Chemometrics applied to quantitative analysis of ternary mixtures by Terahertz spectroscopy

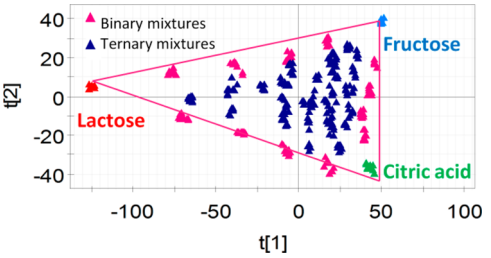
Josette El Haddad, Frédérick de Miollis, Joyce Bou Sleiman, Lionel Canioni, Patrick Mounaix, Bruno Bousquet

Chemometrics Applied to Quantitative Analysis of Ternary Mixtures by Terahertz Spectroscopy

Josette El Haddad, Frederick de Miollis, Joyce Bou Sleiman, Lionel Canioni, Patrick Mounaix, and Bruno Bousquet*

Université de Bordeaux, CNRS, LOMA, UMR 5798, Talence, F 33400, France

ABSTRACT: Chemometrics was applied to qualitative and quantitative analyses of terahertz spectra obtained in transmission mode. A series of mixtures of three pure analytes, namely, citric acid, D (−)fructose, and α lactose monohydrate under various concentrations, was prepared as pressed pellets with polyethylene as binder. Then, terahertz absorbance spectra were recorded by terahertz time domain spectroscopy and analyzed. First, principal component analysis allowed one to correctly locate the samples into a ternary diagram. Second, quantitative analysis was achieved by partial least squares (PLS) regression and artificial neural networks (ANN). The concentrations were predicted with values of relative mean square error lower than 0.9% for the three constituents. As a conclusion, chemometrics was demonstrated to be very efficient for the analysis of the ternary mixtures prepared for this study.



Terahertz waves (1 THz = 10^{12} Hz) are electromagnetic waves ranging from 0.3 to 10 THz. The low energy interactions and their capacity to propagate through a wide variety of materials[1−3] allow these waves to assess the dielectric properties of the sample and thus advantageously contribute to their understanding in complement to far infrared and Raman spectroscopy. Moreover, the interactions with THz waves are nondestructive, which is compatible with quality control applications for industry[4,5] and/or probing biological materials[6,7] for identification[8] and for medical diagnostics.[9] Otherwise, THz imaging and remote sensing have allowed one to detect hazardous and illicit products.[1,10]

The THz spectroscopy technique is a powerful tool for characterizing vibrational modes, such as rotational, torsional, phonon, and intra and intermolecular modes.[11] THz spectroscopy is commonly considered as being different from conventional far infrared spectroscopy because the terahertz response is coupled to the collective behavior of molecules in their environment. THz spectroscopy can also distinguish polymorphism[12] and chirality[13] between molecules.

Many chemical compounds and biological molecules have already been investigated using THz TDS in the spectral range below 3 THz, such as DNA components,[6] amino acids,[14] and crystalline samples.[15] This illustrates the importance of this spectral range, in providing the so called "fingerprint" of the conformational structure of molecules and a new means to recognize or distinguish some chemical compounds. Today, THz spectroscopy is already utilized to monitor and control pharmaceutical processes.[16]

Recent developments in laser based THz systems provided measurements with the stability that was mandatory for quantitative analysis.

However, THz spectroscopy can also be used for identification, recognition, and sorting. In this case, quantitative analysis is not required and a simple comparison between a few selected spectral features (i.e., absorption peaks or bands) can lead to interesting results.[17] Moreover, in order to improve the ability of THz spectroscopy, some chemometric methods were applied. A review of terahertz pulsed spectroscopy[18] summarized the most common methods applied for processing the THz spectra and, more precisely, quantitative univariate and multivariate methods. The advantage of using a multivariate approach was also reported for process analytical technology (PAT).[19] Moreover, principal component analysis (PCA) has been used to describe and compress the THz data. Watanabe et al.[20] thus employed the PCA method to retrieve the spatial distribution of different chemical compounds in a pellet, and Zeilter et al.[21] also applied PCA to investigate the effects of temperature and hydration on the absorbance spectra of pharmaceutical materials. For quantitative analysis, the partial least square (PLS) regression has been successfully implemented for different types of analyses[12,19,22] since the terahertz spectra under study revealed linear behaviors. PLS is a multivariate method based on a very similar algorithm as the one of PCA. In addition to its application to process THz spectra of pharmaceutical samples,[12] the PLS method was found to also be efficient in the framework of cultural heritage.[23] Another demonstration of the PLS method coupled to THz spectroscopy was reported in the case of nutrition through the analysis of pesticides in rice.[24] Finally, artificial neural networks (ANN) were also applied to the THz data since this chemometric method allows one to take into account possible nonlinear behaviors in the detected signals to identify illicit drugs.[25] In this case, ANN was not utilized directly as a

quantitative method. In previous studies, our group has already applied ANN to quantitative analysis of LIBS spectra[26,27] and has demonstrated the advantages of this technique.

In this paper, we investigated a series of ternary mixtures based on D (−)fructose, α lactose monohydrate, and citric acid in various concentrations. All the samples were mixed with polyethylene playing the role of binder. Previous studies of THz spectroscopy demonstrated that fructose was characterized by well contrasted peaks in the THz frequency range.[28,29] Lactose has been also identified and quantified by THz spectroscopy,[30,31] as well as citric acid.[32] In this work, we applied different methods of chemometrics in order to analyze the THz spectra of theses samples. First of all, we applied PCA in order to manage data compression for qualitative analysis, and second, we performed quantitative analysis based on the PLS and ANN methods. We considered three analytes and an idealized binder such as polyethylene. Thus, this demonstration based on three analytes should be considered as a first step prior to more complex cases, which should be addressed in future works. Indeed, if the parameters presented in this study appear to be moderate for many well established techniques, this is not the case in the framework of terahertz spectroscopy. Moreover, starting with a basic case allows one to demonstrate the advantages of using chemometrics for reaching acceptable performance in THz spectroscopy.

EXPERIMENTAL SECTION

For this study, we prepared a series of 39 samples. The three selected pure analytes, namely, citric acid (Aldrich), D (−)fructose (Sigma Life Science), and α lactose monohydrate (Sigma Life Science), each of them containing less than 0.05% impurities, were mixed together with pure polyethylene (Aldrich), hereafter called PE, as a binder. First of all, for each sample, the analytes were weighed out and then ground into a mortar. Second, they were iteratively mixed together in small amounts and ground again in order to avoid the formation of aggregates and heterogeneous clusters into the sample and then to limit undesired scattering. Indeed, it should be noted that the size of the grains is known to have a very high influence on the porosity of the PE based samples and consequently on scattering. In the present work, the sample preparation allowed one to obtain terahertz time domain spectra free of the Christiansen effect reported by Franz et al.[33] in the case of coarse grained powder. We prepared mixtures containing 80% in mass of PE and 20% in mass of the ternary mixture (fructose−lactose−citric acid) for a total weight of 400 mg per sample. Two replicates of 400 mg each were prepared for each mixture, and the amounts were ground and homogenized. Finally, the samples were prepared as pressed pellets by using a manual press (8 tons/cm^2 during 1 min). Following this accurate sample preparation, THz TDS experiments were conducted as described below. Then, we applied a homemade algorithm based on the work of Duvillaret et al.[34] that was able to retrieve from the etalon effect both the dielectric function and the thickness, simultaneously. On the basis of this approach, the sample thicknesses were obtained within 1% accuracy.

Figure 1 displays the 39 samples inside a ternary diagram with each pure analyte as a pole. The samples were selected as follows: Three of them contained only one pure analyte and were consequently displayed at the three poles of the triangle; then, 12 samples contained different mixtures of two pure analytes among the three and were consequently displayed on

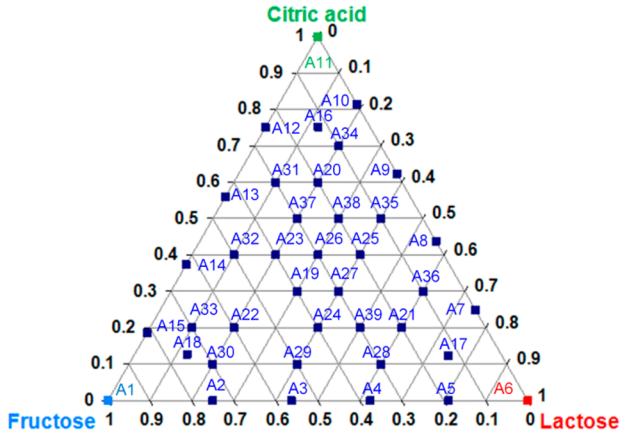


Figure 1. Ternary diagram displaying the 39 samples analyzed by THz spectroscopy. The three poles are related to fructose, lactose, and citric acid, and thus, each sample corresponds to a mixture of these analytes. The value 1 stands for 100%, and each value between 0 and 1 corresponds to the relative percentage in mass of the related analyte.

the three sides of the triangle; and finally, 24 samples contained different mixtures of the three pure analytes and were consequently displayed inside the triangle. Let us consider the sample A30 as an example; this sample contained 70% fructose, 10% citric acid, and 20% lactose, but it should be recalled that these three analytes represented 20% of the weight while the other 80% corresponded to the polyethylene binder. This remark applies to all of the other samples. Finally, an additional sample was prepared in order to be used as the reference sample. This reference sample, not displayed in Figure 1, solely contained 400 mg of polyethylene and was also prepared as a pressed pellet is the same way as the other samples.

Experimental Setup for Terahertz Absorbance Spectroscopy. The experimental setup was a commercial system, namely, the TPS Spectra 3000 from Teraview. Its principle is depicted in Figure 2. Basically, we used a standard THz TDS

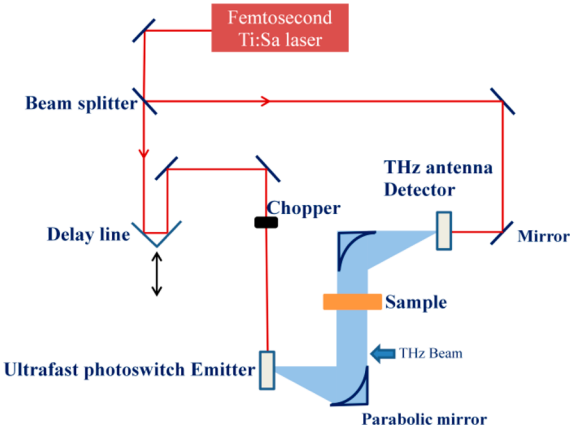


Figure 2. Experimental setup for transmission terahertz spectroscopy.

transmission setup based on a mode locked Ti Sapphire laser providing 80 fs pulses with a 76 MHz repetition rate. The laser output was split into pump and probe beams. The pump was focused onto a photoswitch (left in Figure 2) for the generation of the terahertz field. This terahertz field traveled through the sample and was finally detected with a photoswitch triggered by the probe laser beam (right in Figure 2). Upon its interaction

with the sample, the time resolved field variation was measured using the variation in the photocurrent induced by the probe laser beam into the detector made of a LT GaAs semi conductor.

Finally, the photocurrent induced by the probe laser beam was filtered out and amplified by a lock in digital amplifier at the frequency of the THz emission given by the chopper that modulated the incident laser beam. The time delay line allowed one to sample the signal step by step and then to rebuild the terahertz field by sampling technique.

All the signals processed in this study were the result of averaging over 1000 acquisitions in order to lower the noise. Special attention was paid to evaluate the performance of this THz TDS system, namely, the signal to noise ratio and the reliable spectral range. Typically, the latter was ranging from 0.05 to 3 THz for this instrument, depending on the sample under investigation. The measurements were carried out under dry air. The level of dry air was controlled in real time through the monitoring of the absorption of two spectral lines of water vapor at 1.12 and 1.7 THz, on the reference spectrum, i.e., the spectrum obtained from the pure PE pellet. Moreover, we also studied the reproducibility of the measurements by a series of repetitions. More precisely, the same sample has been analyzed many times during several days. Between two consecutive measurements, the sample was extracted from the chamber and then introduced back in the chamber. We observed that the variations due to this back and forth positioning were negligible compared to the effect of the residual water content in air inside the chamber. By calculating the Fourier transform of the temporal signal recovered after a sampling along the time delay, we obtained a signal displayed on a frequency axis for each of the 40 samples including the reference sample composed of pure PE. Then, for each measurement, the sample signal $S(\omega)$ was divided by the reference signal $R(\omega)$ obtained from the pure PE pellet. Finally, the absorbance was calculated from the mathematical expression: $A(\omega) = -\log(S^2(\omega)/R^2(\omega))$, and for all the samples, the absorbance spectra were treated via multivariate approaches. It should be noted that the spectra were recorded starting with pure PE and then following the index of the samples, i.e., from A1 to A39.

RESULTS AND DISCUSSION

For each sample, the experiment was repeated 10 times in order to get some statistics to allow one to detect outliers. The absorbance of the pure polyethylene pellet was found to be much lower than the one of the other pellets over the entire spectral range under study. Figure 3 displays the absorbance spectra of the three pure analytes, namely, fructose (sample A1), lactose (sample A6), and citric acid (sample A11), prepared with 80% of polyethylene (PE) and also the spectrum of a typical mixture sample (sample A26). It should be noted that the thickness of our pellets was about 3 mm.

Consequently, the etalon effect was about 30 ps after the main peak. Moreover, its amplitude was very small due to the very low reflection of the pellets. The temporal signals were finally recorded on a 25 ps window in order to filter out the etalon effect. Thus, no fringe pattern was observed on the spectra displayed in Figure 3. One can observe in this figure distinctive spectral bands potentially allowing for identification and hopefully quantitative analysis. Indeed, three peaks were observed for D (−)fructose at 1.3, 1.71, and 2.13 THz, four peaks for α lactose monohydrate at 0.53, 1.19, 1.37, and 1.81 THz, and three peaks for citric acid monohydrate at 1.29, 1.7, and 2.4 THz. Since the absorbance is additive when the wave is transmitted through a series of separate samples, one could also expect some additive effect in the case of a ternary mixture.

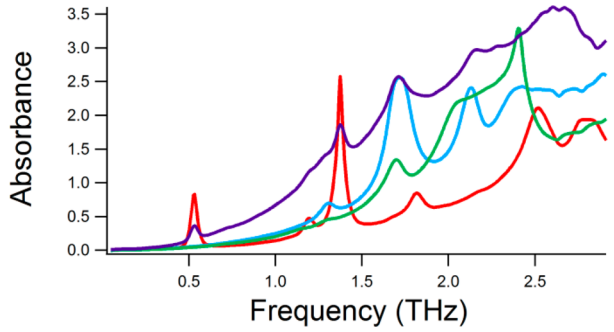


Figure 3. Absorbance spectra of fructose (blue), lactose (red), citric acid (green), and a mixture of the three analytes (purple) in the range of 0−3 THz.

Indeed, the absorbance spectrum of sample A26 represented by the purple curve in Figure 3 naturally contains the spectral features of the three pure analytes. Nevertheless, it was not simply the weighted sum of the three spectra related to the pure analytes. Consequently, direct analysis of unknown mixture samples was difficult to achieve and this was the motivation for the use of chemometrics.

Data Description by PCA. For each sample, two pellets were prepared and five THz absorbance spectra were recorded per pellet. Consequently, ten spectra per sample were taken into account. Considering the number of 39 samples, 390 raw spectra were introduced into the PCA model. The projection of these spectra onto the plane of the two first components of the PCA model revealed that two groups of points should be considered as outliers. They were related to one of the two replicate pellets of the samples A28 and A34. These samples were classified as outliers because they displayed scores very different from the other samples, and thus, they were outside the Hotelling's ellipse representing 95% confidence for a given PCA model. Thus, a new PCA model was calculated without these 10 spectra, namely, 5 from each sample. It should be pointed out that the total number of spectra was consequently changed to 380, corresponding to the 39 samples discussed above. The spectral range of the THz absorption spectra was between 0.05 and 2.90 THz with 457 variables for each spectrum. The first derivative of the absorbance was calculated for each spectrum, and then, the resulting values were mean centered as preprocessing of the PCA. In the case of the THz spectra analysis presented in this paper, such a preprocessing provided a very strong advantage since raw absorbance spectra revealed unexpected offsets that might be due to small variations of the samples' thickness.

Figure 4 displays the scores of the PCA in the plane of the two first components. One can observe the small dispersion of each group of 10 points related to the 10 spectra recorded for each single sample. In addition, one can clearly observe that the points are spread inside a triangle (pink lines in Figure 4). We concluded from Figure 4 that the plane defined by the two first components allowed one to retrieve the ternary diagram given in Figure 1. More precisely, the three poles of the triangle in Figure 4 corresponded to the three pure analytes, namely, citric acid for the sample A11 (green), fructose for the sample A1 (blue), and lactose for the sample A6 (red). The samples related to binary mixtures were correctly displayed on the three

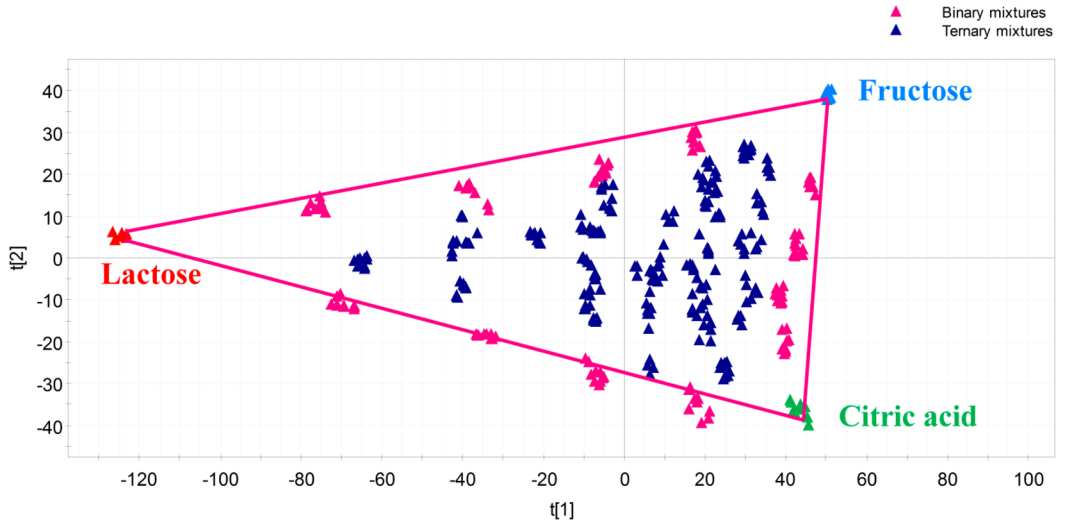


Figure 4. Scores of PCA in the plane of the two first principal components for a data set of 380 spectra (39 samples) containing 457 variables each. Solid lines interconnecting the poles have been added to help visually.

sides of the triangle. It should be noted also that the relative positions of all the samples presented in Figure 1 were correctly retrieved through the scores of the PCA (cf. Figure 4). To better understand the ability of PCA to efficiently describe the THz absorbance spectra, two figures should be compared, respectively, Figure 5 giving the first derivative of the THz absorbance spectra for the three pure analytes and Figure 6 displaying the loadings of the two first components of PCA on the original frequency range. This comparison clearly reveals that the first loading p[1] displayed in red in Figure 6 was clearly anticorrelated to the first derivative of the absorbance spectrum of lactose displayed in red in Figure 5 especially through the peaks at 0.53 and 1.38 THz. For this reason, the samples related to lactose were very well separated by the first component of the PCA and displayed at the extreme left of the scores' graph displayed in Figure 4. Similarly, the second loading p[2] displayed in Figure 6 was clearly correlated to fructose and anticorrelated to citric acid. This was particularly easy to verify at 2.41 THz. This was the reason for the good ability of this PCA model to retrieve the original ternary diagram with only two components. As a first conclusion, PCA was demonstrated to be the ideal tool to describe spectral data from THz absorbance experiments. The ternary diagram was perfectly retrieved, and the relative position between the samples was correct. This demonstrates that the THz absorbance spectra contain very good features to describe the samples, mixtures of three pure analytes. It should also be pointed out that the ternary diagram was retrieved without any knowledge of the data set.

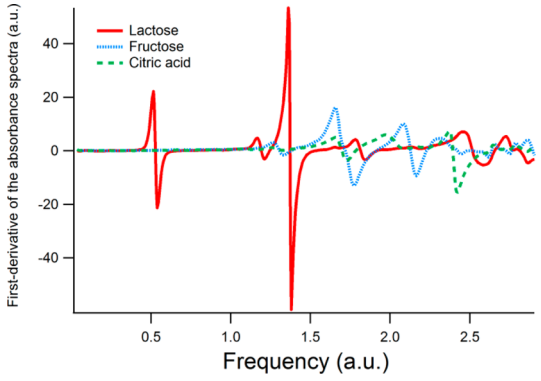


Figure 5. First derivative of the THz absorbance spectra for the three pure analytes, namely, fructose (blue), lactose (red), and citric acid (green).

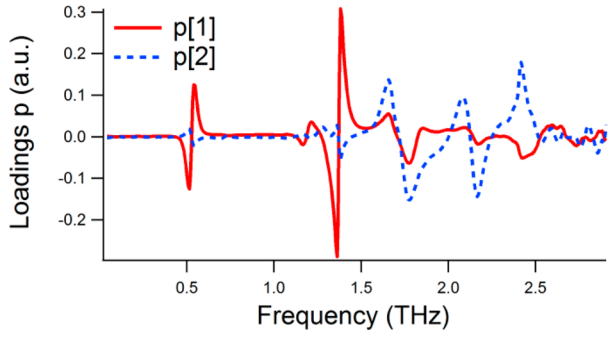


Figure 6. Loadings of the PCA model displayed on the original spectral range. p[1] for the first principal component and p[2] for the second one.

The results presented in Figure 4 allowed one to conclude that the PCA offers the possibility of semiquantitative analysis. By extension, it would be possible to quantify these samples by principle component regression (PCR)[35] which consists of calculating a regression upon the principal components. However, considering the question of quantitative analysis in the following section, we decided to use the partial least squares regression (PLS) instead of the PCR because PLS is recognized as being more efficient in interpreting the loadings and requires a low number of principle component.[35]

Quantitative Analysis by PLS. In this study, the samples contained up to three analytes in addition to the PE binder. Consequently, there were two strategies for quantitative analysis: either the quantification of one single analyte at a time or the simultaneous quantification of the three analytes. The PLS regression was utilized in this work. Thus, the PLS 1 algorithm was dedicated to the prediction of the concentrations of lactose, fructose, and citric acid one at a time, and the PLS 2 algorithm was dedicated to simultaneously predicting the

concentrations of these three analytes. Prior to any quantitative analysis, it is highly recommended for one to prepare the data in order to be able not only to build the best quantitative model but also to evaluate the performance of the model. On the basis of the PCA model, which revealed two outliers, namely, one of the two replicates of the samples A28 and A34, it was decided to finally exclude not only the two replicates of these samples but also more generally any sample displaying a large difference between the two replicates in the PCA model. Thus, the sample A23 was also excluded from the PLS analysis. After removing these three outliers, the original data set was split into three independent subsets namely, the calibration, the validation, and the test sets.

The calibration set was composed of 190 spectra, i.e., 10 spectra per sample for the 19 samples: A1 A2 A4 A5 A6 A7 A9 A10 A11 A12 A14 A15 A16 A17 A18 A21 A26 A38 A39. This calibration set was used to build the regression model. The validation set was made of 110 spectra, i.e., 10 spectra per sample for the 11 samples: A3 A8 A13 A19 A20 A22 A27 A29 A31 A33 A35. This validation set was used for external validation of the model. With these two subsets of data, it was possible to choose the model by minimizing the mean relative error of prediction. Finally, the test set was made of 60 spectra, i.e., 10 spectra per sample for the 6 samples: A24 A25 A30 A32 A36 A37. This test set was used to evaluate the ability of the model to predict, a posteriori, the concentrations of unknown samples. It should be noted that the terms of calibration, validation, and test sets have been adopted by Hamzaçebi et al.[36] In addition, in order to evaluate the performance of the model to achieve quantitative analysis, the root mean square error hereafter called RMSE was calculated. The definition that we adopted for RMSE is given by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \quad (1)$$

where $\hat{y}_i$ corresponds to the reference value of concentration of the sample i, y_i is the value predicted by PLS, and N is the number of samples.

In order to present values of RMSE that could be easily compared to the ones obtained from other studies, it is necessary to properly describe the samples. In this work, each 400 mg pellet contained 80 mg of analytes (pure or mixture), i.e., 20% of the weight of the pellet. Consequently, we have built quantitative models from values of concentrations given in %. Thus, the sample A6 for instance is related to 20% lactose and 80% PE while the sample A31 corresponds to 2% lactose, 12% citric acid, 6% fructose, and 80% PE. Consequently, since the predicted values of concentrations are calculated in %, RMSE is also given in %. However, the value of RMSE should be discussed in weight, i.e., in mg instead of % for a better understanding. As a consequence, RMSE = X% corresponds to RMSE = X mg for 80 mg of analytes (or 20%) and for 400 mg of pellet (or 100%). Finally, if interested in the total amount of the analytes instead of the complete pellet made of the mixture, analytes (20%) + binder (80%), the RMSE values would be multiplied by a factor of 5.

The values of RMSE are given in Table 1 in the case of different PLS models applied to the series of data described above. As a preprocessing step, the first derivative was applied and then the data were mean centered. In Table 1, one can find in the first column the type of model, here PLS 1, then the analyte (F: fructose; L: lactose; and CA: citric acid) in column 2, and then the spectral range of the absorbance spectra in THz. It also reports the number of variables K and the number of principal components A considered for the model. The data were introduced into the models following the index of the samples, namely, from A1 to A39. Finally, it reports the values of RMSE for the three sets of data (calibration set: C, validation set: V, and test set: T). The results presented in Table 1 correspond to three individual PLS 1 models applied to the 457 variables in the spectral range of 0.05−2.90 THz. It should be noted that the RMSE values for lactose were found to be very low, i.e., around 0.27% or less. This result was in good agreement with the previous study based on PCA, which revealed that lactose was very well described along the axis of the first component of PCA with a very good separation of the points related to their concentrations.

Table 1. Values of RMSE Calculated for Different PLS 1 Models[a]

					RMSE (%)		
model	analyte	range (THz)	K	A	C	V	T
PLS-1	F	0.05 2.90	457	3	0.52	1.06	0.78
PLS-1	CA	0.05 2.90	457	7	0.46	0.89	0.80
PLS-1	L	0.05 2.90	457	5	0.17	0.27	0.21

[a]The analytes are fructose (F), lactose (L), and citric acid (CA). C, V, and T stand for calibration (190 samples), validation (110 samples), and test (60 samples) sets, respectively. A designates the number of PLS components, and K is the number of variables per sample.

Quantitative Analysis by ANN. We also studied the advantage of using artificial neural networks for the quantitative analysis of the samples presented above. ANN is a well known nonlinear method of chemometrics.[37] In this study, we selected a 3 layer network composed of an input layer, a hidden layer, and an output layer. For the input layer, each neuron received one value per sample selected from the THz spectra. The output layer contained only one neuron giving the predicted value of concentration of the analyte as output value. The hidden layer was composed of an adjustable number of neurons interconnecting the neurons from the input and output layers. For each neuron, the activation function was the sigmoid function providing an output value ranging between 0 and 1. The learning step consisted of applying the data from the calibration set. Iteratively, the feed forward and the back propagation of the error algorithms were applied.

Practically, it was not possible to introduce all of the data from each spectrum into the ANN due to a dimensionality issue. Moreover, THz absorbance spectra often do not contain well defined peaks, and thus, it could be difficult to select a reduced number of significant data points. Consequently, we decided to apply first a PCA model in order to compress the original data to only a few significant data points. Indeed, PCA transformed the hundreds of spectral variables of the THz absorbance spectra into a few scores of the principal components. Finally, input values of the ANN models were the scores of the PCA model.

In the case of fructose, the spectral data from the range of 0.05−2.6 THz were processed by PCA. For each value of A, the number of principal components, three ANN models were calculated, one for each analyte, and the smallest values of RMSE determined the best models. Thus, for fructose, the best ANN model was obtained for A = 5; for lactose, the optimal ANN model was related to A = 3, and for citric acid, the best model was obtained for A = 3. The results of the three ANN

models are presented in Table 2. For each model, the number of neurons in the hidden layer is given, as well as the learning rate, momentum, and number of iterations optimized via external cross validation. For fructose, the values of RMSE were found around 0.7%. This result was obtained in the case of the spectral range of 0.05−2.60 THz. The analysis of lactose was identically based on a preliminary PCA model in order to use the scores of the PCA as input data for the ANN. In this case, the spectral data from the range of 0.05−2.90 THz were processed by PCA for the samples contained in the calibration set. We verified that the best results for predicting the concentration of lactose were obtained with the three first scores of PCA as input values of the ANN. The values of RMSE were found to be smaller than 0.34% for lactose. These values were two times smaller in the case of lactose than in the case of fructose. This result is in good agreement with the results that were obtained earlier by PCA and PLS. For citric acid, the spectral data from the range of 0.05−2.90 THz were processed by PCA; the scores of the first three components of PCA were utilized as input data of the ANN model, and the results are given in Table 2. In the case of citric acid, the RMSE values were lower than 0.9%. This result was not as good as the one obtained in the case of lactose but was still satisfactory in regards to the 400 mg of total weight of the samples corresponding to the value of 100% and to the 80 mg of analytes corresponding to the value of 20%. The results are given in Table 2.

Table 2. Results of the Three Optimized ANN Models Dedicated to the Quantitative Analysis of Fructose, Lactose, and Citric Acid[a]

		fructose	lactose	citric acid
spectral range (THz)		0.05 2.60	0.05 2.90	0.05 2.90
input data for ANN		5	3	3
neurons in the hidden layer		2	3	3
learning rate		0.05	0.1	0.05
momentum		0.2	0.1	0.1
number of iterations		12 000	54 000	18 000
calibration set	RMSE (%)	0.78	0.22	0.54
validation set	RMSE (%)	0.70	0.34	0.89
test set	RMSE (%)	0.49	0.33	0.47

[a]The number of input data is the number of principal components of the PCA model.

DISCUSSION

It should be noted that RMSE is a mean value, and consequently, it should be more informative to display for each sample the values of concentrations predicted by ANN and by PLS versus the reference values obtained by weighing out the powder of analytes and binder, as reported in Figure 7. From Figure 7, one can observe a very similar predictive ability between PLS and ANN, in the case of data processing of THz absorbance spectra for the mixtures prepared for this study. This reveals that no significant nonlinear behavior was present in this case. In the case of fructose (Figure 7a), the values of concentration predicted by ANN (blue) were found to be slightly closer to the curve $y = x$ (dashed lines) than the values predicted by PLS (red) especially for the sample A30 corresponding to 14% of fructose, which was underestimated by PLS. Figure 7 also displays error bars that help in the interpretation of the results. Regarding ANN and PLS, the error bars describe the relative standard deviation obtained after the analysis of the 10 replicates of the measurement per sample. However, in the case of fructose, only one sample was analyzed by the value of concentration, preventing the observation of possible matrix effect. In the case of lactose, the predicted values were very close to the reference ones and the error bars were very small compared to the ones obtained for the other analytes. Moreover, it should be emphasized that six samples composed the test set and that the concentration values of the samples A30 and A37 were correctly predicted to be close to 4% while those of the samples A24 and A25 were correctly predicted to be close to 8%. Finally, in the case of citric acid,

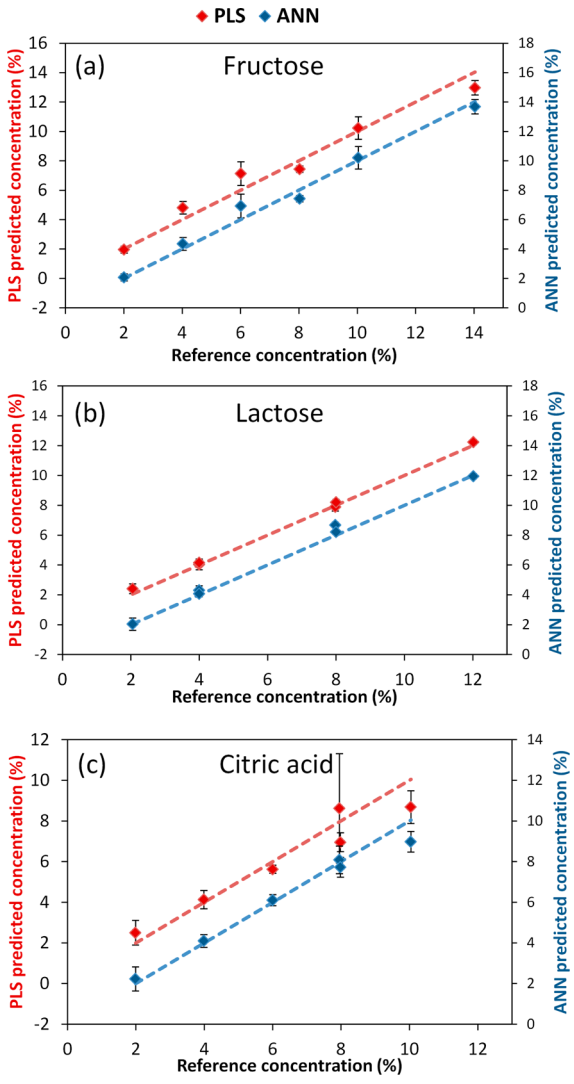


Figure 7. Predicted concentrations of fructose (a), lactose (b), and citric acid (c) by ANN (blue, right vertical axis) and PLS (red, left vertical axis) versus the reference values obtained by weighing out the powders. The results are given for samples belonging to the test set. The error bars represent the standard deviation over the 10 measurements per sample. Dashed lines depict the lines defined by the equation $y = x$.

one should notice that for the sample A37 corresponding to 10% of citric acid, both PLS and ANN underestimated the value of concentration. In addition, one should notice that two samples corresponding to 8% citric acid, namely, A25 and A32, revealed larger discrepancy between the predicted and reference values. This effect was more visible in the case of PLS analysis with large error bars for the sample A32 revealing possible coupling between citric acid and fructose.

Finally, no advantage for ANN was evidenced probably because the absorption spectra were driven by linear behaviors. The performance obtained by ANN strongly depends on the preliminary compression achieved by PCA. It is very interesting to observe that only 3 to 5 data, namely, the scores of PCA, were enough to build a very good ANN model dedicated to quantitative analysis.

CONCLUSION

We have successfully applied chemometrics to the analysis of ternary mixtures of fructose, lactose, and citric acid measured by transmission terahertz time domain spectroscopy. PCA was utilized with a very good efficiency to analyze terahertz data of ternary mixtures. Indeed, PCA allowed one to visualize the ternary diagram of the three initial analytes. Then, artificial neural network was used as a quantitative method and compared to the PLS. Both of these methods provided predicted values of concentrations for lactose, fructose, and citric acid characterized by RMSE values lower than 0.9% (4 mg in the case of pressed pellets of 400 mg containing 320 mg of polyethylene as binder and 80 mg of mixture). Finally, this work demonstrated the advantages of using chemometrics to treat terahertz absorbance spectra. Further work will be dedicated to the analysis of more complex samples by applying chemometrics. More generally, chemometrics should have a larger role in future developments and transfer of terahertz spectroscopy for chemical analysis as a new tool for industrial applications.

AUTHOR INFORMATION

Corresponding Author

*E mail: bruno.bousquet@u bordeaux.fr. Tel:+33 (0)5 40 00 28 70. Fax: +33 (0)5 40 00 69 70.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was sponsored by the "Agence Nationale de la Recherche" (ANR), within the framework of the InPoSec project (www.inposec.com)

REFERENCES

(1) Mittleman, D. M., Ed. *Sensing with Terahertz Radiation*; Springer: New York, 2003.
(2) Ferguson, B.; Zhang, X. C. *Nat. Mater.* **2002**, *1*, 26−33.
(3) Dexheimer, S. L., Ed. *Terahertz Spectroscopy: Principles and Applications*; CRC Press: Boca Raton, FL, 2008.
(4) May, R. K.; Evans, M. J.; Zhong, S.; Warr, I.; Gladden, L. F.; Shen, Y.; Zeitler, J. A. *J. Pharm. Sci.* **2010**, *100*, 1535−1544.
(5) Jepsen, P. U.; Cooke, D. G.; Koch, M. *Laser Photonics Rev.* **2011**, *5*, 124.
(6) Markelz, A.; Roitberg, A.; Heilweil, E. *Chem. Phys. Lett.* **2000**, *320*, 42−48.
(7) Andrea, M.; Scott, W.; Jay, H.; Robert, B. *Phys. Med. Biol.* **2002**, *47*, 3797.
(8) Oh, S. J.; Kang, J.; Maeng, I.; Suh, J. S.; Huh, Y. M.; Haam, S.; Son, J. H. *Opt. Express* **2009**, *17*, 3469−3475.
(9) Parrott, E. P. J.; Sun, Y.; Pickwell Macpherson, E. *J. Mol. Struct.* **2011**, *1006*, 66−76.
(10) Abraham, E.; Younus, A.; El Fatimy, A.; Delagnes, J. C.; Nguéma, E.; Mounaix, P. *Opt. Commun.* **2009**, *282*, 3104−3107.
(11) Baxter, J. B.; Guglietta, G. W. *Anal. Chem.* **2011**, *83*, 4342−4368.
(12) Strachan, C. J.; Taday, P. F.; Newnham, D. A.; Gordon, K. C.; Zeitler, J. A.; Pepper, M.; Rades, T. *J. Pharm. Sci.* **2005**, *94*, 837−846.
(13) King, M. D.; Hakey, P. M.; Korter, T. M. *J. Phys. Chem. A* **2010**, *114*, 2945−2953.
(14) Ueno, Y.; Rungsawang, R.; Tomita, I.; Ajito, K. *Anal. Chem.* **2006**, *78*, 5424−5428.
(15) Day, G. M.; Zeitler, J. A.; Jones, W.; Rades, T.; Taday, P. F. *J. Phys. Chem. B* **2005**, *110*, 447−456.
(16) Darkwah, J.; Smith, G.; Ermolina, I.; Mueller Holtz, M. *Int. J. Pharm.* **2013**, *455*, 357−364.
(17) Zeitler, J. A.; Kogermann, K.; Rantanen, J.; Rades, T.; Taday, P. F.; Pepper, M.; Aaltonen, J.; Strachan, C. J. *Int. J. Pharm.* **2007**, *334*, 78−84.
(18) El Haddad, J.; Bousquet, B.; Canioni, L.; Mounaix, P. *TrAC, Trends Anal. Chem.* **2013**, *44*, 98−105.
(19) Wu, H.; Heilweil, E. J.; Hussain, A. S.; Khan, M. A. *Int. J. Pharm.* **2007**, *343*, 148−158.
(20) Watanabe, Y.; Kawase, K.; Ikari, T.; Ito, H.; Ishikawa, Y.; Minamide, H. *Opt. Commun.* **2004**, *234*, 125−129.
(21) Zeitler, J. A.; Newnham, D. A.; Taday, P. F.; Threlfall, T. L.; Lancaster, R. W.; Berg, R. W.; Strachan, C. J.; Pepper, M.; Gordon, K. C.; Rades, T. *J. Pharm. Sci.* **2006**, *95*, 2486−2498.
(22) Nguyen, K. L.; Friščić, T.; Day, G. M.; Gladden, L. F.; Jones, W. *Nat. Mater.* **2007**, *6*, 206−209.
(23) Trafela, T.; Mizuno, M.; Fukunaga, K.; Strlič, M. *Appl. Phys. A: Mater. Sci. Process.* **2013**, *111*, 83−90.
(24) Hua, Y.; Zhang, H. *IEEE Trans. Microwave Theory Tech.* **2010**, *58*, 2064−2070.
(25) Ting, H.; Jingling, S.; Meiyan, L. *Measurement* **2011**, *44*, 391−398.
(26) Sirven, J. B.; Bousquet, B.; Canioni, L.; Sarger, L.; Tellier, S.; Potin Gautier, M.; Hecho, I. L. *Anal. Bioanal. Chem.* **2006**, *385*, 256−262.
(27) El Haddad, J.; Villot Kadri, M.; Ismaël, A.; Gallou, G.; Michel, K.; Bruyère, D.; Laperche, V.; Canioni, L.; Bousquet, B. *Spectrochim. Acta, Part B: At. Spectrosc.* **2013**, *79−80*, 51−57.
(28) Jun ichi, N.; Ken, S.; Tetsuo, S.; Tadao, T.; Tomoyuki, K. *J. Phys. D: Appl. Phys.* **2003**, *36*, 2958.
(29) Zheng, Z. P.; Fan, W. H.; Liang, Y. Q.; Yan, H. *Opt. Commun.* **2012**, *285*, 1868−1871.
(30) Cogdill, R.; Short, S.; Forcht, R.; Shi, Z.; Shen, Y.; Taday, P.; Anderson, C.; Drennen, J. *J. Pharm. Innovation* **2006**, *1*, 63−75.
(31) Fischer, B.; Hoffmann, M.; Helm, H.; Modjesch, G.; Jepsen, P. U. *Semicond. Sci. Technol.* **2005**, *20*, S246.
(32) Newnham, D. A.; Taday, P. F. *Appl. Spectrosc.* **2008**, *62*, 394−398.
(33) Franz, M.; Fischer, B. M.; Walther, M. *Appl. Phys. Lett.* **2008**, *92*, 021107.
(34) Duvillaret, L.; Garet, F.; Coutaz, J. L. *Appl. Optics* **1999**, *38*, 409−415.
(35) Wentzell, P. D.; Vega Montoto, L. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 257−279.
(36) Hamzaçebi, C.; Akay, D.; Kutay, F. *Expert Syst. Appl.* **2009**, *36*, 3839−3844.
(37) Marini, F.; Bucci, R.; Magrì, A. L.; Magrì, A. D. *Microchem. J.* **2008**, *88*, 178−185.