Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2016, **8**, 8498

Classification of individual cotton seeds with respect to variety using near-infrared hyperspectral imaging

Sófacles Figueredo Carreiro Soares,[ab] Everaldo Paulo Medeiros,[c] Celio Pasquini,[d] Camilo de Lelis Morello,[c] Roberto Kawakami Harrop Galvão[e] and Mário César Ugulino Araújo*[a]

Received 21st October 2016
Accepted 7th November 2016

DOI: 10.1039/c6ay02896a

www.rsc.org/methods

This paper proposes the use of Near Infrared Hyperspectral Imaging (NIR-HSI) as a new strategy for fast and non-destructive classification of cotton seeds with respect to variety. A total of 807 seeds of four different cotton varieties are employed in this study. For classification purposes, each seed is represented by an average spectrum obtained by coaveraging the pixel spectra of the NIR-HSI image. Conventional NIR and VIS-NIR spectra are also employed for comparison. By using Partial-Least-Squares Discriminant Analysis (PLS-DA), correct classification rates of 98.0%, 89.7% and 91.7% were achieved in the NIR-HSI, conventional NIR and conventional VIS-NIR datasets. The superiority of the NIR-HSI system can be ascribed to a more comprehensive scan of the seed area, as compared to the conventional VIS-NIR spectrometer.

1. Introduction

Cotton is an agricultural commodity of worldwide importance, with more than 60 producer countries and annual revenues of approximately USD 12 billion.[1] Therefore, technological innovations that lead to productivity gains may have a significant economic and social impact. In particular, the development of genetically superior cultivars is a key factor for increasing the plantation yield. Important aspects include increased tolerance to herbicides[2] and pest attacks,[3,4] as well as better germination, vigour and cotton fibre characteristics.[5]

New cultivars can be registered in order to obtain legal protection of intellectual property rights.[6] Plant characteristics such as germination, vigour, productivity, leaf shape, flower color and plant size[7,8] can be used to identify seeds of registered cultivars in cases of dispute or possible contamination of batches (such as the mixture of conventional and genetically modified seeds). However, these features can only be determined by planting the seed and waiting for the germination and development of the plant, which is a destructive and time-consuming method.[9] An alternative consists of using molecular biology tools to detect specific markers such as proteins, enzymes, specific amino acid sequences and genomic DNA.[10–14] Again, these techniques do not preserve the sample and may only be available in a few laboratories far from production centres.

In view of these drawbacks, there is a growing interest in the use of near-infrared (NIR) and VIS-NIR diffuse reflectance spectroscopy with appropriate chemometric modelling. Examples reported in the literature include the classification of *Glycyrrhiza uralensis Fisch*,[15] soybeans,[16] pistachios,[17] and castor seeds.[9] However, in conventional NIR or VIS-NIR spectrometers the analysis is restricted to a relatively small area of the seed, which may lead to sampling problems in the monitoring of chemical compounds and diseases that occur in specific areas of the seed.[18] Larger areas can be scanned by adjusting the presentation of the sample, but this procedure may be too cumbersome and time consuming for routine use. In view of this inconvenience, a more practical approach may consist of the use of imaging techniques.

NIR hyperspectral imaging (NIR-HSI) combines the advantages of digital imaging and spectroscopy to acquire spatial and spectral information simultaneously.[19,20] The hyperspectral image is formed by a spectrum of each pixel and an image of each spectral bin (wavelength). NIR-HSI has been successfully applied in several qualitative studies involving seeds, such as classification of seed types,[21–23] identification of geographical origins,[19] characterization of individual seeds,[24] detection of impurities,[25] insect damage[26–28] and colour defects.[29]

In this context, the present work proposes the use of NIR-HSI coupled with Partial-Least-Squares Discriminant Analysis

[a]*Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, CEP 58051-970 – João Pessoa, PB, Brazil. Fax: +55 83 3216-7437; Tel: +55 83 3216-7438. E-mail: laqa@quimica.ufpb.br; mariougulino@gmail.com*

[b]*Universidade Federal da Paraíba, CT, Departamento de Engenharia Química, CEP 58051-900 – João Pessoa, PB, Brazil*

[c]*Embrapa, Centro Nacional de Pesquisa de Algodão, 58428-095, Campina Grande-PB, Brazil*

[d]*Universidade Estadual de Campinas, Instituto de Química, CEP 13083-970, Campinas-SP, Brazil*

[e]*Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900 São José dos Campos, SP, Brazil*

(PLS-DA) for the classification of cotton seeds with respect to variety. The results are compared with those obtained with a conventional VIS-NIR spectrometer, in order to investigate the possible benefits of using an imaging system.

2. Experimental

2.1. Cotton seed samples

Cotton seeds of four different varieties were provided by Embrapa Algodão (Campina Grande, Paraíba, Brazil). The seeds were conditioned at room temperature (20 °C) and relative humidity (65%) for at least one hour before spectral recording. No chemical treatment was employed. Examples of the four seed varieties are presented in Fig. 1. Despite the visual similarity of the seeds, these varieties differ in terms of average fiber yield, cycle duration (short: 120 to 140 days; long: 150 to 180 days), plant size (short or tall), and resistance to disease, among other morphological characteristics.

In this study, 409 and 398 seeds were randomly selected for the acquisition of NIR-HSI images and conventional VIS-NIR spectra, respectively. In what follows, the terms "Class" and "Sample" will be employed to indicate each variety and each individual seed, respectively.

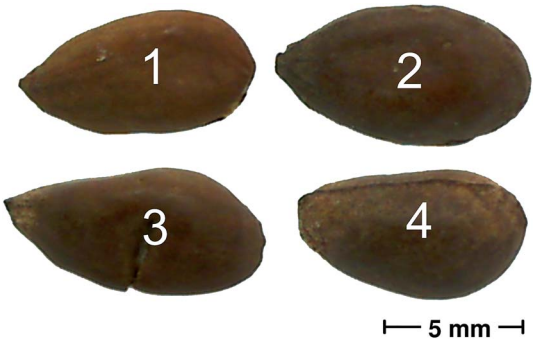

Fig. 1 Examples of the four types of seeds under study.

2.2. Hyperspectral image acquisition and pre-processing

A Teflon support of 10 × 20 cm was used to accommodate the samples. The NIR-HSI images were acquired using a Sisu-CHEMA SWIR chemical imaging workstation (Specim, Oulu, Finland) with 30 mm fore lens (50–100 mm field of view) and a line-scan system, operating at a wavelength range of 928–2524 nm with 6 nm intervals. This wavelength range was segmented into 256 slices, resulting in an NIR-HSI image cube with 256 images (one image per wavelength slice). The images outside the range of 1100–2500 nm were discarded due to the low intensity of the detector signal.

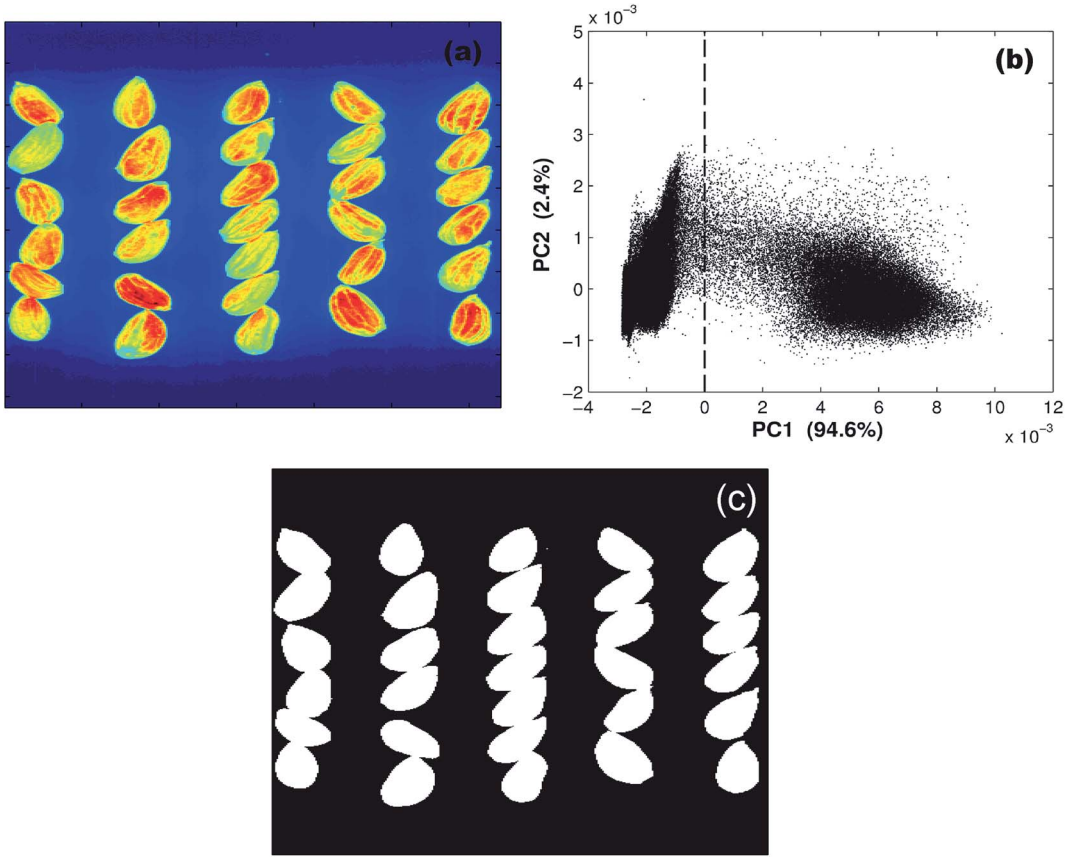

Fig. 2 (a) Example of the image obtained after pre-processing. The colour scale indicates the average reflectance value at each pixel. (b) PC1 × PC2 scores plot of the image pixels, with a vertical line indicating the threshold for separation of the seed images from the background. (c) Binary image with black and white regions corresponding to the pixels with negative and positive PC1 scores, respectively.

The NIR-HSI data were transformed into pseudo-absorbance values using a white reference standard and internal dark reference (obtained with the light source off and the camera lens completely covered with its opaque cap).[19] For this purpose, a logarithmic function was applied to the ratio between the total reflectance spectrum and the sample spectrum, both corrected for the dark reference. To remove baseline features,[30] a second-derivative spectrum of each image pixel was calculated by using a Savitzky–Golay filter with a second order polynomial and 21-point window.

Fig. 2a presents an image obtained after the pre-processing procedures. In order to separate the individual seed images from the Teflon background, Principal Component Analysis (PCA)[31] was carried out. For this purpose, the NIR spectrum of each pixel was regarded as an object. The resulting PC1 × PC2 scores plot shown in Fig. 2b reveals the presence of two clusters, corresponding to negative and positive values along the PC1 axis. By using a threshold of zero for the PC1 score, the binary image in Fig. 2c was formed. As can be seen, the seeds were clearly separated from the background. A manual selection was carried out to outline the region of interest corresponding to each seed.[32]

The average spectrum of the NIR-HSI pixels in each region of interest was employed to represent the corresponding seed. Since this coaveraging process tends to reduce the noise level in the spectrum, the second-derivative Savitzky–Golay calculations were repeated with a smaller window of 17 points. The resulting spectra are presented in Fig. 3.

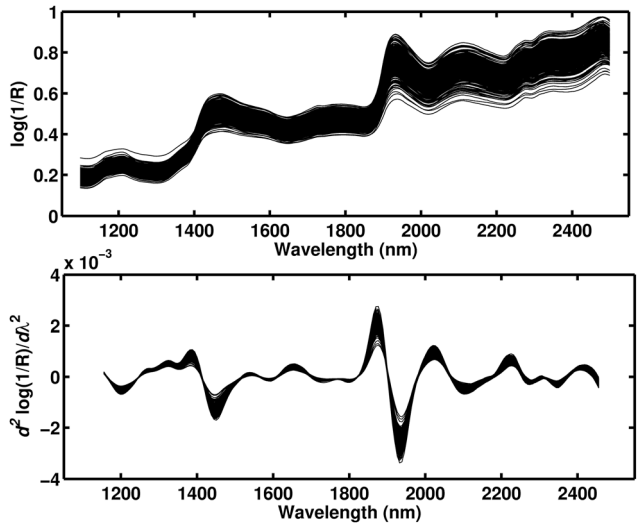

Fig. 3 NIR-HSI spectra before (top) and after (bottom) the second derivative calculation.

2.3. Acquisition and pre-processing of conventional VIS-NIR spectra

Diffuse reflectance spectra were obtained by using a XDS Rapid ContentTM Analyzer VIS-NIR spectrophotometer (Foss Analytical, Hogans, Sweden), fitted with a circular quartz cell of 3 cm diameter. Each spectrum was acquired as the average of 32 scans in the range 400–2500 nm, with 0.5 nm intervals. A boxcar smoothing window of 12 points was employed in order to obtain a spectral resolution similar to the NIR-HSI spectra. Finally, second-derivative spectra were calculated by using a Savitzky–Golay filter with a second-order polynomial and 19-point window. The results are shown in Fig. 4. It is worth noting that the VIS-NIR spectra display a discontinuity at 1100 nm, due to a change of detector in the spectrometer. For this reason, the boxcar and second-derivative transformations were applied separately to the spectral regions below and above 1100 nm.

In order to assess the full potential of the conventional VIS-NIR system for the analytical problem under consideration, the chemometric calculations were carried out by using the NIR range 1100–2500 nm (as in the NIR-HSI system) and also the full VIS-NIR range 400–2500 nm.

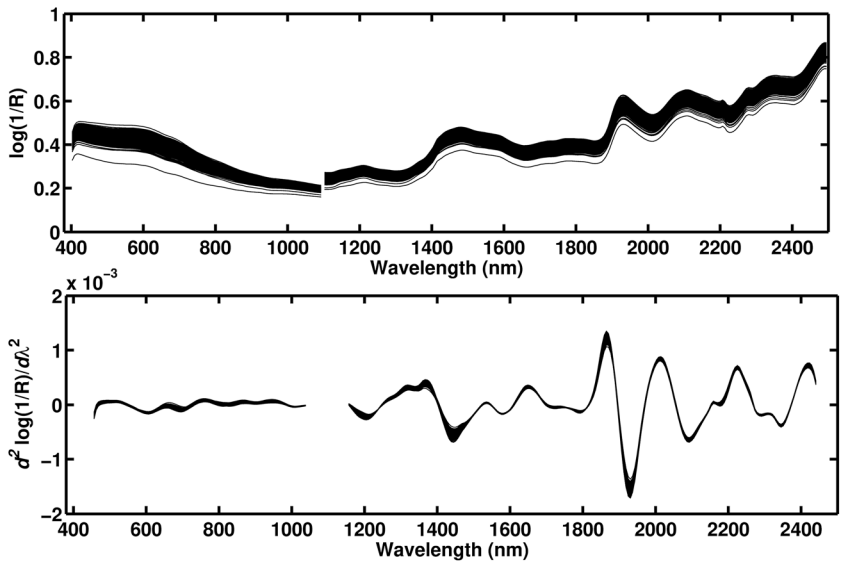

Fig. 4 VIS-NIR spectra before (top) and after (bottom) the boxcar and second-derivative calculations.

Table 1 Number of training, validation and test samples in each class

Sets	Class	NIR-HSI	Conventional NIR	Conventional VIS-NIR
Training	1	50	50	51
	2	54	51	51
	3	51	50	51
	4	49	51	49
	Total	**204**	**202**	**202**
Validation	1	24	25	24
	2	26	24	24
	3	25	24	24
	4	24	24	24
	Total	**99**	**97**	**96**
Test	1	24	25	24
	2	26	24	24
	3	25	24	24
	4	24	24	24
	Total	**99**	**97**	**96**
Total number of samples		**402**	**396**	**394**

2.4. Removal of outliers

The values of *leverage* and residual sum of squares (Q) obtained by Principal Component Analysis of the spectra were employed to detect outliers.[33,34] As a result, seven samples in the NIR-HSI dataset, four samples in the conventional VIS-NIR dataset and two samples in the conventional NIR dataset were removed.

Table 2 Confusion matrix for the classification of the test samples in the NIR-HSI, conventional NIR and conventional VIS-NIR datasets. The number of samples in each class is denoted by N

	NIR-HSI					Conventional NIR					Conventional VIS-NIR				
		Assigned class					Assigned class					Assigned class			
True class	N	1	2	3	4	N	1	2	3	4	N	1	2	3	4
1	24	23	1	—	—	25	24	1	—	—	24	23	1	—	—
2	26	—	26	—	—	24	2	18	—	4	24	1	21	2	—
3	25	—	1	24	—	24	—	1	23	—	24	—	1	22	1
4	24	—	—	—	24	24	—	2	1	21	24	—	2	—	22

2.5. Training, validation and test sets

After the removal of the outliers, the samples were divided into training, validation, and test sets by applying the classic Kennard–Stone (KS)[35] uniform sampling algorithm to each class separately. The number of samples in each set is presented in Table 1.

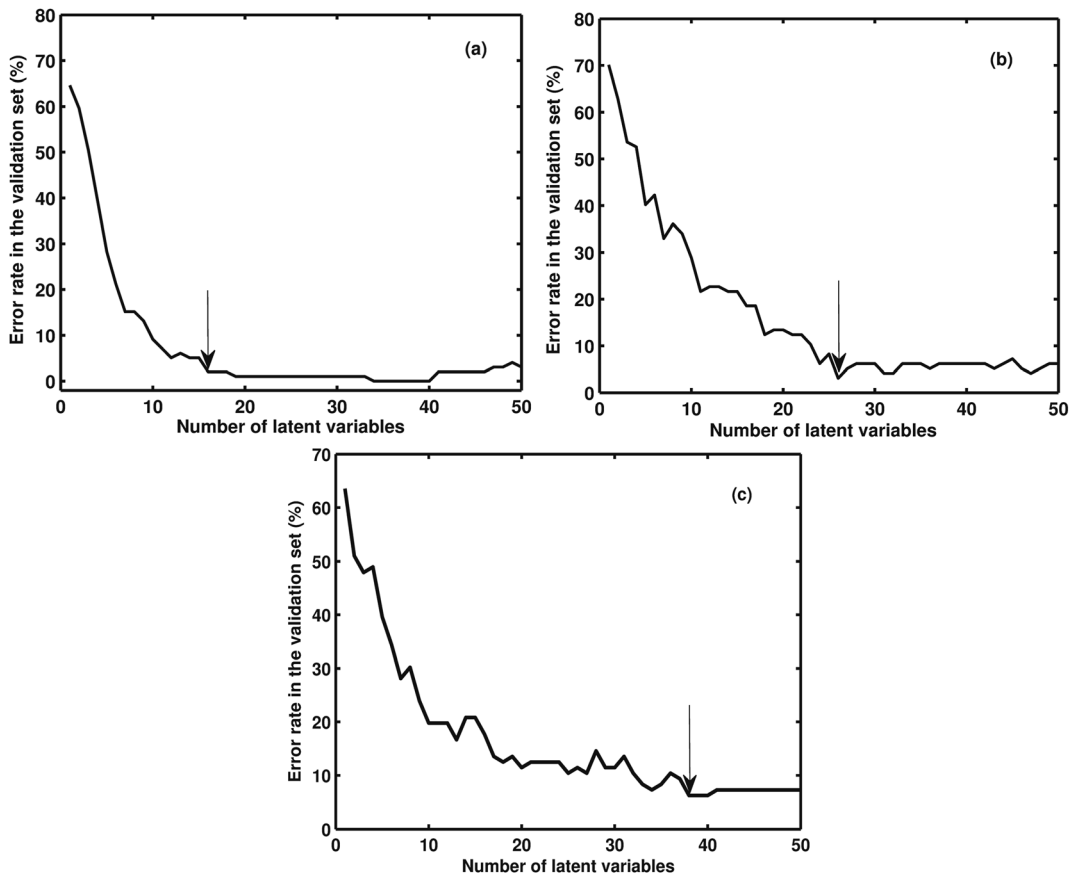

Fig. 5 Classification error rate in the validation set as a function of the number of latent variables in the PLS-DA model: (a) NIR-HSI, (b) conventional NIR, and (c) conventional VIS-NIR datasets. The arrows indicate the selected point in each plot.

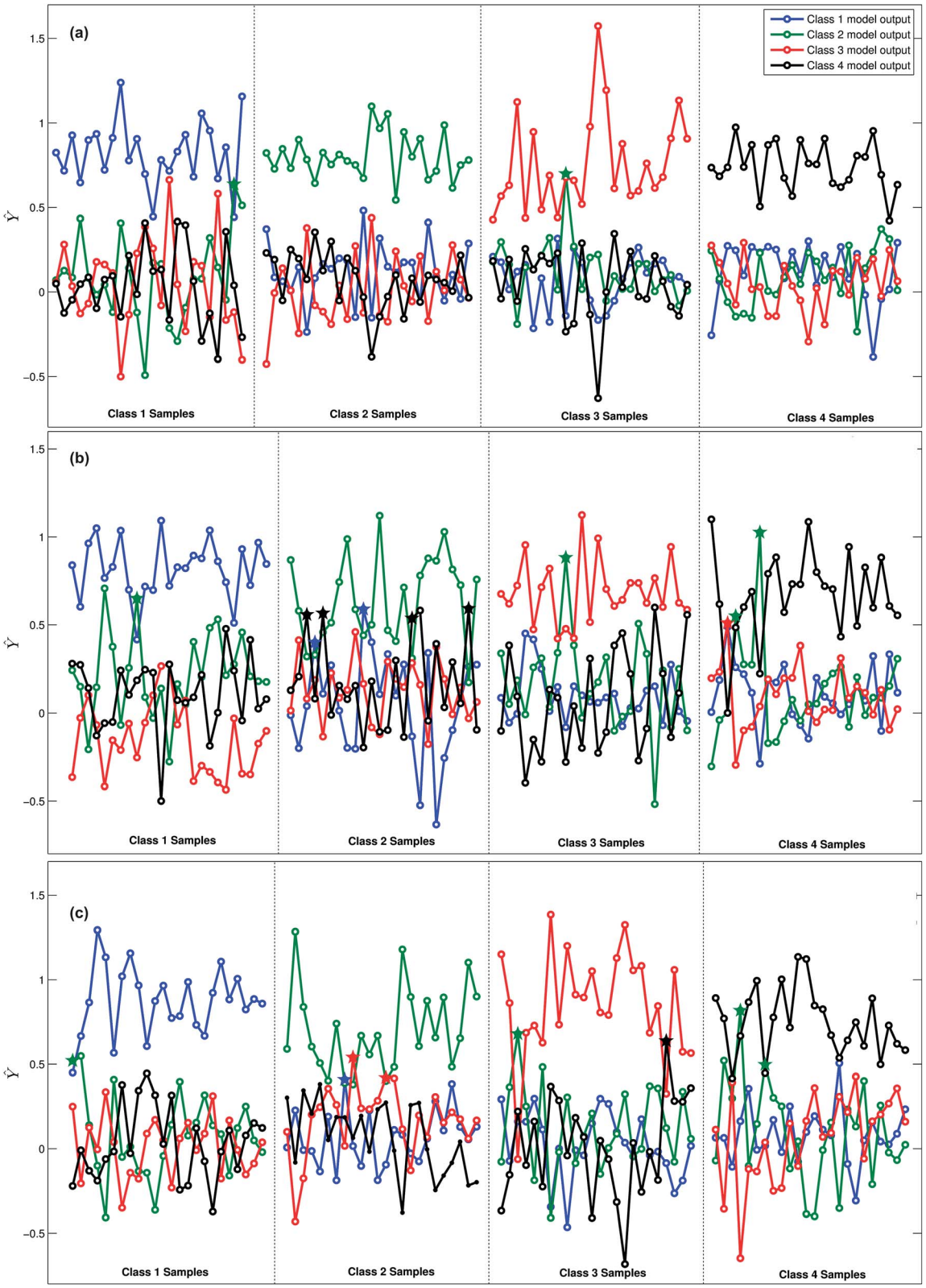

Fig. 6 Output of the PLS-DA models: (a) NIR-HSI, (b) conventional NIR and (c) conventional VIS-NIR datasets. The output values that resulted in classification errors are indicated by star markers.

2.6. Modelling procedures and software

In order to build a PLS-DA model, the class index of each sample in the training set is encoded as a vector $\mathbf{y} = [y_1 y_2 y_3 y_4]$ of four binary variables. If the sample belongs to class i, the value of y_i is set to one and the remaining y_j variables ($j \neq i$) are set to zero. As a result, the class indexes of the N_{train} training samples are

encoded in a matrix $\mathbf{Y}_{train}$ of dimensions ($N_{train} \times 4$). The PLS2 algorithm[36,37] is then applied to this $\mathbf{Y}_{train}$ matrix and the corresponding $\mathbf{X}_{train}$ matrix of spectra. The resulting PLS-DA model can be used to classify new samples on the basis of the predicted y-values.[38] In the present work, the sample is assigned to the class corresponding to the largest predicted y-value.

The number of latent variables was chosen by evaluating the classification performance of the PLS-DA model in the validation set. The test set was only employed in the final comparison of the NIR-HSI, conventional NIR and conventional VIS-NIR results.

The PLS-DA calculations were carried out using the MATLAB® 2010b software by using functions from the classification toolbox package[38] available at http://michem.disat.unimib.it/chm/download/classificationinfo.htm.

3. Results and discussion

3.1. PLS-DA results

Fig. 5 presents the scree plots of the classification error rate that were employed in the choice of an appropriate number of latent variables in the PLS-DA models of the (a) NIR-HSI, (b) conventional NIR and (c) conventional VIS-NIR datasets. As shown in Fig. 5a, 16 latent variables were selected (error rate of 2.0%) because the decrease in the error rate is very small beyond that point. In Fig. 5b, the minimum point of the scree plot was selected (26 latent variables, corresponding to an error rate of 3.1%). In Fig. 5c, 38 latent variables were selected (error rate of 6.2%) because the error rate does not decrease beyond that point.

The **X** and **Y** variances explained by the latent variables were 99.84% and 72.68% in the NIR-HSI dataset, 99.97% and 70.95% in the conventional NIR dataset, and 99.96% and 78.23% in the conventional VIS-NIR dataset.

The resulting PLS-DA models were employed to classify the samples in the test set. Table 2 presents the classification outcome in the form of a confusion matrix. As can be seen, only 2 samples were incorrectly classified in the NIR-HSI dataset, whereas 11 classification errors were obtained in the conventional NIR dataset and 8 classification errors were obtained in the conventional VIS-NIR dataset. Such results are better visualized in Fig. 6, which presents the PLS-DA model outputs of the

Table 3 Sensitivity and specificity values obtained in the classification of the test samples using the NIR-HSI, conventional NIR and conventional VIS-NIR datasets

	Sensitivity				Specificity			
Dataset	Class 1	Class 2	Class 3	Class 4	Class 1	Class 2	Class 3	Class 4
HSI-NIR	0.958	1	0.960	1	1	0.973	1	1
Conventional NIR	0.960	0.750	0.958	0.875	0.972	0.945	0.986	0.945
Conventional VIS-NIR	0.958	0.875	0.917	0.917	0.986	0.944	0.972	0.986

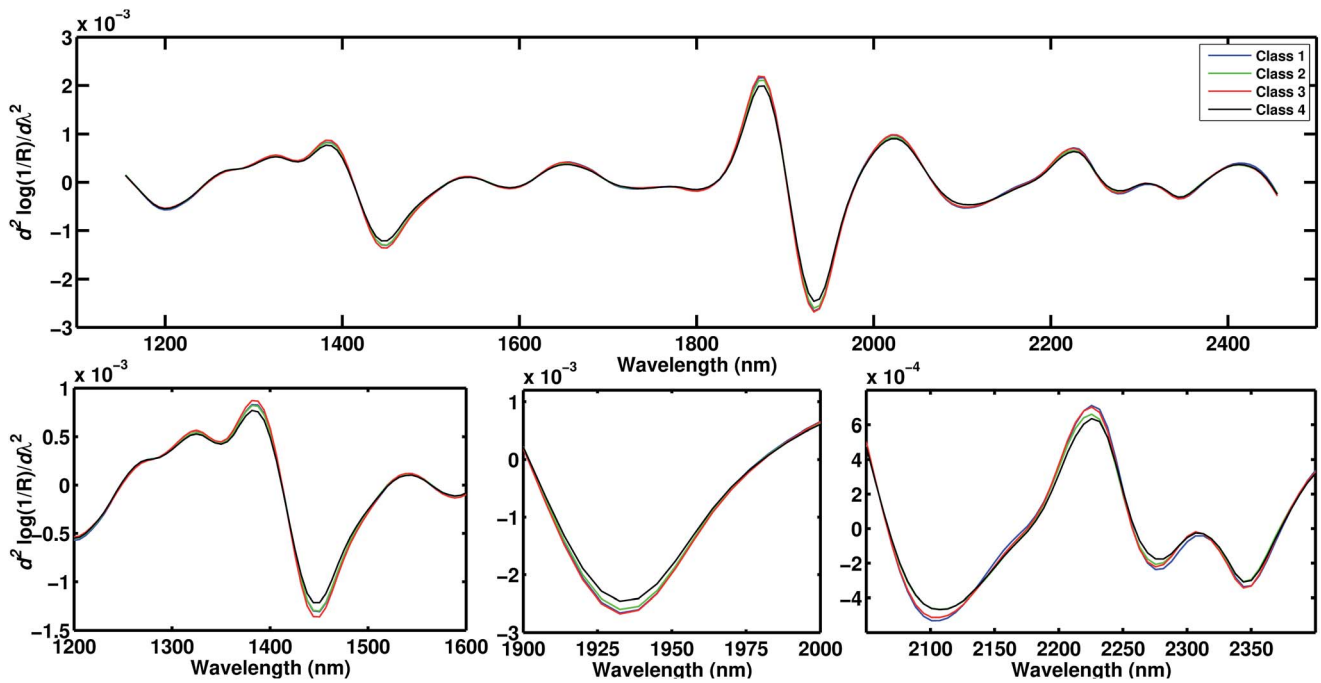

Fig. 7 Average NIR-HSI spectra obtained for each of the four classes, with enlarged views of regions in which the class separation is more apparent.

Table 4 Interpretation of spectral bands

Wavelength (nm)	Vibration bond	Structure
1440	C–H combination	$^{\cdot}CH_2$
1450	O–H first overtone	Starch
	C═O stretch third overtone	C═O
1460	Sym N–H stretch first overtone	Urea
1930	O–H stretch/HOH deformation combination	Starch Cellulose H_2O
1940	O–H bend second overtone	H_2O
1950	C═O stretch second overtone	$-CO_2R$
2090	O–H combination	Oil $^{\cdot}OH$
2100	O–H bend/C–O stretch combination	Starch
	Asym C–O–O stretch third overtone	Starch or cellulose
2180	N–H bend second overtone	Protein
	C–H stretch C═O stretch combination	Oil
	C═O stretch/amide combination	
2282	C–H stretch/CH_2 deformation	Starch
2344	C–H stretch/CH deformation	Cellulose
	CH_2 bend second overtone	

(a) NIR-HSI, (b) conventional NIR and (c) conventional VIS-NIR datasets. Indeed, the outputs in Fig. 6a provide a more clear indication of the correct class for each sample, as compared to Fig. 6b and c.

The superiority of the NIR-HSI results is also reflected in larger sensitivity and specificity values, as shown in Table 3. Overall, correct classification rates of 98.0%, 89.7% and 91.7% were obtained with the NIR-HSI, conventional NIR and conventional VIS-NIR datasets, respectively.

3.2. Discussion

In order to interpret the spectral information involved in the classification process, the average NIR-HSI spectra of the four classes were compared, as shown in Fig. 7. The plots in the bottom of this figure present enlarged views of spectral regions in which the class separation was more apparent. Table 4 provides an interpretation in terms of the associated NIR bands, as discussed elsewhere.[39] Some of these spectral regions have indeed been correlated with oil and protein content in cotton seeds, as reported elsewhere.[40]

4. Conclusion

This paper presented a new strategy for fast and non-destructive classification of cotton seeds with respect to variety, on the basis of NIR-HSI images or conventional NIR and VIS-NIR spectra. Good accuracy was obtained in the classification of the test samples, with correct classification rates of 98.0%, 89.7% and 91.7% in the NIR-HSI, conventional NIR and conventional VIS-NIR datasets, respectively. The superiority of the NIR-HSI system can be ascribed to the more comprehensive scan of the seed area, as compared to the conventional NIR spectrometer.

In view of the better classification rate obtained by using the full VIS-NIR range compared to the NIR range in the conventional spectrometer, it may be argued that the NIR-HSI results could be further improved by using wavelengths below 1100 nm. In the present work, such a possibility could not be exploited due to the low intensity of the detector signal outside the range of 1100–2500 nm.

In future investigations, the Teflon support for the presentation of the samples could be divided into individual cells in order to better separate the seeds. Such a modification would greatly simplify the image segmentation process, dispensing with the need for manual delineation of the regions of interest.

Acknowledgements

The authors acknowledge the support of CNPq (MSc scholarship, research fellowships and Universal grant 475204/2004-2), INCTAA (CNPq Proc. no. 573894/2008-6 and FAPESP Proc. no. 2008/57808-1) and Embrapa (03.15.00.051.00.00).

References

1 ABRAPA, http://www.abrapa.com.br/en/estatisticas/Paginas/Algodao-no-Mundo.aspx, accessed July 2016.
2 C. Bayley, N. Trolinder, C. Ray, M. Morgan, J. E. Quisenberry and D. W. Ow, *Theor. Appl. Genet.*, 1992, **83**, 645.
3 A. M. Showalter, S. Heuberger, B. E. Tabashnik and Y. Carrière, *J. Insect Sci.*, 2009, **9**, 1.
4 J. B. Torres, J. R. Ruberson and M. Whitehouse, in *Organic Farming, Pest Control and Remediation of Soil Pollutants*, ed. E. Lichtfouse, Sustainable Agriculture Reviews 1, Dordrecht, 2010, ch. 4, pp. 15–53.
5 J. Han, J. Tan, L. Tu and X. Zhang, *Plant Biotechnol. J.*, 2014, **12**, 861.
6 S. I. C. Carvalho, L. B. Bianchetti and F. J. B. Reifschneider, *Hortic. Bras.*, 2009, **27**, 135.
7 E. J. Oliveira, N. L. P. Dias and J. L. L. Dantas, *Euphytica*, 2011, **185**, 253.
8 G. G. Brito, N. D. Suassuna, V. N. Silva, V. Sofiatti, V. Diola and C. L. Morello, *Acta Sci., Agron.*, 2014, **36**, 335.
9 M. B. H. Santos, A. A. Gomes, W. T. S. Vilar, P. B. A. Almeida, M. Milani, M. B. M. Nóbrega, E. P. Medeiros, R. K. H. Galvão and M. C. U. Araújo, *J. Braz. Chem. Soc.*, 2014, **25**, 969.
10 J. Wang, L. W. Pembleton, R. C. Baillie, M. C. Drayton, M. L. Hand, M. Bain, T. I. Sawbridge, G. C. Spangenberg, J. W. Forster and N. O. I. Cogan, *Mol. Breed.*, 2014, **33**, 435.
11 I. G. Mylonas, A. Georgiadis, A. P. Apostolidis, K. Bladenopoulos and M. Koutsika-Sotiriou, *Rom. Biotechnol. Lett.*, 2014, **19**, 9421.
12 W. Dong, T. Cheng, C. Li, C. Xu, P. Long, C. Chen and S. Zhou, *Mol. Ecol. Resour.*, 2014, **14**, 336.
13 M. Simon, A. Simon, F. Martins, L. Botran, S. Tisné, F. Granier, O. Loudet and C. Camilleri, *Plant J.*, 2012, **69**, 1094.

14 A. Levi, A. H. Paterson, V. Barak, D. Yakir, B. Wang, P. W. Chee and Y. Saranga, *Mol. Breed.*, 2008, **23**, 179.
15 L. Yang and Q. Sun, *Anal. Methods*, 2016, **8**, 1914.
16 J. H. Lee and M. G. Choung, *Food Chem.*, 2011, **126**, 368.
17 R. Vitale, M. Bevilacqua, R. Bucci, A. D. Magrì, A. L. Magrì and F. Marini, *Chemom. Intell. Lab. Syst.*, 2013, **121**, 90.
18 L. Esteve Agelet and C. R. Hurburgh Jr, *Talanta*, 2014, **121**, 288.
19 J. Gao, X. Li, F. Zhu and Y. He, *Comput. Electron. Agr.*, 2013, **99**, 186.
20 M. Huang, Y. Ma, Y. Li, Q. Zhu, G. Huang and P. Bub, *Anal. Methods*, 2014, **6**, 7793.
21 S. Mahesh, A. Manickavasagan, D. S. Jayas, J. Paliwal and N. D. G. White, *Biosystems Eng.*, 2008, **101**, 50.
22 W. Kong, C. Zhang, F. Liu, P. Nie and Y. He, *Sensors*, 2013, **13**, 8916.
23 X. Zhang, F. Liu, Y. He and X. Li, *Sensors*, 2012, **12**, 17234.
24 F. J. Rodríguez-Pulido, D. F. Barbin, D. W. Sun, B. Gordillo, M. L. González-Miret and F. J. Heredia, *Postharvest Biol. Technol.*, 2013, **76**, 74.
25 S. Serranti, D. Cesare, F. Marini and G. Bonifazi, *Talanta*, 2013, **103**, 276.
26 C. Ridgway and J. Chambers, *J. Near Infrared Spectrosc.*, 1998, **6**, 115.
27 C. B. Singh, D. S. Jayas, J. Paliwal and N. D. G. White, *J. Stored Prod. Res.*, 2009, **45**, 151.
28 C. B. Singh, D. S. Jayas, J. Paliwal and N. D. G. White, *Biosystems Eng.*, 2010, **105**, 380.
29 M. Berman, P. Connor, L. Whitbourn, D. Coward, B. Osborne and M. Southan, *J. Near Infrared Spectrosc.*, 2007, **15**, 351.
30 Å. Rinnan, F. v. d. Berg and S. B. Engelsen, *TrAC, Trends Anal. Chem.*, 2009, **28**, 1201.
31 R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812.
32 M. Vidal and J. M. Amigo, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 138.
33 T. Næs, T. Isakson, T. Fearn and T. D. Hardback, *A User-friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, 2002.
34 R. G. Brereton, *Chemometrics for Pattern Recognition*, John Wiley & Sons, Chichester, 2009, ch. 6, pp. 233–287.
35 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137.
36 M. R. de Almeida, D. N. Correa, W. F. C. Rocha, F. J. O. Scafi and R. J. Poppi, *Microchem. J.*, 2013, **109**, 170.
37 Y. Lu, C. Du, C. Yu and J. Zhou, *Anal. Methods*, 2014, **6**, 1412.
38 D. Ballabio and V. Consonni, *Anal. Methods*, 2013, **5**, 3790–3798.
39 J. S. Shenk, J. J. Workman and M. O. Westerhaus, in *Handbook of Near-infrared Analysis*, ed. D. A. Burns and E. W. Ciurczak, CRC press, Boca Raton, 3rd edn, 2007, ch 17, pp. 347–386.
40 Z. Huang, S. Sha, Z. Rong, J. Chen, Q. He, D. M. Khan and S. Zhu, *Ind. Crops Prod.*, 2013, **43**, 654–660.