Chemometrics and Intelligent Laboratory Systems 93 (2008) 63–69

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Mathematical models for prediction of active substance content in pharmaceutical tablets and moisture in wheat

C.S. Soh [a], P. Raveendran [a,*], R. Mukundan [b]

[a] Electrical Engineering Department, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia
[b] Department of Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 15 March 2007
Received in revised form 6 March 2008
Accepted 11 April 2008
Available online 20 April 2008

Keywords:
Discrete cosine transform
Pharmaceutical tablets
Wheat
NIR spectra

ABSTRACT

In the prediction of active substance content in pharmaceutical tablets and moisture in wheat, a very large number of wavelengths were used. Hence, a method to identify a limited number of wavelengths was developed. We introduce a novel approach that uses the discrete cosine transform (DCT) for this purpose. The data was obtained using near infrared spectrometer. From the DCT coefficients, a limited number was chosen as predictor variables to be used in partial least square (PLS) regression. Likewise, a limited number of DFT coefficients were also used in the PLS regression. The performance of combining the DCT with PLS was compared with that of the PLS model using the full spectral data and with the discrete Fourier transform (DFT). The results showed that the PLS model using DCT coefficients produced lower root mean square error than using the full NIR spectral data with the PLS and also the DFT.

1. Introduction

Near infrared (NIR) spectral data set has more number of variables or wavelengths than the number of samples or objects. Absorbance at different wavelengths, which are used as predictor variables in multivariate calibration of NIR data, are highly correlated. Therefore, multivariate calibration methods such as partial least square (PLS) regression [1] and principal component regression (PCR) are commonly used.

Near infrared spectra typically have smooth continuous curves and broad peaks. This is related to the fact that groups of neighbouring wavelengths are highly collinear. Thus, certain features of an NIR spectrum are spread across a few neighboring wavelengths. In view of this, NIR spectra are suitable for compression and feature extraction. The ability to reduce NIR spectra into few variables without degrading its useful information content is an important preprocessing step in parsimonious multivariate calibration of NIR spectroscopic data.

Many studies had been performed to reduce the number of predictor variables used in multivariate calibration of NIR spectra by extracting features from the spectra [2,3,4,5,6]. Some of these studies utilized linear transformation methods such as Fourier transform [2] and wavelet transform [3,4]. Linear transformation methods when applied to NIR spectra have the advantage of extracting features from NIR spectra by transforming the spectra into Fourier or wavelet domain. High frequency variations in the NIR spectra such as noise are easily separated from the rest of the spectra by identifying coefficients associated with high frequencies in the Fourier or wavelet domain. Similarly, low frequency variations such as baseline variation are separated into coefficients at low frequencies. This filtering ability is an advantage of wavelet, which is commonly used in baseline removal and denoising of NIR spectra [7,8].

Due to the presence of redundant information in NIR spectra, linear transformation can be used to concentrate spectral information into a small number of coefficients or variables [4]. Although compression could result in slight loss of information, it does not necessarily affect the prediction accuracy of multivariate calibration model significantly. In fact, theoretical study by Nadler et al. had demonstrated that feature extraction and selection prior to multivariate calibration could achieve near-optimal prediction error [9]. Their study adopted adaptive wavelet feature selection algorithm to perform dimensional reduction in order to achieve near-optimal prediction errors.

With the advantages of Fourier and wavelet transform applied to multivariate calibration of NIR spectra, it is interesting to explore and investigate other types of linear transformation that can be used for similar purpose. In this study, the DCT was used to predict active substance content in pharmaceutical tablet and moisture in wheat. A limited number of DCT coefficients ranging from 10 to 30 were used as predictor variables in the PLS regression. Their performances were compared with the PLS using the full spectral data and also with varying DFT coefficients combined with the PLS.

This paper is organized as follows. In Section 2, a brief review of the discrete cosine transform is presented. The experimental study, which

* Corresponding author. Tel.: +603 7967 5253.
E-mail address: ravee@um.edu.my (P. Raveendran).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.04.004

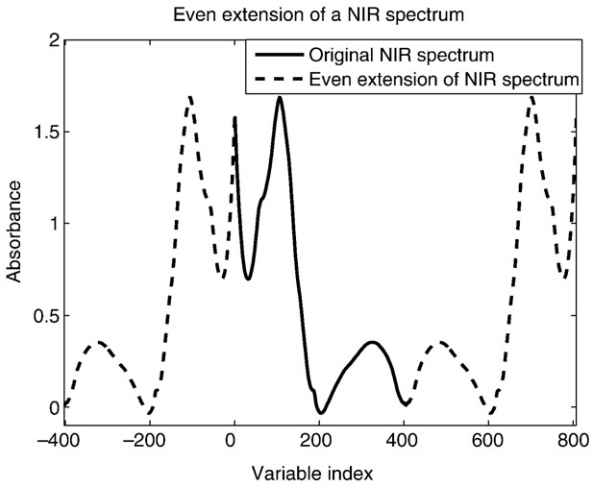

Fig. 1. Even extension of NIR data.

includes the data set as well as multivariate calibration of NIR data is discussed in Section 3. Section 4 consists of the performance results and discussion. Finally, Section 5 concludes the study.

2. Overview

In this section, a brief theory of the discrete cosine transform (DCT) is discussed and comparison is drawn between the DCT and the DFT. In the subsequent section, a brief theory on NIR spectra is presented.

2.1. Discrete cosine transform

The discrete cosine transform (DCT) [10] is closely related to the discrete Fourier transform (DFT). While the DFT has the cosine and sine functions as its basis functions, the DCT uses only cosine functions. Therefore, a function is expressed as a sum of sine and cosine functions with different frequencies and amplitudes using the DFT. The DCT expresses a function as a sum of cosines with different frequencies and amplitudes.

These differences are due to the underlying boundary conditions assumed by both transforms. While the DFT assumes periodic repetition of the finite interval function, the DCT implies even symmetry at the boundary of the function interval.

Let $x(n)$ denote an N-point sequence that is zero outside $0 \le n \le N-1$. The DCT of $x(n)$ is defined as,

$$X(k) = \sum_{n=0}^{N-1} 2x(n) \quad \cos\left[\frac{\pi}{2N}k(2n+1)\right], \tag{1}$$

for $0 \le k \le N-1$, and,

$$X(k) = 0 \tag{2}$$

otherwise.

The inverse DCT is defined as,

$$x(n) = \frac{1}{N}\left\{\frac{X(0)}{2} + \sum_{k=1}^{N-1} X(k) \quad \cos\left[\frac{\pi}{2N}k(2n+1)\right]\right\}, \tag{3}$$

for $0 \le n \le N-1$, and,

$$x(n) = 0, \tag{4}$$

otherwise.

The DCT has excellent energy compaction for correlated data [11]. Most of the information of a signal is concentrated in a few low frequency components of DCT. This contributes to the application of the DCT as a compression tool. The DCT possesses decorrelation characteristics, that is the ability to remove redundancies in data. Highly correlated data points can be transformed into uncorrelated transform coefficients.

Note that the DFT also possesses good energy compaction and decorrelation properties. However, the DFT being a complex transform, requires two quantities, the magnitude and phase, to represent the original data in the Fourier domain. For the DCT, the transformation of a sequence of real numbers is another sequence of real numbers. In addition, the DFT assumes periodic extension of finite interval function, and therefore, introduces discontinuities at the boundaries. These discontinuities require more high frequency components to be represented. In contrast, the DCT extends a function with even symmetry and therefore, avoid discontinuity at the boundary of the function as shown in Fig. 1.

The difference in boundary condition for both the DFT and the DCT has a profound impact on the performance of both transforms in applications such as signal compression and coding. When using the DFT, discontinuity at the boundary causes abrupt transition from one period of the signal to another. The discontinuity contributes energy

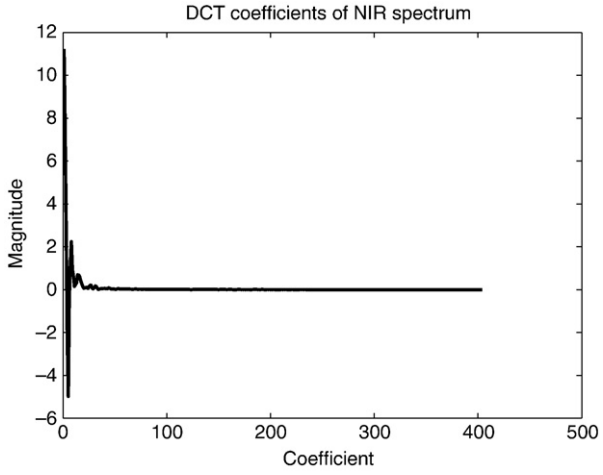

Fig. 2. Magnitude plot of DCT coefficients of NIR data.

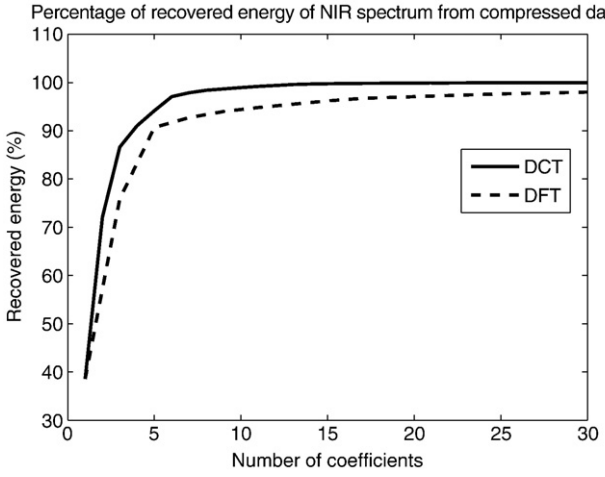

Fig. 3. Percentage of recovered energy for the DCT and the DFT compressed NIR data.

to high frequency components. Energy gets leaks from one frequency region to another region [12]. Therefore, energy scattering results in more coefficients are needed to contain substantial energy of the signal. In the DCT, even symmetry at the boundary of the signal allows continuous transition from one interval of the signal to the next. There is no need for high frequency components to describe discontinuity. Energy is confined in its own respective frequency regions [12]. Less number of coefficients is sufficient to contain a large portion of the signal energy. This contributes to excellent energy compaction property of the DCT. This property is evident when reconstruction of a signal using the same number of coefficients for both Fourier transform and the DCT yields smaller reconstruction error but higher percentage of recovered energy for the DCT.

Fig. 2 shows the magnitude plot of DCT coefficients for one NIR spectrum. It can be observed from the figure that only the first few coefficients at the left side of the plot have high values while the rest of the coefficients have values close to zero. This shows that the DCT can compact most of the energy into the first few coefficients.

2.2. NIR spectra and boundary effects

In wavelet analysis of NIR spectra, multi resolution analysis is employed, in which the signal to be analyzed is of length 2^n , n is an integer [13]. Signal length not of 2^n is extended by zero padding the signal. Zero padding the signal will introduce edge effect. In order to avoid this, linear padding is employed to extend the signal [3]. Information lies close to the boundary of NIR spectra may be subjected to artifacts when feature extraction methods such as wavelets are employed without proper consideration on the boundary effects [14].

Boundary effects commonly associated with NIR spectra transformation can be reduced by careful selection of spectra extension techniques. Empirical studies showed that symmetric padding performs better than zero padding in terms of compression and denoising, although boundary effects cannot be eliminated completely [15]. Depczynski et al. proposed using Sturm–Liouville wavelets to eliminate completely the boundary effects [15,16,17].

3. Experimental study

In this section, the organization of data set for pharmaceutical tablet and wheat is described. All simulations were performed in MATLAB 7.0.1. [22] running on Microsoft Windows XP operating system. The detailed description of the compression of NIR spectra using the DCT and the DFT is discussed. Finally, the multivariate calibration of NIR spectra using the PLS regression and the performance parameters used are presented.

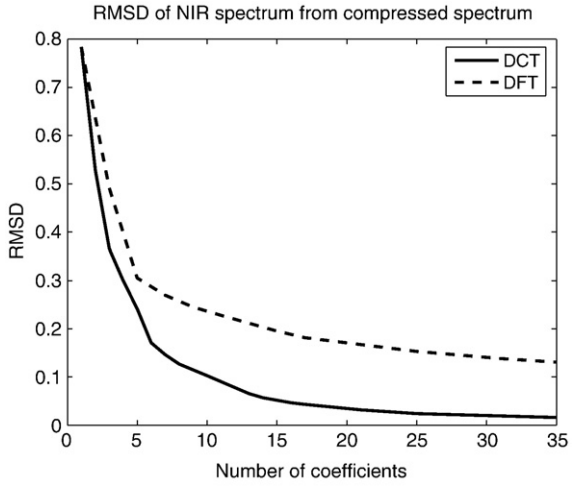

Fig. 4. Comparison of NIR spectrum reconstruction performances for the DCT and the DFT.

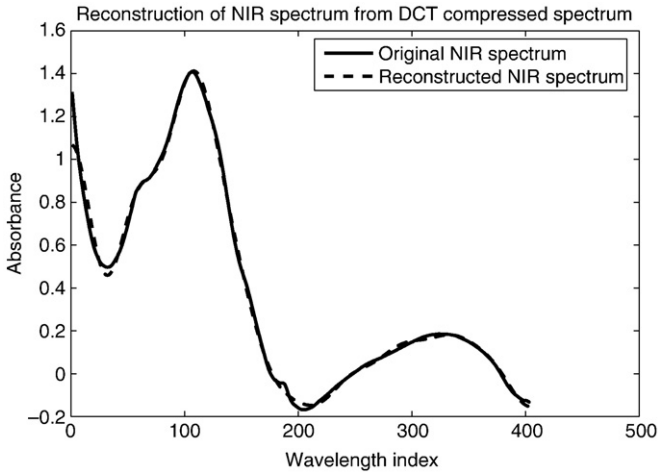

Fig. 5. Reconstructed NIR spectrum from 17 DCT coefficients compared with the original NIR spectrum.

3.1. Data set

We had utilized NIR transmittance spectral data for pharmaceutical tablet in this study [18]. This data set was downloaded from the web site of the Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen [19]. The NIR transmittance spectral data contains 310 spectra and 404 wavelengths. We divided the data set into 232 spectra for calibration while the remaining 78 spectra for testing the calibration model. The separation of data set into training set and test set is performed using the Kennard–Stone algorithm [20]. There is only one property to be predicted, that is the active substance content.

The second data set we used is the wheat data set [21]. The data set was downloaded from the ftp site ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas. This data set consists of 100 spectra and 701 wavelengths with moisture and protein content as properties to be determined. We only made use of the moisture property in this study. The data set is divided into 75 spectra for calibration and 25 spectra for testing using the Kennard–Stone algorithm [20].

3.2. Compression of NIR spectra

We tested the performance of the DCT in compression of NIR spectra and compared it with that of the DFT. We attempted to reconstruct the NIR compressed spectrum by hard thresholding. The first few transform coefficients with the highest magnitude were selected. The rest of the low magnitude coefficients were assigned values zero. Thus, coefficients with magnitude above certain threshold value contributed to the reconstruction of the NIR spectrum. Compression and reconstruction of NIR spectrum is performed in the following steps:

1. Transform the original NIR spectral data using the DCT or the DFT
2. Identify k number of coefficients with the highest magnitudes, or specify a threshold value, in which coefficients with magnitudes higher than the threshold are retained
3. The rest of the low magnitude coefficients are replaced by zero.
4. Perform inverse transform on the transformed spectral data to reconstruct NIR spectral data

Let $x=[x_1\ x_2\cdots\ x_M]^T$ be the original NIR spectrum with M wavelengths, and $\hat{x}=[\hat{x}_1\ \hat{x}_2\cdots\ \hat{x}_M]^T$ is the reconstructed NIR spectrum.

The reconstruction performance is calculated based on relative mean square difference (RMSD), as follows [23],

$$\mathrm{RMSD} = \frac{\mathrm{norm}\ (x - \hat{x})}{\mathrm{norm}\ (x)} \tag{5}$$

where norm(x) is the Euclidean norm of x given as,

$$\mathrm{norm}\ (x) = ||x||_2 = \left(\sum_{i=1}^{M} |x_i|^2\right)^{1/2} \tag{6}$$

The total energy of a signal transformed into its coefficients is $\Sigma_i|c_i|^2$, where c_i is the coefficient and i is the coefficient index. Let the coefficients with magnitude greater than threshold ε denoted by c_i^*, that is, $|c_i^*|>\varepsilon$. Thus, the percentage of recovered energy E_r by retaining the coefficients with magnitude higher than threshold ε, is given as [15],

$$E_r = \frac{\Sigma_i|c_i^*|^2}{\Sigma_i|c_i|^2} \times 100 \tag{7}$$

3.3. Multivariate calibration of NIR spectra

DCT coefficients obtained from transformation of NIR data were used as predictor variables in PLS regression. The NIR spectra were mean centered and variance scaled as a preprocessing step. Mean centering is performed by subtracting the mean spectrum of the calibration data set from every spectrum in the data set. Variance scaling is performed by taking the nth wavelength of every spectrum and divide by the standard deviation of the nth wavelength calculated from the calibration set. We selectively chose subset of the coefficients as variables to build calibration model. Selection was based on the first 10 to 30 coefficients. Prediction error for calibration set is calculated as root mean square error of calibration (RMSEC) as follows,

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I - K - 1}} \tag{8}$$

where y_i is the actual active substance content, $\hat{y}_i$ is the predicted active substance content, I is the number of samples in the calibration set and K is the number of PLS factors used in the model.

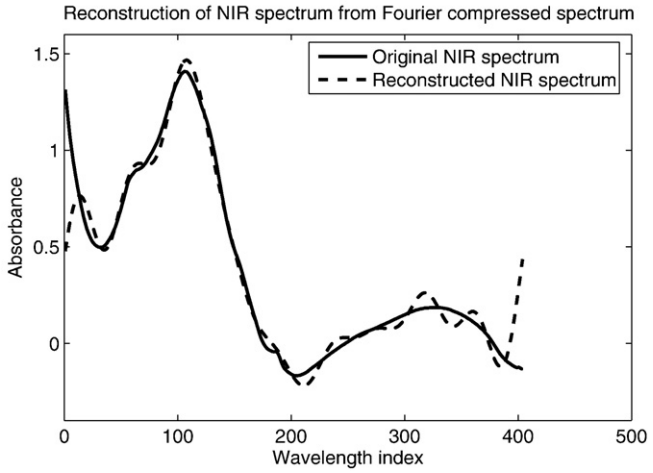

Fig. 6. Reconstructed NIR spectrum from 17 DFT coefficients compared with the original NIR spectrum.

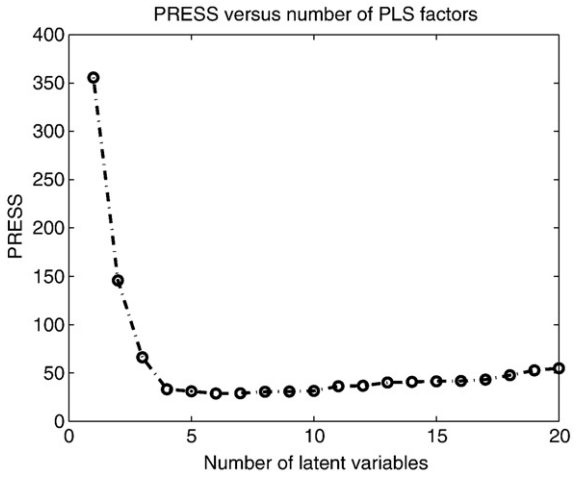

Fig. 7. PRESS values versus number of PLS factors.

Prediction performance of the PLS models were compared in terms of root mean square error of prediction (RMSEP) for test set.

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{J}\left(y_i - \hat{y}_j\right)^2}{J}} \tag{9}$$

where y_j is the actual active substance content, $\hat{y}_i$ is the predicted active substance content and J is the number of samples in the test set.

The optimum number of PLS factors was determined from cross-validation of calibration set by leave-g-out cross-validation. Predicted residual error sum of squares (PRESS) was calculated from the cross-validation,

$$\mathrm{PRESS} = \sum_{i=1}^{I}(y_i - \hat{y}_i)^2 \tag{10}$$

where y_i is the actual active substance content, $\hat{y}_i$ is the predicted active substance content and I is the number of samples in the cross-validation set.

Although NIR spectroscopic data has high dimensions due to high number of absorption measurements at multiple wavelengths, the underlying information can be described by a reduced number of variables called latent variables [24,25]. These latent variables can be

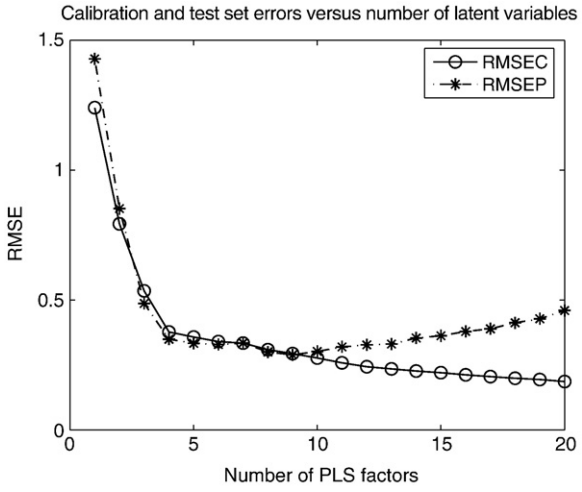

Fig. 8. Root mean square errors for calibration and test set versus number of latent variables.

Table 1
Prediction results for PLS model using original spectra for pharmaceutical tablet data set

No. of PLS factors	6
RMSEC	0.3402
RMSEP	0.3288
Test set	
>15%	0
10%–15%	4
5%–10%	18
2.5%–5%	20
1%–2.5%	23
0%–1%	13

derived from the original variables by linear combination of the absorption values at different wavelengths [24,26]. PLS and PCR are successful due to the ability to extract these latent variables from NIR data [24].

We adopted Wold's R criterion [27] to identify the optimum number of latent variables. According to Wold's R criterion, the optimum number of latent variables is obtained when the ratio of PRESS value at $(A+1)$ latent variables to PRESS value at A latent variables exceeds unity, in which A is taken as the optimum number of latent variables [27].

$$R = \frac{\mathrm{PRESS}(A+1)}{\mathrm{PRESS}(A)} \tag{11}$$

where PRESS$(A+1)$ is the PRESS at $(A+1)$ latent variables, PRESS(A) is the PRESS at A latent variables and R is the ratio of both PRESS values.

4. Results and discussion

Since two experiments were carried out using two different data sets, we first present the results for pharmaceutical tablet data set using the DCT and the DFT combined with PLS regression, and followed by the results for the wheat data set using the DCT and the DFT combined with PLS regression.

4.1. Pharmaceutical tablet

The recovered energy E_r from the DCT and the DFT is illustrated in Fig. 3. This graph demonstrates the high energy compaction property of the DCT. It can be observed from the graph that although using the same number of transform coefficients for both the DCT and DFT, the DCT stores higher percentage of spectral energy than the DFT.

The results for compression and reconstruction of NIR spectrum are illustrated in Figs. 4, 5 and 6. In Fig. 4, the reconstruction error based on RMSD shows that the errors resulting from using the DCT are much lower than the DFT for the same number of coefficients. As the number of coefficients are increased, the RMSD error for the DCT approaches towards zero unlike in the case of the DFT.

Table 2
Prediction results for PLS models using DCT coefficients for pharmaceutical tablet data set

No. of DCT coefficients	10	15	20	22	25	30
No. of PLS factors	10	8	11	10	2	3
RMSEC	0.3437	0.3420	0.3250	0.3266	0.3618	0.3569
RMSEP	0.3145	0.3175	0.2967	0.2904	0.3230	0.3198
Test set						
>15%	0	0	0	0	0	0
10%–15%	3	3	2	1	2	2
5%–10%	17	23	19	17	19	18
2.5%–5%	24	20	23	28	23	24
1%–2.5%	24	16	19	17	20	17
0%–1%	10	16	15	15	14	17

Table 3
Prediction results for PLS models using DFT coefficients for pharmaceutical tablet data set

No. of DFT coefficients	10	15	20	22	25	30
No. of PLS factors	6	6	6	7	7	6
RMSEC	0.3553	0.3440	0.3473	0.3439	0.3457	0.3496
RMSEP	0.3344	0.3187	0.3189	0.3167	0.3161	0.3144
Test set						
>15%	0	0	0	0	0	0
10%–15%	4	3	3	5	4	1
5%–10%	17	15	15	13	16	17
2.5%–5%	17	23	22	20	17	25
1%–2.5%	23	21	23	23	24	21
0%–1%	17	16	15	17	17	14

As the number of transform coefficient increases, the reconstruction error decreases because additional spectral information contained in additional coefficients is included in the spectrum reconstruction. However, up to a certain number of transform coefficient, the reduction of RMSD is no longer significant. This shows that there is a limit at which the number of transform coefficient is sufficient for reconstruction of the spectrum. The number of coefficient chosen for the spectrum reconstruction depends on the RMSD requirement. If lower RMSD is desired, then higher number of transform coefficient is used for spectrum reconstruction.

Figs. 5 and 6 show an example of using a fixed number of coefficients for NIR spectrum reconstruction. Fig. 5 shows the reconstruction of NIR spectrum using 17 DCT coefficients while Fig. 6 shows the same NIR spectrum reconstruction using 17 DFT coefficients. In both figures, although using the same number of coefficients, we notice that by using the DCT, the original NIR spectrum can be obtained. This is not the case when the DFT is used as shown in Fig. 6. This demonstrates the superiority of the DCT for compression of NIR data compared to the DFT. Thus, the DCT requires fewer number of coefficients than the DFT to compress the NIR data. Note that in Figs. 5 and 6, we used only 17 coefficients as one example of spectrum reconstruction. Similar result can be obtained using different number of coefficients.

Fig. 7 shows the changes in PRESS values with increasing number of latent variables included in the PLS model using the original spectral data without the DCT. Using Wold's R criterion, the optimum number of latent variables is six.

Fig. 8 illustrates the changes of root mean square errors for calibration set and test set with increasing number of latent variables, for PLS model using the original NIR spectra without the DCT. It can be

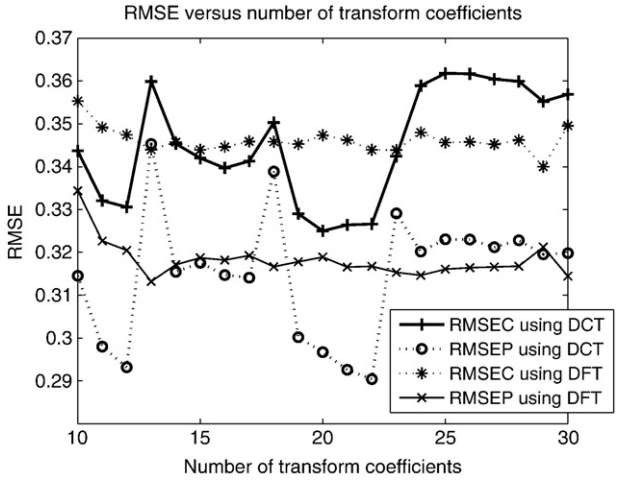

Fig. 9. Root mean square error versus number of transform coefficients for pharmaceutical tablet data set.

Table 4
Prediction results for PLS model using original spectra for wheat data set

No. of PLS factors	4
RMSEC	0.2259
RMSEP	0.2137
Test set	
>15%	0
10%–15%	0
5%–10%	0
2.5%–5%	2
1%–2.5%	12

observed from the graph that as number of latent variables increases, RMSEC decreases while RMSEP decreases up to a certain optimum number of latent variables and subsequently increases thereafter. This indicates overmodeling of the calibration data set.

Table 1 shows the result of PLS model using the optimum number of latent variables, that is six. The RMSEC obtained from the PLS model is 0.3402, while the RMSEP is 0.3288. The percentage values are the distribution of samples in the calibration set and the test set according to individual sample percentage error. For example, there are 13 samples in the test set that fall within 0% to 1% error.

Table 2 summarizes the PLS modeling results using DCT coefficients obtained from the transformation of NIR spectral data using the DCT. The first 10 to 30 coefficients were used for PLS modeling. Table 2 only shows results for 10, 15, 20, 22, 25 and 30 coefficients. Due to energy compaction property of the DCT, higher coefficients have values close to zero and therefore can be safely excluded from PLS modeling. The close to zero values of higher DCT coefficients can be observed from Fig. 1. Generally, PLS models using DCT coefficients recorded lower RMSEP values than PLS model using original spectra as can be seen from the comparisons between Tables 1 and 2. PLS model using 22 DCT coefficients resulted in the lowest RMSEP at 0.2904, as shown in Table 2.

Table 3 summarizes the PLS modeling results using DFT coefficients obtained from transformation of NIR spectral data using the DFT. Similar to the previous method using the DCT, only 10 to 30 coefficients were used in PLS modeling. Table 3 only shows results for 10, 15, 20, 22, 25 and 30 coefficients. The graph in Fig. 9 plots RMSEC and RMSEP values for 10 to 30 coefficients. Refering to this graph, the lowest RMSEP value for the DFT model is achieved using 13 DFT coefficients with RMSEP of 0.3132. This is higher than the minimum RMSEP value recorded in Table 2.

4.2. Wheat

For the wheat data set, Table 4 shows the PLS model result using the original spectra without transformation. Table 5 summarizes the results for PLS models using different number of DCT coefficients. Although we experimented with 10 to 30 coefficients, only the results for 10, 15, 20, 22, 25 and 30 coefficients are shown here. Similarly, the results for PLS

Table 5
Prediction results for PLS models using DCT coefficients for wheat data set

No. of DCT coefficients	10	15	20	22	25	30
No. of PLS factors	8	7	4	6	6	6
RMSEC	0.2297	0.2101	0.2382	0.2186	0.2056	0.2325
RMSEP	0.2145	0.2138	0.2000	0.1822	0.2052	0.1712
Test set						
>15%	0	0	0	0	0	0
10%–15%	0	0	0	0	0	0
5%–10%	0	0	0	0	0	0
2.5%–5%	2	1	1	0	0	1
1%–2.5%	12	16	12	14	15	10
0%–1%	11	8	12	11	10	14

Table 6
Prediction results for PLS models using DFT coefficients for wheat data set

No. of DFT coefficients	10	15	20	22	25	30
No. of PLS factors	5	5	8	7	8	8
RMSEC	0.2288	0.2360	0.1981	0.1964	0.1939	0.1949
RMSEP	0.2373	0.2197	0.2321	0.2469	0.2441	0.2417
Test set						
>15%	0	0	0	0	0	0
10%–15%	0	0	0	0	0	0
5%–10%	0	0	0	0	0	0
2.5%–5%	2	2	3	3	5	5
1%–2.5%	13	14	13	12	10	10
0%–1%	10	9	9	10	10	10

models using DFT coefficients are summarized in Table 6. Only the results for 10, 15, 20, 22, 25 and 30 DFT coefficients are shown here. A complete plot of RMSEC and RMSEP values for all 10 to 30 coefficients for both DCT and DFT based PLS models are shown in the graph in Fig. 10. By comparing the results in Tables 4, 5 and 6, the PLS model with the lowest RMSEP value was obtained from the DCT model using 30 coefficients which give the RMSEP value of 0.1712. From the RMSE plot in Fig. 10, it can be observed that the RMSEP values for the DCT based PLS models are generally lower than the RMSEP values of DFT based models for different number of transform coefficients.

From the results above, it is clear that selection of the optimum number of transform coefficients is crucial since too few coefficients selected will result in important coefficients being excluded from PLS modeling. However, too many coefficients selected leads to false information being introduced into the PLS model because most of the transform coefficients have values very close to zero. Introducing coefficients with values very close to zero will affect the accuracy of the PLS model. Therefore, an important task in calibration is identifying the optimum number of coefficients to effectively capture sufficient information in the NIR spectra for the purpose of building an optimum PLS model. This can be done by experimenting with a range of number of coefficients as performed in this study.

5. Conclusion

Prediction of active substance content in pharmaceutical tablets and moisture content from wheat using the discrete cosine transform coupled with partial least squares regression is demonstrated in this work. The objective of this study is to investigate the use of discrete cosine transform for compression and reconstruction of NIR spectra and also for calibration and prediction of substance properties. The performance was compared with using the same number of DFT

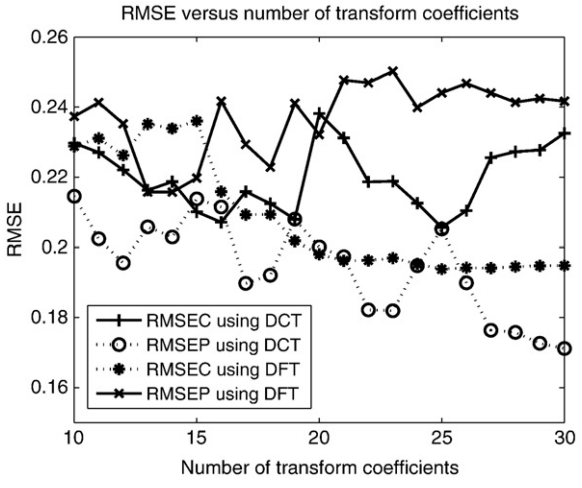

Fig. 10. Root mean square error versus number of transform coefficients for wheat data set.

coefficients with the original data set using the PLS regression. The number of coefficients used for PLS modeling is determined by experimenting with different numbers of coefficients and selecting the PLS model with the lowest root mean square error of prediction. We showed that using DCT and DFT coefficients as predictor variables not only reduced the number of predictor variables used in the training or calibration, it also leads to improved prediction accuracy if an appropriate number of coefficients is selected.

References

[1] Soft modeling, The basic design and some considerations, in: K.-G. Jöreskog, H. Wold (Eds.), Systems Under Indirect Observation, vols. I and II, North-Holland, Amsterdam, 1982.
[2] L. Pasti, D. Jouan-Rimbaud, D.L. Massart, O.E. de Noord, Application of Fourier transform to multivariate calibration of near-infrared data, Anal. Chim. Acta 364 (1998) 253–263.
[3] J. Trygg, S. Wold, PLS regression on wavelet compressed NIR spectra, Chemometr. Intell. Lab. Syst. 42 (1998) 209–220.
[4] C.E.W. Gributs, D.H. Burns, Parsimonious calibration models for near-infrared spectroscopy using wavelets and scaling functions, Chemometr. Intell. Lab. Syst. 83 (2006) 44–53.
[5] A. Höskuldsson, Variable and subset selection in PLS regression, Chemometr. Intell. Lab. Syst. 55 (2001) 23–38.
[6] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Comparison of different variable selection methods conducted on NIR transmission measurement on intact tablets, Chemometr. Intell. Lab. Syst. 69 (2003) 3–12.
[7] B.K. Alsberg, A.M. Woodward, D.B. Kell, An introduction to wavelet transforms for chemometricians: a time frequency approach, Chemometr. Intell. Lab. Syst. 37 (1997) 215–239.
[8] D. Chen, X. Shao, B. Hu, Q. Su, A background and noise elimination method for quantitative calibration of near infrared spectra, Anal. Chim. Acta 511 (2004) 37–45.
[9] B. Nadler, R.R. Coifman, The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration, J. Chemometr. 19 (2005) 107–118.
[10] N. Ahmed, T. Natarajan, K.R. Rao, Discrete cosine transform, IEEE Trans. Comput. C-23 (1) (1974) 90–93.
[11] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall International, Inc., 1989, p. 153.
[12] S.V. Narasimhan, M. Harish, Spectral estimation based on discrete cosine transform and modified group delay, Signal Processing 86 (2006) 1586–1596.
[13] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Machine Intell. 11 (7) (1989) 674–692.
[14] K. Jetter, U. Depczynski, K. Molt, A. Niemöller, Principles and applications of wavelet transformation to chemometrics, Anal. Chim. Acta 420 (2000) 169–180.
[15] U. Depczynski, K. Jetter, K. Molt, A. Niemöller, The fast wavelet transform on compact intervals as a tool in chemometrics II. Boundary effects, denoising and compression, Chemometr. Intell. Lab. Syst. 49 (1999) 151–161.
[16] U. Depczynski, K. Jetter, K. Molt, A. Niemöller, The fast wavelet transform on compact intervals as a tool in chemometrics I. Mathematical background, Chemometr. Intell. Lab. Syst. 39 (1997) 19–27.
[17] U. Depczynski, Sturm–Liouville wavelets, Appl. Comput. Harm. Anal. 5 (2) (1998) 216–247.
[18] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg Nielsen, Chemometric quantitation of the active substance in a pharmaceutical tablet using near infrared (NIR) transmittance and NIR FT Raman spectra, Appl. Spectrosc. 56 (5) (2002) 579–585.
[19] www.models.life.ku.dk/research/data/tablets. Web site.
[20] R.W. Kennard, L.A. Stone, Computer aided designs of experiments, Technometrics 11 (1969) 137–148.
[21] J.H. Kalivas, Two data sets of near infrared spectra, Chemometr. Intell. Lab. Syst. 37 (1997) 255–259.
[22] The MathWorks, Inc., Natick, Massachusetts.
[23] S. Ren, L. Gao, Simultaneous quantitative analysis of overlapping spectrophotometric signals using wavelet multiresolution analysis and partial least squares, Talanta 50 (2000) 1163–1173.
[24] D. Jouan-Rimbaud, B. Walczak, D.L. Massart, I.R. Last, K.A. Prebble, Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data, Anal. Chim. Acta 304 (1995) 285–295.
[25] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.
[26] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, Chemometr. Intell. Lab. Syst. 58 (2001) 131–150.
[27] S. Wold, Cross-validation estimation of the number of components in factor and principal component analysis, Technometrics 24 (1978) 397–405.