Rapid non-destructive analysis of lignin using NIR spectroscopy and chemo-metrics

Xin Wu1,2 | Guanglin Li1 | Xuwen Liu1 | Fengyun He1

1College of Engineering and Technology, Southwest University, Chongqing, China
2Chongqing College of Electronic Engineering, Chongqing, China

Abstract
Lignin plays an important role in the formation of stone cells in pears. However, the accumulation of lignin had adverse impact on the flavor and quality of the fruit. A rapid and accurate method for measuring the lignin content of pears is therefore required.

An improved variables selection method called ‘the bootstrapping soft shrinkage combined with frequency and regression coefficient of variables (FRCBOSS)’ was therefore developed based on ‘the bootstrapping soft shrinkage (BOSS)’ technique, to identify the characteristic wavelengths of near-infrared (NIR) spectra for non-destructive and rapid analysis of lignin. Sub-models were generated by weighted bootstrap sampling (WBS) in the FRCBOSS method. For the BOSS method, the new weights of variables were determined only as the absolute values of regression coefficients of variables in each iteration. In contrast, the FRCBOSS method also considers the frequency of variables in variable space. Moreover, the FRCBOSS algorithm overcomes the disadvantage of BOSS in selecting variables, which could incorporate useful wavelengths that would otherwise be removed by the BOSS method. In addition, a range of different pre-treatment methods were used for comparison in the detection of lignin in the Snow pears. These include Savitzky–Golay Smoothing (SG), Normalization (NORM), Standard Normal Variate (SNV), and 1st Derivative (D1), as well as a combination of these methods and the different variables selection method (SiPLS, SiPLS-SPA, SiPLS-CARS, SiPLS-BOSS, and SiPLS-FRCBOSS).

The number of variables selected by FRCBOSS was a little larger than that selected by BOSS. The partial least square regression (PLSR) model based on the 19 variables selected by SiPLS-FRCBOSS method had the best prediction ability, with a $R_p$ value of 0.880 and a RMSEP value of 1.004%. We conclude that NIR diffuse reflectance spectroscopy technology combined with FRC-BOSS is an accurate and useful tool for the non-destructive and rapid determination of pear lignin contents.

KEYWORDS
FRCBOSS, lignin content, NIR spectroscopy, pear, variables selection
1 | INTRODUCTION

Snow pear, a common and popular fruit in China, has been demonstrated as traditional remedies for relieving respiratory symptoms, constipation and alcoholism in traditional Chinese medicine (TCM) (Li et al., 2012) for over 2000 years. Near-Infrared (NIR) spectroscopy (Ibañez et al., 2019; Song et al., 2019; Zhang et al., 2018), combined with processing analytical technology have proven to be a useful and non-destructive spectral techniques (Ibañez et al., 2019; Ozdemir et al., 2019) for rapid quantitative analysis of internal chemical components in fruits and food (Xia et al., 2018).

Studies have shown that stone cell content plays a critical role in determining fruit quality of pears, but not the case in other fruits. More specifically, higher concentration and larger size of stone cells lead to thicker pulp and deterioration of the taste. Although genetic variation of the present stone cells is a key factor (Tao et al., 2009) that modulates the stone cell content, it has also been found that Lignin content plays a direct, positive role to modulate the size and content of stone cells (Cai et al., 2010; Xue et al., 2019) in snow pears. The more lignin (Xue et al., 2020) is synthesized, the more stone cells are formed. Accumulation of stone cell mass increases the pulp density, leading to the worsening of the taste. Therefore, the lignin content can be used as a distinguishing indicator for internal quality of ‘Snow’ pear.

In recent research, Lignin can be extracted, purified, and analyzed by ultraviolet (UV) spectroscopy, Fourier-transform infrared (FTIR) spectroscopy and 1H nuclear magnetic resonance (1H NMR) spectroscopy (Cai et al., 2010). The lignin is then dissolved in 25% solution of acetyl bromide in glacial acetic acid by the acetyl bromide soluble lignin (ABSL) method (Fukushima et al., 2000), and was read and analyzed at 280 nm. Ultraviolet (UV) Spectrophotometry is a simple and inexpensive method for the determination of lignin. Ionization Differential Ultraviolet Spectrophotometry ($\Delta\varepsilon$ - Method) can be used to analyze lignin dissolved at neutral and alkaline conditions by using a UV spectrophotometer (Goldmann et al., 2017). Hydrogen thermolysis (HyPy) (Beramendi-Orosco et al., 2004) has been described as a potential rapid method for obtaining lignin enriched residues from samples, which can be used for simple estimation of lignin content. Lignin extracted from cordgrass, switchgrass, and corn stover can also be analyzed and characterized by using ethyl acetate-ethanol-water organic solvent pretreatment method (Cybulksa et al., 2012). Another analysis method of lignin is CuO oxidation (De la Cruz et al., 2015), lignin is oxidatively hydrolyzed into phenol monomers, which are quantified by high-performance liquid chromatography or gas chromatography–mass spectrometry. Klason method is another classical method for analyzing lignin content (Assis et al., 2017; Bunzel et al., 2011). The above traditional measurement of lignin content is a chemical method, and a large number of chemical reagents such as H$_2$SO$_4$, HCL and other acid reagents are required during the process, which requires extra safety precautions and hazard waste treatment in the measurement process. These methods generally can be highly accurate but are time-consuming, and it is impractical to analyze a large sample size of lignin content quickly. Therefore in this study, we propose the use of feature wavelength analysis as a non-destructive and in particular, a rapid analysis of lignin content in ‘Snow’ pears by using NIR spectroscopy.

NIR spectroscopy has become a simple, rapid, and non-destructive analytical technique (Assis et al., 2017; Fan et al., 2016; Lee & Han, 2016), to measure the chemical composition, physical and chemical properties of samples effectively, and has been widely used in detection of fruits, vegetables, meat, and other foods. Recently, much attention has been paid on the rapid detection of soluble solid content (SSC), total acid (TA), firmness and other internal quality of fruits, such as apples, pears and watermelons, etc., by using NIR spectroscopy combined with chemometric methods has been reported. Camila Assisi’s team used NIR Spectroscopy with the range of 10,000–4000 cm$^{-1}$, Ordered Predictors Selection (OPS), and Partial Least Squares (PLS) to build the multivariate calibration models for estimating the lignin content in different parts of sugarcanes (Assis et al., 2017). And the OPS algorithm selected fewer variables, but achieved greater predictive capacity, with the values of RMSEP, Rp, and RPD for the middle stalk is 0.61, 0.95, and 3.24, respectively. An inexpensive and portable NIR spectrometer is employed to rapidly determine the lignin in wood combined with multiplicative scatter correction (MSC), particle swarm optimization (PSO) and the kernel extreme learning machine (KELM) algorithm by Hao Yang’s team (Yang et al., 2019). There have been extensive studies on the determination of lignin in Rice Straw (Hu et al., 2018), Fine Root (Elle et al., 2019), Pinus radiate wood (Fahey et al., 2018), Pinus pinaster (Pinb) wood (Alves et al., 2020), Cryptomeria japonica (Horikawa et al., 2019), pulp wood feedstock (Liang et al., 2020), flax fiber (Huang & Yu, 2019), etc., by using NIR spectroscopy combined with variable selection algorithms and modeling approaches. However, there are much fewer studies, and in fact, we only found one such case about the determination of lignin content in fruits. Xiaohui sheng’s team used NIR spectroscopy and the uninformative variable elimination (UVE) algorithm to non-destructively detect lignin content in Korla fragrant pear (Sheng et al., 2020). After the SNV pretreatment, the SEP of the UVE-PLS model was 1.36%, $R_p^2$ was 0.87, and RPD was 2.03. In this study, more and deeper methods and algorithms will be applied to determination of lignin content in snow pears by using NIR spectroscopy.
In general, a large number of spectra data could be acquired by a near infrared spectrometer, which have serious collinearity problems, and inevitably contain the interference and useless information. If the partial least square regression (PLSR) model is established based on the full wavelength, it will not only increase the complexity, but even reduce the predictive performance of the model. Therefore, it is important to identify effective variables before the PLSR model established.

In order to improve the predictive ability and efficiency of the PLSR model, a series of variable selection methods are studied and classified into interval selection and individual selection. Synergy interval partial least squares (SiPLS) (Li et al., 2018), interval partial least square (iPLS) (Liu et al., 2019), moving window partial least squares (MWPLS) (Li et al., 2009), changeable size moving window partial least squares (CSMWPLS) (Luo et al., 2018), backward interval partial least square (BiPLS) (Assis et al., 2017), interval combination optimization (ICO), etc., which belong to the interval selection methods. The collinearity problem among wavelength variables in an interval is not considered in this method. Generally, the selected variables are concentrated in several intervals, and the variables need to be further simplified. Some typical individual selection methods include Monte Carlo non-information variable elimination (MCUVE) (Li et al., 2014; Yan et al., 2019), competitive adaptive reweighted sampling (CARS) (Bai et al., 2019; Jiang et al., 2015; Yan et al., 2019), genetic algorithm (GA) (Du et al., 2019), successive projections algorithm (SPA) (Li et al., 2014; Liu et al., 2017) and bootstrapping soft shrinkage (BOSS) (Deng et al., 2016). The MCUVE selects the useful variables according to the stability of variables. The number of variables selected is usually large and the prediction result is poor by the MCUVE method. CARS (Li et al., 2009) algorithm will remove variables with small absolute value of regression coefficient by force. However, the absolute value of regression coefficient can be affected by the change of the sample space, resulting in the possible elimination of useful variables. GA algorithm has a high risk of over fitting, which will produce local optimal solution, resulting in low computational efficiency and poor prediction performance. SPA algorithm can reduce the collinearity problem among variables extent, but the selected variables contain less useful variables or even interference variables. BOSS algorithm is a wavelength selection method based on variable space, which selects feature wavelength according to the absolute value of regression coefficient in variable space by using weighted bootstrap sampling (WBS). But frequency of variables as be an importance feature is ignored and the selected variable may not be optimal in the BOSS algorithm.

On account of the existing problem above the wavelength selection methods, a bootstrapping soft shrinkage combined with frequency and regression coefficient of variables (FRCBOSS) approach is proposed in this paper. FRCBOSS algorithm inherits the advantages of the weighted bootstrapping sampling (WBS) in the BOSS method. The frequency and regression coefficient of variables are chosen as evaluation indexes, and the optimal variable can be selected. In this study, the PLSR model will be established based on the feature wavelength selected by the FRCBOSS method for rapid and non-destructive analysis of lignin content in ‘Snow’ pears. Therefore, the research processes of this work are as follows: (1) A Fourier-Transform Near-infrared (FT-NIR) spectrometer is used to acquire NIR spectra of samples; (2) The pears are grounded into dry powder and then the lignin content reference value is measured by the Klason method; (3) Savitzky–Golay smoothing (SG), Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), First Derivative (D1) and their combination are applied in the NIR spectra preprocessing; (4) SiPLS, CARS, SPA, BOSS, FRCBOSS, and combination are used to select the useful wavelength, PLSR models for predicting lignin content in ‘Snow’ pears are established and compared; (5) The PLSR model is established based on 19 variables selected by the improved variables selection method FRCBOSS, which has the best prediction ability of lignin content in ‘Snow’ pears.

2 MATERIALS AND METHODS

2.1 Sample description

In July 2020, a total of 195 ‘Snow’ pears were purchased from a reliable local fruit market in Chongqing, China. In order to ensure the accuracy of the experiment, pears were washed,
numbered, and then stored in an ice box (4°C, 75%RH) for 24 h before the experiment was carried out (Yuan et al., 2020). Figure1 showed that three measuring points (P1, P2, and P3) around the equator of pear were set as a 120° angular distance.

2.2 | Spectra acquisition

The NIR diffuse reflectance spectra of ‘Snow’ pears were acquired by a FT-NIR spectrometer (Nicolet iS50, ThermoFisher scientific, U.S.A). The spectrometer was consisted of a high sensitivity InGaAs detector, a near infrared light source (a 20 W tungsten lamp), a processor, a Michelson interferometer, a beam splitter, some optical fibers, and probes, with resolution of 4 cm⁻¹ was used to record NIR spectra in the range of 10,000–4000 cm⁻¹. The average reflectance spectra from three positions around the equator of the sample were acquired.

2.3 | Lignin content reference value measurement

When the spectra acquisition of each ‘Snow’ pear was completed, the lignin content reference value was measured immediately (see Table 1). The intact pear was made into dry powder firstly, as follows (Figure 2): (1) An aliquot of 200 g fresh pear flesh (1 cm outside the core to 2 mm under the pericarp) was obtained from one intact pear; (2) A hydraulic press was used to press the 200 g pear flesh for eliminating the juice and the residues was obtained; (3) The residues were dried at 50°C for 5 days and crushed by a miniature pulverizer (IkaA11, Germany); (4) The residues were separated on a 100-mesh sieve; (5) The above powder (50 g) after separating was extracted in water and ethanol for 10 h and 5 h, respectively, with Soxhlet extraction device; (6) The final dry pear powder was obtained after drying in a cabinet (65 °C).

Then, the lignin content was measured by using a traditional Klason method, as follows (Figure 3): (1) Dry pear powder (500 mg) was put into a 250 ml beaker, and 30 ml of 72% H2SO4 was added and then external agitation was applied until the mixed solution reached complete homogeneity; (2) The mixed solution was stirred and set at 30°C for 2 h; (3) The beaker with mixed solution was sampled in the boiling bath of water for 2 h; (4) The mixed solution was diluted with deionized water to reduce the sulfuric acid concentration and autoclaved at 120°C and 15 psi for 1 h; (5) The treated solution was poured into the sand core funnel (2.5 cm diameter, 1.6 μm particle retention, dried, and weighed), filtrated and washed acid free by distilled water during filtrating. (6) The funnel containing lignin and acid insoluble residue particles was put into a drying cabinet for 5 h (65°C) until the weight was constant. The compounds without structural biomass should be removed, which may interfere with the measurement of lignin content in the Klason method (Assis et al., 2017; Bunzel et al., 2011). Formula (1) shows the calculation method of lignin content in Klason.

\[ W_L = W_1 - W_2 \]

where \( W_L \) is the weight of lignin content in ‘Snow’ pears, the \( W_1 \) is the weight of the funnel containing lignin, the \( W_2 \) is weight of the sand core funnel.

The lignin content of 195 pears covered 67.75 to 84.57 mg/g. As shown in Figure 4, the concentration of the sample was approximately in accordance with the normal distribution. 146 pear samples were selected as calibration model samples, and the remaining 49 pear samples were used for prediction model samples by using Kennard stone (KS) algorithm. The lignin distribution of the calibration set and the prediction set were shown in Table 1.

2.4 | Spectra data pretreatment

The NIR diffuse reflectance spectra of ‘Snow’ pears existed baseline shifts, transformation and noise due to the different light propagation distance and shape of samples. The spectra were preprocessed before modeling and analysis. The following methods were applied in the NIR spectra preprocessing, in order to getting better pretreatment results (Yang et al., 2018): spectra preprocessed by Savitzky–Golay Smoothing (11 points), spectra preprocessed by Normalize (area), spectra preprocessed by Standard Normal Variate (SNV), spectra preprocessed by 1st Derivative (D1), spectra preprocessed by Savitzky–Golay Smoothing (11 points) combined with Standard Normal Variate (SNV), spectra preprocessed by Savitzky–Golay Smoothing (11 points) combined with

| Data set      | Component | No. of samples | Unit | Range      | Mean ± SD  | SEL  |
|---------------|-----------|----------------|------|------------|------------|------|
| All samples   | Lignin    | 195            | mg/g | 67.75–4.57 | 77.27 ± 3.02 | 0.121 |
| Calibration set |          | 146            |      | 67.75–84.57| 77.54 ± 2.90 | 0.240 |
| Prediction set |          | 49             |      | 75.67–84.22| 79.33 ± 2.19 | 0.312 |

Abbreviations: No. of Samples, Number of Samples; SD, standard deviation; SEL, standard error of laboratory.
Normalize (area), spectra preprocessed by Savitzky–Golay Smoothing (11 points) and Normalize (area) combined with 1st Derivative (D1).

A set of dataset consists of 195 samples of ‘Snow’ pears measured on a FT-NIR spectrometer. The average reflectance spectrum from three positions contains 3112 wavelength points at intervals of 4 cm⁻¹ with the range of 10,000–4000 cm⁻¹. The dataset was divided into a calibration set (146 samples) and a prediction set (49 samples) by using Kennard stone (KS) algorithm (refer Table 1), respectively.
2.5 | FRCBOSS algorithm

2.5.1 | Weighted bootstrap sampling (WBS)

Weighted bootstrap sampling (WBS) is similar to Bootstrap sampling (BSS) method. It is a statistical technique for random sampling with replacement, which hires different weights for different objects. In each trail, the objects with larger weights are much easier to be selected. In actual, the weight vector of the variable will be normalized, therefore, the value of weight of each variable is between 0 and 1.

The probability of each variable being selected is:

\[
P_i = \left(1 - \frac{\omega_i}{\sum \omega_i}\right)^R
\]

where \(n\) is the number of variables, \(\omega_i\) is the weight of the \(i\) variable, \(R\) is the number of variables retained for the previous iteration. A variable with a larger weight have a greater probability to be selected in the sampling process (refer Formula (2)). If the same with the number of objects \(R\) is the number of replacements, the ratio of objects been selected is approaching 0.632 when \(R\) goes to infinity. The number of selected variables will reduce gradually in each iteration, and the variables space will shrink softly.

2.5.2 | New weights by FRC

Assuming that the NIR spectra data matrix is \(X_{nxp}\), and \(n\) is number of samples in row, \(p\) is number of variables in columns. The vector \(y_{nx1}\) is the measured property of interest. According to beer's law, the relationship between spectral matrix and concentration vector is:

\[y_{nx1} = X_{nxp}\beta_{nxp} + e\]

where \(e\) is the random error vector, \(\beta\) is the regression coefficient vector. The \(N\) subsets are generated on variables space by using weighted bootstrap sampling (WBS), and the \(N\) PLS sub-models are established using subsets obtained. The value of root mean squared error of cross validation (RMSECV) between regression coefficient matrix \(\beta_{nxp}\) and \(N\) sub-models will be calculated in each iteration. The \(N\) optimal models are selected from the \(N\) models, and the subsets with lowest value of root mean squared error of cross validation (RMSECV) are obtained.

2.5.3 | FRC-BOSS wavelength selection method

The regression coefficient and frequency of the variable are used as evaluation indexes of weight of the variable in FRC-BOSS algorithm. The larger the weight, the larger probabilities the variable is selected in the next replacement. The iteration repeats until the number of variables in the subsets equals to 1.

Specific implementation steps of FRCBOSS as follows.

Step 1. \(N\) subsets are generated according to the weight of variables by using weighted bootstrap sampling (WBS), and the \(N\) PLS sub-models are calculated. The value of root mean squared error of cross validation (RMSECV) between regression coefficient matrix \(\beta_{nxp}\) and \(N\) sub-models will be calculated in each iteration. The \(N\) optimal models are selected from the \(N\) models, and the subsets with lowest value of root mean squared error of cross validation (RMSECV) are obtained.

Step 2. The root mean squared error of cross validation (RMSECV) of the \(N\) subsets are calculated by using partial least square (PLS) algorithm.

Step 3. \(N\) optimal models are selected in the \(N\) subsets, and the subsets with lowest value of root mean squared error of cross validation (RMSECV) are obtained.

Step 4. The weight of each variable is recalculated according to the above formula (5) in each iteration process.
Step 5. When the number of remaining variables \( p_1 > 1 \) in each iteration, return to step 1 for the next iteration, otherwise, execute step 6.

Step 6. The subset with lowest RMSECV during the iteration is chosen as the optimal variable set.

### 2.6 Model building and evaluation

In this paper, the correlation coefficient of calibration \((R_c)\), the root mean square error of cross validation \((\text{RMSECV})\), the correlation coefficient of prediction \((R_p)\) and the root mean square error of prediction \((\text{RMSEP})\) are chosen as indicators to evaluate the performance of the calibration and prediction models (Tian et al., 2018). \( R_c \) and \( \text{RMSECV} \) are defined as follows:

\[
R_c = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_{i,a} - y_{i,p})^2}{\sum_{i=1}^{n} (y_{i,a} - y_{i,cm})^2}}
\]

\[
\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,a} - y_{i,p})^2}{n}}
\]

where \( n \) is the number of samples in the calibration set, \( y_{i,a} \) is the measured value of the \( i \) sample in the calibration set, \( y_{i,p} \) is the predicted value of \( i \) sample in the calibration set, \( y_{i,cm} \) is the mean of the measured value in the calibration set.

\( R_p \) and \( \text{RMSEP} \) are defined as follows:

\[
R_p = \sqrt{1 - \frac{\sum_{j=1}^{m} (y_{j,a} - y_{j,p})^2}{\sum_{j=1}^{m} (y_{j,a} - y_{j,pm})^2}}
\]

\[
\text{RMSEP} = \sqrt{\frac{\sum_{j=1}^{m} (y_{j,a} - y_{j,p})^2}{m}}
\]

where \( m \) is the number of samples in the prediction set, \( y_{j,a} \) is the measured value of the \( j \) sample in the prediction set, \( y_{j,p} \) is the predicted value of \( j \) sample in the prediction set, \( y_{j,pm} \) is the mean of the measured value in the prediction set.

After preprocessing, the SiPLS, CARS, SPA, BOSS, and FRCBOSS and their combinations were used to select the useful variables, and PLSR models for predicting lignin content in ‘Snow’ pears were established and compared. The evaluation indexes of models included the coefficient of determination of calibration \((R_c)\), the root mean square error of cross validation \((\text{RMSECV})\), the coefficient of determination of prediction \((R_p)\) and the root mean square error of prediction \((\text{RMSEP})\).

### RESULTS AND DISCUSSION

#### 3.1 Effect of different pre-processing on the PLS model for lignin

Figure 5 shows the raw NIR diffuse reflectance spectra of all pear samples, and the abscissa is wavenumber with range of 10,000–4000 cm\(^{-1}\) and the ordinate is absorbance. It can be seen that the trends of NIR spectra were similar, and mainly included the frequency-doubling and frequency-combining information of C-H, O-H and other chemical bond stretching. The two strong absorption peaks around 5155 and 6944 cm\(^{-1}\) are due to the third overtone of the O-H. There is a small absorption peak around 8403 cm\(^{-1}\), which is relevant to the third overtone of the C-H functional group (Ma et al., 2018). These NIR spectra of pear samples not only contain the hidden information of different components of fruit, but also include baseline shifts, transformation, and noise. Therefore, the raw NIR spectra need to be preprocessed before modeling and analysis.

PLS models are established using full spectrum data of calibration sets coupled with corresponding reference values for lignin in pears. The Unscrambler X version 10.4 software is applied to preprocess the raw NIR spectra by different methods. As shown in Figure 6, the characteristics of the spectrum are more obvious after preprocessing based on the five different methods. SG and NORM preprocessing are used to remove both the random noise and the adverse effects in spectra. SG can improve the signal-to-noise ratio of the sample signal, but fail to eliminate the additive effect. NORM eliminate the adverse effects caused by too large-scale differences of data.

Effect of different pretreatments on the full-spectrum PLSR models based on the raw spectra and are provided in Table 2. The results show the performance of the PLSR model was improved by using different pretreatments. For the prediction of lignin, the preprocessed spectra show lower RMSECV (ranged from 1.730% to 1.893%), higher \( R_c \) (ranged from 0.760% to 0.806%). NORM spectra get better calibration statistics than SG spectra, and the combination of NORM and SG preprocessing can further improve the accuracy and robustness of full-spectrum PLSR model. Notably, the NORM and SG spectra provided the stronger calibration with the lower RMSECV. Therefore, the PLSR model based on spectra data by SG and NORM methods is optimal model after comparison in Table 2 (Bold font with *).

#### 3.2 Optimization of PLS models by variable selection

In this study, Synergy interval Partial Least Squares (SiPLS) (Jiang et al., 2012) is an optimization algorithm of interval
partial least squares (iPLS) and firstly used to select effective regions from full wavelength in order to simplify the PLSR model. In SiPLS variable selection methods, the full wavelengths is divided into 19 subintervals, and the 2, 9, 17 and 19 synergy subintervals are selected as the effective regions, with 655 variables in total (see Figure 7 and Figure 8). Then CARS is applied based on the variables selected by SiPLS (SiPLS-CARS), the sampling times is set to 70, and the 10-folds cross validation are used to calculate and select the optimal number of sampled variables. Figure 8 shows that an optimal variables subset is obtained and the corresponding number of sampled variables (effective wavelengths) is 28. The SPA combined with SiPLS method (SiPLS-SPA) can further reduce the number of effective variables and overcome the limitations of SPA algorithm with low S/N or insufficiency in multivariate calibration. The final number of effective wavelengths is 19 by SiPLS-SPA method in Figure 8. The BOSS method is used to choose effective variables based on the variables selected by SiPLS (SiPLS-BOSS). In BOSS, the max number of latent variables is set to 10, the time of fold is 5, and the number of iteration is 1000, respectively. After running SiPLS, the redundant variables selected by BOSS are removed and 16 variables are retained (see Figure 8). In the application of FRCBOSS, the number of sub-models N is 2300, the optimal model δ is 0.05 and the fusion coefficient of frequency and regression coefficient α is 0.4, respectively. Then SiPLS-FRCBOSS is applied and the optimum solution is obtained, which is consisted of 19 variables (see Figure 8). The SiPLS, SiPLS-CARS, SiPLS-SPA, SiPLS-BOSS and SiPLS-FRCBOSS are applied to select characteristic wavelength related to quantitative analysis of lignin in pears, and a comparison of five variables selection algorithms is conducted for NIR model optimization.

The less correlated variables are removed. More concise and effective variable subsets are provided by the methods mentioned above. According to different optimization strategies, the numbers of eliminated variables by each method could be arranged in the order SiPLS-BOSS > SiPLS-FRCBOSS > SiPLS-SPA > SiPLS-CARS > SiPLS. The number of selected variables selected by SiPLS-SPA and SiPLS-FRCBOSS have some similarity, but those selected by SiPLS-FRCBOSS are more concentrated. The variables selected by SiPLS-CARS are a few larger than that by SiPLS-SPA, SiPLS-BOSS and SiPLS-FRCBOSS methods. BOSS method combines the strategies of model population analysis (MPA) and weighted bootstrap sampling (WBS) to extract useful information according to the weight of each variable (Deng et al., 2016). The weight of each variable is only determined by the absolute values of regression coefficients. And the method follows the principle of soft shrinkage, in which the unimportant variables are not removed directly but are assigned smaller weights. The variable set with the lowest RMSECV is selected in BOSS method. Although the number of selected variables is the least by the BOSS, the important variables are may eliminated at the same time. The number of variables selected by FRCBOSS is a little larger than that by BOSS, because the method selects the useful information according the absolute values and stability value of regression coefficients, and the method can reselect the variables removed by BOSS method. SPA minimizes the NIR spectra collinearity according to the projection operation, and only select the variables with minimum redundant information for multivariate (Xiaobo et al., 2010). CARS method can remove a large number of variables according to the absolute value of regression coefficient by force in the early stage of variable selection (Li et al., 2009). However, the absolute value of regression coefficient will change with the change of the sample space, resulting in the eliminating variables may contain useful variables.

The effective wavelengths on ‘Snow’ pears datasets are extracted by using SiPLS, SiPLS-SPA, SiPLS-CARS, SiPLS-BOSS, and SiPLS-FRCBOSS methods, which have similar characteristics (see Figure 8). The SiPLS method select 4-synergy subintervals wavelengths, and the number of variables is larger. The comparison of five variable selection methods shows that SiPLS-FRCBOSS has the highest selection efficiency in the selection of characteristic absorption which is closely related to various chemical structures in the near infrared spectrum.
The effective spectral regions mainly around 4000 cm\(^{-1}\), 4002 cm\(^{-1}\), 4007 cm\(^{-1}\), 4009 cm\(^{-1}\), 4048 cm\(^{-1}\), 4050 cm\(^{-1}\), 4262 cm\(^{-1}\), 4264 cm\(^{-1}\), 4312 cm\(^{-1}\), 4785 cm\(^{-1}\), 7176 cm\(^{-1}\), 7232 cm\(^{-1}\), 7234 cm\(^{-1}\), 7239 cm\(^{-1}\), 7342 cm\(^{-1}\), 7367 cm\(^{-1}\), 7376 cm\(^{-1}\), 7459 cm\(^{-1}\), and 7469 cm\(^{-1}\) are selected by SiPLS-FRCBOSS method (see in Table 3 and Figure 8).

For lignin, absorbance bands between about 4000 cm\(^{-1}\) and 4350 cm\(^{-1}\) and around at 4785 cm\(^{-1}\) are assigned to combination bands of O-H stretching and C-O stretching, as well as C-H stretching and C=C stretching (Baillèresa et al., 2002; Yang et al., 2016). The strong absorption signal from 7342 cm\(^{-1}\) to 7450 cm\(^{-1}\) corresponds to first overtone of O-H stretching from phenolic groups present in lignin (Yonenobu & Tsuchikawa, 2003). These characteristic absorptions can be preserved in the variable subset based on SiPLS-FRCBOSS. On the contrary, due to partial collinearity, some overlapping peaks, such as 9367 cm\(^{-1}\), 9623 cm\(^{-1}\), and 9664 cm\(^{-1}\) selected by other methods are easily removed from variable subset by SiPLS-FRCBOSS, which ultimately affects the accuracy and robustness of the PLS model. On the other hand, the water content in pears also has stronger absorption peaks in the spectral ranges of

### Table 2: Cross validation results based on different pretreatment methods

| Model       | WR               | NVs | Pretreatment | NFs | Cross Validation |
|-------------|------------------|-----|--------------|-----|------------------|
|             | 4000–10,000 cm\(^{-1}\) |     | NONE         | 17  | 0.796 1.787%     |
| PLS         |                  | 3112| SG(11)       | 19  | 0.801 1.760%     |
|             |                  |     | NORM         | 16  | 0.804 1.730%     |
|             |                  |     | SNV          | 12  | 0.767 1.861%     |
|             |                  |     | D1           | 9   | 0.760 1.893%     |
|             |                  |     | SG-SNV       | 10  | 0.760 1.887%     |
|             |                  |     | SG-NORM\(^*\) | 19  | 0.806 1.742%     |
|             |                  |     | SG-NORM-D1   | 11  | 0.767 1.868%     |

Abbreviations: NFs, number of factors; NVs, number of variables; WR, wavelength range.
7176–7250 cm\(^{-1}\) selected by five variable selection methods, which mainly include the first overtone O-H stretching and combination band (stretching and deformation) (Schwanninger et al., 2011). The broad and overlapping O-H bands due to water attenuated the stability of near infrared analysis model for pear properties.

3.3 | Modeling and comparison

PLS models based on variables selected by SiPLS, SiPLS-SPA, SiPLS-CARS, SiPLS-BOSS, and SiPLS-FRCBOSS are established to non-destructively measure lignin content in ‘Snow’ pears. The results are listed in Table 3. Besides,
the predictive results (predicted) obtained using with NIRS technology are compared with the ones (reference) obtained using the traditional Klason method, as shown in Figure 9. The PLS models optimized by selected variables show lower RMSECV (1.458%–1.580%) and higher Rc (0.838–0.863) to full spectrum models in cross validation, which preliminarily proved the feasibility of variable selection for model optimization. SiPLS-SPA and SiPLS-CARS can simplify the PLS model successfully in Figure 9b and c, but the prediction ability of the PLS model based on SPA and CARS is also reduced, as the $R_p$ of prediction is 0.826 and 0.807, the RMSEP is 1.248 and 1.303 (refer in Table 3), respectively. The PLS model based on the 16 variables selected by SiPLS-BOSS is established (SiPLS-BOSS-PLS model) in Figure 9d, which has a good prediction ability, as the $R_p$ of prediction is 0.850 and the RMSEP is 1.154 in Table 3. However, the prediction ability of SiPLS-BOSS-PLS model isn't the best compared with SiPLS-FRBOSS-PLS model, probably because the BOSS algorithm selects feature wavelength only according to the absolute value of regression coefficient and neglecting the frequency of variables in variables space.
and resulting in some useful wavelengths are removed. The results of PLS modeling based on 19 variables selected by SiPLS-FRCBOSS are encouraging in Figure 9e, and the prediction ability of the SiPLS-FRCBOSS-PLS model is the best (refer in Table 3), with \( R_p \) of prediction is 0.880 and RMSEP is 1.004%. Although the number of variables selected by SiPLS-FRCBOSS is a little larger than that selected by SiPLS-BOSS, FRCBOSS algorithm could reselect the useful wavelength removed by BOSS method. It indicates that the FRCBOSS algorithm proposed in this paper can overcome the disadvantage of BOSS in selecting variables, the latter algorithm only considering the absolute value of regression coefficient and neglecting the frequency of variables in variable space. Klason is a traditional method to measure lignin content, which is destructive and slow measurement method, while NIRS technology is a non-destructive and fast measurement method. Therefore, compared with the traditional measurement method, NIRS technology has more advantages for analysis of lignin content in ‘Snow’ pears.

In recent years, there are studies on determination of lignin in other materials (such as wood, rice straw, fine root, flax fiber etc.) For lignin in sugarcane, the obtained values are in the range of 16.23–33.85% with a mean value of 23.44%. The OPS method combined with PLS regression allows the building of more simple, interpretable, and predictive models for determination of lignin in sugarcane, with the \( R_c \), RMSECV, \( R_p \) and RMSEP for bagasse-with-juice equal to 0.94, 0.71, 0.94, and 0.65, respectively (Assis et al., 2017). The lignin content in pulp wood is from 14.82% to 34.2%, with an average of 26.43%. NIR spectra obtained by a portable instrument combined with KELM, PSO and MSC methods were employed to predict the lignin contents in various types of pulp wood, with \( R_p = 0.981 \), RMSEP = 0.958 and RPD = 5.015, respectively (Liang et al., 2020). The mean value of three major lignin monomers (H, S, and G) in the rice are 0.202%, 0.821% and 1.33%, respectively. The average of total lignin (H+S+G) is 2.353%. Four modified partial squares methods for rapid predicting of H, S, G, and H+S+G are built. An acceptable determination coefficient for calibration and external validation that ranged from 0.85 to 0.93, and 0.82 to 0.88, respectively (Hu et al., 2018). The range of lignin content in this fine root samples (10–43%) exceeds that typically reported in the literature, lignin content of grass species fine roots suggest a mean of 20.5% ±7.7%. The authors used CARS-PLS to select the most relevant wavelengths for root lignin prediction based on full spectrum information, with the \( R_c \), RMSECV and RPD are 0.86, 2.82%, and 2.67, respectively (Elle et al., 2019). Lignin contents of flax samples ranged from 1.27% to 7.06% with a SD of 1.37, and the average of 3.35%. For the lignin PLS model, the reduced range (6900–5600 cm\(^{-1}\)) had significantly better model fitness than the range (12,500–4000 cm\(^{-1}\)) model, with \( R_c \) of 0.936, \( R_p \) of 0.769, RMSEC of 0.351%, RMSEP of 0.455% and RPD of

| Materials | Mean (w/w, %) | VS method | Models | Full | nVars | nLV | Rc | RMSECV | Rp | RMSEP |
|-----------|--------------|-----------|--------|------|-------|-----|----|--------|----|--------|
| Sugarcane | 23.44%       | OPS       | PLS    | 1038 | 265   | 10  | 0.94| 0.71%  | 0.94| 0.65%  |
| Pulp wood | 26.43%       | PSO       | KELM   | 1108 | 265   | 14  | 0.995| 0.553% | 0.981| 0.98%  |
| Rice (H)  | 1.33%        | PLS       | PLS    | 1100 | 2500 | 6   | 0.85| 0.19%  | 0.83| 0.2%   |
| Rice (S)  | 1.33%        | PLS       | PLS    | 780  | 2500 | 8   | 0.88| 0.69%  | 0.82| 0.65%  |
| Rice (G)  | 1.33%        | PLS       | PLS    | 400  | 2500 | 6   | 0.92| 1.14%  | 0.86| 1.2%   |
| Rice (H+S+G) | 2.353% | PLS       | PLS    | 800  | 2500 | 16  | 0.93| 1.28%  | 0.88| 1.91%  |
| Flax      | 3.35%        | PLS       | PLS    | 800  | 2500 | 162 | 0.936| 0.351% | 0.769| 0.455% |
| ‘Snow’ pear | 7.727%     | SiPLS-BOSS | PLS    | 3112 | 19    | 10  | 0.863| 1.485% | 0.880| 1.004% |

Abbreviations: Full, full spectrum; Mean, the average of lignin content; nLV, the number of latent variable; nVars, the number of variables; VS method, variable selection method; WR, wavelength range.

| Materials | Mean (w/w, %) | VS method | Models | Full | nVars | nLV | Rc | RMSECV | Rp | RMSEP |
|-----------|--------------|-----------|--------|------|-------|-----|----|--------|----|--------|
| Sugarcane | 23.44%       | OPS       | PLS    | 1038 | 265   | 10  | 0.94| 0.71%  | 0.94| 0.65%  |
| Pulp wood | 26.43%       | PSO       | KELM   | 1108 | 265   | 14  | 0.995| 0.553% | 0.981| 0.98%  |
| Rice (H)  | 1.33%        | PLS       | PLS    | 1100 | 2500 | 6   | 0.85| 0.19%  | 0.83| 0.2%   |
| Rice (S)  | 1.33%        | PLS       | PLS    | 780  | 2500 | 8   | 0.88| 0.69%  | 0.82| 0.65%  |
| Rice (G)  | 1.33%        | PLS       | PLS    | 400  | 2500 | 6   | 0.92| 1.14%  | 0.86| 1.2%   |
| Rice (H+S+G) | 2.353% | PLS       | PLS    | 800  | 2500 | 16  | 0.93| 1.28%  | 0.88| 1.91%  |
| Flax      | 3.35%        | PLS       | PLS    | 800  | 2500 | 162 | 0.936| 0.351% | 0.769| 0.455% |
| ‘Snow’ pear | 7.727%     | SiPLS-BOSS | PLS    | 3112 | 19    | 10  | 0.863| 1.485% | 0.880| 1.004% |

Abbreviations: Full, full spectrum; Mean, the average of lignin content; nLV, the number of latent variable; nVars, the number of variables; VS method, variable selection method; WR, wavelength range.
2.366 (Huang & Yu, 2019). In this study, the lignin content mean value is 7.727%, the SiPLS-FRCBOSS-PLS model for determination of lignin in pears is built, with Rc is 0.863, Rp is 0.880, RMSECV is 1.458% and RMSEP is 1.004%. All the data above are refer in Table 4. The prediction ability for lignin in sugarcane, wood and root is better than that in rice and pear based on the NIR spectroscopy technology. The content of lignin in rice and pear is very low, which has adverse effect on the determination lignin reference value by Klason method and the acquisition of NIR spectrum by spectrometer. Table 4 shows that the models for measuring lignin in ‘Snow’ pear and rice straw have a considerable level. It is feasible for non-destructive and rapid analysis of lignin content in ‘Snow’ pears by using NIR Spectroscopy combine with SiPLS-FRCBOSS-PLS model, which has a better result.

4 | CONCLUSIONS

Our experimental results show that the FRCBOSS algorithm proposed in this paper provide advantages over BOSS method in selecting more effective variables, as the BOSS algorithm only considers the absolute value of regression coefficient and neglects the frequency of variables in variable space. The PLS model based on variables selected by FRCBOSS method is applied to the non-destructive testing of the lignin content in Snow’ pears with promising results. The extremely low level of lignin content in rice and pear presents a particular challenge in determining lignin reference value by Klason method and the acquisition of NIR spectrum by spectrometer. Our presented models have been demonstrated as an effective and viable option for measuring the low level of lignin content in ‘Snow’ pear and rice straw. We have successfully shown that the application of NIR spectroscopy combined with SiPLS-FRCBOSS-PLS algorithm is feasible for rapid and non-destructive analysis of lignin content in Snow’ pears.

ACKNOWLEDGEMENTS

We are very thankful to Max Wu who checked and revised the use of English of this manuscript. We are thankful to all infants and their parents who took part in the study.

CONFLICTS OF INTEREST

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

AUTHOR CONTRIBUTIONS

Conceptualization, Xin Wu and Guanglin Li.; methodology, Xuwen Liu and Fengyun He.; software, Xin Wu.; validation, Xin Wu and Xuwen Liu.; formal analysis, Xin Wu; investigation, Xin Wu and Xuwen Liu; resources, Guanglin Li.; data curation, Xin Wu; writing—original draft preparation, Xin Wu.; writing—review and editing, Xin Wu and Guanglin Li.; visualization, Fengyun He.; supervision, Guanglin Li.; project administration, Guanglin Li.; funding acquisition, Guanglin Li. All authors have read and agreed to the published version of the manuscript.

ORCID

Guanglin Li @ https://orcid.org/0000-0003-0123-8822

REFERENCES

Alves, A., Simões, R., Lousada, J. L., Lima-Brito, J., & Rodrigues, J. (2020). Predicting the lignin H/G ratio of Pinus sylvestris L. wood samples by PLS-R models based on near-infrared spectroscopy. *Holzforschung*, 74(7), 655–662. https://doi.org/10.1515/hf-2019-0186

Assis, C., Ramos, R. S., Silva, L. A., Kist, V., Barbosa, M. H. P., & Teofilo, R. F. (2017). Prediction of lignin content in different parts of sugarcane using near-infrared spectroscopy (NIR), ordered predictors selection (OPS), and partial least squares (PLS). *Applied Spectroscopy*, 71(8), 2001–2012. https://doi.org/10.1177/0003702817704147

Bai, Y., Xiong, Y., Huang, J., Zhou, J., & Zhang, B. (2019). Accurate prediction of soluble solid content of apples from multiple geographical regions by combining deep learning with spectral fingerprint features. *Postharvest Biology and Technology*, 156, 110943. https://doi.org/10.1016/j.postharvbio.2019.110943

Bailleresa, H., Davrieuxa, F., & Ham-Pichavant, F. (2002). Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. *Annals of Forest Science*, 59, 479–490.

Beramendi-Orosco, L. E., Castro-Díaz, M., Snape, C. E., Vane, C. H., & Large, D. J. (2004). Application of catalytic hydropyrolysis for the rapid preparation of lignin concentrates from wood. *Organic Geochemistry*, 35(1), 61–72. https://doi.org/10.1016/j.orggeochem.2003.07.001

Bunzel, M., Schussler, A., & Tchetseub Saha, G. (2011). Chemical characterization of Klasson lignin preparations from plant-based foods. *Journal of Agriculture and Food Chemistry*, 59(23), 12506–12513. https://doi.org/10.1021/jf2031378

Cai, Y., Li, G., Nie, J., Lin, Y., Nie, F., Zhang, J., & Xu, Y. (2010). Study of the structure and biosynthetic pathway of lignin in stone cells of pear. *Scientia Horticulturae*, 125(3), 374–379. https://doi.org/10.1016/j.scienta.2010.04.029

Cybulska, I., Brudecki, G., Rosentrater, K., Julson, J. L., & Lei, H. (2012). Comparative study of organosolv lignin extracted from prairie cordgrass, switchgrass and corn stover. *Bioresource Technology*, 118, 30–36. https://doi.org/10.1016/j.biortech.2012.05.073

De la Cruz, F. B., Dittmar, T., Niggemann, J., Osburn, C. L., & Barlaz, M. A. (2015). Evaluation of copper oxide oxidation for quantification of lignin in municipal solid waste. *Environmental Engineering Science*, 32(6), 486–496. https://doi.org/10.1089/ees.2014.0402

Deng, B.-C., Yun, Y.-H., Cao, D.-S., Yin, Y.-L., Wang, W.-T., Lu, H.-M., Luo, Q.-Y., & Liang, Y.-Z. (2016). A bootstrapping soft shrinkage approach for variable selection in chemical modeling. *Analytica Chimica Acta*, 908, 63–74. https://doi.org/10.1016/j.aca.2016.01.001
Du, X.-L., Li, X.-Y., Liu, Y., Zhou, W.-H., & Li, J.-L. (2019). Genetic algorithm optimized non-destructive prediction on property of mechanically injured peaches during postharvest storage by portable visible/shortwave near-infrared spectroscopy. *Scientia Horticulturae*, 249, 240–249. https://doi.org/10.1016/j.scienta.2019.01.057

Elle, O., Richter, R., Vohland, M., & Weigelt, A. (2019). Fine root lignin content is well predictable with near-infrared spectroscopy. *Scientific Reports*, 9(1), 6396. https://doi.org/10.1038/s41598-019-42837-z

Fahey, L. M., Nieuwoudt, M. K., & Harris, P. J. (2018). Using near infrared spectroscopy to predict the lignin content and monosaccharide compositions of *Pinus radiata* wood cell walls. *International Journal of Biological Macromolecules*, 113, 507–514. https://doi.org/10.1016/j.ijbiomac.2018.02.105

Fan, S., Guo, Z., Zhang, B., Huang, W., & Zhao, C. (2016). Using Vis/NIR diffuse transmission spectroscopy and multivariate analysis to predict soluble solids content of apple. *Food Analytical Methods*, 9(5), 1333–1343. https://doi.org/10.1007/s12161-015-0313-5

Fukushima, R. S., Garippo, G., Habitante, A. M. Q. B., & Lacerda, R. S. (2000). Extração da lignina e emprego da mesma em curvas de calibração para a mensuração da lignina em produtos vegetais.pdf. *Revista Brasileira De Zootecnia*, 61(5), 1302–1311.

Goldmann, W. M., Ahola, J., Mankinen, O., Kantola, A. M., Komulainen, S., Telkkä, V.-V., & Tanskanen, J. (2017). Determination of phenolic hydroxyl groups in technical lignins by ionization difference ultraviolet spectrophotometry (Δε-IDUS method). *Periodica Polytechnica Chemical Engineering*, 61(2), 93–101. https://doi.org/10.3311/PP.ch9269

Horikawa, Y., Hirano, S., Mihashi, A., Kobayashi, Y., Zhai, S., & Sugiyama, J. (2019). Prediction of lignin contents from infrared spectroscopy: Chemical digestion and lignin/biomass ratios of *Cryptomeria japonica*. *Applied Biochemistry and Biotechnology*, 188(4), 1066–1076. https://doi.org/10.1007/s12100-019-02965-8

Hu, Z., Zhang, G., Chen, Y., Wang, Y., He, Y., Peng, L., & Wang, L. (2018). Determination of lignin monomer contents in rice straw using visible and near-infrared reflectance spectroscopy. *Bioresource Technology*, 132, 3284–3299. https://doi.org/10.1016/j.biortech.2013.02.0289

Huang, J., & Yu, C. (2019). Determination of cellulose, hemicellulose and lignin content using near-infrared spectroscopy in flax fiber. *Textile Research Journal*, 89(23–24), 4875–4883. https://doi.org/10.1117/0040517519843464

Ibáñez, G., Cebolla-Cornejo, J., Martí, R., Roselló, S., & Valcárcel, M. (2019). Non-destructive determination of taste-related compounds in tomato using NIR spectra. *Journal of Food Engineering*, 263, 237–242. https://doi.org/10.1016/j.jfoodeng.2019.07.004

Jiang, H., Liu, G., Mei, C., Yu, S., Xiao, X., & Ding, Y. (2012). Measurement of process variables in solid-state fermentation of wheat straw using FT-NIR spectroscopy and synergy interval PLS algorithm. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 97, 277–283. https://doi.org/10.1016/j.saa.2012.06.024

Jiang, H., Zhang, H., Chen, Q., Mei, C., & Liu, G. (2015). Identification of solid state fermentation degree with FT-NIR spectroscopy: Comparison of wavelength variable selection methods of CARS and SCARS. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 149, 1–7. https://doi.org/10.1016/j.saa.2015.04.024

Lee, Y., & Han, S. H. (2016). Feasibility of nondestructive sugar content analysis of Korean pears by using near-infrared diffuse-reflection spectroscopy.pdf. *Bulletin of the Korean Chemical Society*, 37(2), 166–173. https://doi.org/10.1002/bkcs.10648

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, 648(1), 77–84. https://doi.org/10.1016/j.aca.2009.06.046

Li, J., Huang, W., Chen, L., Fan, S., Zhang, B., Guo, Z., & Zhao, C. (2014). Variable selection in visible and near-infrared spectral analysis for noninvasive determination of soluble solids content of ‘Ya’ pear. *Food Analytical Methods*, 7(9), 1891–1902. https://doi.org/10.1007/s12161-014-9832-8

Li, J., Tian, X., Huang, W., Zhang, B., & Fan, S. (2016). Application of long-wave near infrared hyperspectral imaging for measurement of soluble solid content(ssc) in pear. *Food Analytical Methods*, 9(11), 3087–3098. https://doi.org/10.1007/s12161-016-0498-2

Li, X., Zhang, J., Gao, W., & Wang, H. (2012). Study on chemical composition, anti-inflammatory and anti-microbial activities of extracts from Chinese pear fruit (Pyrus bretschneideri Rehd.). *Food and Chemical Toxicology*, 50(10), 3673–3679. https://doi.org/10.1016/j.fct.2012.07.019

Li, Y., Li, F., Yang, X., Guo, L., Huang, F., Chen, Z., Chen, X., & Zheng, S. (2018). Quantitative analysis of glycated albumin in serum based on ATR-FTIR spectrum combined with SiPLS and SVM. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 201, 249–257. https://doi.org/10.1016/j.saa.2018.05.022

Liang, L., Wei, L., Fang, G., Xu, F., Deng, Y., Shen, K., Tian, Q., Wu, T., & Zhu, B. (2020). Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 225, 117515. https://doi.org/10.1016/j.saa.2019.117515

Liu, D., Li, Q., Li, W., Yang, B., & Guo, W. (2017). Discriminating forchlorfenuron-treated kiwifruits using a portable spectrometer and Vis/NIR diffuse transmission spectroscopy technology. *Analytical Methods*, 9(28), 4207–4214. https://doi.org/10.1039/c7ay00832e

Liu, X., Wu, X., & Li, G. (2019). Optimized prediction of sugar content in ‘Snow’ pear using near-infrared diffuse reflectance spectroscopy combined with chemometrics. *Spectroscopy Letters*, 52(7), 376–388. https://doi.org/10.1080/00387010.2019.1648297

Luo, X., Ye, Z., Xu, H., Zhang, D., Bai, S., & Ying, Y. (2018). Robustness improvement of NIR-based determination of soluble solids in apple fruit by local calibration. *Postharvest Biology and Technology*, 139, 82–90. https://doi.org/10.1016/j.posth.2018.01.019

Ma, T., Li, X., Inagaki, T., Yang, H., & Tsuchikawa, S. (2018). Noncontact evaluation of soluble solids content in apples by near-infrared hyperspectral imaging. *Journal of Food Engineering*, 224, 53–61. https://doi.org/10.1016/j.jfoodeng.2017.12.028

Ozdemir, I. S., Bureau, S., Ozturk, B., Seyhan, F., & Aksoy, H. (2019). Effect of cultivar and season on the robustness of PLS models for soluble solid content prediction in apricots using FT-NIRS. *Journal of Food Science and Technology*, 56(1), 330–339. https://doi.org/10.1007/s13197-018-3493-3
Schwanninger, M., Rodrigues, J. C., & Fackler, K. (2011). A review of hand band assignments in near infrared spectra of wood and wood components. *Journal of Near Infrared Spectroscopy, 19*(5), 287–308. https://doi.org/10.1255/jnirs.955

Sheng, X., Li, Z., Li, Z., Dong, J., Wang, J., & Yin, J. (2020). Non-destructive determination of lignin content in Korla fragrant pear based on near-infrared spectroscopy. *Spectroscopy Letters, 53*(4), 306–314. https://doi.org/10.1080/003870010.2020.1740276

Song, J., Li, G., & Yang, X. (2019). Optimizing genetic algorithm-based partial least squares model of soluble solids content in Fukumoto navel orange based on visible-near-infrared transmittance spectroscopy using discrete wavelet transform. *Journal of the Science of Food and Agriculture, 99*(11), 4898–4903. https://doi.org/10.1002/jsfa.9717

Tao, S., Khanizadeh, S., Zhang, H., & Zhang, S. (2009). Anatomy, ultrastructure and lignin distribution of stone cells in two Pyrus species. *Plant Science, 176*(3), 413–419. https://doi.org/10.1016/j.plantsci.2008.12.011

Tian, X., Wang, Q., Li, J., Peng, F., & Huang, W. (2018). Non-destructive prediction of soluble solids content of pear based on fruit surface feature classification and multivariate regression analysis. *Infrared Physics & Technology, 92*, 336–344. https://doi.org/10.1016/j.infrared.2018.06.019

Xia, Y., Huang, W., Fan, S., Li, J., & Chen, L. (2018). Effect of fruit moving speed on online prediction of soluble solids content of apple using Vis/NIR diffuse transmission. *Journal of Food Process Engineering, 41*(8), https://doi.org/10.1111/jfpe.12915

Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M., & Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta, 667*(1–2), 14–32. https://doi.org/10.1016/j.aca.2010.03.048

Xu, X., Xu, H., Xie, L., & Ying, Y. (2019). Effect of measurement position on prediction of apple soluble solids content (SSC) by an on-line near-infrared (NIR) system. *Journal of Food Measurement and Characterization, 13*(1), 506–512. https://doi.org/10.1007/s11694-018-9964-4

Xue, C., Yao, J. L., Qin, M. F., Zhang, M. Y., Allan, A. C., Wang, D. F., & Wu, J. (2019). PbrmiR397a regulates lignification during stone cell development in pear fruit. *Plant Biotechnology Journal, 17*(1), 103–117. https://doi.org/10.1111/pbi.12950

Xue, Y.-S., Xu, S.-Z., Xue, C., Wang, R.-Z., Zhang, M.-Y., Li, J.-M., Zhang, S.-L., & Wu, J. (2020). Pearprocess: A new phenotypic tool for stone cell trait evaluation in pear fruit. *Journal of Integrative Agriculture, 19*(6), 1625–1634. https://doi.org/10.1016/s2095-3119(20)63193-8

Yan, H., Song, X., Tian, K., Gao, J., Li, Q., Xiong, Y., & Min, S. (2019). A modification of the bootstrapping soft shrinkage approach for spectral variable selection in the issue of over-fitting, model accuracy and variable selection credibility. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 210*, 362–371. https://doi.org/10.1016/j.saa.2018.10.034

Yang, H., Liu, Y., Xiong, Z., & Liang, L. (2019). Rapid determination of holocellulose and lignin in wood by near infrared spectroscopy and kernel extreme learning machine. *Analytical Letters, 53*(7), 1140–1154. https://doi.org/10.1080/00032719.2019.1700267

Yang, X., Li, G., Song, J., Gao, M., & Zhou, S. (2018). Rapid discrimination of Notoginseng powder adulteration of different grades using FT-MIR spectroscopy combined with chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 205*, 457–464. https://doi.org/10.1016/j.saa.2018.07.056

Yang, Z., Li, K., Zhang, M., Xin, D., & Zhang, J. (2016). Rapid determination of chemical composition and classification of bamboo fractions using visible-near infrared spectroscopy coupled with multivariate data analysis. *Biotechnology for Biofuels, 9*, 35.

Yonenobu, H., & Tsuchikawa, S. (2003). Near-infrared spectroscopic comparison of antique and modern wood. *Applied Spectroscopy, 57*, 1451.

Yuan, L.-M., Mao, F., Chen, X., Li, L., & Huang, G. (2020). Non-invasive measurements of ‘Yunhe’ pears by vis-NIRS technology coupled with deviation fusion modeling approach. *Postharvest Biology and Technology, 160*, 111067. https://doi.org/10.1016/j.postharvbio.2019.111067

Zhang, D., Xu, L., Liang, D., Xu, C., Jin, X., & Weng, S. (2018). Fast determination of sugar content in dangshan pear (Pyrus spp.) using hyperspectral imagery data. *Food Analytical Methods, 11*(8), 2336–2345. https://doi.org/10.1007/s12161-018-1212-3

---

**How to cite this article:** Wu X, Li G, Liu X, He F. Rapid non-destructive analysis of lignin using NIR spectroscopy and chemo-metrics. *Food Energy Secur. 2021;00:e289*. [https://doi.org/10.1002/fes3.289](https://doi.org/10.1002/fes3.289)