010). “The Biodiesel Handbook”, Academic Press and AOCS Press.
- [14] Goldberger, A. S. (1964). "Classical Linear Regression". Econometric Theory. New York: John Wiley & Sons. pp. 158.
- [15] Pencheva. T ;Atanassov. K and Shannon. A. Anthony ; Tenth Int. Workshop on Generalized Nets Sofia, 5 December 2009, 1-7.
- [15] <http://www.gnu.org/software/octave/doc/interpreter/>, last accessed in November 2019.
- [16] Marguerite. F. and Wolfe P.; "An Algorithm for Quadratic Programming." Naval Research Logistics Quarterly 3 (1956): 95-110. Web. 4 June 2015.
- [17] Wülfert.F; Kok. T. W; de Nood. O. E.andSmilde A. K.(2000), Chemon. Intell. Lab. Sys., Vol. 24 (2), pp189-200.
- [18] Box, G. E. P. (1953). "Non-Normality and Tests on Variances". Biometrika. Vol.40 (3/4), pp318– 335.
- [19] Gajjar, S.; Kulahci, M. and Palazoglu. A., (2016), “Use of Sparse Principal Component Analysis (SPCA) for Fault Detection”, IFAC,pp 693-698, 11th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems June 6-8, 2016. NTNU, Trondheim, Norway