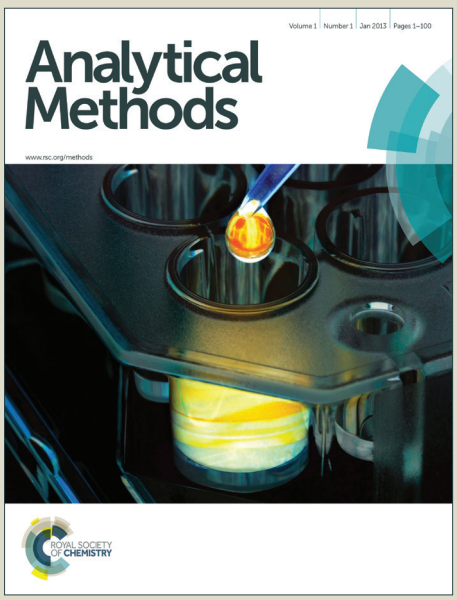
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Ultraviolet spectroscopy and supervised pattern recognition method to authentication of transgenic and non-transgenic soybean oils

F. C. G. B. S. Alves[a] and P. Valderrama[a*]

[a]Universidade Tecnológica Federal do Paraná (UTFPR), P.O. Box 271, 87301-899, Campo Mourão – Paraná – Brazil

* **Corresponding author**: +55 (44) 3518-1400; patriciav@utfpr.edu.br or pativalderrama@gmail.com

Abstract

A methodology was developed to authentication of transgenic from non-transgenic soybean oils samples by using Ultraviolet (UV) spectroscopy coupled with Partial Least Squares Discriminant Analysis (PLS-DA). The accuracy, represented by RMSEC and RMSEP, was 0.223 and 0.278, respectively. For the model, the sensitivities were 0.875 and 1.000 for transgenic and non-transgenic classes, respectively. Based on these results, the model was able to classify the non-transgenic soybean oil samples. The transgenic class present specificity equal to 1, this result means that any non-transgenic sample was classified in the transgenic class. The non-transgenic class present specificity equal 0.875 due the prediction of two samples of transgenic class in the non-transgenic class. The ability of UV spectroscopy for soybean oil authentication can be assigned to the bathochromic shift, probably due to the differences in the chromophore group of genotypic structure present in the transgenic and non-transgenic samples.

Keywords: soybean oil, UV spectroscopy, authentication, chemometrics, bathochromic shift

1. Introduction

Food authentication could be considered a further guarantee for the quality and safety of a foodstuff.[1] Although of the genetically modified food (GMF) offer some advantages such as better nutritional value or resistance against insects and diseases. In Brazil and European Union the GMF is still considered undesirable and the implementation of any labelling policy will require the development of reliable detection methods.[2]

Polymerase chain reaction (PCR) is usually accepted as an analytical method to detect GMF.[3] The PCR method is used to amplify target DNA fragments,[4] however, to achieve successful results in DNA amplification methods there is a dependency of the efficiency of DNA extraction protocols.[3] DNA extraction is considered a critical point in the analysis of complex samples and very processed food matrices.[5]

The authentication of transgenic and non-transgenic foods by using infrared spectroscopy and chemometric methods has been explored. As an example, Alcantara et al.[6] used a Fourier Transform Mid- Infrared (FT-MIR) spectroscopy coupled with Principal Component Analysis (PCA) and K-Nearest Neighbor (KNN) to discriminate transgenic from non-transgenic soybean grains. Moreover, a review on the identification of transgenic foods using Near Infrared (NIR) spectroscopy was published by Alishahi et al.[7] Luna et al.[3] proposed a rapid characterization of transgenic and non-transgenic soybean oils by using NIR spectroscopy, PCA, Support Vectors Machine-Discriminant Analysis (SVM-DA) and Partial Least Squares with Discriminant Analysis (PLS-DA). Besides this, an approach to discriminate transgenic from non-transgenic soybean oil was proposed by using FT-MIR, SVM-DA, PLS-DA and Soft Independent Modeling of Class Analogies (SIMCA).[2]

Therefore, the combination of spectroscopy with chemometrics is a powerful tool for quality control in food science because it is a fast technique that needs no pretreatment (or minimal pretreatment) of the sample, and it offers results for classification and/or authentication of edible oils.[2,3] Due to these advantages, and considering that no application of ultraviolet (UV) spectroscopy for this purpose was found on the literature, the objective of this paper is developing a methodology for authentication of transgenic from non-transgenic soybean oil samples using UV spectroscopy coupled with PLS-DA chemometric tool. For liquid samples, UV spectroscopy can be used since the region contains information about the chemical structures of the compounds due to chromophore absorptions.[8] Furthermore, there are several papers combining UV and chemometrics to analyzes food samples.[8-13]

2. Experimental

One hundred and five soybean oil samples (65 transgenic and 40 non-transgenic) of different brands and different lots were purchased in various local supermarkets at Campo Mourão, Brazil. All the bottles were labeled by own the manufacturer as being transgenic or non-transgenic.

UV spectra were collected at 1 nm intervals over the 200–400nm spectral region on an Ocean Optics spectrometer USB-650-UV-VIS model by using a quartz cuvette with 1mm of the optical path. The soybean oil was analyzed directly, without further preparation. The data were processed with MATLAB R2007b, were the spectra was organized into a matrix and the supervised pattern recognition method PLS-DA was performed with application of the PLS Toolbox 5.2 from Eigenvector Research.

3. Chemometric methods

3.1. PLS-DA

PLS-DA is a supervised pattern recognition method,[14] and has its foundation on the Partial Least Squares Regression (PLSR).[15] Both, PLS-DA and PLSR, are based on PCA, an unsupervised pattern recognition method.[16] These methods can be applied to first order data as UV spectra, were one vector is obtained for each sample. The vectors are organized in a matrix (**X**) that is decomposed, by principal components (PCs), in a product of two matrices: scores and loadings matrices.[13]

In PLSR the matrix **X** is related to another matrix, **Y** (or vector **y**), that contain the response of an interest properties (acidity or vitamin C, for example) obtained by a reference method (titration, for example). The **X** and **Y** matrices are decomposed simultaneity in scores and loadings. The PCs, which are orthogonal in PCA, in the PLSR suffers modifications to choose the maximum covariance between **X** and **Y** then the PCs receives the terminology of Latent Variables (VLs) in PLSR.[15]

On PLS-DA method, the **Y** matrix contain information about sample class and, due this, they are a supervised pattern recognition method. The **Y** values are 'zero' or 'one' and these codes indicate if the sample is from one class or another class. For example, consider four classes and, a sample in the second class, the **y** value for this sample is **y** = {0 1 0 0}.[14] The predicted results from PLS-DA must be 'zero' or 'one', however, in practice these values are close of these. It is calculated a threshold value between the predicted values, and values above this threshold value indicates that the sample belongs to the modeled class. On the other hand, predicted values below this threshold limit indicate that the sample does not belong to the modeled class. For threshold estimation the distribution of the prediction values obtained from a

PLS-DA model in the calibration samples are needed to find a threshold value which will best split those classes with the least probability of false classifications of future predictions. It is assumed that the predicted values for each class are approximately normally distributed and the calculation is performed by Bayesian statistic.[17]

The optimum PLS-DA model dimension can be determined by the minimum root mean square error of cross-validation (RMSECV) for the calibration samples, obtained by the leave-one-out or contiguous block procedure. PLS-DA method has been employed, for example, in food analysis,[2,3,13] for authentication of wood samples,[17] tree species,[18] pharmaceutical analysis,[19,20] authentication of geographical origin,[21] and forensic analysis.[22,23]

3.2. Why PLS-DA method was chosen for this proposal?

Regarding the chemometric methodologies that were proposed in previous research about authentication of transgenic and non-transgenic soybean oils,[2,3] if a linear method promote a suitable model, then the use of a non-linear method like SVM-DA is not justified. About PCA, it is an unsupervised pattern recognition method. Unsupervised pattern recognition suggests that the method should be employed only in exploratory analysis and not to make predictions. The KNN, SIMCA, and PLS-DA are linear methods for supervised pattern recognition.

KNN is a method based on distance calculation. Then, in the samples prediction this method will always classify the samples in a modeled class. This means that if a sample does not belong to any of classes that have been modeled, i.e., if the sample is an outlier, by calculating the shortest distance, it is classified as belonging to one of the modeled classes. Therefore, KNN method does not identify outliers. SIMCA is a method based on the PCA where a PCA model is

built for each present class in the system. The advantage of SIMCA over KNN method is that SIMCA can identified outliers.[24]

SIMCA presents considerable success as classification tool when the variation between groups is larger than the variation within the groups.[25] However, when the variability within the group is greater than the variability among groups, the SIMCA method cannot distinguish between the groups and, in such cases, PLS-DA has been an alternative. Considering the stated in this work PLS-DA method was chosen because: 1) can be able to identify outliers; 2) we had no knowledge if the variability within the group would be greater than the variability among groups; 3) the data fit on a linear model.

4. Results and discussion

Figure 1 shows the UV spectra after baseline correction and smoothing by using savgol algorithm[26] (first order polynomial applied on each five spectra point). It is possible to note a slight difference in the spectra in the region around 310 nm. However, due to the lack of selectivity in UV spectroscopy is difficult to draw conclusions only by regarding the spectra and a statistical methodology, as PLS-DA, can contribute to the reliability of the results.

The calibration and validation data sets were composed of 75 (50 transgenic and 25 non-transgenic) and 30 (15 transgenic and 15 non-transgenic) samples, respectively, selected by the Kennard–Stone algorithm.[27] In this algorithm, the first sample selected is that with the largest distance from the center of the data. The next sample again has the largest distance from the last point, and so on, until the number of samples for the calibration set is complete.

The optimum PLS-DA model dimension was determined by the minimum root mean square error of cross-validation (RMSECV) for the calibration samples, obtained by the

contiguous block procedure with eight samples. This procedure result the choice of eight latent variables for mean-centered model development.

The next step was outlier identification. Outliers can be defined as observations showing some type of departure from the bulk of the data. They may occur for many different reasons, for example, laboratory error, objects from another population or instrument error.[28] In this work, the outlier identification was performed by leverage and Q Residuals analysis on the calibration and validation samples. Leverage represents how much one sample is distant from the center of the data and, Q Residuals represent the unmodeled residuals in spectra. According to the Figure 2, three samples from transgenic calibration set present a high leverage (on the top). However, these samples present a low Q Residuals. It is possible to observe also that one sample from transgenic calibration set with a high Q Residual (in the right side). Nonetheless, this sample present a low leverage. Samples can be considered certainly outliers when it have both high leverage and high Q Residuals and then, the calibration and validation data sets have no outliers since no sample presents high leverage value and Q Residuals, simultaneously.

Figure 3 presents the distribution of the estimated class values, for both calibration and validation data sets, for transgenic and non-transgenic soybean oils, of the authentication model. For both types of samples, a clear separation between the estimated class values for the transgenic and non-transgenic can be observed. It is also important to note that the agreement between the RMSEC and RMSEP, 0.223 and 0.278, respectively confirms the absence of overfitting.

Sensitivity and specificity were determined from data of Figure 3. Sensitivity is the model ability to classify the validation samples belonging to a particular class. If the model classify all samples in a given class correctly, then the sensitivity to this class is equal to 1. For

the model, the sensitivities were 0.875 and 1.000 for transgenic and non-transgenic classes, respectively. Based on these results, the model was able to classify the non-transgenic soybean oil samples.

The specificity is related to the incorrect prediction validation samples of other classes in a particular class. Thus, if the model does not present error in predicting a sample, this model presents specificity equal to 1. The transgenic class present specificity equal to 1, this result means that any non-transgenic sample was classified in the transgenic class. The non-transgenic class present specificity equal 0.875 due the prediction of two samples of transgenic class in the non-transgenic class. A similar result was achieved by Luna et al.[3] using PLS-DA and SVM-DA models to discriminant transgenic and non-transgenic soybean oil by NIR spectroscopy.

The scores plot for PLS-DA model is presented in Figure 4. A separation between transgenic and non-transgenic samples can be observed, indicating that the non-transgenic samples were discriminated by the positive part of LV3, while the transgenic samples were discriminated by the negative part of LV3.

Loadings plot in Figure 5 shows that the region between 300-340 nm contributes to the differentiation between classes. By analyzing this figure, it is possible note that the peak in 300-310nm contributed to the transgenic samples classification because they have negative loadings for LV3. For non-transgenic samples classification the peak around 330nm is the most important because they have positive loadings for LV3.

By comparing the spectra of transgenic and non-transgenic soybean sample in the Figure 1B with the loadings plot it is possible assign that differentiation between sample classes to a bathochromic shift, probably due to the differences in the chromophore group of genotypic structure[7] present in the transgenic and non-transgenic samples.

Conclusions

UV spectroscopy associated with PLS-DA chemometric method showed to be a powerful tool to authenticate soybean oil samples as transgenic or non-transgenic. Furthermore, it enables a fast and nondestructive analysis of soybean oil without any sample preparation. Even though the UV spectroscopy is not a selective technique when coupled with the supervised chemometric method PLS-DA, the technique can promote the authentication of transgenic and non-transgenic soybean oils. The ability of UV spectroscopy for soybean oil authentication can be assigned to the bathochromic shift, probably due to the differences in the chromophore group of genotypic structure present in the transgenic and non-transgenic samples.

Acknowledgements

Fundação Araucária (Infrastructure Program for Young Researchers – First Projects Program, 14/2011 – process 223/2013).

References

1 M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, R. Nescatelli and F. Marini, *Chemometrics in Food Chemistry,* ed. F. Marini, Elsevier, Amsterdam, 2013, chapter 5, 171-233.

2 A.S. Luna, A.P. Silva, J.S.A. Pinho, J. Ferré and R. Boqué, *Food Research International*, 2015, **67**, 206-211.

3 A.S. Luna, A.P. Silva, J.S.A. Pinho, J. Ferré and R. Boqué, *Spectrochimica Acta Part A*, 2013, **100**, 115-119.

4 P. Bogani, M. Minunni, M.M. Spiriti, M. Zavaglia, S. Tombelli, M. Buiatti and M. Mascini, *Food Chemistry*, 2009, **113**, 658-664.

5 J. Costa, I. Mafra, J.S. Amaral and M.B.P.P. Oliveira, *Food Research International*, 2010, **43**, 301-306.

6 G.B. Alcantara, A. Barison, M.S. Santos, L.P.S. Santos, J.F.F. Toledo and A.G. Ferreira, *Orbital – The Electronic Journal of Chemistry,* 2010, **2**, 41-52.

7 A. Alishahi, H. Farahmand, N. Prieto and D. Cozzolino, *Spectrochimica Acta Part A,* 2010, **75**, 1-7.

8 D.A. Santos, P.H. Março and P. Valderrama, *Brazilian Journal of Analytical Chemistry*, 2013, **12**, 495-498.

9 M.J. Martelo-Vidal and M. Vázquez, *Food Chemistry*, 2014, **158**, 28-34.

10 R.P. Gonçalves, P.H. Março and P. Valderrama, *Food Chemistry,* 2014, **163**, 83-86.

11 C.S. Morais, L.M. Leme, P. Valderrama and P.H. Março, *Brazilian Journal of Food Research*, 2015, **6**, 70-79.

12 D.A. Santos, K.P. Lima, P.H. Março and P. Valderrama, *Brazilian Journal of Food Research*, 2015, **6**, 1-8.

13 L. Valderrama, R.P. Gonçalves, P.H. Março and P. Valderrama, *Brazilian Journal of Food Research*, 2014, **5**, 32-40.

14 M. Barker and W. Rayens, *Journal of Chemometrics*, 2003, **17**, 166-173.

15 P. Geladi and B.R. Kowalski, *Analytica Chimica Acta,* 1986, **185**, 1-17.

16 S. Wold, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37-52.

17 T.C.M. Pastore, J.W.B. Braga, V.T.R. Coradin, W.L.E. Magalhães, E.Y.A. Okino, J.A.A. Camargo, G.I.B. Muniz, O.A. Bressan and F. Davrieux, *Holzforschung*, 2011, **65**, 73-80.

18 K.Y. Peerbhay, O. Mutanga and R. Ismail, *Journal of Photogrammetry and Remote Sensing*, 2013, **79**, 19-28.

19 R.S. Fernandes, F.S.L. Costa, P. Valderrama, P.H. Março and K.M.G. Lima, *Journal of Pharmaceutical and Biomedical Analysis*, 2012, **66**, 85-90.

20 D.C. Maltas, K. Kwok, P. Wang, L.S. Taylor and D. Bem-Amotz, *Journal of Pharmaceutical and Biomedical Analysis*, 2013, **80**, 63-68.

21 M. Bassbasi, M. De Luca, G. Ioele, A. Oussama and G. Ragno, *Journal of Food Composition and Analysis*, 2014, **33**, 210-215.

22 V.A.G. Silva, M. Talhavini, I.C.F. Peixoto, J.J. Zacca, A.P. Maldaner and J.W.B. Braga, *Microchemical Journal*, 2014, **116**, 235-243.

23 M.R. Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi and R.J. Poppi, *Microchemical Journal*, 2013, **109**, 170-177.

24 R.E. Bruns and J.F.G. Faigle, *Química Nova*, 1985, **8**, 84-99.

25 D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a textbook*, Elsevier, Amsterdam, 1988, chapter 23, 385-413.

26 A. Savitsky and M.J.E. Golay, *Analytical Chemistry*, 1964, **36**, 1627-1639.

27 R.W. Kennard and L.A. Stone, *Technometrics*, 1969, **11**, 137-148.

28 H. Martens and T. Naes, *Multivariate calibration*, Wiley, New York, 1996, chapter 5, 267-296.

Figure captions

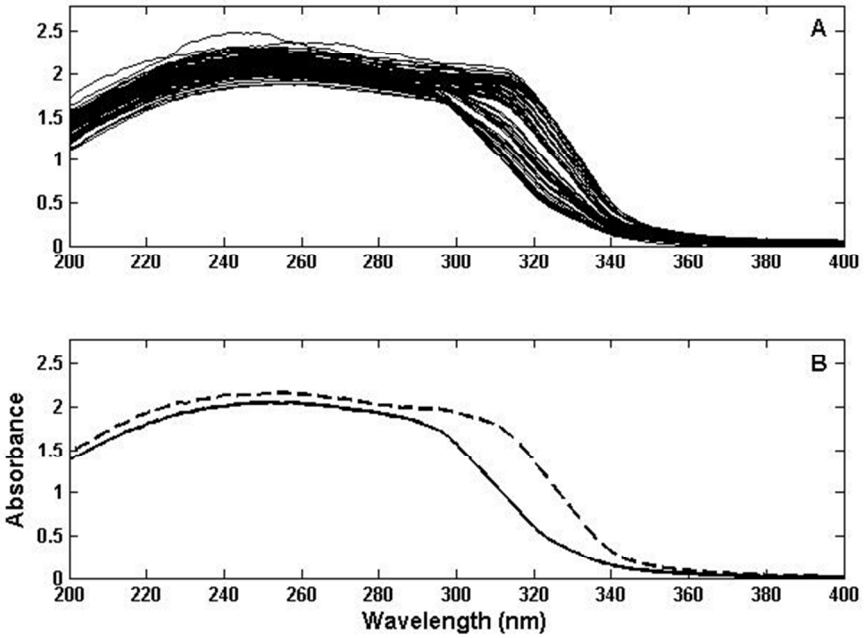
Figure 1. UV spectra of soybean samples (A) and UV spectra of soybean oil non-transgenic (—) and transgenic (-----).

Figure 2. Q Residuals against Leverage for PLS-DA model. (●) transgenic calibration samples. (**o**) transgenic validation samples. (■) non-transgenic calibration samples. (□) non-transgenic validation samples.

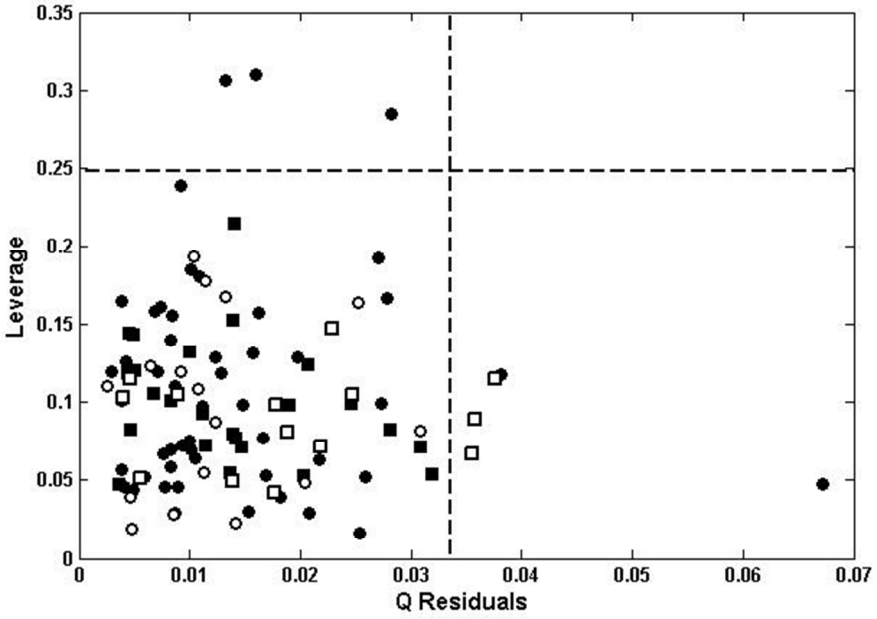
Figure 3. Estimated class values for calibration and validation sets for discrimination between transgenic (A) and non-transgenic (B) soybean oils. (●) transgenic calibration samples. (**o**) transgenic validation samples. (■) non-transgenic calibration samples. (□) non-transgenic validation samples.

Figure 4. Scores plot of PLS-DA model. (●) transgenic calibration samples. (**o**) transgenic validation samples. (■) non-transgenic calibration samples. (□) non-transgenic validation samples.

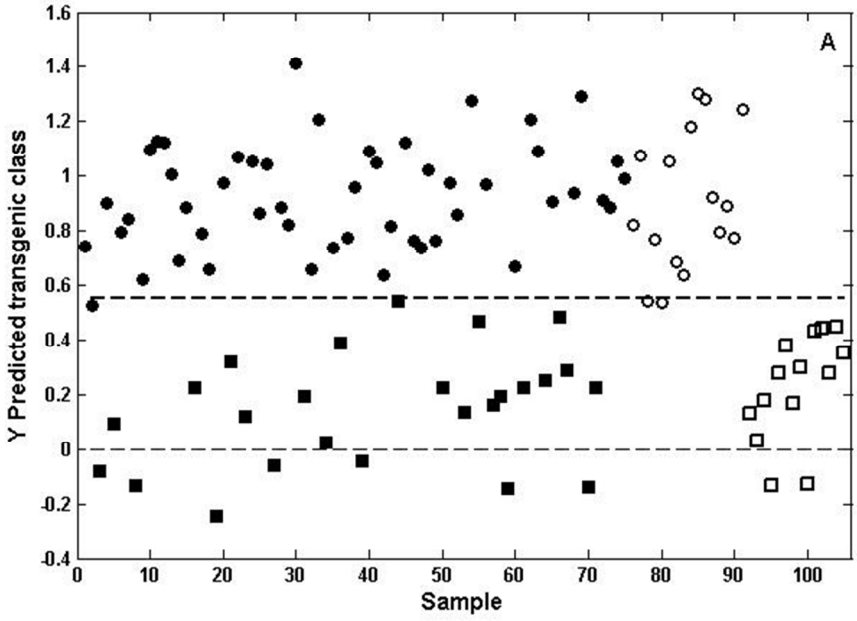
Figure 5. Loadings plot of third latent variable for PLS-DA model.



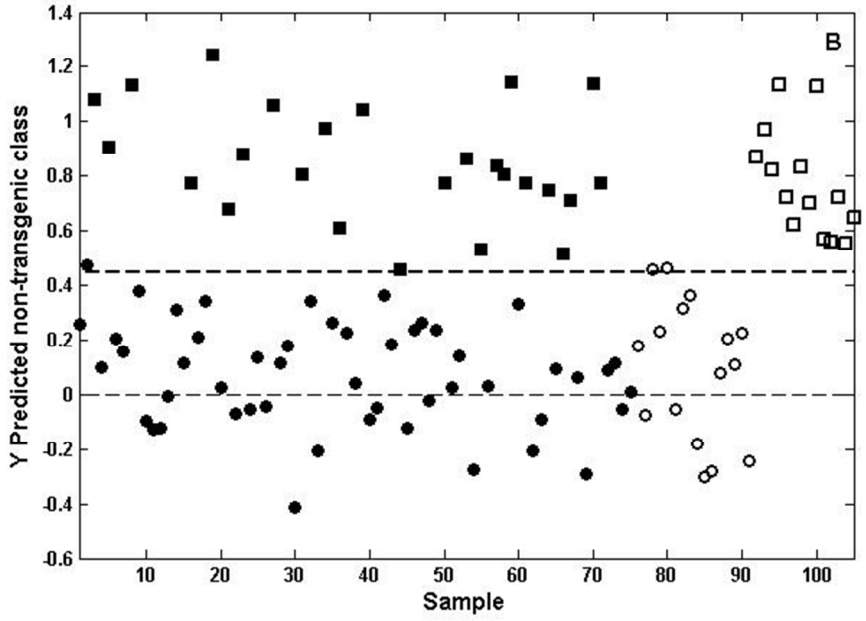
244x172mm (72 x 72 DPI)



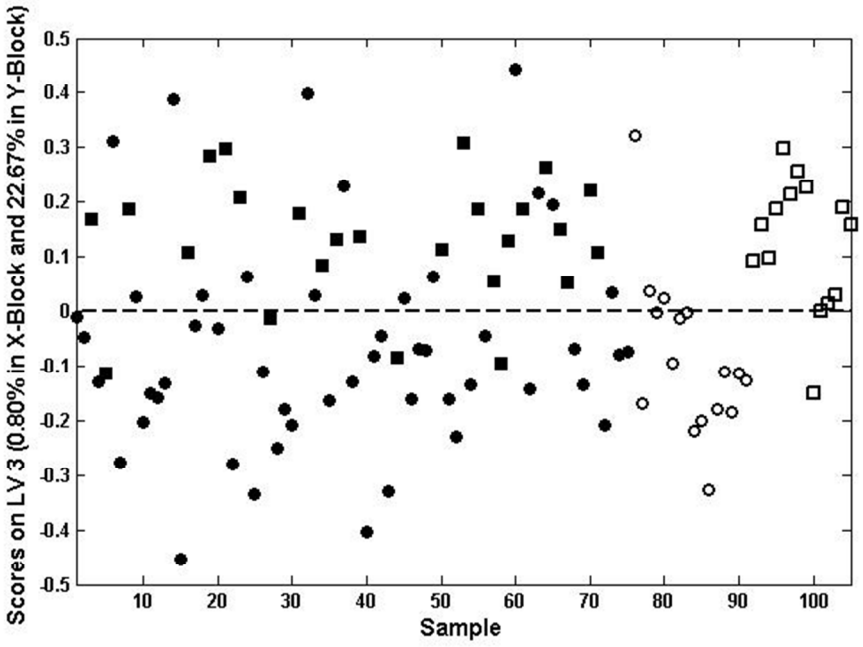
244x164mm (72 x 72 DPI)



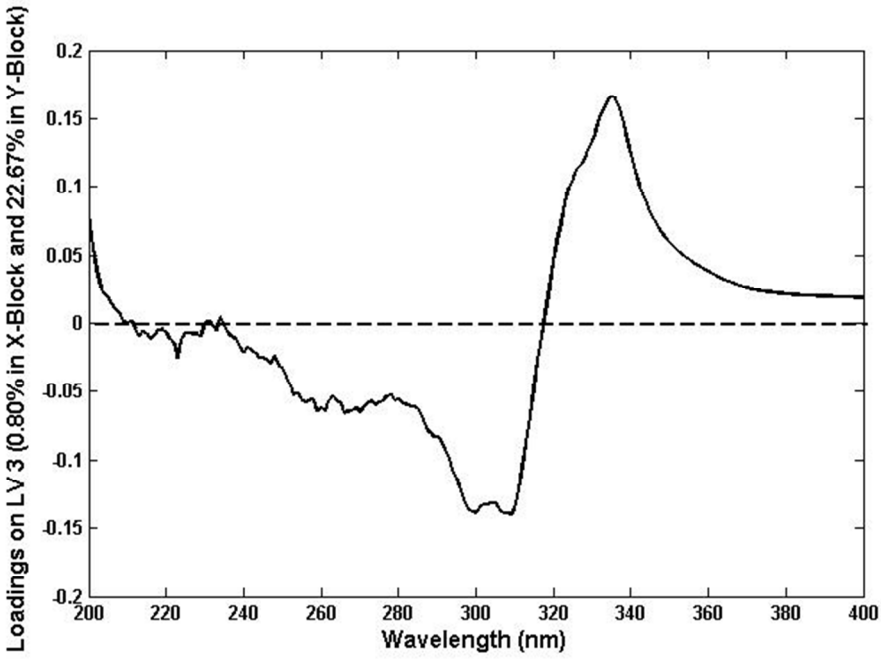
244x164mm (72 x 72 DPI)



244x164mm (72 x 72 DPI)



244x164mm (72 x 72 DPI)



244x164mm (72 x 72 DPI)