Preventing over-fitting in hyperspectral image regression

A.A. Gowen,[a,*] J. Burger,[b] C. Esquerre,[c] G. Downey[c] and C.P. O'Donnell[a]

[a]*Biosystems Engineering, School of Agriculture, Food Science and Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland.*
E-mail: aoife.gowen@ucd.ie
[b]*BurgerMetrics SIA, Jelgava, Latvia*
[c]*Teagasc, Ashtown Food Research Centre, Ashtown, Dublin 15, Ireland*

Introduction

Many different approaches are available for the development of regression models (e.g. partial least squares regression (PLSR), principal components regression (PCR), stepwise linear regression), all of which require representative calibration sets containing spectra with corresponding measured variables (e.g. fat content, protein content). This poses a problem in hyperspectral imaging (HSI). It is practically impossible to measure the exact concentration of components in a sample at the pixel scale and therefore not possible to provide reference values for each pixel spectrum. To overcome this limitation, HSI regression models may be built using mean, or pixel spectra obtained from some representative region of a sample on which the reference value was obtained.[1] The regression models developed can be applied to each pixel spectrum of the hyperspectral image, resulting in a prediction image in which the spatial distribution of the predicted component(s) is easily interpretable. HSI regression using PLS has been used for predictive mapping distribution of chemical components in a variety of sample types. However, PLSR models are known to be prone to over-fitting, especially in the absence of statistically-independent datasets for model validation. The term 'over-fitting' in this respect usually means inclusion of too many latent variables in the prediction model. Selection of latent variables is therefore a critical step in PLSR model building. Various metrics can be extracted from hyperspectral imaging data to estimate the correct number of latent variables in PLSR models. The D-metric[1] estimates accuracy and precision from HSI prediction images by combining the pooled root mean squared error of prediction ($RMSEP_G$) and standard deviation (S_G) from different image regions (ROIs) as follows [Figure 1(a)]:

$$\text{D-metric}=[RMSEP_G^2+ S_G^2]^{1/2}$$

The D-metric does not take into account spatial distribution of pixel values in the prediction image. One relatively straightforward image analysis technique that does is Adjacent Pixel Intensity Difference Quantisation (APIDQ).[3] This technique considers the intensity difference in horizontal (dI_x) and vertical (dI_y) irections of an input image using subtraction [Figure 1(b)]:

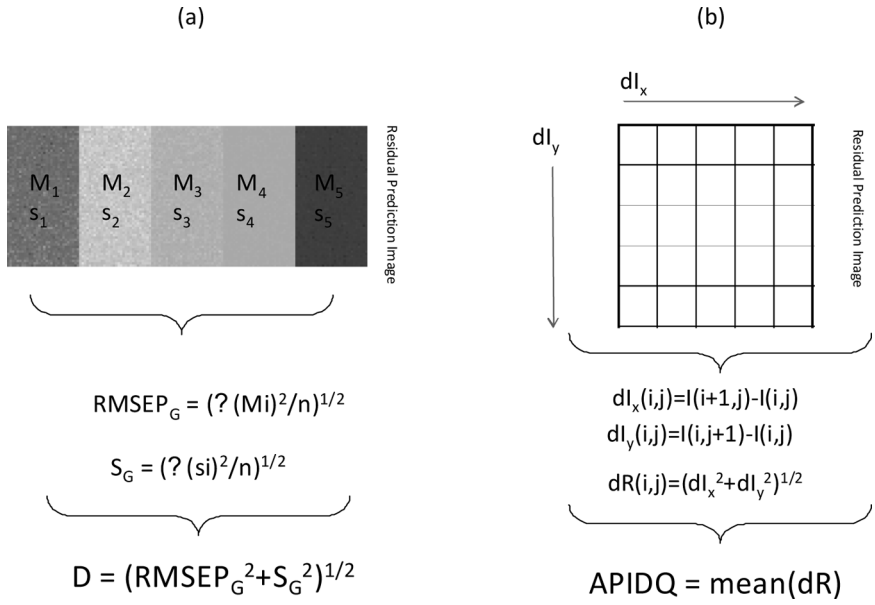

Figure 1. (a) Calculation of D-metric from a set of residual prediction images. (b) Calculation of APIDQ from prediction image.

$$dI_x(i,j)=I(i+1,j)-I(i,j);\ dI_y(i,j)=I(i,j+1)-I(i,j)$$

The azimuth and radial (dR) coordinates of dI_x and dI_y represent the direction and extent respectively of intensity variation in an image.

The aim of this work was to investigate the performance of PLSR models for predicting attributes from HSI data and in doing so examine the usefulness of the D-metric and APIDQ for preventing over-fitting in HSI regression.

Materials and methods

The HSI system employed in this research consisted of a high performance CCD camera (580 × 580 pixels) and a spectrograph (Specim V10E) attached to the camera covering the spectral range 400 nm to 1000 nm. In order to simulate a simple system of uniform samples with varying spectral response, a paint reference colour sheet (code: RAI 17-21, Fleetwood Ltd) with five different green levels was imaged (Figure 2).

The Hunter L-, *a*- and *b*-values of each green level were measured using a colorimeter (CR-400, Minolta Corp., Japan). A rectangular region of 1,800 spectra was defined for each green level [Figure 2(a)] and PLSR models were built to predict Hunter L-, *a*-, *b*-values. Standard normal variate (SNV) preprocessing was applied to the spectral data.

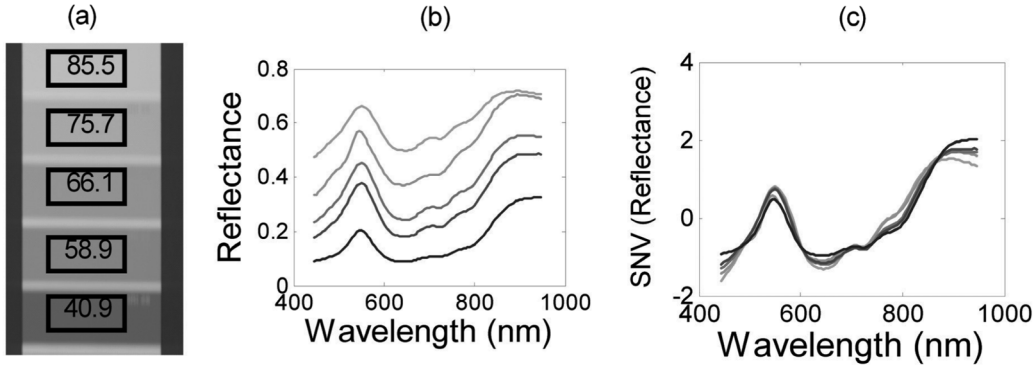

Figure 2. (a) Paint colour reference sheet with different green levels showing regions used in hyperspectral image regression model building and corresponding *L*-values measured in these regions. (b) Mean reflectance spectra from each region shown in (a). (c) standard normal variate (SNV) pre-processed spectra from each green level.

Data analysis

A Monte-Carlo re-sampling strategy was employed for model building and the optimal number of latent variables was estimated using the method described by Martens and Dardenne.[2] This method uses only spectral data as follows: 100 spectra were randomly-selected from each green level to make a training set and the remaining spectra (1800 minus 100 spectra) were used for model tuning. This was repeated 100 times; each time the optimal number of latent variables (A_{opt}) was estimated using the root mean squared error (*RMSE*) of cross-validation (training set) and prediction (tuning set) as follows:

$$A_{opt} = RMSE(A) + A*y*s$$

where A = number of LVs; y = penalty factor between (0.01,0.1); s = max [*RMSE*(*A*)]

Regression coefficients were estimated using the full set of calibration data (i.e. 1800 × 5 spectra) and applied to the hyperspectral image to create prediction images. The D-metric and APIDQ were calculated from residual prediction images for PLSR models with different numbers of latent variables.

Results and discussion

The optimal numbers of latent variables (A_{opt}) for the prediction of *L*-value, estimated from the Monte-Carlo re-sampling strategy, are shown in Table 1.

It is evident that the selection of penalty factors influences A_{opt}. In addition, the training and tuning sets suggest different A_{opt} (i.e. two or three latent variables). Prediction images for *L*-value applied to the calibration set are shown in Figure 3.

Visual inspection of these images indicated that three latent variables were sufficient for prediction of *L*-value; addition of further latent variables did not seem to improve the predictive

Table 1. Optimal numbers of latent variables (A_{opt}) for the prediction of *L*-value, estimated by Monte-Carlo resampling using different penalty values. Each number is the mean of 100 random sampling runs.

Penalty	A_{opt} training	A_{opt} tuning
0.05	1.8	3
0.03–0.05	2	3
0.01–0.1	1.7	2.6

performance of the model. The two-latent variable model is clearly inadequate, resulting in a poor prediction image. This simple example shows how visual interpretation of prediction images can be a useful step in model evaluation. The D-metric and APIDQ for prediction of *L*-, *a*- and *b*-value, calculated from the prediction images of the calibration set, are shown in Figure 4.

In the case of *L*-value prediction [Figure 4(a)], the D-metric exhibited no clear minimum but did not change appreciably after three latent variables (LVs); The APIDQ increased to a local maximum at three LVs but did not seem to change appreciably after that. The D-metric and APIDQ plots for prediction of *a*- and *b*-value were remarkably similar in shape [Figures 4(b)-(c)], indicating correlation between these two variables. For both *a*- and *b*-prediction the D-metric, while not reaching a minimum, seemed to reach a stable value after five LVs; the APIDQ reached

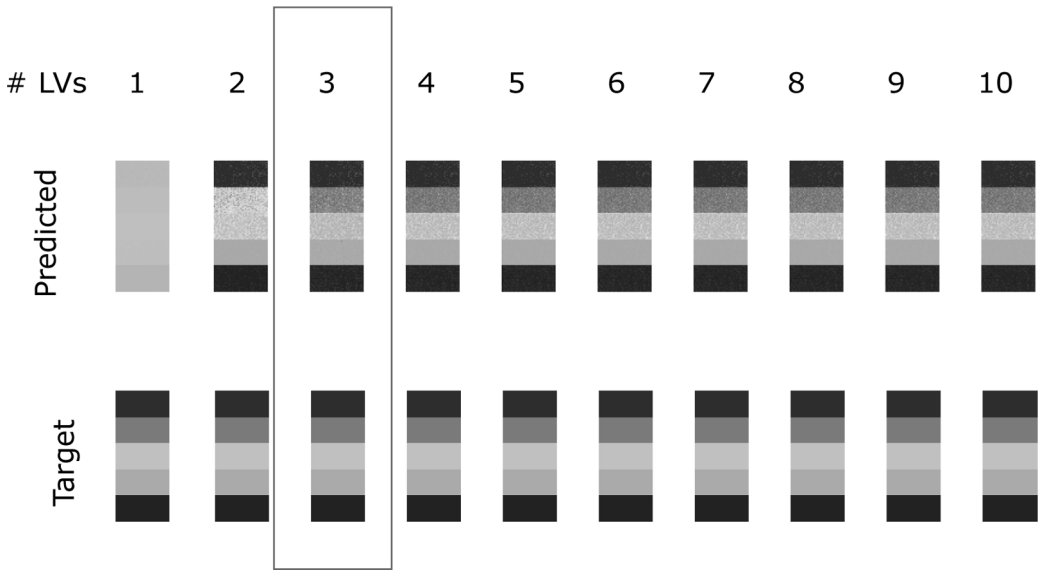

Figure 3. Predicted and Target images for prediction of *L*-value of paint colour reference sheet for different numbers of latent variables (#LVs). The prediction image for the optimal number of latent variables is indicated with a red rectangle.

Table 2. Optimal numbers of latent variables for the prediction of *a*- and *b*-value, estimated by Monte-Carlo resampling ("A_{opt} training" and "A_{opt} tuning" in which each number is the mean of 100 random sampling runs), observation of prediction images ('Pred Image').

	a	*b*
A_{opt} training	4	4
A_{opt} tuning	4.5	4.6
Pred. image	4	4
D & APIDQ	4 or 5	4 or 5

a maximum at four LVs after which changes were minimal. Based on this visual analysis of the D-metric and APIDQ curves, it seems that a three LV model is suitable for the *L*-value prediction while a four or five LV model is appropriate for *a*- and *b*-value prediction on the dataset examined. It is evident that observation of the prediction images D-metric and APIDQ agreed with the Monte-Carlo re-sampling strategy which suggested an A_{opt} of four or five.

Conclusions

This work highlights one of the major advantages of hyperspectral imaging not possible with traditional point spectroscopy; the ability to construct prediction images. Such images enable direct visualisation of model performance and can aid in the selection of an appropriate number of latent variables for any given regression model, thus preventing over-fitting. Numerous metrics may be calculated from prediction images to make the selection of latent variable dimension more objective. The examples given here, D-metric and APIDQ provide promising results compared to standard strategies such as Monte-Carlo re-sampling. Based on the findings of this study, a

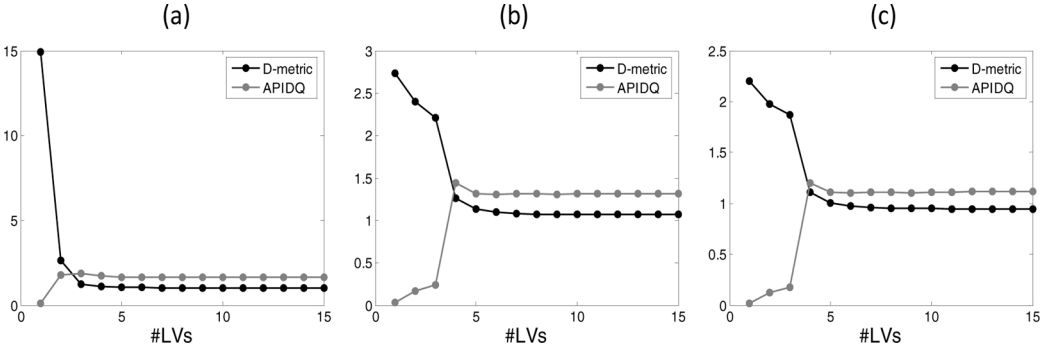

Figure 4. D-metric and APIDQ calculated from residual prediction images of calibration dataset for prediction of: (a) *L*-values; (b) *a*-values and (c) *b*-values.

combination of strategies, including visual analysis of prediction images, would be advisable for optimisation of hyperspectral image regression models. In future work, the methods presented here will be applied to new datasets in order to further test their performance.

References

1. J. Burger and P.Geladi, *J. Chemometr.* **20,** 106 (2006).
2. H. Martens and P. Dardenne, *Chemometr. Intell. Lab. Syst.* **44,** 99 (1998).
3. Lee *et al*., *Int. J. Comp Sci. Network Security* **9,** 147 (2009).