Using Orthogonal Signal Correction to Correct the Near Infrared Reflectance Spectroscopy for Determination of Protein Content in Brown Rice

Wen-Shin Lin¹, Shin-Chun Cheng², Chien-Pang Lee³, Mei-Chu Hong⁴ and Bo-Jein Kuo²,*

¹ Department of Plant Industry, National Pingtung University of Science and Technology, Pingtung 91201, Taiwan ROC
² Department of Agronomy, National Chung Hsing University, Taichung City 40227, Taiwan ROC
³ Department of Information Management, Da-Yeh University, Changhua 51591, Taiwan ROC
⁴ Taichung District Agricultural Research and Extension Station, Council of Agriculture, Changhua 51544, Taiwan ROC

ABSTRACT

When determining the protein content of brown rice (unpolished grain of rice), one usually grinds the rice grains into flour and then adopts the near infrared reflectance spectroscopy (NIRS) to construct the calibration model. However, the NIRS is easily influenced by interference among different constituents. To effectively eliminate interference in NIRS on brown rice, this study employed several data pre-treatment methods, including normalization, standard normal variate (SNV), multiplicative signal correction (MSC), and two orthogonal signal correction (OSC) approaches, which were OSC-CV and JSoSc. Then, the partial least squares regression (PLSR) was used to construct a calibration model to determine the protein content of brown rice. Results showed that the PLSR model constructed by using OSC-CV to correct the NIRS of brown rice had the better statistical indicators in terms of the root mean squared error and coefficient of determination ($R^2$) than the other pre-treatment methods. The performance of the PLSR model was similar to those constructed on the original NIRS of the brown rice flour. Moreover, the number of latent variables required in the PLSR model was less than that with the original NIRS of brown rice flour. In conclusion, the OSC-CV method was a useful tool for correction and interpretation when constructing the PLSR model for determining the protein content of brown rice with NIRS data.

Key words: Pre-treatment, NIR data, Partial least squares regression, Orthogonal signal correction.
Using OSC to Correct NIRS for Determining of Protein Content

立検量模式。然而，NIRS 容易受到不同組成的干擾影響。為了有效降低在使用 NIRS 時的干擾影響，本研究以 normalization、standard normal variate (SNV)、multiplicative signal correction (MSC)及兩種直交訊息校正法 (orthogonal signal correction; OSC)，包括 OSC-CV 及 JSosc 等方法，對經由糙米粒之 NIRS 資料進行前處理，再利用淨最小平方回歸 (partial least squares regression; PLSR) 建立檢量模式，以進行蛋白質含量的預測。研究結果顯示，經過 OSC-CV 前處理方法轉換後，在均方根誤差及 $R^2$ 等統計指標的表現上，均較其他的前處理方法為佳。此外，OSC-CV 修正糙米粒之 NIRS 後建立的檢量模式，其分析能力與利用糙米粉之 NIRS 所獲得的結果相近。並且，利用 OSC-CV 修正糙米粒之 NIRS 後建立的檢量模式時，其潛在變數之數目較糙米粉之 NIRS 所建立的檢量模式為少，因而使得所建立的模式更容易解釋。

關鍵詞：資料前處理、近紅外光譜資料、淨最小平方迴歸、直交訊息校正法。

INTRODUCTION

Recently, more emphasis has been placed on the food quality. Food policy which had been focusing on the volume of production has gradually begun to value improvement in quality. Taking rice as an example, the protein content of rice is one of the key factors affecting its nutritional value and eating quality (Pang et al. 1997). The protein content influences the time required for cooking, the texture of the cooked rice, and the nutritional value (Delwiche et al. 1996). Rice varieties which contain higher levels of protein require longer cooking durations, and the cooked rice is yellower in color. In addition, due to the harder texture, the eating quality is worse. Therefore, the level of protein is one of the important factors in determining the quality of rice.

Both the traditional destructive Kjeldahl Method (Miller and Houghton 1945) and the non-destructive near infrared spectroscopy (NIRS) method can be used to determine the protein content of rice. The traditional chemical method, which is destructive, prohibits the reuse of experimental samples. In crop breeding research, this method may cause insufficiency in the breeding materials. However, the NIRS method uses various wavebands in the near infrared region of to obtain the reflectance spectral data for the tested rice. Then, actual protein content in the rice was regressed on the NIRS spectra to construct an optimal calibration model. Because the NIRS method does not require destructive chemical analysis, it is widely applied in crop breeding research.

When applying the NIRS method to analyze the protein content of brown rice, one usually grinds the rice grains into flour, as larger rice grains demonstrate greater absorption toward the NIRS (Hruschka 1990). Therefore, the varied grain sizes influence the accuracy of the NIRS data, limiting the prediction of protein content. However, data pre-treatment methods which adjust for the effects of interference could help to obtain NIRS data, which further improves the capability and applicability of the calibration model (Blanco et al. 2001). For instance, Delwiche et al. (1996) performed second difference pre-treatment for the NIRS data of whole-grain milled rice to reduce the problem of baseline shift. The information of NIRS wavebands for amylose and protein contents in whole-grain milled rice could be used to improve the capability of the model prediction.

Thenadil and Martin (2005) indicated that when using the NIRS method in estimating the quantity of components in industrial products, such as slurry, or biological products, such as suspending agents, varied particle sizes of the testing sample usually led to errors. As a result, pre-treatment methods such as the MSC (multiplicative signal correction), SNV (standard normal variate), OSC (orthogonal signal correction), OPLS (orthogonal projection to latent structures), EMSC (extended multiplicative signal correction), and ISC (inverse signal correction) are necessary for data transformation adjusting for the influences of interference.

Blanco et al. (2001) used the NIRS method and adopted different pre-treatment methods to test the marketed products and laboratory
samples of nimesulide. Then, partial least squares regression (PLSR) was used to construct the calibration models of the original NIRS data and the data transformed with different pre-treatments. Results of the study indicated that the calibration models based on the NIRS data transformed by SNV and MSC methods showed significant improvements in several indicators, e.g., selection of the number of optimal latent variables, root mean square error of calibration (RMSEC), and root mean square error of prediction (RMSEP). Svensson et al. (2002) suggested that data corrected by the OSC pre-treatment method facilitated the analysis and interpretation of characteristics with fewer components, and the removed orthogonal data could also be analyzed.

Therefore, this study adopted pre-treatment methods including the normalization, SNV, MSC, and two OSC approaches (OSC-CV and JSosc) to correct NIRS data of brown rice, and used the PLSR to construct the calibration models for predicting the protein content of brown rice. In addition, this study made a comparison of performance of predicted protein content among calibration models constructed by data transformed with different pre-treatments.

**MATERIALS AND METHODS**

1. **Experimental materials**

The materials of this study were brown rice produced from the first and second cropping seasons in the experimental fields of Taichung District Agricultural Research and Extension Station, Council of Agriculture located in Central Taiwan in 1997. There were 210 rice samples, including 88 varieties of the indica rice subspecies, 80 of japonica rice subspecies, and 42 of waxy rice subspecies. Rice grains were first de-hulled to obtain brown rice samples and were then grinded and milled to obtain brown rice flour samples. Each sample was scanned from a fixed angle using the Bran+Luebbe InfraAlyzer 500 near infrared reflectance spectrophotometer three times to obtain the mean spectra value (range of the spectrum: 1,100-2,500 nm with a resolution of 4 nm). Log (1/R) transformation was then performed on the spectra values (R) to generate the absorbance spectra values, which provided the NIRS data for brown rice and brown rice flour. Finally, the semi-micro Kjeldahl method was adopted to identify protein content of brown rice flour samples. The rice samples were divided into calibration and validation sets systematically by actual protein content. There were 132 rice samples in calibration set and 78 rice samples in validation set, respectively.

2. **Partial least squares regression (PLSR)**

PLSR is an improved method over PCR (principal components regression). Unlike PCR, which simply uses the information of X to calculate the loading matrix $\hat{p}_a$, PLSR incorporates the information of both X and Y to obtain $\hat{p}_a$, which improves the predictive ability of the model (Garthwaite 1994, Naes and Martens 1985). The following details the procedures for constructing a one-response variable PLSR (PLS1), which involves the orthogonal scores algorithm. The calculation process is listed below.

**Step 1.** Center the matrix of explanatory variables (X-matrix) and that of the response variables (Y-vector), and determine the number of latent variables $A_{max}$ ($a = 1, \ldots, A_{max}$).

**Step 2.** Calculate weight $w_a$ for Y.

Have $X_{a+1} = y_{a+1} w_y + E$ and $w_y w_y = 1$.

Then, $\hat{w}_a = c X_{a+1} y_{a+1}$,

where $c = (y_{a+1} X_{a+1} y_{a+1})^{-0.5}$.

**Step 3.** Calculate the score vector $i_a$ for $X_{a+1}$ and $y_{a+1}$.

$X_{a+1} = t_a \hat{w}_a + E$, and $w_y w_y = 1$.

Hence, $i_a = X_{a+1} \hat{w}_a$.

**Step 4.** Calculate the loading vector $\hat{p}_a$ for $X_{a+1}$.

$X_{a+1} = t_a \hat{p}_a + E$

$\hat{p}_a = X_{a+1} i_a / i_a i_a$

**Step 5.** Calculate the loading vector $\hat{q}_a$ for $y_{a+1}$.

$y_{a+1} = i_a q_a + f$

$\hat{q}_a = y_{a+1} i_a / i_a i_a$

**Step 6.** Calculate the residual matrix $\hat{E}_a$ and $\hat{f}_a$ for $X_{a+1}$ and $y_{a+1}$, respectively.
\[
\hat{E}_s = x_{s,t} - t \hat{p}_s
\]
\[
\hat{f}_s = y_{s,t} - t \hat{q}_s
\]

Take \( \hat{E}_s \) as \( X_s \) and \( \hat{f}_s \) as \( y_s \) in the next iteration and re-calculate Step 2. Repeat the above steps until calculation of the \( A_{\text{max}} \) latent variables is completed. PLS1 is then complete, which yields a regression model between \( y \) and \( A_{\text{max}} \) latent variables. The latent variables used in PLSR indicate the results of evaluation and weighting of each wavelength. Therefore, the coefficient of each latent variable can be shifted back to the regression coefficient of each wavelength using score and loading vectors. The common regression equation (Helland 1988) is generated as follows.

\[
y = \hat{X}_b^{PLS\text{orth}} + \hat{X}^{PLS\text{orth}}_b \hat{b}_1 \hat{q}^{\text{top}}((\hat{p} \hat{w})^{-1})
\]

A characteristic of PLSR is to integrate the scattered information of the explanatory variables into several latent variables. In fact, using latent variables to describe the internal structure of data somehow helps to reduce the dimensions of the data. The latent variables created by specific weights and linear combinations possess the following characteristics: (1) latent variables are orthogonal to each other; (2) the first latent variable explains the most amount of variance of the data and the last latent variable explains the least amount of the variance and so forth; and (3) the PLSR score plot of \( t_2 \) (score vectors of the second latent variable) against \( t_1 \) (score vectors of the first latent variable) can be used to examine whether multiple sources of variation exist in the data. If a scatter plot shows several obvious clusters, it means that special samples are contained in the data, or the data samples are from different populations. Conversely, a scatter plot without apparent clusters implies high similarity of the samples. Furthermore, for a PLSR with satisfactory fitting ability, there should be an excellent linear relationship between the score vectors of the explanatory variables and that of the response variable. Therefore, if the PLSR score plot of \( u_1 \) (score vectors of the response variable of the first latent variable) against \( t_1 \) (score vectors of the explanatory variables of the first latent variable) shows an excellent linear relationship, the model possesses excellent fitting ability.

### 3. Pre-treatment Methods

#### (1) Normalization

Normalization is the most common pre-treatment strategy. Taking the NIRS data of brown rice grain as an example, if the absorption of near infrared is influenced by grain size or water content, which alters the spectra value from \( A_1 \) to \( A_C \), one could divide \( A_C \) by the spectra value \( A_{\text{normalization}} \) of a specific wavelength or waveband to remove the interference coefficient \( C \) and thus reduce interference. In actual practice, all of the spectral values obtained are divided by the spectral value of a specific wavelength or waveband to transform the NIRS data into values between 0 and 1, which reduces interference and increases the fitting ability of the constructed calibration model (Bran+Luebbe 1996).

#### (2) Standard normal variate (SNV)

SNV proposed by Barnes et al. (1989) is considered one of the pre-treatment methods for standardizing the NIRS data. The principle of this method is to transform the original NIRS data into a new data set. The mean and standard deviation of the new data set are equal to 0 and 1, respectively. Through auto-scaling, the method performs centering and scaling to adjust problems with baseline shift, model sensitivity, and the scatter effect associated with the NIRS data (Bertran et al. 2001, Fernández-Cabanás et al. 2007), and reduces the influence of outliers on the constructed model (Bertran et al. 2001, Naes et al. 1990). The following Eq. (1) shows the original NIRS data containing \( q \) samples and \( p \) wavelengths in a matrix format:

\[
X = \begin{bmatrix}
  x_{11} & \cdots & x_{1b} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{q1} & \cdots & x_{qb} & \cdots & x_{qp}
\end{bmatrix}
\]

where \( x_{ab} \) denotes the \( a \)th sample and the \( b \)th wavelength of the original NIRS data, and \( a = 1, \ldots, q, \ b = 1, \ldots, p \).
To transform NIRS data, the means and standard deviations of each sample should first be obtained. Then, one will subtract the mean spectral value of the samples from the original spectral value of each sample, and divide the obtained values by the standard deviation of the spectral values of each sample, as shown in Eq. (2).

\[
X_{\text{STD}} = \begin{bmatrix}
\frac{x_{11} - \bar{x}_1}{s_1} & \frac{x_{1b} - \bar{x}_1}{s_1} & \ldots & \frac{x_{1p} - \bar{x}_1}{s_1} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{x_{q1} - \bar{x}_q}{s_q} & \frac{x_{qb} - \bar{x}_q}{s_q} & \ldots & \frac{x_{qp} - \bar{x}_q}{s_q}
\end{bmatrix}
\]

The \(\bar{x}_q\) in this equation indicates the mean of the \(p\) spectral values of the \(q\)th sample; \(s_q\) denotes the standard deviation of the \(p\) spectral values of the \(q\)th sample.

(3) Multiplicative signal correction (MSC)

Geladi et al. (1985) indicated that when conducting NIRS analyses, one should consider the light scatter’s wavelength dependency issue, which implies that the spectral values of each waveband of a sample are influenced by an identical scatter effect. Therefore, the MSC pre-treatment strategy is necessary for transforming the NIRS data. Assuming that the mean spectral value of all samples is an ideal sample, this ideal sample could be used to estimate the scatter coefficient of the individual samples and then transform the spectral value for each sample (Wold et al. 1998), which eliminates the multiplicative scatter effect due to the varied particle sizes in the samples (Martens and Næs 1989). In addition to effectively minimizing the influence of the multiplicative scatter effect on the constructed calibration model, the MSC, comparable to the SNV, reduces the effects of outliers on the model (Bertran et al. 2001, Naes et al. 1990).

Eq. (1) shows the original NIRS data containing \(q\) samples and \(p\) wavebands in a matrix format. One first calculates the mean spectral value (\(\bar{x}_q\)) of the \(q\) samples under the \(b\)th waveband. It is assumed that the ideal sample \(= [\bar{x}_1, \ldots, \bar{x}_q, \ldots, \bar{x}_p]\). Then, intercept \(\hat{\beta}_{0,a}\) and slope \(\hat{\beta}_a\) of the simple linear regression model are estimated using least squares method by regressing on the ideal sample. One then subtracts \(\hat{\beta}_{0,a}\) from the original spectral value, and then divides the result by \(\hat{\beta}_a\), yielding the spectral value \(X_{\text{MSC}}\), after MSC transformation [Eq. (3)].

\[
X_{\text{MSC}} = \begin{bmatrix}
\frac{x_{11} - \hat{\beta}_{0,1}}{\hat{\beta}_1} & \frac{x_{1b} - \hat{\beta}_{0,1}}{\hat{\beta}_1} & \ldots & \frac{x_{1p} - \hat{\beta}_{0,1}}{\hat{\beta}_1} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{x_{q1} - \hat{\beta}_{0,q}}{\hat{\beta}_q} & \frac{x_{qb} - \hat{\beta}_{0,q}}{\hat{\beta}_q} & \ldots & \frac{x_{qp} - \hat{\beta}_{0,q}}{\hat{\beta}_q}
\end{bmatrix}
\]

(4) Orthogonal signal correction (OSC)

The OSC pre-treatment method, which could remove the influence caused by noise (Gao and Ren 2010, Niazi and Goodarzi 2008, Wold et al. 1998), was proposed by Wold et al. (1998). The principle of this method is similar to that of the partial least squares (PLS) method. However, in PLS, one adopts the score vector that has the greatest variance with the response variable vector as the first score vector, which means using the internal component containing the most information to construct a relationship between the explanatory variables and the response variable. Conversely, in OSC, the score vector that has the smallest variance with the response variable \(y\) is chosen as the first score vector. This score vector is made orthogonal to \(y\). Next, explanatory variables \(X\) can be made orthogonal to \(y\) by subtracting the product of the score vector and the loading vector from the original data. This is a complete OSC adjustment which removes noise, i.e. the information unrelated to \(y\), from the data (Wold et al. 1998). The OSC employs several algorithms, and the literature has recognized JSosc as the most reliable one (Svensson et al. 2002).

i. The JSosc algorithm

JSosc is one of the OSC pre-treatment methods proposed by Sjöblom et al. (1998), and its algorithm is displayed as follows.

Step 1. Standardize the original explanatory
variables $X$ and the response variable $y$. Using the non-linear iterative partial least squares (NIPALS) algorithm of the PCA, the first score vector $t$ of the explanatory variables $X$ can be obtained. The $t$ can then be centered based on $\bar{t}$.

$$t = t - \bar{t}$$

Step 2. Make the score vector $t$ orthogonal to the response variable $y$ ($y$ is centered).

$$t_\perp = t - \frac{y^T t}{y^T y} y$$

Add $t$, mean of the score vector $t$, to $t_\perp$.

$$t^* = t_\perp + \bar{t}$$

Step 3. Calculate the weighted vector $w$.

$$w = t^* X / (t^* t^*)$$

Normalize $w$ to get the standardized weighted vector.

$$w = w / \sqrt{w^T w}$$

Step 4. Calculate a new score vector $t$ using the explanatory variable $X$ and the standardized weighted vector $w$.

$$t = X w$$

Repeat the above steps until the obtained score vector $t$ becomes stable. If the score vectors stabilize, one can perform the next step. Otherwise, the obtained score vector $t$ is put into Step 2 until stable score vectors are derived.

Step 5. Build a PLS model with the explanatory variable $X$ against the stable score vector $t$ as generated in Steps 1 to 4 based on the 15 internal components of the PLS, one can obtain the weighted matrix $W$, loading matrix $P$, and score vector $\hat{q}$, which can then be used to calculate $w^*$.

$$w^* = \hat{b} = W (P W)^{-1} \hat{q}$$

However, both Bertran et al. (2001) and Svensson et al. (2002) reported that the calculation in Step 5, which used the 15 internal components of PLS for calculation, could lead to problems of overfitting in the model. Therefore, they recommended using cross-validation to determine the number of internal components to be used.

Step 6. Calculate $t^{**}$ using explanatory variables $X$ and $w^*$.

$$t^{**} = X w^*$$

Step 7. Calculate the loading vector $p$ of $X$.

$$p = X t^{**} / (t^{**} t^{**})$$

Step 8. Complete one OSC adjustment by subtracting the orthogonal parts from the explanatory variable $X$.

$$X_{osc} = X - \bar{t} p$$

Step 9. Repeat the process of placing the $X_{osc}$ derived from Step 8 into Step 1 to perform several repetitions of OSC pre-treatment. Hence, the OSC (orthogonal signal correction) can be shown as following:

$$X_{osc} = X - \sum_{i=1}^{n} t_i p_i = X - TP$$

In the above equations, $X$ denotes the original NIRS data, $n$ means the number of OSC components, $t_i$ means the $i$th OSC score vector, $p_i$ represents the $i$th OSC loading vector, and $X_{osc}$ is the NIRS data corrected by OSC pre-treatment.

When using the OSC to adjust for a new data set to serve as the validation set, one first uses $\bar{X}_{cal}$, the mean of the calibration set data, to center data in the validation set ($\bar{X}_{cal}$) to obtain the centered data of the validation set ($\bar{X}_{cal-center}$). These data are multiplied by the weighted vector $w^*$ derived from Step 5 to obtain new score vector $t_{val}$. Then, the loading vector $p$ generated in Step 7 and the $t_{val}$ can be used to adjust the data of the validation set.

The OSC algorithm for the validation set is as follows:

Step 10. Center the data of the validation set.

$$\bar{X}_{val-center} = X_{val} - \bar{X}_{cal}$$

Step 11. Calculate new score vectors $t_{val}$ for the validation set.

$$t_{val} = X_{val-center} \times w^*$$

Step 12. Subtract the product of the new score vectors $t_{val}$ obtained from Step 11 and the loading vector $p$ obtained from Step 7 from the data of the centered validation set to complete one OSC adjustment.
\[ \mathbf{X}_{\text{Val-OSC}} = \mathbf{X}_{\text{Val-center}} - \mathbf{t}_{\text{Val}} \mathbf{p} \]

**ii. The OSC-CV algorithm**

Svensson et al. (2002) indicated that when adopting the OSC pre-treatment method for correcting data, one should not only decrease the number of internal components but also reduce the complexity of the calibration model, so that the model can be easily interpreted. Therefore, from this perspective, the present study adopted JSosc as the pre-treatment method, and used PLSR to construct the calibration model. In JSosc, the number of internal components was set between 1 and 25, and the number of latent variables used in PLSR for constructing models was set between 1 and 20, yielding 25 × 20 combinations. Each combination was performed with leave-one-out cross-validation (LOOCV) to conduct JSosc pre-treatment. PLSR was then used to construct the calibration model. Next, the validation set, which was reserved during cross-validation, was incorporated into the calibration model to calculate the error. After completing the calculation of the 25 × 20 combinations as stated above, one can obtain an error matrix containing 25 × 20 errors. Finally, the number of PLS internal components in JSosc corresponding to the smallest error found in the error matrix was selected as the optimal number of internal components. The number of adjustments depended on the root mean square error of cross validation (RMSECV) of the calibration models derived from each adjustment. The number of adjustments corresponding to the model with the smallest RMSECV was adopted as the optimal number of adjustments. This approach was called OSC-CV in this manuscript.

### 4. Performance Measures

This study adopted 5 statistical indicators to evaluate the fitting and predictive abilities of the constructed models, including RMSECV, RMSEP, RMSEC, \( R^2 \), and \( Q^2 \). The descriptions of these indicators are introduced in the following.

The RMSECV assesses the calibration models using the LOOCV approach. A smaller RMSECV value indicates a smaller difference between the predicted value obtained from the model and the observed value, suggesting a better predictive ability. The equation is shown below [Eq. (4)].

\[
RMSECV = \sqrt{\frac{\sum_{i=1}^{N_c} (y_{i(o)} - \hat{y}_{i(o)})^2}{N_c}} \tag{4}
\]

where \( y_{i(o)} - \hat{y}_{i(o)} \) denotes the difference between the original observed value and the predicted value, respectively, obtained using the fitted calibration model of the \( i \)th sample that is deleted. \( N_c \) denotes the sample size of the calibration set.

RMSEC is an internal validation method which examines whether the model constructed from calibration data sets correctly predicts the calibration data. A smaller RMSEC value suggests a smaller difference between the predicted value and the observed value. The equation for RMSEC is shown as follows [Eq. (5)].

\[
RMSEC = \sqrt{\frac{\sum_{i=1}^{N_c} (y_{i(o)} - \hat{y}_{i(o)})^2}{N_c}} \tag{5}
\]

where \( \hat{y}_{i(o)} \) denotes the predicted value of the \( i \)th sample of the calibration set, \( y_{i(o)} \) denotes the observed value of the \( i \)th sample of the calibration set, and \( N_c \) is the sample size of the calibration set.

RMSEP is also an external validation method which evaluates whether the model constructed from calibration data sets correctly predicts the validation data. The equation is shown below [Eq. (6)].

\[
RMSEP = \sqrt{\frac{\sum_{i=1}^{N_v} (y_{i} - \hat{y}_{i})^2}{N_v}} \tag{6}
\]

where \( \hat{y}_{i} \) denotes the predicted value of the \( i \)th sample of the calibration set, \( y_{i} \) denotes the observed value of the \( i \)th sample of the validation set, and \( N_v \) is the sample size of observed values of the validation set.

\( R^2 \) indicates the percentage of explainable variance; where a \( R^2 \) value closer to 1 indicates a better model fitting ability (Golbraikh and Tropsha 2002). Below is the equation for calculating \( R^2 \) [Eq. (7)].

\[
R^2 = \frac{\sum_{i=1}^{N_v} (\hat{y}_{i} - \bar{y})^2}{\sum_{i=1}^{N_v} (y_{i} - \bar{y})^2} = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{7}
\]
where \( \hat{y}_i \) denotes the estimate of the \( i \)th sample, \( y_i \) denotes the actual value of the \( i \)th sample, and \( \bar{y} \) is the mean. SSR refers to the regression sum of squares, SSTO refers to the total sum of squares, and SSE means the error sum of squares. In actual practice, the coefficient of determination for the calibration set is shown as \( R^2_{\text{cal}} \), and \( N = N_C \). The coefficient of determination for the validation set is shown as \( R^2_{\text{val}} \), and \( N = N_V \) (Kutner et al. 2004).

\( Q^2 \) based on LOOCV method, may also be used to evaluate model predictive ability. A \( Q^2 \) value closer to 1 indicates better predictive ability of the model (Golbraikh and Tropsha, 2002).

\[
Q^2 = 1 - \frac{\sum_{i=1}^{N_C} (y_{i} - \hat{y}_{i})^2}{\sum_{i=1}^{N_C} (y_i - \bar{y})^2}
\]

(8)

where \( (y_{i} - \hat{y}_{i}) \) respectively denotes the difference between the original observed value and the predicted value, and was obtained using the fitted calibration model of the \( i \)th sample that is deleted. \( N_C \) is the sample size of the calibration set.

**RESULTS**

**Comparison between brown rice flour and brown rice grain**

Table 1 shows the results of descriptive analyses regarding the protein content of brown rice in the calibration and validation sets. As shown in Table 1, in both the calibration and validation sets, the waxy rice contained the highest level of protein content, and the japonica rice had the lowest level. In general, the protein content of brown rice was between 5.6899% and 12.3067%.

The PLSR was used to construct calibration models with the NIRS data of brown rice flour and the original spectral data of brown rice to predict protein content, and the results are shown in Table 2. As indicated, when compared to the calibration model constructed using brown rice flour, the model built from data on brown rice without pre-treatment demonstrated a poorer RMSEC, RMSEP, and RMSECV. Furthermore, 14 latent variables were required to construct a calibration model from the brown rice data, increasing the difficulty of the model interpretation. Conversely, for the brown rice flour data, the calibration model required only 7 latent variables.

Fig. 1 plots the absorbance spectral value of different levels of protein at different NIRS wavelengths. As shown, the dispersion among absorbance spectral values in brown rice flour was smaller than that for brown rice (grains). As the rice grains were not ground, the greater difference in particle size as compared to brown rice flour was assumed to be the source of noise. As a result, the absorbance spectral values for

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Indica</th>
<th>Japonica</th>
<th>Waxy</th>
<th>Indica</th>
<th>Japonica</th>
<th>Waxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. value (%)</td>
<td>6.9899</td>
<td>5.6899</td>
<td>6.9921</td>
<td>7.3185</td>
<td>5.8178</td>
<td>7.6971</td>
</tr>
<tr>
<td>Mean (%)</td>
<td>8.8000</td>
<td>7.6599</td>
<td>9.2147</td>
<td>8.7611</td>
<td>7.0397</td>
<td>8.9043</td>
</tr>
<tr>
<td>SD (%)</td>
<td>0.8659</td>
<td>1.1956</td>
<td>1.2604</td>
<td>0.7072</td>
<td>0.9010</td>
<td>1.0530</td>
</tr>
</tbody>
</table>

**Table 2. Modelling results using original spectral data of brown rice flour and brown rice.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>( LVs )</th>
<th>RMSEC</th>
<th>( R^2_{\text{cal}} )</th>
<th>RMSECV</th>
<th>( R^2_{\text{val}} )</th>
<th>RMSECV</th>
<th>( Q^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown rice flour</td>
<td>7</td>
<td>0.4616</td>
<td>0.8688</td>
<td>0.4298</td>
<td>0.8910</td>
<td>0.5017</td>
<td>0.8839</td>
</tr>
<tr>
<td>Brown rice grain</td>
<td>14</td>
<td>0.4854</td>
<td>0.8524</td>
<td>0.4909</td>
<td>0.8310</td>
<td>0.6223</td>
<td>0.8490</td>
</tr>
</tbody>
</table>

\( LVs \) denotes the number of latent variables.
brown rice could be affected by the multiplicative scatter effect, which influenced the corresponding calibration model. According to the results, predicting protein content using brown rice is ineffective. Hence, the following sections describe how pre-treatment methods were used to transform the brown rice. These methods not only helped remove noise and improved the calibration models, but also saved on time and costs associated with grinding.

**The performance of different pre-treatments**

Table 3 shows a performance comparison among calibration models for brown rice using 5 different pre-treatment strategies. It was found that transformed data using the normalization method did not significantly improve the performance and predictive ability of the model. A possible reason was that the absorbance spectral values of the brown rice samples were not influenced by a fixed constant. In addition, compared to the model using data without pre-treatment, the models constructed from data with the SNV and MSC methods did not demonstrate significant improvement in terms of RMSEC and RMSEP. Regarding orthogonal signal correction (OSC), compared to the model using data without pre-treatment, the calibration model constructed from data treated with the JSoC pre-treatment required significantly fewer optimal latent variables. Therefore, the complexity of the model was reduced, facilitating the interpretability of the model.

Furthermore, the JSoC pre-treatment largely improved the performance of the model in terms of its $R^2_{c}$, RMSEC, and RMSECV. However, the RMSEP was 0.6955, which was not as good as expected. A possible reason for this was that the model incorporated too much information of the calibration data set and led to problems of over-fitting. Compared to the data without pre-treatment and that with JSoC pre-treatment, the data managed with the OSC-CV method also largely decreased the required number of optimal latent variables, reduced the complexity of the model and facilitating interpretability. Moreover, compared to the JSoC method, the calibration model handled with the OSC-CV method did not have problem of over-fitting, and the RMSECV of the model could be significantly reduced.

**Comparison between JSoC and OSC-CV**

To compare the appropriateness between the JSoC and OSC-CV methods in correcting the NIRS data of the brown rice, more thorough examinations were carried out. Table 4 and Table 5 show the model fitting ability according to the number of adjustments in the JSoC and OSC-CV methods, respectively. “One JSoC component”
and “one OSC-CV component” indicate performing one time of adjustment based on the JSosc and OSC-CV methods, respectively (Table 4 to 5). As shown in Table 4, every additional JSosc adjustment decreased RMSEC and RMSECV while increasing RMSEP, suggesting a probable data over-fitting. However, in OSC-CV, the RMSEC, RMSEP, and RMSECV remained stable regardless of the number of adjustments, and none of these indicators increased unexpectedly (Table 5). This finding makes the authors value the OSC-CV over the JSosc. In addition, as both the JSosc and OSC-CV could be influenced by the number of adjustments, we recommend that the number of adjustments are determined based on RMSECV to avoid an over-fitting by the excessive adjustments. When the $n^{th}$ adjustment leads to the smallest RMSECV of the model, $n$ could be adopted as the optimal number of adjustments. Taking Table 4 and Table 5 as examples, the best number of adjustments for JSosc and OSC-CV were 3 and 2, respectively, as they corresponded to the smallest RMSECV (0.3808 and 0.4978).

To summarize the above results, though the data transformed with OSC-CV did not produce better RMSEC and RMSECV than that of the data transformed with JSosc, OSC-CV did not lead to over-fitting, as the JSosc method did. As a result, OSC-CV is recommended as the best pre-treatment method for NIRS data of brown rice grain.

Table 3. Modelling results of brown rice obtained by five different pre-treatment methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Lvs</th>
<th>RMSEC</th>
<th>R2Cal</th>
<th>RMSEP</th>
<th>R2Val</th>
<th>RMSECV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalization</td>
<td>11</td>
<td>0.5318</td>
<td>0.8228</td>
<td>0.4827</td>
<td>0.8389</td>
<td>0.6384</td>
</tr>
<tr>
<td>SNV</td>
<td>11</td>
<td>0.5146</td>
<td>0.8341</td>
<td>0.5094</td>
<td>0.8235</td>
<td>0.6260</td>
</tr>
<tr>
<td>MSC</td>
<td>9</td>
<td>0.5355</td>
<td>0.8204</td>
<td>0.5122</td>
<td>0.8200</td>
<td>0.6304</td>
</tr>
<tr>
<td>JSosc</td>
<td>1</td>
<td>0.3763</td>
<td>0.9113</td>
<td>0.6955</td>
<td>0.3808</td>
<td></td>
</tr>
<tr>
<td>OSC-CV</td>
<td>1</td>
<td>0.4922</td>
<td>0.8482</td>
<td>0.4959</td>
<td>0.8286</td>
<td>0.4978</td>
</tr>
</tbody>
</table>

$Lvs$ denotes the number of latent variables.

Table 4. Modelling results obtained by applying the JSosc pre-treatment with different number of components.

<table>
<thead>
<tr>
<th>Lvs</th>
<th>One JSosc component</th>
<th>Two JSosc component</th>
<th>Three JSosc component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSEC</td>
<td>RMSEP</td>
<td>RMSECV</td>
</tr>
<tr>
<td>1</td>
<td>0.4929</td>
<td>0.4477</td>
<td>0.5032</td>
</tr>
<tr>
<td>2</td>
<td>0.4683</td>
<td>0.4904</td>
<td>0.4797</td>
</tr>
<tr>
<td></td>
<td>0.3724</td>
<td>0.5413</td>
<td>0.4010</td>
</tr>
<tr>
<td>4</td>
<td>0.3720</td>
<td>0.5397</td>
<td>0.4003</td>
</tr>
</tbody>
</table>

$Lvs$ denotes the number of latent variables.

Table 5. Modelling results obtained by applying the OSC-CV pre-treatment with different number of components.

<table>
<thead>
<tr>
<th>Lvs</th>
<th>One OSC-CV component</th>
<th>Two OSC-CV components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSEC</td>
<td>RMSEP</td>
</tr>
<tr>
<td>1</td>
<td>0.5147</td>
<td>0.4396</td>
</tr>
<tr>
<td>2</td>
<td>0.4905</td>
<td>0.4779</td>
</tr>
</tbody>
</table>

$Lvs$ denotes the number of latent variables.
Comparison between original and OSC-CV transformed data

The results above have concluded that OSC-CV is the best pre-treatment method for the NIRS data of brown rice. The following displays a comparison of OSC-CV transformed brown rice and the original brown rice NIRS data. Fig. 2 shows the scatter plots containing score vectors of the first and the second latent variables of the PLSR-based calibration models using the original NIRS data of brown rice and the data pre-treated with OSC-CV, respectively. As shown in Fig. 2(a), spectral variation existed in the waxy rice subspecies, which separated its distribution from those of the indica and japonica rice subspecies. Nevertheless, the variations attributed to rice subspecies disappeared in the data transformed with OSC-CV method, which therefore improved fitting ability of the calibration model [Fig. 2(b)]. Fig. 3 displays the scatter plot between the score vectors of the explanatory variable (X-score) and the response variable (y-score) of the first latent variable in the PLSR-based calibration model. As shown in Fig. 3(a), the distribution of the data without pre-treatment was disorderly. Conversely, the scatter plot [Fig. 3(b)] of the NIRS data transformed with OSC-CV showed a linear relationship, revealing excellent fitting ability. Fig. 4 shows the confidence interval for the amount of the protein content obtained from the PLSR calibration model using the NIRS data transformed with OSC-CV. The results indicate that the PLSR-based calibration model based on the NIRS data transformed with OSC-CV indeed has the excellent ability in fitting and predicting.

Fig. 5 shows the scatter plot of PLSR regression coefficients at different wavelengths. A larger absolute value for the regression coefficient suggests a greater influence of its corresponding wavelength on the calibration model. As shown in Fig. 5(a), given the potential noise in the raw data, great variations among the regression coefficients existed; therefore, the influence of individual wavelengths could not be easily identified. However, the variations in regression coefficients based on the data transformed with OSC-CV method were less dramatic [Fig. 5(b)]. As wave crests at 1208 nm, 1404 nm, 1648 nm, 1844 nm, 2024 nm, and 2232 nm corresponded to greater absolute values for the regression coefficient, these wavebands posed larger influences on the model. This finding was similar to that in the literature, where the wave crests for absorbance in protein were at 1187 nm, 1422 nm, 1696 nm, 1848 nm, 2055 nm, and 2265 nm (Li and Shaw 1996, Williams and Norris 1990).

Comparison of the NIRS data between brown rice flour and OSC-CV transformation of brown rice in the whole data set

As the above tests had recognized OSC-CV as the best pre-treatment strategy for brown rice, the following shows comparisons of

(a) Original

(b) OSC-CV

Fig. 2. The scatter plot of PC 2 X-score vs. PC 1 X-score: (a) original brown rice spectral data; and (b) brown rice spectral data with OSC-CV pre-treatment.
Using OSC to Correct NIRS for Determining of Protein Content

Fig. 3. The scatter plot for PC 1 Y-score vs. PC1 X-score: (a) original brown rice spectral data; and (b) brown rice spectral data with OSC-CV pre-treatment.

Fig. 4. The scatter plot with confidence interval for the amount of the protein content obtained from the PLSR model using OSC-CV for the NIRS. (a) calibration set; (b) validation set.

Fig. 5. Regression coefficient plots for the PLSR model: (a) original brown spectral rice data; and (b) brown rice spectral data with OSC-CV pre-treatment.
models constructed using the OSC-CV transformation of brown rice, and the brown rice flour in the whole data set. Two outlier detection methods, the Studentized deleted residuals (Kutner et al., 2004) and the Hat matrix leverage values (Kutner et al., 2004), identified some outliers in the data; therefore, we removed these outliers and used the remaining 199 samples to construct models (Table 6). As shown in Table 6, the RMSECV of the brown rice data transformed with OSC-CV method was better than that of brown rice flour data; the other statistical indicators (RMSEC, $R^2_{\text{Cal}}$, and $Q^2$) between two data sets were similar. Therefore, the fitting ability of the model was also similar between models constructed using these two data sets. It is worth noting that OSC-CV transformed data required only 1 latent variable to construct the model, while the data of brown rice flour required 7 latent variables. As a result, the model constructed using the data transformed with OSC-CV method could be easily interpreted. According to these results, we recommend that a calibration model can be constructed with the data of brown rice transformed with OSC-CV when studying the protein content of brown rice. This approach not only creates a model similar to that obtained using NIRS data of brown rice flour, but also saves the time required for grinding and reduces costs.

**DISCUSSION**

To predict the protein content of brown rice, this study adopted the Normalization, SNV, MSC, JSosc, and OSC-CV pre-treatment methods to transform the NIRS data of brown rice and constructed the calibration models using PLSR. According to the results, both JSosc and OSC-CV significantly improved the RMSECV of the PLSR models, and lowered the required number of latent variables for PLSR. However, as JSosc could cause problem of over-fitting, this study recognizes OSC-CV as the best pre-treatment method for transforming the NIRS data of brown rice.

When conducting thorough examinations of OSC-CV, the cross-validation of the NIRS data performed with the JSosc and OSC-CV adjustments obtained results consistent with the literature (Svensson et al. 2002). The fact indicated that the JSosc strategy could cause problem of over-fitting because of the number of PLS internal components and the number of adjustments. However, OSC-CV was relatively stable, as the RMSEC, RMSEP, and RMSECV of the model did not increase unexpectedly when the number of adjustments increased. This study further recommends to use RMSECV to determine the number of adjustments for the JSosc and OSC-CV to avoid the over-fitting issue. In addition, OSC-CV minimized variations due to different rice subspecies and shortened the distance among the first and the second score vectors of the PLSR explanatory variables. In OSC-CV, the score vectors of the PLSR within response variables and explanatory variables formed an excellent linear relationship, and it required only 1 latent variable to create an appropriate PLSR calibration model.

The results of comparisons among pre-treatment methods are consistent with those of the literature (Sjöblom et al. 1998). The OSC-CV method is appropriate in transforming NIRS data to reduce variations in the data attributed to different rice groups. In addition, the model constructed using the NIRS brown rice data pre-treated with OSC-CV was similar to models constructed using data from brown rice flour, and data transformed with OSC-CV required only 1 latent variable to create a PLSR model. Therefore, it is recommended that when analyzing protein

<table>
<thead>
<tr>
<th>Method</th>
<th>$N$</th>
<th>$L_{LVs}$</th>
<th>RMSEC</th>
<th>$R^2_{\text{Cal}}$</th>
<th>RMSECV</th>
<th>$Q^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown rice flour</td>
<td>199</td>
<td>7</td>
<td>0.4209</td>
<td>0.8857</td>
<td>0.4464</td>
<td>0.8713</td>
</tr>
<tr>
<td>OSC-CV transformed brown rice grain</td>
<td>199</td>
<td>1</td>
<td>0.4224</td>
<td>0.8712</td>
<td>0.4255</td>
<td>0.8693</td>
</tr>
</tbody>
</table>

$L_{LVs}$ denotes the number of latent variables.
content of brown rice, OSC-CV should be adopted to correct the NIRS data to remove irrelevant information. This facilitates the PLSR-based calibration model constructed for protein content of brown rice that has the stable fitting ability and the minimal latent variables and possesses good interpretability.

REFERENCES


