Quantitative near infrared analysis of orange juices using partial least squares method

Weijie Li, Michel Foulon and Marc Meurens
UCL, AGRO/BNUT, Place Croix du Sud 2/8, B-1348, LLN, Belgium.

Benoit Moreau
UCL, AGRO/BIOM, Place Croix du Sud, 2/16, B-1348, LLN, Belgium.

Introduction

Fruit juices, especially orange juices, are extremely important commercial commodities in Europe and the United States. A rapid and accurate quantitative method is necessary for quality control of orange juices. Near infrared (NIR) offers many advantages and may be an ideal choice for quality control of large numbers of samples. The aims of our investigation were to study the transmittance near infrared spectra of large numbers of dry extract of orange juice samples, and then to use the data to develop calibration models for quantitative analysis.

Partial least squares (PLS) modelling is a powerful new multivariate statistical tool, which has been successfully applied to the quantitative analysis of NIR data.[1] This modelling can provide extensive possibilities for extracting information either from the variables or from the objects in the calibration modelling. The optimal number of principle components for the calibration models to determine sugars, organic and amino acids were studied in our investigation.

NIR data, however, pose some special problems, with interference from other chemical constituents, physical phenomenon and the measurement itself. The mathematical pretreatment, based on transformation of the spectral data: derivatives and multiplicative scatter correction (MSC), has been used to enhance the qualitative understanding of spectra and the predictive ability of calibration models.

Materials and methods

Commercial orange juices (218 samples), both single strength and concentrates, were collected from different countries in Europe, Africa and America. These samples had been previously analyzed by Schutzgemeinschaft der Fruchsafindustrie (Zornheim, Germany) for contents of glucose, fructose, sucrose, citric and malic acids by enzymatic methods, and amino acids by HPLC. These results were used as reference data in contrast with NIR spectral data for the PLS calibration.

Before NIR spectroscopic analysis, the concentrates were diluted to 11.18°Brix (w/w) with distilled water, and the single strength juices were homogenized by a *Kontes* homogenizer. Each diluted concentrate or homogenized single strength juice (0.6 mL) was dried on a fiberglass disk in a DESIR (dry extract system for infrared) unit designed by NIRSystems (Perstorp Analytical Co., Belgium), and then presented in transmission mode to a Pacific Scientific spectrometer

(NIRSystems 6250 monochromator). The spectrum of each orange juice sample was calculated from an average of six spectra, which were obtained from two fiberglass disks, and each disk could give three positions for measurement by manual 120° rotation of the sample cup inside the spectrometer. Every measurement was separately scanned 10 times from 1100 to 2500 nm. The reference spectrum of a blank fiberglass disk was scanned before the sample.

Data analysis by the PLS algorithm was performed in two steps with the software package Unscrambler 5.5 (CAMO A/S, Trondheim, Norway). The calibration step estimates the relationship (called calibration model) between the spectral and chemical data for each component from a calibration set of 150 samples. These were randomly selected from the total sample collection, then followed by the validation step, in which the calibration models were used to predict the component concentrations in a test set of remaining spectra (68 samples). The validity of these calibration models in terms of residual variance and correlation coefficient was detected by a comparison between the predicted values and the chemical data of the test set. Samples in the test set were chosen within the constituent ranges of the calibration set for an efficient validation.

Results and discussion

Mathematical pretreatment of NIR spectra

The interest to perform some mathematical pretreatment of NIR spectra is to make all major interference vary as independently as possible of each other in the samples and to reduce the influence of light scattering. The pretreatment of NIR spectra can lead to some improvement of final calibration results. The following two techniques were used to improve the initial spectral data of the orange juice samples: (i) first and second derivatives, which were used to glean additional information and to avoid some band interference and overlap between components, and (ii) MSC transformation, in which each sample's spectrum was corrected according to the major water absorption band (1900–2000 nm). This major water band was not influential on the rest of

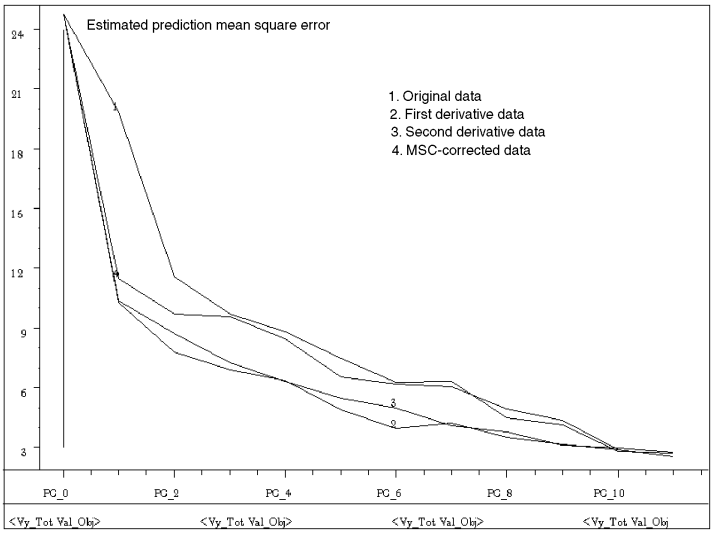

Figure 1. Predicting ability of the spectral data by different mathematical pretreatment for the determination of glucose.

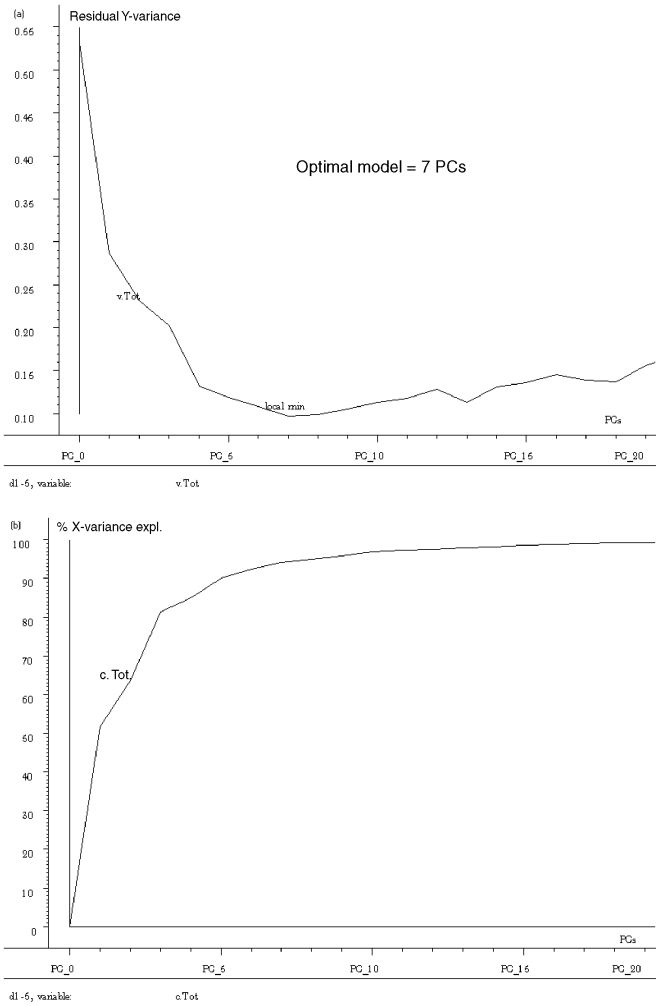

Figure 2. Selection of optimal number of principal components for glucose determination: (a) Validation variance and (b) calibration variance.

the bands of components of interest due to the elimination of water. After correction, all samples had the same scatter level.

We have constructed the calibration models for each component of interest on the basis of different transformed spectral data of orange juices. For example, in Figure 1, the validated estimated prediction error was shown against the number of principal components (PCs) by PLS calibration for the calibration models to determine glucose. From this plot, we found that the original spectral data had given a higher square error than that of the transformed spectral data. However, the results for first and second derivatives and MSC standardized spectral data were approximate. Therefore, mathematical pretreatment of NIR spectral data is necessary prior to calibration. The selection of mathematical pretreatment of spectral data obviously depends on the

final calibration results. In the above example, the first derivative was chosen to reach the lowest prediction error.

Calibration and validation

In analyzing real samples, the PLS1 algorithm, in which the calibration analysis is performed on one component each time, more often exhibits better predictive properties than a global algorithm PLS2.[2] Therefore, we only applied the PLS1 calibration method to all the spectral pretreatment of orange juice samples to measure the contents of sugars, organic and amino acids. In addition, the PLS in Unscrambler is a powerful tool for the efficient detection of outliers. Outliers were eliminated by combining information about residuals and hence identifying points which are both suspect and influential in the modeling.

Selection of an optimal number of PCs allows us to establish the calibration models as much as the complexity of the system without overfitting the data sets. To accomplish this goal, we used the separate test set to examine how good the model was and how accurate the prediction of new data could be. One reasonable choice for an optimal number of PCs would associate with the first local minimum value in the plot of validation variance as a function of the number of PCs. For example, the plot of validation variance to determine glucose is shown in Figure 2(a) with the best mathematical pretreatment. The first minimum value achieved at seven of PCs. The optimal number of PCs was chosen as seven, where about 95% of the original variance in the calibration set data has been accounted for [Figure 2(b)]. Using lower order PCs can leave important NIR structure unmodelled. With higher PCs, a risk in increasing the prediction error is greater and this may result in overfitting. The optimal number of PCs shown in Table 1 was obtained in this way for each component of interest with, respectively, about 95% of explained x-variance in the calibration set.

Table 1. Calibration and prediction results of PLS-1 calibration models for the orange juice analysis.

	Range	*RMSEP*	*SEP*	*r*	PCs	Method
Glucose (g L^{-1})	4.10–39.30	1.57	1.57	0.92	7	D1OD[a]
Fructose (g L^{-1})	4.70–44.00	1.53	1.54	0.93	12	MSC[b]
Sucrose (g L^{-1})	6.50–95.20	3.20	3.02	0.96	9	D2OD[c]
Citric acid (g L^{-1})	0.80–23.40	0.92	0.91	0.98	14	MSC
Malic acid (g L^{-1})	0.28–3.41	0.20	0.21	0.90	22	D2OD
Proline (mmol L^{-1})	2.74–16.29	1.57	1.52	0.88	13	D2OD
γ-Aminobutyric acid (mmol L^{-1})	0.78–4.50	0.47	0.47	0.83	18	D1OD
Arginine (mmol L^{-1})	1.26–5.77	0.61	0.58	0.88	11	D2OD
Asparagine (mmol L^{-1})	0.83–5.59	0.58	0.58	0.73	16	D2OD

[a]First derivative.
[b]Multiplicative scatter correction.
[c]Second derivative.

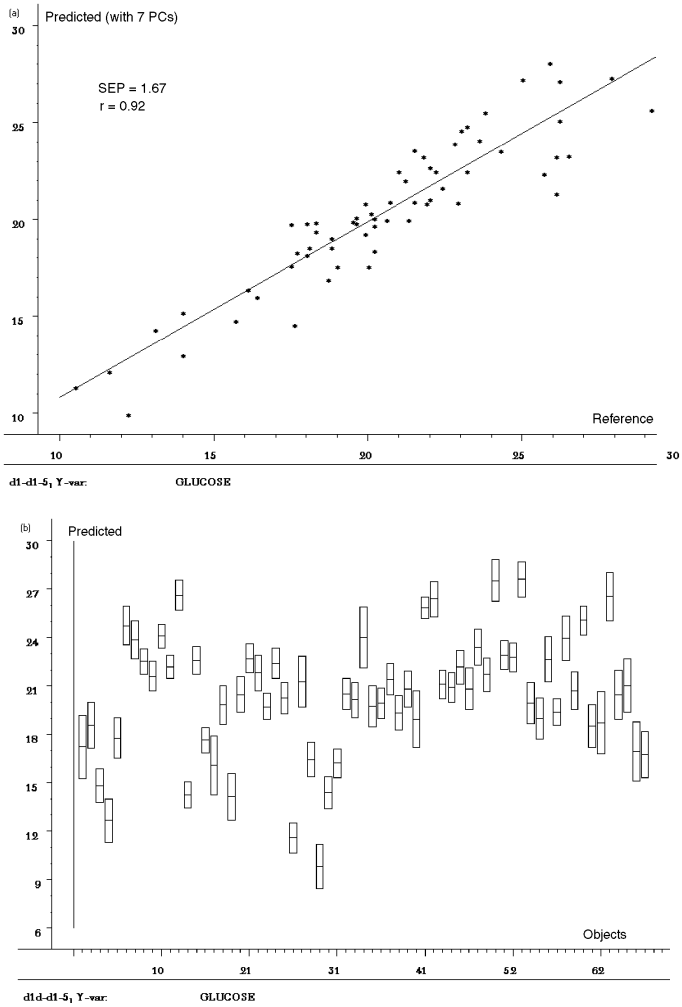

Figure 3. Prediction abilities for glucose determination in test set: (a) predicted against reference values and (b) predicted values of glucose with deviations for each sample.

Three statistics:[3] root mean square error of prediction (*RMSEP*), standard error of prediction (*SEP*) and persons correlation coefficient (*r*), were used to evaluate the prediction ability of the calibration models. The final calibration and prediction results of the calibration models with the best mathematical pretreatment are shown in Table 1. The calibration models with high constituent concentrations like sugars and citric acid have given better prediction abilities than those with low concentrations like malic and amino acids. The relatively high *r* with the worse *RMSEP* and *SEP* is possible because the *r* only measures the degree to which the calibration models fit the data. It does not give an indication of how reliable those predicted values are.

The prediction abilities of the calibration models for the determination of glucose and malic acid are plotted in Figures 3 and 4. Although a good regression line is shown in Plots 3(a) and 4(a) for both of the components in the test set, the deviations of predicted values of malic acid shown

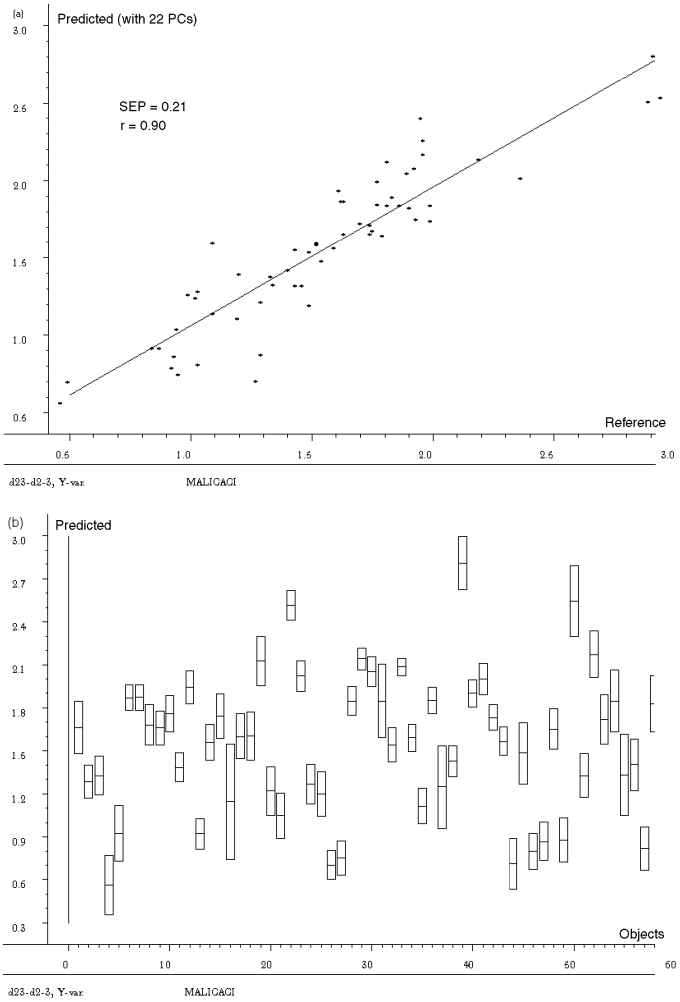

Figure 4. Prediction abilities for malic acid determination in test set: (a) predicted against reference values and (b) predicted values with deviations for each sample.

in Figure 4(b) were relatively worse than those of glucose shown in Figure 3(b). This indicates that the calibration model of malic acid was less reliable than that of glucose. Therefore, an important aspect of further investigation is to improve the sensitivity and accuracy of NIR spectrometry.

References

1. M. Martens and H. Martens, *Appl. Spectrosc.* **40,** 303 (1986).
2. D.M. Haaland and E.V. Thomas, *Anal. Chem.* **60,** 1193 (1980).
3. *Unscrambler User's Guide.*Version 5.5. Program package for multivariate calibration. Marketed by CAMO A/S, Trondheim, Norway (1994).