

Identification of Unknown Substances by Terahertz Spectroscopy and Multivariate Data Analysis

Andreas Pohl^{1,2} · Nils Deßmann^{1,2} · Katja Dutzi² · Heinz-Wilhelm Hübers^{1,2}

Received: 15 April 2015 / Accepted: 12 October 2015 /
Published online: 27 October 2015
© Springer Science+Business Media New York 2015

Abstract The identification of various substances by multivariate data analysis of terahertz transmittance spectra is demonstrated. Transmittance spectra were obtained by the use of a Fourier transform infrared spectrometer. By means of principal component analysis and partial least squares regression, the spectral data were analyzed in order to identify substances and mixtures of several substances. With only three principal components, detection and identification of substances are possible with high accuracy. Using these methods, concentration ratios of substances in mixtures of two substances can be determined with an accuracy of 10 %. It is shown that the method is robust against disturbances in the spectra such as standing waves. This is particularly important for practical applications.

Keywords Substance identification · Spectroscopy · Multivariate data analysis · Principal component analysis · Partial least squares regression · Terahertz

1 Introduction

Terahertz (THZ) radiation has much promise for security applications in terms of sensing and imaging. For example, the detection of hidden threats or harmful substances can prevent terrorist attacks at places where a large number of people congregate, such as airports. Due to many chemical compounds that have characteristic absorptions in the THz frequency range and the fact that common packaging materials are transparent throughout a large portion of the THz frequency range, various studies were made regarding the applicability of THz radiation to security. In addition, THz radiation is supposed to be not harmful for human beings due to

✉ Andreas Pohl
a.pohl@dlr.de

¹ Department of Physics, Humboldt University zu Berlin, Newtonstr. 15, 12489 Berlin, Germany

² Institute of Optical Sensor Systems, German Aerospace Center (DLR), Rutherfordstraße 2, 12489 Berlin, Germany

its low photon energy. This means photons cannot break chemical bonds or ionize atoms or molecules [1, 2]. The potential offered by THz radiation is examined for example by imaging through packaging materials [3, 4], standoff detection of explosives [4–11], chemical sensing of drugs [4, 9, 12], substance identification [13, 14], or textile identification [15]. Moreover, it leads to medical applications like the detection of cancerous tissue caused by a locally higher water content [16].

In recent years, multivariate analysis (MVA) techniques such as principal components analysis (PCA) or partial least squares regression (PLS) have been applied to THz spectra of a variety of substances. These spectra were obtained with time-domain spectrometers (TDS). Burnett et al. employed PCA in order to investigate the feasibility of automatic spectral recognition of illicit materials by THz TDS [14]. Bardon et al. applied PCA to differentiate between different historically informed black inks [17]. Ermolina et al. used PLS for estimating the degree of crystallinity in sucrose [18], and El Haddad et al. applied PCA to the quantitative analysis of ternary mixtures of citric acid, fructose, and lactose [19]. Neumaier et al. applied MVA techniques for the first time to THz spectra of non-solid samples, namely, to pure gasses and gas mixtures. Instead of TDS, they used a THz spectrometer with coherent transmitter and receiver [20].

In the study presented in this paper, we focus on the analysis of THz transmittance spectra using PCA and PLS. In total, nine substances and three mixtures were measured with a Fourier transform infrared (FTIR) spectrometer. In particular, we investigate the robustness of the methods against disturbances such as standing waves in the spectrum.

2 Experimental

2.1 Sample Preparation and Experimental Setup

In this study, nine substances (KCl, NaCl, NH_4CO_3 , K_2CO_3 , sugar, fertilizer, acetylsalicylic acid, chocolate, fruit chew) and three mixtures of acetylsalicylic acid and sugar with ratios of 1:1, 1:3, and 3:1 were investigated. The samples were chosen in order to represent explosives as well as food products, which are commonly carried along by persons. All substances were ground to fine powders except chocolate and fruit chew, which were prepared as thin layers. For each spectrum, a new sample was prepared. The mixtures were prepared in a large amount from which the samples for the measurements were prepared.

The transmittance spectra were measured using a BrukerVertex 80v FTIR spectrometer. During the experiment, the FTIR spectrometer was evacuated to a residual pressure below 2 hPa. A water-cooled mercury arc lamp was used as radiation source along with a broadband Mylar multilayer beam splitter of 6 μm thickness. The interferograms were recorded with a sensitive liquid helium-cooled silicon bolometer from Infrared Laboratories, Inc. The frequency resolution in the experiments was 1 cm^{-1} or approximately 30 GHz. As acquisition mode, we used double-sided, forward-backward scanning that creates a full interferogram in the forward scanning direction and a second mirrored interferogram for the backward direction. The scanning mirror frequency was set to 40 kHz. One hundred twenty-eight scans were averaged to one single interferogram resulting in a measurement time of around 100 s. Spectra were obtained in the range from 0.45 to 3 THz (15 to 100 cm^{-1}) in order to cover the most common frequency range for spectral fingerprints [8, 9, 12, 15] and to be able to penetrate clothing and packaging materials. A 100 cm^{-1} optical lowpass filter inside the detector cryostat was used to block higher frequencies. The spectra were obtained by Fourier transformation

with a Blackman-Harris 3-term apodization function and by using the power spectrum. During the measurement, the samples were fixed in a dedicated sample holder between two high-density polyethylene (HDPE) plates with a thickness of 1 mm. The amount of material used for the measurement of one substance was approximately 0.1 g. Thus, a sample thickness of approximately 0.05 mm across the area of the window of the sample holder ($30 \times 40 \text{ mm}^2$) was given. The holder was positioned in the center of the sample compartment of the spectrometer with a tilt angle of 30° to the incident radiation. The sample holder was tilted with respect to the optical axis in order to reduce standing waves.

2.2 Spectra and Fundamental Characteristics

The measurements resulted in transmittance spectra like the example shown in Fig. 1. As reference, a spectrum of the empty sample holder was collected separately. The transmittance spectra were obtained from the sample spectra divided by the reference spectra. Figure 1 shows the transmittance of sugar in the range from 15 to 100 cm^{-1} . Despite the tilt of the sample holder, the transmittance spectrum suffers from strong interference fringes, which partially mask the absorption features of the sugar. However, they are much less pronounced as if the sample holder was perpendicular to the incident radiation. The fringes are caused by the thin HDPE windows, where the radiation exhibits multiple internal reflections. It should be noted that calculating a spectral ratio against an empty absorption cell did not remove the fringes satisfactorily. In Fig. 2, the corresponding interferogram is shown. It has two pronounced sidebursts, left and right to the centerburst. To remove the fringes in the spectrum, the corresponding sidebursts were substituted by their mean values, which is usually close to zero [21]. The cut-out window was selected for only one side of the interferogram while the other side was cut symmetrically. By monitoring the resultant spectra as a function of size and position, the cut-out window was chosen as the best trade-off between maximum fringe elimination and minimum data removal. The whole cut-out procedure was then repeated automatically for all spectra via a self-programmed algorithm in LabVIEW. The sideburst removal goes along with a slight reduction of the spectral resolution, because some information from the centerburst is lost as well. However, the fringes can be almost completely removed by this technique (Fig. 1) and only some fringes remain below

Fig. 1 Comparison of the transmission spectra derived from the original and the corrected data using the sugar sample

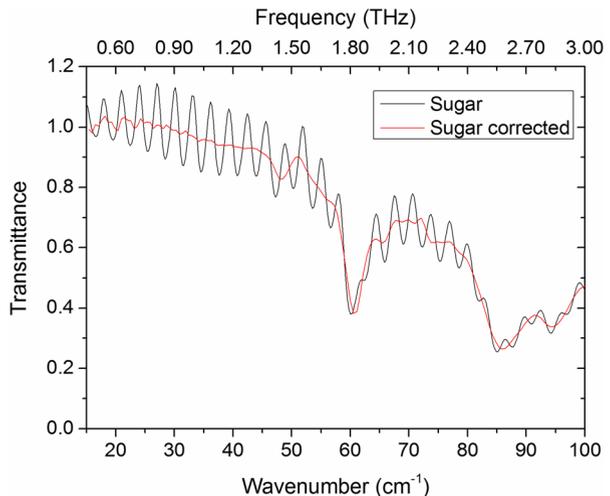
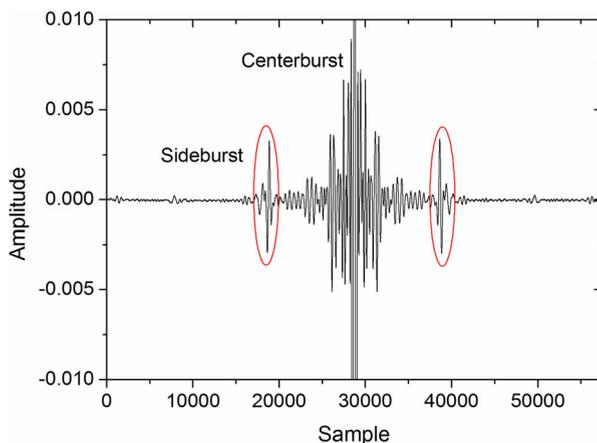


Fig. 2 The interferogram of sugar is shown. By removing the sidebursts, the interference fringes in the transmittance spectra are eliminated



30 cm^{-1} . In the further analysis, all data were treated by this method, i.e., exactly the same part of the interferogram was replaced by its average. This ensures that the procedure has no influence on the multivariate data analysis. No further data manipulation was done. In the following, datasets with removed fringes are referred to as “corrected data”. It is worth noting that although this study was done under laboratory conditions, this technique should also lead to good results in field studies with standing waves of unknown origin. Other methods of fringe removal are the grafting procedure, modeling the interference fringe and subtract from the spectral data, or by using a polarized beam source and mount the sample at the Brewster angle [22]. Modeling the interference by a sinusoidal function was tested, but could not provide as good results as the sideburst-removal did.

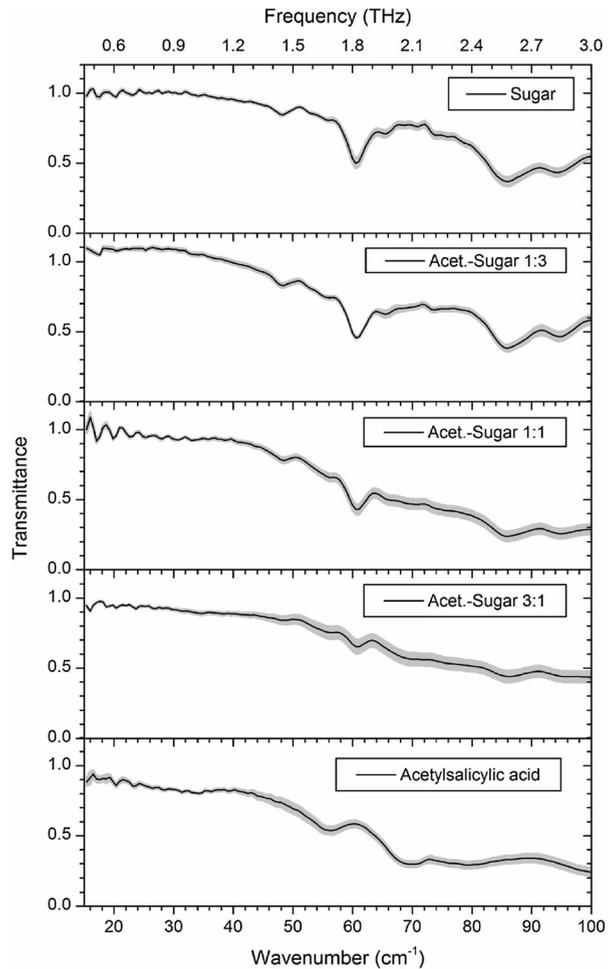
In Fig. 3, a comparison between the transmittance spectra of sugar, acetylsalicylic acid, and its three mixtures (1:3, 1:1, 3:1) is shown. The spectrum of sugar has clearly visible features at 48 , 60 , 85 , and 95 cm^{-1} whereas the spectrum of acetylsalicylic acid has only two representative features at 56 and 68 cm^{-1} . Within the three different mixtures, the features are combined according to the amount of the pure sugar and acetylsalicylic acid. This is due to the absence of chemical reactions between the two substances. For substances that would cause some kind of chemical reaction, the obtained spectral data would not contain information about the reactants but about the reaction product.

In order to provide information about the reproducibility of the spectra of a specific substance, the averaged spectrum is shown in black with a two-sided 95 % Student’s *t* confidence band (gray) for the 10 measured samples. Transmittance spectra of the other samples are shown in Fig. 4. With the exception of chocolate and NH_4CO_3 , there are no distinct absorption features in the spectra of these samples. One feature, which all samples have in common, is that the transmittance decreases with increasing frequency. This is caused by increasing absorption and increasing scattering losses of the transmitted THz wave [8, 23, 24].

3 Multivariate Analysis

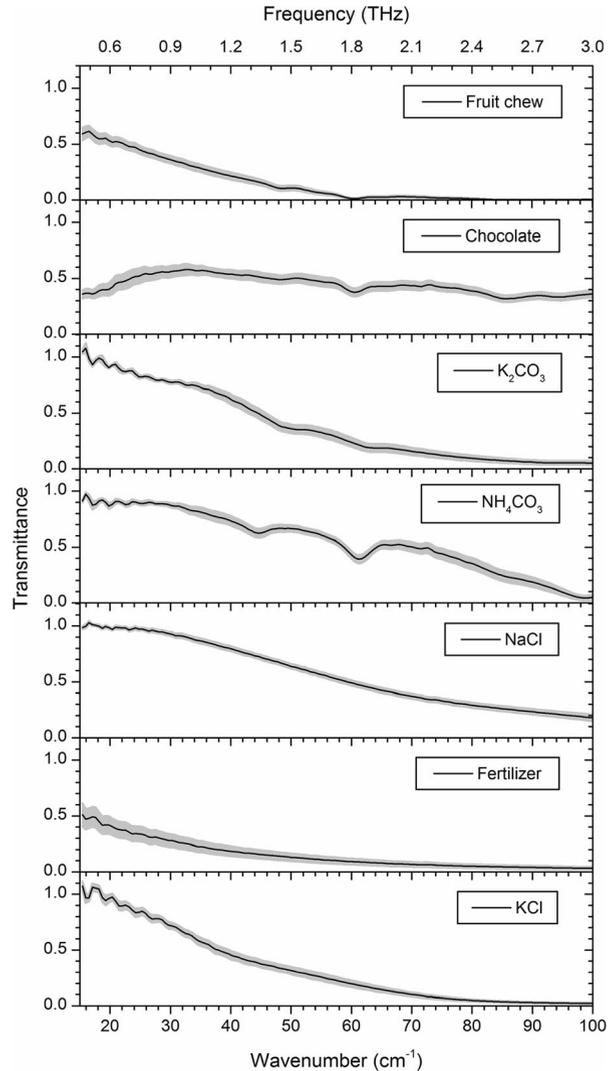
The data analysis was done by using the multivariate analysis (MVA) tool The Unscrambler X (Camo Software AS) [25, 26]. Prior to any analysis, all measured spectral data were merged into a single data file, which contains all information in a form of a matrix. A spectrum resulted

Fig. 3 Averaged transmittance spectra (corrected, black line) of pure sugar (top), three mixtures of acetylsalicylic acid and sugar (1:3, 1:1, 3:1) and pure acetylsalicylic acid (bottom) within the two-sided 95 % Student's t confidence band. The appearance of absorption features from both substances in the mixtures is clearly visible



in 155 data points in the range of 15 to 100 cm^{-1} . Each of the 12 samples was measured 10 times. Hence, the resulting data matrix has 120 rows and 155 columns. The transmission spectra were used in raw format without any pre-processing except for eliminating the sidebursts in the interferograms as described above. This ensures minimal influence on the forthcoming MVA. By using one or more pre-processing steps, the inherent apparatus function would be distorted and therefore a misinterpretation of the data is possible. The data were analyzed by PCA [27] and PLS. PCA is an MVA method, which contains only the spectral information of the measurements. It represents the solution of an eigenvalue problem. The approach is a projection of the input data to a new orthogonal basis by determining new basis vectors (principal components, PCs) in correlation to the maximum variance in the data. Mathematically, it is an algorithm for orthogonal linear transformation of data according to their variance (see for example reference [25]). The transformation is defined as $X=TP^T+R$, with X the data matrix, T the scores matrix, P^T the transposed loadings matrix, and R the residual matrix. The transformation itself does not produce a data reduction but a subspace of the new coordinate system contains the relevant information, which is necessary to separate

Fig. 4 Transmittance spectra (*corrected*) of the samples that were used in this study besides the sugar and acetylsalicylic acid mixtures shown in Fig. 3



between the samples. Higher-order PCs contain the remaining information. The second method is the PLS, which is a supervised regression technique that links spectral data X (a matrix containing the spectra) with associated dependent variables Y (a matrix containing for example class membership or the percentage composition of samples) for classification. PLS utilizes the information in the Y data to find latent variables (LVs) in the X data that will best predict the Y data. Therefore, the data matrix Y contains the sample class membership in binary form for non-members (0) and members (1) or any other representative measurable characteristics like the concentration of substances in a mixture. In comparison to the PCA, the derived LVs are similar to the PCs but describe only spectral characteristics and variations, which are necessary for identification of the samples. Spectral variations that have no effect on the Y data are neglected. After establishing a PLS model, a sample prediction based on this model was done for a set of unknown spectra.

3.1 Principal Component Analysis

The PCA was done with both the raw data and the corrected dataset of 10 measurements for 12 substances and mixtures. We used a maximum of seven PCs for the calculation but with already three PCs nearly 98 % of the data variance can be explained by the PCA model within the corrected data and nearly 97 % within the original data. Figures 5 and 6 show the two-dimensional score plots of the PCA for the first three PCs of the raw data and the corrected

Fig. 5 Results of the PCA analysis of the raw dataset. Shown are the scores of the first and second PC (*top*), first and third PC (*middle*) and second and third PC (*bottom*). The variance in the data explained by the appropriate PC is given in *brackets*

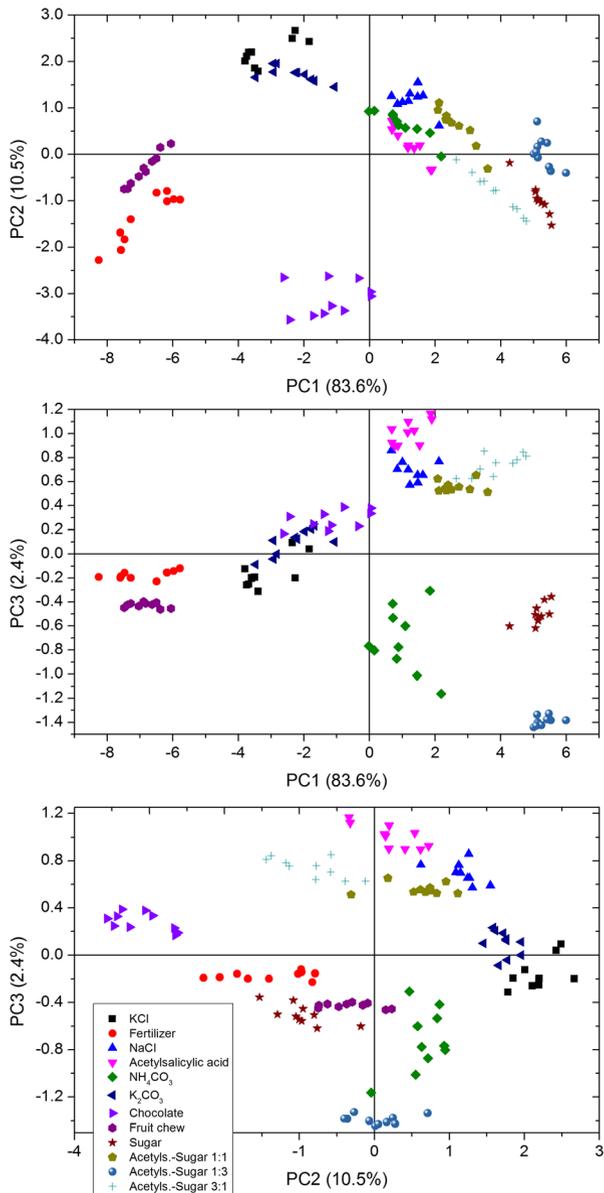
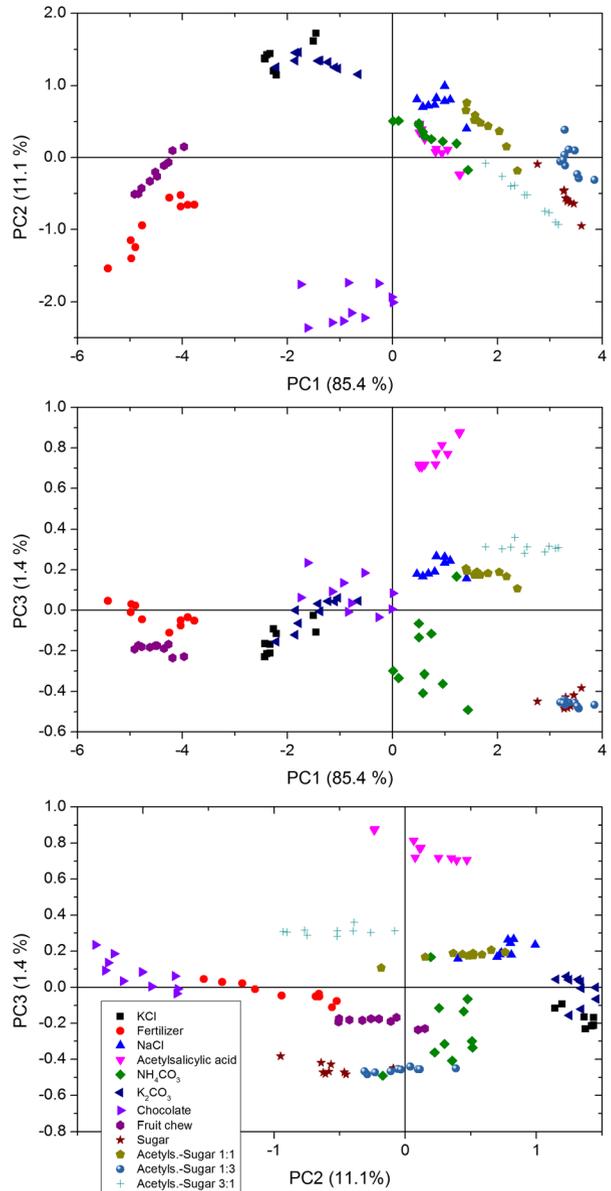


Fig. 6 Results of the PCA analysis of the corrected dataset without any other data manipulation. Shown are the scores of the first and second PC (*top*), first and third PC (*middle*) and second and third PC (*bottom*). The variance in the data explained by the appropriate PC is given in brackets



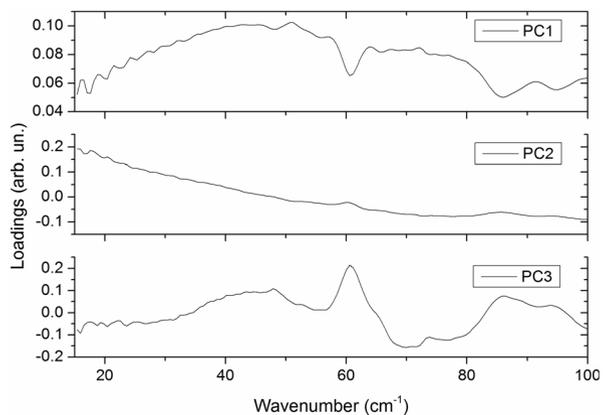
data. Each point represents a single spectrum. As can be seen, spectra of the same substance cluster in the score plots of the first three PCs. It is worth noting that despite the strong standing waves in the spectra obtained from the uncorrected interferograms, the differences between the two PCA models are very small, and the clustering is almost identical. Also, the variance is almost the same independently whether the original data or the corrected data were used for the PCA. This demonstrates that the standing waves in the spectrum do not have a significant influence on the MVA and the identification of the substance. To better explain the

model results and to identify which spectral characteristics mostly influence the clustering, we refer to the corrected data. The loadings for the corrected data are shown in Fig. 7. In PC1, the sample separation is achieved mainly due to the strong absorption features in the spectrum of sugar. These explain 85 % of the variance. In the score plots PC1 vs. PC2 and PC1 vs. PC3, the spectra that contain these features appear on the right side due to their positive correlation. Spectra that do not correlate or are anticorrelated appear in the middle or on the left, respectively. The separation in PC2 is primarily based on the frequency-dependent transmittance. PC2 basically resembles the decreasing transmittance of the samples with increasing wavenumber. The absorption features of sugar and acetylsalicylic acid are the main features that form the separation in PC3. Besides that, the sugar features are inverted which leads to an anticorrelation. Therefore, samples that contain sugar yield negative score values. The samples containing acetylsalicylic acid result in positive score values. Higher dimensions of the PCs lead to a slightly better explanation of the remaining substances but with minor influence in the total spectral variation. The use of seven PCs in the PCA model results in an overall explanation of 99.8 %, which is only a marginal improvement compared to the 97.9 % explanation with three PCs. Despite the small improvement of the model explanation that occurs by increasing the dimension of the PCA, the separation of individual substances may occur in the higher PCs.

The explanation based on scores alone does not provide any information about the localization and separation of clusters containing the same samples with respect to other clusters or samples. Therefore, a cluster analysis of the PCA results is necessary. To calculate a scalar measure of the statistical cluster spreading, we assume that the scattering of the appropriate measurement parameters like temperature, pressure, sample thickness, is described by a normal distribution. Therefore, the score clusters are expected to be described by a normal distribution as well. We determined the distances of each score value $S_{A,k;m}$ of every substance A to the cluster centroid $S_{A;k}$, with m the number of measured spectra and k the PC dimension. The standard deviation $\sigma_{A;k}$ of these distances is then calculated according to

$$\sigma_{A,k} = \sqrt{\frac{1}{m} \sum_m (S_{A,k,m} - S_{A,k})^2}. \tag{1}$$

Fig. 7 The loadings of the first three PCs indicate those spectral features which are responsible for the separation in the appropriate PC



We then calculated the amount of overlap $I_{AB,k}$ (the so-called Bhattacharyya coefficients [28]) of the normal distributions $N_{A,k}$ and $N_{B,k}$ for each substance A to all other substances B and each PC k :

$$I_{AB,k} = I_{BA,k} = \int_{-\infty}^{+\infty} \sqrt{N_{A,k}(x)N_{B,k}(x)} dx. \quad (2)$$

This is a measure of the cluster overlap and therefore defines the probability that the measured spectrum of a substance, which belongs to cluster A , cannot be separated from the cluster B in the corresponding PC dimension. The probability of a cluster separation \mathbf{S} is defined by:

$$\mathbf{S} = 1 - \prod_k I_{A,B,k}. \quad (3)$$

The matrix \mathbf{S} is symmetric due to the reversibility of S_{AB} from cluster A to B . Also, diagonal elements are zero, because same clusters are not separable. The separation matrix \mathbf{S} is shown in Table 1.

With three PCs, the separability is very good and reaches 100 % for most of the substance combinations (Table 1). Due to the fact that independent spectral features are described by different PCs, the separability of two clusters (or substances) does not increase continuously by increasing the dimensionality. This means that substances with similar spectral features separate mostly in the PC that describes their spectral differences. For example, KCl and K_2CO_3 are mainly separated by PC5 (total separation 100 %), whereas three PCs can only achieve a cluster separation of 64 %, and 4 PCs reach a total separation of 75 %.

Table 1 Cluster separation probability (in %) calculated for three PCS used in the PCA model. The substances are numbered from 1 to 12 with the assignment below

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	100	100	100	100	64	100	100	100	100	100	100
2	100	0	100	100	100	100	100	95	100	100	100	100
3	100	100	0	100	92	100	100	100	100	80	100	100
4	100	100	100	0	100	100	100	100	100	100	100	100
5	100	100	92	100	0	100	100	100	100	94	100	100
6	64	100	100	100	100	0	100	100	100	100	100	100
7	100	100	100	100	100	100	0	100	100	100	100	100
8	100	95	100	100	100	100	100	0	100	100	100	100
9	100	100	100	100	100	100	100	100	0	100	63	100
10	100	100	80	100	94	100	100	100	100	0	100	100
11	100	100	100	100	100	100	100	100	63	100	0	100
12	100	100	100	100	100	100	100	100	100	100	100	0

1 KCl, 2 fertilizer, 3 NaCl, 4 acetylsalicylic acid, 5 NH_4CO_3 , 6 K_2CO_3 , 7 chocolate, 8 fruit chew, 9 sugar, 10 acetylsalicylic acid–sugar 1:1, 11 acetylsalicylic acid–sugar 1:3, 12 acetylsalicylic acid–sugar 3:1

3.2 Partial Least Squares Regression

A PLS model was built by using eight spectra of the dataset for each substance as reference data. The remaining two spectra are used for the prediction process as “unknown” substances. The matrix Y was defined in non-discrete form, which means that in addition to the reference spectra also ratios of pure substances that build the components of mixtures were specified. Therefore, the value 0 is assigned when a substance was not in the mixture, and the value 1 is assigned to a pure substance. Sugar for example is part of chocolate (0.48 for 48 %) and fruit chew (0.60 for 60 %). The sugar-acetylsalicylic acid mixtures are 0.25 (sugar) and 0.75 (acetylsalicylic acid) for the 1:3 mixture, 0.5 and 0.5 for the 1:1, 0.75 and 0.25 for 3:1. In analogy to the PCA, the PLS model was built with seven LVs. The scores and loadings obtained from the PLS are similar to those of the PCA except that eight instead of 10 spectra were used to generate the PLS model. The corresponding cluster analysis based only on the first three LVs or PCs shows the similarities between the PLS (Table 2) and the PCA (Table 1). It is worth noting that despite the good cluster separation when only the first three LVs are used, the PLS explains only 64.7 % of the data variance. This means that information in higher dimensions needs to be considered in order to generate a good PLS model. On the other hand, too many dimensions would lead to a so-called overfitting. That means by starting with an optimal model, additional dimensions would only describe remaining noise in the data and therefore the prediction error increases. To obtain an optimal model, the prediction error should be minimal (namely, root mean square error for calibration (RMSEC), [26]). The RMSEC is calculated from the reference Y values and the corresponding predicted values. In our case, a mean RMSEC value of 17 % is given within all substances whereas the maximum RMSEC is given for NaCl with 24 %. To reduce the prediction error, another PLS model with 12 LVs was built. This equals the number of samples in the study and resulted in an explanation of 86.6 % and a mean RMSEC of 10 % (maximum RMSEC of 14 % in the case of K_2CO_3).

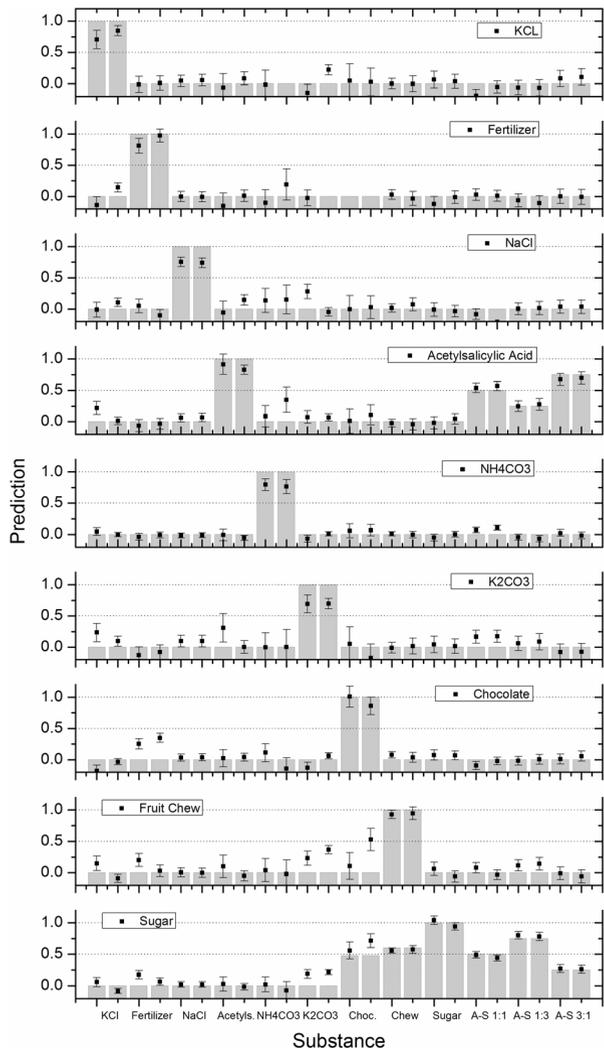
Table 2 PLS cluster separation probability for three LVS. The substances are numbered from 1 to 12 with the assignment below

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	100	100	100	100	56	100	100	100	100	100	100
2	100	0	100	100	100	100	100	100	100	100	100	100
3	100	100	0	100	100	100	100	100	100	100	100	100
4	100	100	100	0	100	100	100	100	100	100	100	100
5	100	100	100	100	0	100	100	100	100	100	100	100
6	56	100	100	100	100	0	100	100	100	100	100	100
7	100	100	100	100	100	100	0	100	100	100	100	100
8	100	100	100	100	100	100	100	0	100	100	100	100
9	100	100	100	100	100	100	100	100	0	100	74	100
10	100	100	100	100	100	100	100	100	100	0	100	100
11	100	100	100	100	100	100	100	100	74	100	0	100
12	100	100	100	100	100	100	100	100	100	100	100	0

1 KCl, 2 fertilizer, 3 NaCl, 4 acetylsalicylic acid, 5 NH_4CO_3 , 6 K_2CO_3 , 7 chocolate, 8 fruit chew, 9 sugar, 10 acetylsalicylic acid–sugar 1:1, 11 acetylsalicylic acid–sugar 1:3, 12 acetylsalicylic acid–sugar 3:1

Predictions for the tested samples are shown in Fig. 8. The first result of this analysis is that the sample identification works well for the pure substances, even though not all pure substances are predicted with the ideal value of 1. Samples that do not contain the substance are predicted with values lower than 0.5. The second result is that the PLS model can identify concentration ratios of substances in powder or grainy mixtures. For example, sugar is identified in the three sugar-acetylsalicylic acid mixtures with a mean value of about 26 %± 6 %, 47 %±5 %, and 79 %±6 % (nominal concentrations: 25 %, 50 %, and 75 % sugar, a concentration error of 5 % is expected due to inhomogeneities in the mixture). Likewise, the prediction for acetylsalicylic acid in the mixtures is about 68 %±9 %, 55 %±7 %, and 26 %± 8 % (nominal concentrations: 75 %, 50 %, and 25 %, 5 % error). In chocolate, sugar is identified with a predicted concentration of about 63 %±12 % (48 % according to the manufacturer specification), in chew of about 57 %±5 % (60 % manufacturer specification).

Fig. 8 PLS prediction based on 12 LVs for two spectra of each. Grey bars show the nominal fraction of a substance in the samples. Scatters give the predicted values with corresponding prediction errors. An ideal prediction has a probability of 1 for a pure substance whereas predictions of samples not containing the substance have a probability of 0. In case of a mixture, the prediction yields the concentration of a particular substance in the mixture. For example, the fraction of sugar in fruit chew is 0.6. In our model, the average deviation of the predicted substance concentrations compared to the real value is 9.7 %



4 Summary and Conclusion

We investigated a set of nine substances and three mixtures of two pure substances by utilizing FTIR spectroscopy. Two MVA techniques, PCA and PLS, were used to analyze the spectral data in the range from 0.45 to 3 THz. Despite the weak spectral THz features and strong standing waves in the spectra, the substances can be distinguished from each other by applying PCA. A cluster separation based on the overlap between normal distributions (Bhattacharyya coefficient) was applied to calculate the separation probability of substance clusters from each other. Based only on the first three PCs, the separation is already excellent, and differentiation of different substances and mixtures is easily possible. With PLS, a model prediction where all substances can be assigned to the correct class was obtained. Beyond that, concentration ratios of pure substances within mixtures can be determined with an average deviation from the real value of 9.7 %. The results demonstrate the applicability of FTIR spectroscopy for the identification of various substances even if they have only weak THz absorption features and if the spectra are strongly distorted by artifacts such as standing waves.

Acknowledgments This work was founded by the German Federal Ministry of Education and Research (Grant No. 13 N12022). A. Pohl, N. Deßmann, and K. Dutzi acknowledge support by the Helmholtz Research School on Security Technologies.

References

1. S. J. Jo, S.-Y. Yoon, J. Y. Lee, K.-T. Kim, S. Jung, J. Park, G.-S. Park, W.-Y. Park, and O. Kwon, “Biological effects of femtosecond-terahertz pulses on C57BL/6 mouse skin,” *Ann. Dermatol.*, vol. 26, no. 1, pp. 129–132, Feb. 2014.
2. M. R. Scarfi, M. Romanò, R. Di Pietro, O. Zeni, a Doria, G. P. Gallerano, E. Giovenale, G. Messina, a Lai, G. Campurra, D. Coniglio, and M. D’Arienzo, “THz exposure of whole blood for the study of biological effects on human lymphocytes,” *J. Biol. Phys.*, vol. 29, no. 2–3, pp. 171–6, Jun. 2003.
3. N. Rothbart, H. Richter, M. Wienold, L. Schrottke, H. T. Grahm, and H. W. Hubers, “Fast 2-D and 3-D terahertz imaging with a quantum-cascade laser and a scanning mirror,” *IEEE Trans. Terahertz Sci. Technol.*, vol. 3, no. 5, pp. 617–624, 2013.
4. J. F. Federici, B. Schulkin, F. Huang, D. Gary, R. Barat, F. Oliveira, and D. Zimdars, “THz imaging and sensing for security applications—explosives, weapons and drugs,” *Semicond. Sci. Technol.*, vol. 20, no. 7, pp. S266–S280, Jul. 2005.
5. R. Beigang, S. G. Biedron, S. Dyjak, F. Ellrich, M. W. Haakestad, D. Hübsch, T. Kartaloglu, E. Ozbay, F. Ospald, N. Palka, U. Puc, E. Czerwinska, A. B. Sahin, A. Sešek, J. Trontelj, A. Švigelj, H. Altan, A. D. van Rheenen, and M. Walczakowski, “Comparison of terahertz technologies for detection and identification of explosives,” *Proc. SPIE*, 2014, vol. 9102, p. 91020C.
6. K. Choi, T. Hong, K. Ik Sim, T. Ha, B. Cheol Park, J. Hyuk Chung, S. Gyeong Cho, and J. Hoon Kim, “Reflection terahertz time-domain spectroscopy of RDX and HMX explosives,” *J. Appl. Phys.*, vol. 115, no. 2, p. 023105, Jan. 2014.
7. N. Palka, “Identification of concealed materials, including explosives, by terahertz reflection spectroscopy,” *Opt. Eng.*, vol. 53, no. 3, p. 031202, Dec. 2013.
8. M. Ortolani, J. S. Lee, U. Schade, and H.-W. Hübers, “Surface roughness effects on the terahertz reflectance of pure explosive materials,” *Appl. Phys. Lett.*, vol. 93, no. 8, p. 081906, 2008.
9. L. Ho, M. Pepper, and P. Taday, “Terahertz spectroscopy: Signatures and fingerprints,” *Nat. Photonics*, vol. 2, no. September, pp. 541–543, 2008.
10. M. R. Leahy-Hoppa, M. J. Fitch, X. Zheng, L. M. Hayden, and R. Oslander, “Wideband terahertz spectroscopy of explosives,” *Chem. Phys. Lett.*, vol. 434, no. 4–6, pp. 227–230, Feb. 2007.
11. L. Zhang, H. Zhong, C. Deng, C. Zhang, and Y. Zhao, “Terahertz wave reference-free phase imaging for identification of explosives,” *Appl. Phys. Lett.*, vol. 92, no. 9, p. 091117, 2008.

12. K. Kawase, Y. Ogawa, Y. Watanabe, and H. Inoue, “Non-destructive terahertz imaging of illicit drugs using spectral fingerprints,” *Opt. Express*, vol. 11, no. 20, pp. 2549–54, Oct. 2003.
13. W. Xie, J. Li, and J. Pei, “THz-TDS signal analysis and substance identification via the conformal split,” *Sci. China Inf. Sci.*, vol. 55, no. 1, pp. 49–63, Dec. 2012.
14. A. D. Burnett, W. Fan, P. C. Upadhyaya, J. E. Cunningham, M. D. Hargreaves, T. Munshi, H. G. M. Edwards, E. H. Linfield, and A. G. Davies, “Broadband terahertz time-domain spectroscopy of drugs-of-abuse and the use of principal component analysis,” *Analyst*, vol. 134, no. 8, pp. 1658–68, Aug. 2009.
15. M. Naftaly, J. F. Molloy, G. V. Lanskiy, K. a. Kokh, and Y. M. Andreev, “Terahertz time-domain spectroscopy for textile identification,” *Appl. Opt.*, vol. 52, no. 19, pp. 4433–7, Jul. 2013.
16. C. S. Joseph, A. N. Yaroslavsky, V. A. Neel, T. M. Goyette, and R. H. Giles, “Continuous wave terahertz transmission imaging of nonmelanoma skin cancers,” *Lasers Surg. Med.*, vol. 43, no. 6, pp. 457–462, 2011.
17. T. Bardon, R. K. May, P. F. Taday, and M. Strlič, “Systematic study of terahertz time-domain spectra of historically informed black inks,” *Analyst*, pp. 4859–4869, 2013.
18. I. Ermolina, J. Darkwah, and G. Smith, “Characterisation of crystalline-amorphous blends of sucrose with terahertz-pulsed spectroscopy: the development of a prediction technique for estimating the degree of crystallinity with partial least squares regression,” *AAPS PharmSciTech*, vol. 15, no. 2, pp. 253–260, 2013.
19. J. El Haddad, F. De Miollis, J. Bou Sleiman, L. Canioni, P. Mounaix, and B. Bousquet, “Chemometrics applied to quantitative analysis of ternary mixtures by terahertz spectroscopy,” *Anal. Chem.*, vol. 86, no. 10, pp. 4927–4933, 2014.
20. P. F.-X. Neumaier, K. Schmalz, J. Borngräber, R. Wylde, and H.-W. Hübers, “Terahertz gas-phase spectroscopy: chemometrics for security and medical applications,” *Analyst*, vol. 140, no. 1, pp. 213–222, 2015.
21. S. L. Hyland, D. G. Ast, and A. Baghdadi, “Oxygen measurements in thin ribbon silicon,” *J. Cryst. Growth*, vol. 82, no. 1, pp. 191–196, 1987.
22. P. R. Griffiths and J. A. de Haseth, *Fourier Transform Infrared Spectrometry*, 2nd ed. Wiley-Interscience, 2007.
23. M. Kaushik, B. W.-H. Ng, B. M. Fischer, and D. Abbott, “Terahertz scattering by granular composite materials: An effective medium theory,” *Appl. Phys. Lett.*, vol. 100, no. 1, p. 011107, 2012.
24. A. Bandyopadhyay, A. Sengupta, R. B. Barat, D. E. Gary, J. F. Federici, M. Chen, and D. B. Tanner, “Effects of scattering on THz spectra of granular solids,” *Int. J. Infrared Millimeter Waves*, vol. 28, no. 11, pp. 969–978, Aug. 2007.
25. K. H. Esbensen, D. Guyot, F. Westad, and L. P. Houmøller, *Multivariate Data Analysis - In Practice*. 5th edition CAMO Software, 2006.
26. W. Kessler, *Multivariate Datenanalyse - für die Pharma, Bio- und Prozessanalytik*. New York: John Wiley & Sons, 2007.
27. R. Bro and A. K. Smilde, “Principal component analysis,” *Anal. Methods*, vol. 6, no. 9, p. 2812, 2014.
28. A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.