Abstract— Initial-pressed juice is an important intermediate product in cane sugar industry, and sugar brix is a key indicator for evaluating sugar quality. Real-time evaluation of sugar quality requires determining the content of sugar brix in all steps of the cane sugar process. Near-infrared (NIR) spectroscopy is simple, rapid and non-destructive technologies on the analysis of material contents. In this study, the chemometric algorithm of parameter-combined tuning of Savitzky-Golay (SG) smoother and Partial Least Squares (PLS) regression was utilized for NIR analysis of sugar brix contents in sugarcane initial-pressure juice. The algorithms of combined optimization of SG smoother and PLS regression was achieved and the calibration models were optimally established by screening the expanded 540 SG smoothing modes and the 1-30 latent variables (LV). The optimized models have high predictive accuracy. These results confirm that the combined optimization of SG smoothing modes and PLS LVs is effective in the quantitative determination of sugar brix contents in sugarcane initial-pressure juice, and that the NIR spectroscopic technology with its chemometric algorithms have the potential in the analysis of cane sugar intermediates.

Key Words— Initial-Pressure Juice, Sugar Brix, Near-Infrared Spectroscopy, Partial Least Squares regression, Savitzky-Golay smoother.

I. INTRODUCTION

Near-infrared (NIR) spectroscopy is a well-performed technology on analysis of both the structure of matter and material contents. It has the advantages of simple, rapid, non-destructive and reagent-free measurement, multi-component simultaneous determination, etc. With the development of modern science and computation, NIR spectroscopic analytical technology is widely applied to many fields, such as agriculture, food, environment, and biomedicine [1-5]. In recent years, there are preliminary studies on the application of NIR to cane sugar industry [6-8]. In the progress of cane sugar, initial-pressed juice is an intermediate product of much importance, and sugar brix is an important indicator for evaluating sugar quality. Real-time evaluation of sugar quality requires determining the content of sugar brix in all steps of cane sugar process. Conventional chemical analytical methods in laboratory cannot achieve the fast (or online) determinations and this has been a long-standing problem to be solved in cane sugar industry.

Online fast detection of cane sugar intermediates is expected to be achieved by using NIR spectroscopic technology. NIR spectroscopy with its chemometric methods owns the ability to output the perspective detection results in just a few minutes [9-10]. Common chemometric methods include Classical Linear Regression, Multiple Linear Regression, Principal Component Analysis, Partial Least Squares and etc. Partial Least Squares (PLS) is a common chemometric analytical algorithm integrating principal component analysis and multivariate linear regression. It can effectively eliminate spectral collinearity by creating comprehensive latent valuables.

The number of latent valuables (LV) is the key parameter of PLS, to reduce spectral noises and extracting responsive information [11-14]. Model predictive results will be reduced if LV is out of the suitable range, too small or too large. Thus a reasonable LV should be selected by taking it as a tunable parameter.

As the NIR spectroscopy is a reflection of the comprehensive information of all chemical components in the analytes, and, indispensably, including some kinds of noises generated in the detecting process. Therefore it is necessary to study the chemometric methods of data pretreatment to reduce spectral noises [15-16]. Savitzky-Golay (SG) smoother is a famous and widely used method for spectral data pretreatment [17-19]. Its major steps contain smoothing and differential, in which the smoothing mode is quite important for model improvement. There are many smoothing modes, determined by the three parameters of Order of Differential (OD), Degree of Polynomial (DP) and Number of Points (NP), and a specific smoothing mode outputs a corresponding calculation equation with its specific coefficients. Thus it is necessary to select a suitable SG smoothing mode and the best way to find it out is to screen it by tuning the three parameters combined with the optimization of PLS latent valuables.

The aim of this research is to quantitatively determine the sugar brix content in sugarcane initial-pressure juice by using the NIR spectral responses. Savitzky-Golay smoother is employed as the method for data pretreatment and PLS is utilized as the algorithm for establishing calibration models. Model improvement is achieved by the combined optimization of PLS LV and the parameters of SG smoother. For a wide-range optimization, we try to expand the tuning range of the three SG smoothing parameters, and then establish NIR calibration models with SG smoothing pretreatment and PLS regressions. The reasonable smoothing modes and the optimal PLS LV are simultaneously determined according to the model predictive results, in the combined computational algorithm platform. The selected pretreatment and modeling methods are examined by the prediction sample set, to have the potential of NIR modeling enhancement.

II. EXPERIMENT AND METHODS

A. Materials and Instruments

Eighty-six samples of sugarcane initial-pressure juice were collected. The sugar brix content of each sample was measured using the traditional chemical methods and the measured values were used as the modeling reference values for NIR quantitative analysis. The sugar brix content ranges from 19.0 to 22.2 (%Bx).

We detected the NIR absorption spectra of initial-pressure juice samples by using Foss Rapid Liquid grating spectrometer...
with a 1-mm pathlength quartz cuvette. The scanning range is set as 800-2500 nm. Because the rotation of the cuvette cell can effectively reduce the unevenness, and multiple scans can effectively reduce the influence of background noise, we designed to measure the spectra while the cuvette cell is rotating. The temperature was controlled at 25 ± 1°C and the relative humidity was at 46 ± 1% RH throughout the spectral scanning process.

B. Partitioning of Calibration and Prediction Samples

NIR spectroscopic analysis requires partitioning the samples into calibration set and prediction set. Calibration samples are used for model establishment and prediction samples for model evaluation. A suitable partition will lead to perspective modeling results. The method of sample partitioning based on joint x-y distances (SPXY) is common used for sample partitioning in the spectroscopic field [20-21]. SPXY process is the performance of Kernel-Stone (KS) algorithm on both the spectral data and the chemical data. The classic KS algorithm is aimed at selecting a representative subset from the sample pool (N samples). In order to ensure a uniform distribution of such a subset along the x matrix (spectral response), KS follows a stepwise procedure in which new selections are taken in regions of the space far from the samples already selected. For this purpose, the algorithm employs the Euclidean distances $d(j, k)$ between the x-vectors of each pair (j, k) of samples calculated as

$$d_j(k) = \sum_{i=1}^{m} (x_i(p) - x_i(p))^2 \cdot j, k \in [1, N].$$

(1)

For spectral data, $x_i(p)$ and $x_i(p)$ are the instrumental responses at the p-th wavelength for samples j and k, respectively. $P$ denotes the number of wavelengths in the spectra.

The algorithm selects the sample that exhibits the largest minimum distance with respect to any sample already selected. The proposal of the present paper consists of augmenting the distance defined in Eq. (1) with a distance in the dependent variable y vector (the contents of the target component in each sample) for the parameter under consideration. Such a distance $d_{ij}(j, k)$ can be calculated for each pair of samples j and k as

$$d_{ij}(j, k) = \sqrt{(y_j - y_k)^2} \cdot j, k \in [1, N].$$

(2)

In order to assign equal importance to the distribution of the samples in the x matrix and y vector, distances $d_{x}(j, k)$ and $d_{y}(j, k)$ are divided by their maximum values in the data set. In this manner, a normalized xy distance is calculated as

$$d_{xy}(j, k) = \frac{d_{ij}(j, k)}{\max_{j \neq k} d_{ij}(j, k)} \cdot j, k \in [1, N].$$

(3)

With a stepwise selection procedure similar to the KS algorithm, SPXY can be applied with $d_{xy}(j, k)$ instead of $d_{ij}(j, k)$ alone.

C. The Extended Savitzky-Golay Smoother

SG smoother is a famous and widely-used pretreatment method to eliminate spectral noise. SG smoothing parameters include Order of Differential (OD), Degree of Polynomial (DP) and Number of Points (NP). For convenience, we denoted that the original spectral smoothing is 0th order differential. And NP is usually an odd number, denoted as NP=2m+1. It means that 2m+1 consecutive spectral data as a window, the spectral data in the window were fitted by using polynomial function

$$d_{xy}(j, k) = \sum_{i=1}^{m} (C_{ip} - C_{mp})(C'_{ip} - C'_{mp}) \cdot j, k \in [1, M].$$

(4)

$$d_{xy}(j, k) = \frac{\sum_{i=1}^{M} (C_{ip} - C_{mp})(C'_{ip} - C'_{mp})}{\sum_{i=1}^{M} (C'_{ip} - C'_{mp})},$$

(5)

where $C_{ip}$ and $C_{mp}$ were predictive value and chemical values of the sample i in the prediction set. $C_{ip}$ and $C_{mp}$ were the mean predicted value and mean chemical value of all samples in the prediction set, and $M$ was the sample number in the prediction set.

The value of Rp is in coherent with RMSEP, usually that a higher Rp corresponds to a lower RMSEP. And, RRMSEP is always proportional to RMSEP. Thus, we take Rp and RMSEP as the main indicators for model optimization.
III. RESULTS AND DISCUSSIONS

The NIR spectra of 86 sugarcane initial-pressure juice were showed in Figure 1. The spectral responses contain the absorption information of many hydrocarbon groups in the initial-pressure juice samples, such as sucrose, organic acids, amino acids and etc. As is showed in Figure 1, the NIR spectra of initial-pressure juice reflect severe spectral overlap, and the absorption was weak. The absorbance is quite strong around 1460 nm and around 1940 nm because of water molecules. In order to reduce the interference of the water molecules, it is necessary to use SG smoother to deal with data pretreatment in spectral modeling.

It can be concluded from Table 2 that the model predictive results have high accuracy at each smoothing order of differential, regardless for the NIR data of initial-pressure juice when integrating the optimization of PLS models combined with SG smoother. And the modeling results are significantly superior in the PLS models with SG smoothing than without SG smoothing. The global optimal SG smoothing mode was 2nd order of differential and 4th (or 5th) degree of polynomial, and the optimal number of points was obviously larger than 25. These results indicated (1) the samples partitioning method of SPXY lead to well-done calibration models; (2) the DP, NP of SG smoother and the LV of PLS always altered corresponding to the varied OD; (3) the NP of SG smoother is necessary to be expanded to the range of larger than 25; and (4) the spectroscopic analytical chemometric parameters are somewhat similar for sugarcane initial-pressure juice.

To view insight the pretreatment effect of SG smoother, we sketch the figures showing the RMSEP corresponding to each order of differential and each number of points, optimized from different DP of SG smoother and different LV of PLS (see Figure 2). Figure 2 demonstrated that a lower-than-25 NP of SG cannot reach the minimum values of RMSEP and the expansion of NP would output the much optimal results. On another aspect, the influence of LV of PLS on modeling effect was also investigated. Figure 3 showed the RMSEP values corresponding to the varied LV of PLS, optimized by SG smoothing mode, with 2nd order of differential and 4th or 5th degree of polynomial. The figure confirmed that the optimized LVs were larger than 10.

Table 2. The optimal models corresponding to each order of differential, with its parameters and predictive results

<table>
<thead>
<tr>
<th>Order</th>
<th>DP</th>
<th>NP</th>
<th>LV</th>
<th>Rp</th>
<th>RMSEP (%Bx)</th>
<th>RRMSEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-smoothed</td>
<td>—</td>
<td>—</td>
<td>11</td>
<td>0.915</td>
<td>1.193</td>
<td>5.8%</td>
</tr>
<tr>
<td>0th order</td>
<td>4, 5</td>
<td>53</td>
<td>9</td>
<td>0.945</td>
<td>0.932</td>
<td>4.6%</td>
</tr>
<tr>
<td>1st order</td>
<td>3, 4</td>
<td>77</td>
<td>10</td>
<td>0.949</td>
<td>0.832</td>
<td>4.1%</td>
</tr>
<tr>
<td>2nd order</td>
<td>4, 5</td>
<td>55</td>
<td>12</td>
<td>0.954</td>
<td>0.762</td>
<td>3.7%</td>
</tr>
<tr>
<td>3rd order</td>
<td>3, 4</td>
<td>77</td>
<td>11</td>
<td>0.953</td>
<td>0.776</td>
<td>3.8%</td>
</tr>
<tr>
<td>4th order</td>
<td>2, 3</td>
<td>71</td>
<td>9</td>
<td>0.944</td>
<td>0.917</td>
<td>4.5%</td>
</tr>
<tr>
<td>5th order</td>
<td>5, 6</td>
<td>39</td>
<td>11</td>
<td>0.900</td>
<td>1.212</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

The NIR spectra of 86 initial-pressure juice samples

The partitioning of calibration samples and prediction samples has to be finished before model establishment. Using the SPXY method, we have the 86 initial-pressure juice samples divided into 56 samples (for calibration) and the other 30 (for prediction). The mean value and the standard derivation of the measured values of sugar brix content for all calibration/prediction samples were showed in Table 1.

The computational algorithm platform was built up for establishing NIR quantitative analytical models, by combining the 540 kinds of SG smoothing modes and the PLS LV tuning and optimizing, where the PLS LV was set changing from 1 to 30. The optimized model parameters were selected according to the model predictive results.

The optimal models corresponding to each order of differential, with its parameters and predictive results, were showed in Table 2. The non-smoothed full-range PLS modeling results was also listed in Table 2 for comparison. According to the maximum Rp (or minimum RMSEP), the optimal NIR model output the predictive results of Rp=0.954, RMSEP=0.762%Bx and RRMSEP=3.7%, with the best smoothing mode of OD=2, DP=4 (or 5) and NP=55, and with the optimal LV=12.

Table 1. Mean value and standard derivation of the measured content of sugar brix for calibration/prediction samples

<table>
<thead>
<tr>
<th>(unit: %Bx)</th>
<th>Mean value</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration</td>
<td>20.80</td>
<td>0.63</td>
</tr>
<tr>
<td>Prediction</td>
<td>20.39</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Figure 1. NIR spectra of 86 initial-pressure juice samples

Figure 2. RMSEP corresponding to each order of differential and each number of points for initial-pressure juice (optimized from different DP of SG smoother and different LV of PLS)
IV. CONCLUSIONS

The chemometric algorithm combined parameter-tuning of SG smoother and PLS regression was utilized for NIR spectroscopic analysis of sugar brix contents in sugarcane initial-pressure juice, to establish and screen for the optimized calibration model. SPXY method was used smoothly and seemed much effective in the partition of calibration and prediction samples. The algorithms of combined optimization of SG smoother and PLS regression was achieved and the calibration models were optimally established by screening the expanded 540 SG smoothing modes and the 1-30 LVs. The combined optimized calibration models have high predictive accuracy, and the optimized modeling results were quite appreciated. These results confirm that the expansion of SG parameters is quite necessary, and the combined optimization of SG smoothing modes and PLS LVs is an important method for the quantitative determination of sugar brix contents in sugarcane initial-pressure juice. Our conclusions demonstrated that the NIR spectroscopic technology with its chemometric algorithms has the potential in the analysis of cane sugar intermediates. This rapid, non-destructive and reagent-free technology has practical meanings and is perspective in the online detection for cane sugar industry.

REFERENCES
