International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:7, No:5, 2013

Principal Component Regression in Noninvasive Pineapple Soluble Solids Content Assessment Based On Shortwave Near Infrared Spectrum

K. S. Chia, H. Abdul Rahim, and R. Abdul Rahim

***Abstract*—**The Principal component regression (PCR) is a combination of principal component analysis (PCA) and multiple linear regression (MLR). The objective of this paper is to revise the use of PCR in shortwave near infrared (SWNIR) (750-1000nm) spectral analysis. The idea of PCR was explained mathematically and implemented in the non-destructive assessment of the soluble solid content (SSC) of pineapple based on SWNIR spectral data. PCR achieved satisfactory results in this application with root mean squared error of calibration (RMSEC) of 0.7611 Brix°, coefficient of determination (R2) of 0.5865 and root mean squared error of cross-validation (RMSECV) of 0.8323 Brix° with principal components (PCs) of 14.

***Keywords*—**Pineapple, Shortwave near infrared, Principal component regression, Non-invasive measurement; Soluble solids content

I. Introduction

IN near infrared spectral analysis, principal component analysis (PCA) compresses a complex spectral data matrix from high dimension to low dimension in such a way that correlations among variables can be removed and important information of the data matrix can be extracted and represented by a small matrix [1]. The variables in this small matrix are uncorrelated among each other and able to provide almost all variances of the complex spectral data matrix. As a result, redundancy and collinearity problems in spectral analysis can be avoided by using uncorrelated principal components (PCs) as the input variables of a predictive model [2], [3].

In order to enhance the performance of PCA, different type of modified PCA methods were proposed in the literature, e.g. nonlinear principal component analysis (NLPCA) [4]-[8] and kernel principal component analysis (K-PCA) [9], [10]. But, K-PCA is infeasible and time consuming for a large number of data. Online and nonlinear PCA, on the other hand, was proposed to counter these problems by incremental eigen space update approach with a feature mapping function [11]. In 2004, the application of the permutation test was adapted to the framework of PCA to determine principal components based on the significations of both the principal components and the variable contribution [12].

Segmented PCA was proposed to provide the same results with less memory requirements for computing complex data [13]. The authors [13] reported that this approach is not only suitable for parallel calculations and cross-validation purposes, but it also avoids the process of reading the complete matrix into the main memory. However, misinterpretation of PCA, such as assuming that high loading variables are correlated in the same PC, is a critical issue and should be avoided. Recently, Camacho *et al*. (2010) [14] proposed Structural and Variance Information (SVI) plots to avoid this misinterpretation in PCA. Besides, the proposed method is possible to be applied for variable selection in spectral analysis.

PCA is very popular for its easier interpreting characteristic. The main application of PCA in spectral analysis is classification, such as classification of vegetable oils [15]. But, it is worth noting that the utilization of principal components (PCs) as the input variables of multiple linear regression (MLR) or artificial neural network (ANN) improves the predictive performance and reduce the redundancy problems in most cases. In 2009, Sven and Tim [2] presented an expectation robust algorithm for principal component regression (RA-PCR) to reduce the effects of outliers and missing elements in the data.

Besides combining PCA and MLR to perform PCR, it is also quite popular to use PCA to compress spectral data first before using ANN. This kind of approach is so-called principal components – artificial neural network (PCs-ANN). In this paper, however, we only discuss the implementation of PCR in shortwave near infrared (SWNIR) spectral analysis. The purpose of this paper is to provide a step-by-step explanation for the utilization of PCR in SWNIR spectral analysis application. The procedure of the experiment is summarized in Section II. General idea of PCR is described in Section III. Section IV denotes an application of the PCR in spectral analysis in soluble solid content (SSC) of pineapple assessment. Lastly, conclusions are drawn in Section V.

II. Material and Methods

In this work, 155 pineapple samples were used to investigate the performance of PCR model in SSC prediction.

K. S. Chia is with the Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM, Skudai, Malaysia (e-mail: kschia2@live.utm.my, k_s.chia@yahoo.co.uk).

H. Abdul Rahim is with the Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM, Skudai, Malaysia (Corresponding authors; phone: +6075535434; e-mail: herlina@fke.utm.my).

R. Abdul Rahim is with the Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM, Skudai, Malaysia (e-mail: ruzairi@fke.utm.my).

Reflectance spectra of each pineapple were scanned according to its label by using a USB4000 Miniature Fiber Optics Spectrometer (650-1000nm) (Ocean Optics, USA). A tungsten halogen light (LS-1, Ocean Optics, USA) was used as the energy source of SWNIR. A diffuse reflectance standard (WS-1, Ocean Optics, USA) was utilised as the optical reference standard to calibrate the spectroscopy. All spectral data were stored in a computer and processed via MATLAB simulation software (MATLAB® Version 7.9.0.529 (R2009b)). The SSC reference acquisition was started immediately after spectra acquisition of each pineapple was completed by using a digital Hand-Held "Pocket" refractometer (ATAGO).

III. Calibration Model

A. Pre-Processing

First order derivative with first order Savitzky-Golay (SG) smoothing filter was implemented to pre-processing the data. There are two steps in this pre-processing approach. The first step is to using SG smoothing filter to filter out high frequency unwanted signal. The second step is to use derivative to remove baseline shift problems.

B. Principal Component Regression

Principal component regression (PCR) is a combination of principal component analysis (PCA) and multiple linear regression (MLR) [16]. Instead of using whole spectrum, the idea of PCR is utilizing the first few principal components (PCs) as the input variables of a MLR model to eliminate the redundancy and collinear problems. Due to the fact that the PCA only decomposes the independent matrix data (spectral data) without including the effects of dependent variable (component of interest), the large loading values are correspond to the spectral regions with large variability [17]. Numerous literatures reported the application of PCR in near infrared spectral analysis improves the accuracy and robustness of a predictive model [18], [19]. In this section, the idea of PCA will be introduced mathematically first, and then followed by the procedure for MLR computation.

C. Principal Component Analysis

The first step in PCR calibration approach is to decompose a spectral data matrix by using PCA. Generally, two types of decomposition techniques can be applied. The first technique is computing eigenvectors and eigenvalues directly. The second technique is using singular-value decomposition (SVD) approach. In this paper, we only implement SVD to decompose the spectral data. This is because SVD is generally accepted to be the most stable and numerically accurate technique [20]. SVD decomposes a spectral data into column-mode eigenvectors, singular values and row-mode eigenvectors as shown in (1):

$$X_{n\times p} = U_{n\times n} \times S_{n\times p} \times (V_{p\times p})^T \quad (1)$$

where, n is the number of sample, p is the number of input variables, U is the normalised score matrix (column-mode eigenvectors), S is the diagonal matrix (singular values), and V is the un-normalised (loading matrix or row-mode eigenvectors). In this work, the function of *svd( )* in MATLAB (R2007a) was applied for the decomposition process. The algorithm of this function is based on LAPACK [21] routines. In fact, the product of column-mode eigenvectors, U, and singular values, S, is the so-called Principal Components (PCs), which are the input variables of a PCR calibration model. Therefore, (1) can be re-written as (2):

$$X_{n\times p} = PC_{n\times p} \times (V_{p\times p})^T \quad (2)$$

Next, if the number of first few PCs used is r, which is much smaller than p, then (2) can be represented as (3):

$$X_{n\times p} = PC_{n\times r} \times (V_{p\times r})^T \quad (3)$$

In order to validate the performance of a predictive model, loading matrix, V, is retained to transform the validation data into new PC. This transformation can be done by multiplying the validation spectral data, $X_{(validation)}$ to the loading matrix, V, as stated in (4):

$$X_{n\times p(validation)} \times (V_{p\times r}) = PC_{n\times r(validation)} \quad (4)$$

Obviously, the size of input variable has been reduced from p to r by using PCA. In this work, the value of p is 601 (total variables of a SWNIR spectrum) and the value of r is normally less than 20 only.

IV. Results and Discussion

Fig. 1 depicts that the root mean square errors of both calibration and cross-validation of PCR model versus the number input (i.e. PCs). Since the number input variables more than 14 does not improve the RMSECV significantly, the best PCs used for PCR is 14 to avoid both under-fitting and over-fitting problems. Under-fitting problem occurs when a predictive model does not have sufficient complexity (e.g. the number of input variables) for calibration between dependent and independent variables, i.e. soluble solids content and PCs, respectively.

Over-fitting, on the other hand, exists when a predictive model is trying to fit noises and unwanted signals in its calibration process, which will ultimately produce outstanding calibration results but worse validation outcomes. However, over-fitting problem in terms of worse validation results does not illustrated in Fig. 1. This unexpected observation may due to the fact the variance from PCs is decreasing subsequently from the first PC, whereby the PCs with small variance may have relatively less influence compared to the first few PCs in the calibration. In addition, it is possible that high frequency noises and unwanted signals were eliminated by using first

order Savitzky-Golay (SG) smoothing filter. Even thought the aforementioned over-fitting problem does not exist, including extra PCs without improving validation results is risky and burden for any calibration model. Therefore, optimal PCs number selection is vital when using PCR model.

Although the first PC contains most of the variance of a spectral data, the prediction results from PCR with only the first PC suggest that the first PC, individually, may not contain useful information to establish the relationship between spectral data and soluble solids contain of pineapple, i.e., R^2 is around zero. However, by using multiple regression model, the first PC coupled with its following PCs can predict the soluble solids content of pineapple with reasonable accuracy, i.e. RMSECV less than one Brix°.

In addition, Fig. 1 also shows that signification improvements occur at the second, seventh and fourteenth PCs. This observation once again indicate that PCs with high variance does not necessary contain sufficient relevant information for the components of interest. PCs with low variance, on the other hand, does not necessary contain unwanted signal or noises.

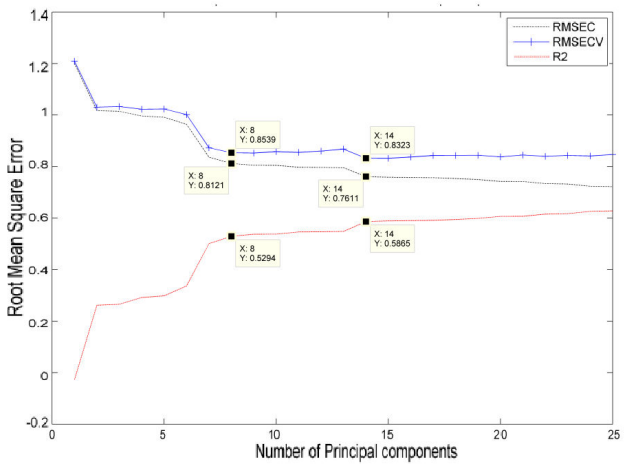

Fig. 1 Root mean square error of PCR against the number of PCs

In this work, by using the first 14 PCs as the input variables, PCR model achieved root mean square error of calibration (RMSEC) of 0.7611 Brix°, coefficient of determination (R2) of 0.5865 and root mean square error of cross-validation (RMSECV) of 0.8323 Brix°. These results are only slightly better than PCR with 8 PCs. In order to determine the optimal model complexity, Akaike's Information Criterion (AIC) and Bayes Information Criterion (BIC) values are computed and depicted in Fig. 2. It is worth noting that the values of AIC and BIC alone are meaningless, except for comparing. In this work, BIC values suggest that the optimal PCs number is 8. AIC values, on the other hand, indicate that the optimal PCs number is 14. The difference between AIC and BIC is that BIC consider the size of samples in its computation, i.e., BIC will penalty more than AIC when the number of sample is more than 100, and vice versa. Since the number of sample used in this work is more than 100, BIC suggests less complex model than AIC.

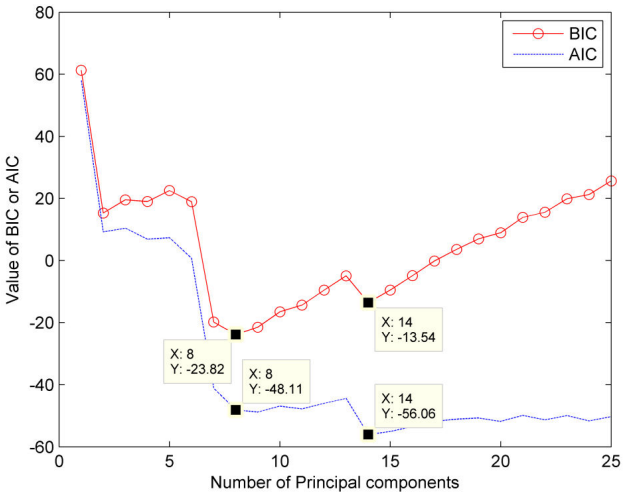

Fig. 2 Akaike's Information Criterion (AIC) and Bayes Information Criterion (BIC) against the number of PCs

V. Conclusion

In this paper, the development of principal component analysis was reviewed briefly in Section I. The implementation of principal component regression in shortwave near infrared (SWNIR) spectral analysis was described in Section II and III. The problems of under-fitting and over-fitting in PCR model were highlighted. In addition, the application of principal component regression (PCR) in nondestructive pineapple soluble solids content (SSC) assessment based on SWNIR data was demonstrated. Although PCR only achieved satisfactory performance in this work due to its limitation (i.e. PCR is not capable to handle non-linear information), it is still an important calibration model to be applied as a reference for comparison in the development of other calibration models. For future work, nonlinear model, e.g. artificial neural network, will be implemented to improve the prediction performance.

Acknowledgment

This work is financed by Zamalah Scholarship provided by University Teknologi Malaysia and Ministry of Higher Education of Malaysia. The authors also gratefully acknowledge to MOHE, UTM and VOTE No. 4L606.

References

[1] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, pp. 433-459, 2010.

[2] S. Serneels and T. Verdonck, "Principal component regression for data containing outliers and missing elements," Computational Statistics & Data Analysis, vol. 53, pp. 3855-3863, 2009.

[3] T. Næs and B.-H. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," Journal of Chemometrics, vol. 15, pp. 413-426, 2001.

[4] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," AIChE Journal, vol. 37, pp. 233-243, 1991.
[5] W. W. Hsieh, "Nonlinear principal component analysis by neural networks," Tellus A, vol. 53, pp. 599-615, 2001.
[6] A. H. Monahan, "Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System," Journal of Climate, vol. 13, pp. 821-835, 2000.
[7] M. Scholz, *et al*., "Non-linear PCA: a missing data approach," Bioinformatics, vol. 21, pp. 3887-3895, 2005.
[8] U. Kruger, *et al*., "Developments and Applications of Nonlinear Principal Component Analysis – a Review," ed, 2007, pp. 1-43.
[9] W. Wu, *et al*., "Kernel-PCA algorithms for wide data Part II: Fast cross-validation and application in classification of NIR data," Chemometrics and Intelligent Laboratory Systems, vol. 37, pp. 271-280, 1997.
[10] C. Ding and X. He, "K-means clustering via principal component analysis," presented at the Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada, 2004.
[11] B. J. Kim and I. K. Kim, "Incremental Nonlinear PCA for Classification," in Knowledge Discovery in Databases: PKDD 2004. vol. 3202, J.-F. Boulicaut, *et al*., Eds., ed: Springer Berlin / Heidelberg, 2004, pp. 291-300.
[12] S. Ledauphin, *et al*., "Simplification and signification of principal components," Chemometrics and Intelligent Laboratory Systems, vol. 74, pp. 277-281, 2004.
[13] A. S. Barros and D. N. Rutledge, "Segmented principal component transform-principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 78, pp. 125-137, 2005.
[14] J. Camacho, *et al*., "Data understanding with PCA: Structural and Variance Information plots," Chemometrics and Intelligent Laboratory Systems, vol. 100, pp. 48-56, 2010.
[15] T. Sato, "Application of principal-component analysis on near-infrared spectroscopic data of vegetable oils for their classification," Journal of the American Oil Chemists' Society, vol. 71, pp. 293-298, 1994.
[16] R. De Maesschalck, *et al*., "The Development of Calibration Models For Spectroscopic Data Using Principal Component Regression," Internet Journal of Chemistry, vol. 2, 1999.
[17] T. Næs, *et al*., A User-Friendly Guide to Multivariate Calibration and Classification: NIR Publications, 2002.
[18] B. Park, *et al*., Principal component regression of near-infrared reflectance spectra for beef tenderness prediction vol. 44. St. Joseph, MI, ETATS-UNIS: American Society of Agricultural Engineers, 2001.
[19] C. W. Chang, *et al*., Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties vol. 65. Madison, WI, ETATS-UNIS: Soil Science Society of America, 2001.
[20] P. J. Gemperline, "Principal Component Analysis," in Practical Guide To Chemometrics, Second Edition, ed: CRC Press, 2006, pp. 69-104.
[21] E. Anderson, *et al*., LAPACK Users' Guide, Third Edition ed.: Society for Industrial and Applied Mathematics, 1999.