THz Spectroscopic Decomposition and Analysis in Mixture Inspection Using Soft Modeling Methods

Chen Xie, Zhangwei Huang, Yiwen E, Xi-Cheng Zhang, Xusheng Kang, Yehao Ma, Pingjie Huang & Guangxin Zhang

ONLINE FIRST

Journal of Infrared, Millimeter, and Terahertz Waves

ISSN 1866-6892

J Infrared Milli Terahz Waves
DOI 10.1007/s10762-020-00755-6

Springer

International Journal of Infrared and Millimeter Waves
https://doi.org/10.1007/s10762-020-00755-6

THz Spectroscopic Decomposition and Analysis in Mixture Inspection Using Soft Modeling Methods

Chen Xie[1] · Zhangwei Huang[1] · Yiwen E[2] · Xi-Cheng Zhang[2] · Xusheng Kang[1] · Yehao Ma[1,3] · Pingjie Huang[1] · Guangxin Zhang[1]

Received: 9 September 2020 / Accepted: 27 October 2020 / Published online: 23 November 2020

Abstract

At present, the terahertz time-domain spectroscopic information of each component in multi-composition compounds detection is comprehensively combined. The lower resolution level of mixed spectra has posed many difficulties in signal analysis due to the overlapping characteristic information in the mixed ones. In this paper, to compare the performances of Nonnegative Matrix Factorization (NMF), Self-modeling Mixture Analysis (SMMA) and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) on complex systems, a binary mixture and a ternary mixture are employed during THz-TDS testing. The position of the absorption peak (PK) and the correlation coefficient are used to evaluate the decomposition effects. The experimental results show that the component spectra resolved by MCR-ALS demonstrate good consistency with the sample components. Further, MCR-ALS presents excellent results in comparison with NMF and SMMA in terms of decomposing precision and computing speed. MCR-ALS thus appears a promising algorithm to resolve the THz multi-way data, and may be useful for many applications, such as medicine quality assurance and unknown component identification in the general area of terahertz science and technology.

Keywords THz-TDS · MCR-ALS · Multiway data analysis · Characteristic information extraction · Multi-composition compound detection

✉ Pingjie Huang
huangpingjie@zju.edu.cn

[1] State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China

[2] The Institute of Optics, University of Rochester, Rochester, NY, USA

[3] Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China

1 Introduction

Antibiotics and amino acids are components with significant implications for human health and well-being frequently found in chemical materials, food and pharmaceutical. With the emergence of increasingly serious problems regarding adulteration or abuse of ingredients in medicines and foods, it has become urgent to research effective methods for detecting and identifying biomolecules and their mixtures. Thus far various techniques have been adopted to inspect biomolecules, such as chromatography, microwave, visual light, infrared and ultraviolet spectroscopy, and others. These methods all have unique advantages and limitations. For example, chromatography is an accurate approach, with the shortcomings of being labor intensive and time-consuming [1], as well as usually being a destructive detection method. The methods of spectroscopic are much easier and faster to manipulate. For example, the infrared spectroscopy method presents sharp assignable features. However, its results are susceptible to the scattering effect in samples and to thermal radiation from surroundings, which may affect the stability [2, 3]. Ultraviolet spectroscopy is harmful to some materials and the human body [4].

The terahertz (THz) wave typically refers to the frequency ranging from 0.1 to 10 THz, which lies between the microwave regions and infrared regions of the electromagnetic spectrum. Terahertz time-domain spectroscopy has fine time resolution and outstanding bandwidth. The reflection or absorption coefficient calculated from the spectrum can be employed to analyze molecular dynamics and interactions, namely crystalline lattice, inter-molecular vibrational modes, hydrogen bonding stretches, and some torsion vibrations [5]. Many biomolecules have unique absorption characteristics in the THz region [6, 7]. Therefore, Terahertz Time-Domain Spectroscopy (THz-TDS) shows the potential in biomolecular detection. Moreover, THz waves have tiny photon energy, of about one-millionth that of an X-ray, which means that it's a non-destructive detection method. These advantages make THz spectroscopy popular in a broad field of sciences, ranging from material analysis to cancer detection [8–10].

Compared with distinguishing pure materials, detection and identification of mixtures is a challenging task. It is well known that many materials have broad absorption peaks in the THz wave frequency domain, and it is difficult to obtain the "stand-free" absorption bands and map the absorption peaks to each component of the mixture. In addition, the peak shift is another obstacle [11]. Current studies of multicomponent mixture identification using THz requires the spectral signatures of ingredients or a database of potential constituents [12–14]. However, in many situations, the components are unknown or their spectra cannot be easily obtained.

It is well known that resolution is a feasible way of recovering the profiles (spectra, concentration) of more than one component in an unresolved and unknown mixture when there is no further available prior information about the nature and composition of the mixture. There have been a few related studies in terahertz field. Li and co-workers investigated an unmixing method based on hard modeling [15] to extract THz spectra of components from two-way mixture data. Hard modeling method bases the identification process on the extraction of parameters from the raw data according to a real model [16]. However, the approach is very complicated and requires that the spectrum of each component in the mixture has at least one characteristic absorption peak. The method is not always suitable in the terahertz mixture spectrum because of the serious aliasing problem.

Soft modeling approach may be a feasible alternative. Soft modeling approaches need no parameters and only rely on eigenvector-eigenvalue decomposition of a raw data matrix [16]. Among several soft modeling algorithms, NMF is usually susceptible to noise corruption and initialization. Moreover, its cost function is nonconvex, which may lead to local minima and unstable resolution. SMMA is also a widely used deconvolution method, but the baseline of the THz spectrum may cause difficulty in determination of pure variables. Although the two-order derivative is able to deal with this problem to some extent, it is easily affected by system noise.

This study aims to develop an effective and simple soft modeling method for terahertz time-domain data resolution. We found that the MCR-ALS method of decomposition performs better in terms of decomposition accuracy and efficiency than those by NMF and SMMA. The binary mixtures as resolved by MCR-ALS have the largest correlation coefficients, of 0.999 and 0.970 respectively. Regarding the estimation of PK, MCR-ALS also performs better than NMF and SMMA. In addition to the spectra of pure components, relative concentration of components can also be obtained. Similarly, MCR-ALS also achieves satisfactory results. The resolved concentration profiles show close agreement with the actual ones. The correlation coefficients are 0.982 and 0.983 for the binary system, and 0.981, 0.981 and 0.944 for the ternary system.

The structure of this article as follows. In Section 2, the concepts and characteristics of several different algorithms (MCR-ALS, NMF and SMMA) are introduced and discussed; and the terahertz experimental setup, Theophylline, are described. In Section 3, the application testing of MCR-ALS, NMF and SMMA algorithms for binary and ternary mixture unmixing are demonstrated. Experimental results based on various decomposition algorithms are compared and analyzed. The key affecting factors are discussed. Finally, our research work is summarized in Section 4.

2 Procedure and Methodology

The procedure in this study mainly includes mixture sample preparation, THz data acquisition (THz spectra acquisition of a mixture), number determination (determination of the number of components), and mixture unmixing (unmixing THz spectral data of a mixture), as shown in Fig. 1.

2.1 Sample Preparation

Theophylline is a medicine for curing emphysema, bronchitis and asthma; histidine is usually applied to treat anemia and rheumatic arthritis. Besides the two

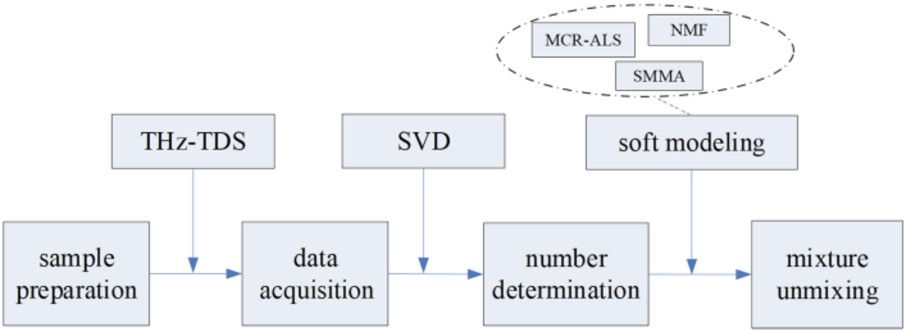

Fig. 1 The procedure of THz spectroscopic decomposition and analysis in mixture detection. SVD, Singular Value Decomposition; NMF, Nonnegative Matrix Factorization; MCR-ALS, Multivariate Curve Resolution-Alternating Least Squares; SMMA, Self-modeling Mixture Analysis

APIs (active pharmaceutical ingredient), excipients are also important ingredients in tablets, which can improve the stability and formability of the drug. This study applies two widely used excipients, mannitol and lactose monohydrate, to prepare samples.

L-Theophylline, L-histidine, L-methionine and mannitol were supplied by Aladdin Chemistry Co. Ltd.; D-lactose monohydrate was purchased from Sino-pharm Chemical Reagent Co. Ltd. The purities of all materials were higher than 98% and no further purification required before use. In this study, two systems were constructed: a binary system containing L-methionine and L-histidine with content ranging from about 15 to 85% (w/w) and a ternary system consisting of L-theophylline, L-lactose monohydrate and mannitol, with content ranging from about 10 to 80% (w/w). The samples (total mass 160 mg) of the two systems were prepared with a certain amount of polyethylene (48 mg) as diluent under 15 MPa pressure, with diameter of 13 mm and thickness of about 1 mm after drying treatment (423 K) for 2 h in a vacuum drying oven [17].

2.2 Experiment Setup

In this paper, samples were measured in transmission with commercial Zomega-3 THz-TDS system (Zomega Terahertz Corp., Troy, USA). The system laser source is a commercial mode-locked Ti: sapphire laser (Coherent Company in the USA), which produces less than 100 fs pulse at wavelength around 800 nm with repetition frequency of about 80 MHz and average power of about 960 mW. The frequency range given by the system is 1–3 THz. More details of this system have previously been reported in our earlier work [15].

During the experiment, in order to avoid the absorption of water vapor, dry nitrogen was used to purge the THz beam path. Every reference and sample signal were recorded five times at room temperature (about 300 K) and humidity below 0.1%.

2.3 Terahertz Optical Parameters Extraction

In the experiment, each sample and reference signal are taken as the average to reduce the random error. The refractive index $n(\omega)$ and absorption coefficient $\alpha(\omega)$ are calculated as follows [18]:

$$n(\omega) = \frac{\varphi(\omega)c}{\omega d} + 1 \tag{1}$$

$$\alpha(\omega) = \frac{2\kappa(\omega)}{c} = \frac{2}{d} ln \frac{4n(\omega)}{A(\omega)(n(\omega)+1)^2} \tag{2}$$

where $\varphi(\omega)$, $A(\omega)$ are the phase and amplitude ratio difference between reference and sample signal, d is the thickness, ω is the frequency, and c represents the speed of light in a vacuum.

2.4 Determination of the Number of Components by Using SVD

According to Beer's law, the absorption of a sample equals the sum of the absorption of its various chemical constituents. Therefore, the mixture spectrum can be considered as the weighted sum of the spectra of the pure ingredients plus the experimental noise [19]. The law can be demonstrated by a bi-linear model as stated in Eq. 3.

$$D = CS^T + E \tag{3}$$

where D is a two-way matrix of mixture spectra, C is the data matrix (concentration profiles) describing how the contributions of the N species change in different rows of the matrix. T represents the transpose of each matrix that includes it. S^T is the data matrix (pure spectral profiles) describing how the responses of these N species change in different rows of the matrix. E is the noise matrix which is caused by equipment deviations, calculation errors and other reasons, which cannot be explained by the chemical ingredients in C and S.

Obviously, the model provides important guiding significance for the analysis of the spectral data of mixtures in order to obtain the qualitative and quantitative information from the raw data D through resolution methods.

Determination of the number of components in various systems is the basic task of the resolution [20]. In this study, the number of components is predicted by Singular Value Decomposition (SVD):

$$D = U \wedge V^T \tag{4}$$

where U $(m \times r)$ is a column-orthogonal matrix, $\wedge(r \times r)$ is a diagonal matrix, and V^T is an $r \times n$ matrix. The relative contribution γ is applied to measure the contribution of each latent variable (LV):

$$\gamma = \frac{\sum_{i=1}^{r^*} \lambda_i^2}{\sum_{l=1}^{r} \lambda_l^2} \times 100\% \tag{5}$$

where λ_i is the ith element of $\wedge$, and r^* represents the number of latent variables (LVs) $(1 \le r^* \le r)$.

2.5 Soft Modeling Algorithms

2.5.1 SMMA

Self-modeling Mixture Analysis (SMMA), firstly proposed by Windig and Guilment [21, 22], aims to determine the spectral variables or selective concentration. The jth purity p_{ij} of a variable x_i is defined as follows:

$$p_{ij} = \omega_{ij} \frac{\sigma_i}{\mu_i + \alpha} \tag{6}$$

where σ_i and μ_i are the standard deviation and mean, respectively, of the variable x_i, α is an offset given in percentage of the maximum μ_i to prevent low-intensity wave numbers achieving high purity, and ω_{ij} is the determinant-dependent constant that assigns the variables less similar to the pure variables having been identified a larger weight.

Once the pure concentration variables C, for all the components have been determined, the spectral profiles S can be calculated through Eq. 7.

$$S = (C^T C)^{-1} C^T D \tag{7}$$

Through normalization of the spectral profiles S, the concentration profiles C can be re-estimated using Eq. 8.

$$C = (DD^T)^{-1} DS \tag{8}$$

2.5.2 NMF

Nonnegative Matrix Factorization (NMF), proposed by Lee and Seung [23], is an effective technique in approximating high dimensional data. The nonnegative constraint ensures the physical meaning of the decomposed result. Thus, NMF is employed to deconvolve the data of mixtures to extract the information of pure components, namely the signal of pure species and relative concentration [24, 25].

The NMF algorithm used in this study is shown in Eq. 9 and Eq. 10. The aim is intended to minimize the error function $f(S, C)$.

$$S_{l+1}^T = S_l^T \frac{(C^T D)_l}{(C^T C S^T)_l} \tag{9}$$

$$C_{l+1} = C_l \frac{(DS)_l}{(CS^T S)_l} \tag{10}$$

$$f(S, C) = \frac{1}{2} \left\| D - CS^T \right\|_F^2 \tag{11}$$

where l is the iteration index. The matrices S and C are randomly nonnegative initializations in the start iteration. Positivity of matrices C and S is ensured throughout the computation. Until the residual error function $f(C, S)$ converges, the percentage of change Q is obtained through Eq. 12. The stop criterion is set to 0.01% in this research. Moreover, NMF was executed 20 times, with the average result used to analyze and compare.

$$Q = 100\% * \left(\frac{f_{l+1} - f_l}{f_l}\right) \tag{12}$$

2.5.3 MCR-ALS

Multivariate Curve Resolution-Alternative Least Square (MCR-ALS) is a method used for the resolution of multiple component responses in unknown mixtures. The method is the determination of the true S and C matrices from analysis of only matrix D. The initial estimates of S and C matrices can be determined by detection of "purest" variables [21] or the techniques based on evolving factor analysis [26]. The initial estimations of C and S are optimized through iteration based on alternating least squares [27]. At each iteration of the optimization the new S and C matrices are obtained according to:

$$S^T = C^+D = C^+CS^T \tag{13}$$

$$C = D(S^T)^+ = C(S^T)(S^T)^+ \tag{14}$$

where D is the original data from experiments and the matrices $(S^T)^+$ and C^+ are the pseudo-inverse of the matrix S^T and C. If the correct number of species has been determined, C and S^T are full-rank column (row) matrices respectively. In the process of iteration, various constraints can be applied to enhance the performance of MCR: (1) selectivity, (2) unimodality, (3) non-negativity, and (4) closure [28]. The iterative optimization is carried out until a preselected number of cycles are reached or convergence is achieved. In this study, non-negativity and closure constraints are employed to enhance the performance of resolution for THz data. The characteristic absorption peaks can also be precisely extracted by MCR-ALS. The numbers of components of various systems were predicted by SVD, relative concentration of components can also be obtained by MCR-ALS.

As we all know, different methods have different characteristics and application scenarios. In order to compare the above three methods vividly, Table 1 lists the advantages and restrictions with them.

3 Experimental Results and Discussion

3.1 Binary Mixtures

To test MCR-ALS, the resolved effect is evaluated by two parameters. The position of the absorption peak (PK) was used to assess the accuracy of the characteristics of the resolved spectrum and the correlation coefficient was employed to test the consistency between resolved spectra and actual spectra [29].

The THz spectra of the binary mixture of L-methionine and L-histidine are presented in Fig. 2. As can be seen there, the absorptions of mixtures strengthen as the content of L-methionine increases from 0.3 to 1.6 THz. Thus, the two components have significantly different absorption strengths. Determination of the number

Table 1 The advantages and restrictions with three methods

Method	Theory	Advantage	Restrictions
SMMA	The spectra are resolved by extracting relative pure variable information of the components.	The results are deterministic, fact satisfactorily, strong anti-interference ability.	Difficulty in separation with overlapping absorption peaks, spectral baseline drift, etc.
NMF	Extracting spectral information by iterative optimization	Applicable to the occasion where the signal to noise ratio is relatively high.	For complex systems it is easy to fall into local optimum. In addition, simultaneous randomization of initialization results in non-uniqueness of decomposition.
MCR-ALS	Approaching the original matrix by alternating minimums based on bilinear stoichiometry.	The test data can form a bidirectional data matrix, and this data set can be interpreted very well by bilinearity.	Intermediate products in the formation of matter can cause undesirable results.

of components of mixtures is an important task for resolution. In this study, SVD was applied to require the number of components. The relative contribution versus the number of LVs is shown in Fig. 3. The relative contribution of two of the LVs is greater than 96.5%, while the contributions of other LVs are all less than 2.5%. The latter are thus considered noise in this study.

MCR-ALS, NMF and SMMA were used to unmix THz spectral data of the binary mixtures. Because of the non-uniqueness performance of NMF, the NMF was executed 20 times for each system, and the average spectra obtained were used as results. Figure 4 shows the results of resolution by the three methods. Compared with NMF

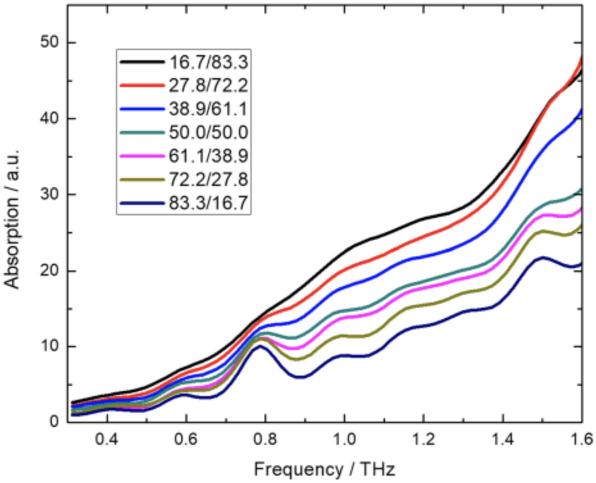

Fig. 2 THz absorption spectra of binary mixtures. Label as L-methionine/L-histidine

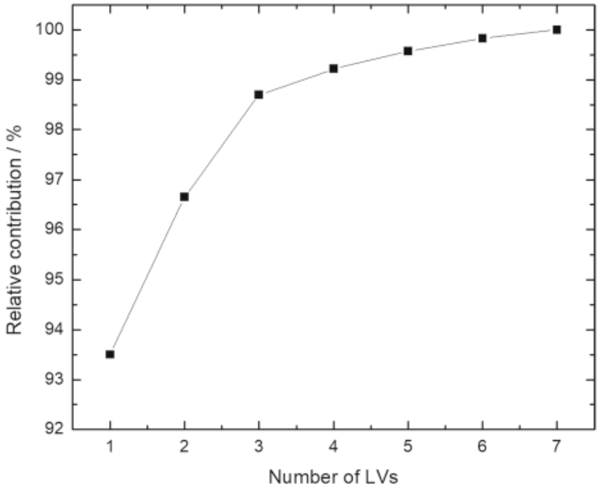

Fig. 3 The relative contribution versus number of latent variables of binary system

and SMMA, the best quality spectra are resolved with the MCR-ALS algorithm, and its resolved spectra of pure components show close agreement with the actual ones. The resolved spectra of L-methionine from NMF and SMMA have a significant error in both spectral profile and absorption peaks.

To provide quantitative evaluation results, Tables 2, 3, and 4 show the details of resolution results with MCR-ALS, NMF and SMMA respectively, involving PKs and correlation coefficient. The spectra of L-methionine and L-histidine as resolved by MCR-ALS have the largest correlation coefficients, of 0.999 and 0.970 respectively. Regarding the PKs estimation, MCR-ALS also performs better than NMF and

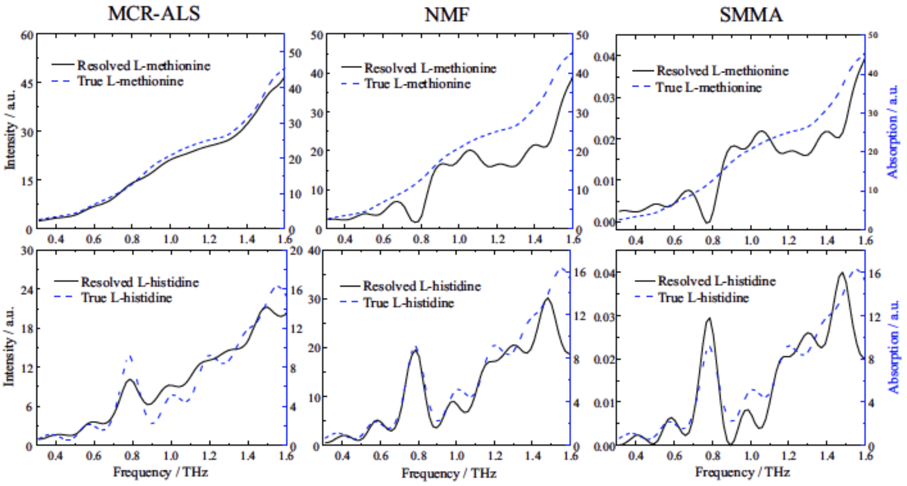

Fig. 4 Resolution result of binary system with MCR-ALS, NMF and SMMA

Table 2 Quantitative resolution results produced by MCR-ALS for binary system

Components	Correlation	Peak position (THz)						
	coefficient	PK1	PK2	PK3	PK4	PK5	PK6	PK7
L-Methionine	0.999	**0.604**		**0.915**	**1.537**			
		0.586	0.787	0.915	1.518			
L-Histidine	0.970	**0.366**	**0.567**	**0.787**	**1.006**	**1.208**		**1.555**
		0.421	0.604	0.787	0.988	1.208	1.317	1.500

In the section of peak position, PK is the absorption peak; true (resolved) absorption peaks are indicated by bold (regular) numbers

Table 3 Quantitative resolution results produced by NMF for binary system

Components	Correlation	Peak position (THz)						
	coefficient	PK1	PK2	PK3	PK4	PK5	PK6	PK7
L-Methionine	0.957		**0.604**	**0.915**				**1.537**
		0.512	0.677	0.915	1.061	1.226	1.409	
L-Histidine	0.939	**0.366**	**0.567**	**0.787**	**1.006**	**1.208**		**1.555**
		0.421	0.586	0.787	0.988	1.189	1.299	1.482

In the section of peak position, PK is the absorption peak; true (resolved) absorption peaks are indicated by bold (regular) numbers

Table 4 Quantitative resolution results produced by SMMA for binary system

Components	Correlation	Peak position (THz)						
	coefficient	PK1	PK2	PK3	PK4	PK5	PK6	PK7
L-Methionine	0.938		**0.604**	**0.915**				**1.537**
		0.512	0.677	0.915	1.061	1.226	1.391	
L-Histidine	0.899	**0.366**	**0.567**	**0.787**	**1.006**	**1.208**		**1.555**
		0.403	0.586	0.787	0.988	1.189	1.299	1.482

In the section of peak position, PK is the absorption peak; true (resolved) absorption peaks are indicated by bold (regular) numbers

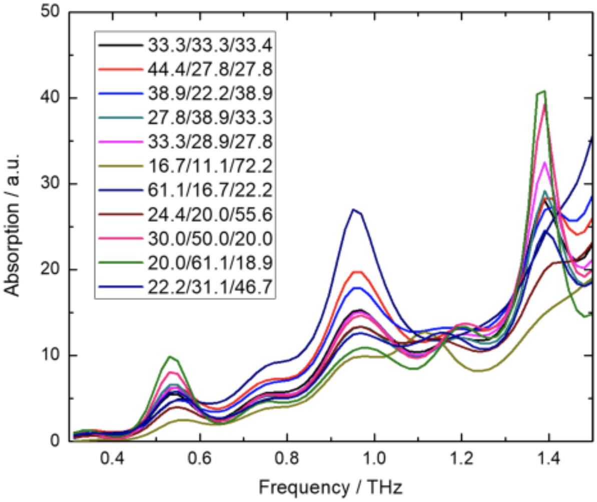

Fig. 5 THz absorption spectra of ternary mixtures

SMMA. The pure spectra extracted by SMMA strongly deteriorate, with the inferior resolution of pure variables resulting from the baseline problem arising from the serious overlap of various PKs. In addition, NMF is susceptible to noise, which makes it prone to fall into local optimum.

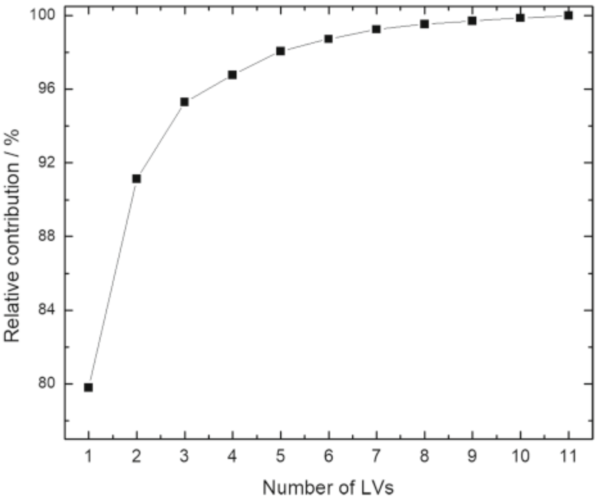

Fig. 6 The relative contribution versus number of latent variables of ternary system

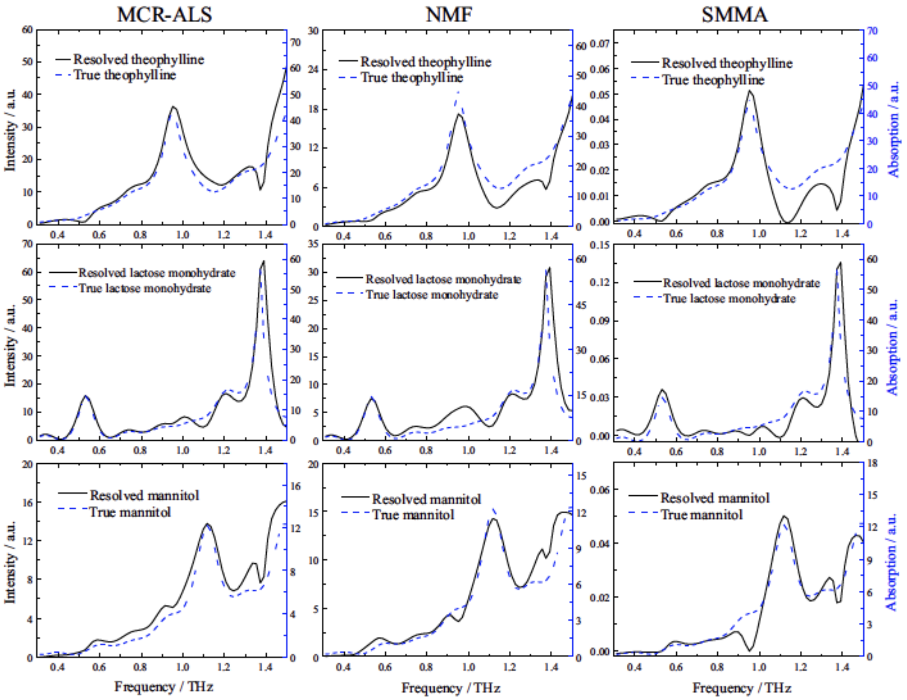

Fig. 7 Resolution result of ternary system with MCR-ALS, NMF and SMMA

3.2 Ternary Mixtures

As mentioned above, MCR-ALS yielded the best estimation of the three methods tested. To test its performance for more complex systems, a ternary system made up of L-theophylline, D-lactose monohydrate and mannitol was built. The THz spectra of the ternary mixtures with various mass ratios are shown in Fig. 5. The spectra of the mixtures have obvious characteristics demonstrating that some components have feature absorption bands. Before unmixing the THz two-way data, SVD was applied to predict the number of components. The relative contribution versus the number of LVs is shown in Fig. 6. The sum of relative contributions of the first three LVs is greater than 95%. In addition, the contributions of the other LVs are all less than 2%. Thus, the method of SVD can accurately predict the number of components.

To unmix the three-way THz data, MCR-ALS, NMF and SMMA were employed, with the resolution results presented in Fig. 7. On the whole, for the ternary system, the three methods are all able to extract the rough spectral profiles of pure components. However, the errors resulting from the shift effect are still present, as is especially obvious in domains containing strong absorption peaks. For example, the spectral profile of theophylline resolved by MCR-ALS in domain 1.3–1.4 THz is affected by the strong absorption band of D-lactose monohydrate in this range. The extracted spectra from other approaches also have similar errors.

Table 5 Quantitative resolution results produced by MCR-ALS for ternary system

Components	Correlation	Peak position (THz)						
	coefficient	PK1	PK2	PK3	PK4	PK5	PK6	PK7
Theophylline	0.969			**0.787**	**0.951**	**1.336**		
		0.439	0.586	0.769	0.951	1.336		
D-Lactose	0.929	**0.348**	**0.531**	**0.750**	**0.933**		**1.226**	**1.372**
monohydrate		0.329	0.531	0.750	0.915	1.006	1.208	1.391
Mannitol	0.986	**0.403**	**0.604**	**0.769**	**0.951**	**1.116**	**1.317**	**1.500**
			0.586	0.769	0.915	1.116	1.336	1.500

In the section of peak position, PK is the absorption peak; true (resolved) absorption peaks are indicated by bold (regular) numbers

Tables 5, 6, and 7 list the quantitative results of resolution produced by MCR-ALS, NMF and SMMA respectively involving absorption peaks and correlation coefficient of components. As a result, the correlation coefficients of all extracted spectra are greater than 0.9, and most PKs can be resolved accurately. Moreover, the resolved PKs are usually greater than the actual ones, as a result of the shift effect, which is induced by the nonlinear absorption of samples.

3.3 Discussion

The results of this study show that MCR-ALS presents better resolution than NMF and SMMA in both binary and ternary systems. MCR-ALS not only produced unique results but also accurately resolved spectra of components. Because of the random initialization and nonconvex cost function, the result of NMF is not unique, and it is liable to fall into local optimum. On the contrary, the initialization of MCR-ALS

Table 6 Quantitative resolution results produced by NMF for ternary system

Components	Correlation	Peak position (THz)						
	coefficient	PK1	PK2	PK3	PK4	PK5	PK6	PK7
Theophylline	0.975			**0.787**	**0.951**	**1.336**		
		0.439	0.604	0.769	0.951	1.317		
D-Lactose	0.916	**0.348**	**0.531**	**0.750**	**0.933**	**1.226**	**1.372**	
monohydrate		0.329	0.531	0.750	0.988	1.208	1.391	
Mannitol	0.977	**0.403**	**0.604**	**0.769**	**0.951**	**1.116**	**1.317**	**1.500**
			0.567	0.769	0.915	1.116	1.354	1.464

In the section of peak position, PK is the absorption peak; true (resolved) absorption peaks are indicated by bold (regular) numbers

Table 7 Quantitative resolution results produced by SMMA for ternary system

Components	Correlation	Peak position (THz)						
	coefficient	PK1	PK2	PK3	PK4	PK5	PK6	PK7
Theophylline	0.953			**0.787**	**0.951**	**1.336**		
		0.421	0.604	0.769	0.951	1.299		
D-Lactose	0.904	**0.348**	**0.531**	**0.750**		**0.933**	**1.226**	**1.372**
monohydrate		0.329	0.531	0.732	0.878	1.006	1.208	1.391
Mannitol	0.965	**0.403**	**0.604**	**0.769**	**0.951**	**1.116**	**1.317**	**1.500**
			0.604		0.897	1.116	1.336	1.464

In the section of peak position, PK is the absorption peak; true (resolved) absorption peaks are indicated by bold (regular) numbers

is based on pure variables, which contain information on the content of components, and MCR-ALS therefore has greater resolution stability. On the other hand, although SMMA tries to find pure variables, the baselines of THz spectra give rise to serious overlap of absorption peaks, which makes it difficult to find enough pure variables, and this poses a severe challenge in the resolution of the binary system. Compared with SMMA, MCR-ALS improves the rough unmixing result through iterative optimization called ALS, to reduce error caused by inferior pure variables. Thus, MCR-ALS can produce accurate and stable resolution.

In addition to the spectra of pure components, relative concentration of components can also be obtained. Similarly, MCR-ALS also achieves satisfactory results. The resolved concentration profiles show close agreement with the actual ones. The correlation coefficients are 0.982 and 0.983 for the binary system, and 0.981, 0.981 and 0.944 for the ternary system. Figure 8 shows the resolved profile of theophylline

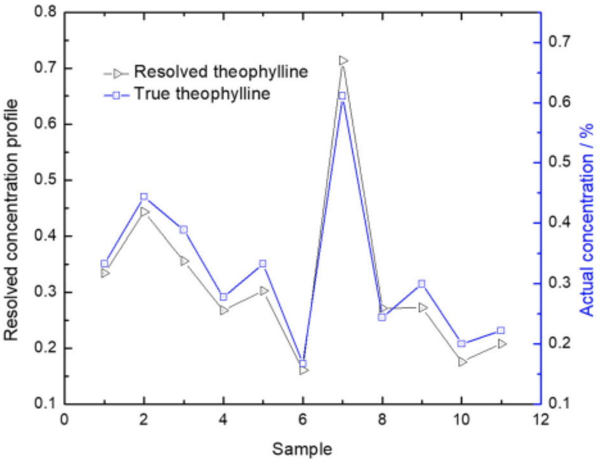

Fig. 8 The resolved profile of theophylline produced by MCR-ALS

with correlation coefficient 0.981. This paper verifies the effectiveness of the soft modeling method for THz data.

Different methods have different applicability, MCR-ALS has its specific advantages in terahertz spectral decomposition. We have tested the binary and ternary mixture in our paper. On the separation level of the more component mixtures, Huang Xinbao et al. proposed a quantitative analysis method for multi-source mixture components based on the vector angle [30]. But it is mainly based on qualitative analysis, and it is necessary to know the specific components. In fact, if the mixture does not have intermediate by-product, the model can also be well explained by the bilinear model. So, there is a certain possibility that the MCR-ALS will be analyzed in the future with a mixture of more components.

Our subsequent research intends to focus on chemical imaging (CL). CL analyses are conducted to estimate spatial, structural and quantitative information about samples. Further, future research will consider the resolution of dynamic processes, such as chemical reactions.

4 Conclusion

This paper demonstrates the application of THz spectroscopy and chemometrics in the characterization of unknown multicomponent systems. Several soft modeling methods, namely MCR-ALS, NMF and SMMA, were employed to analyze the THz data to develop a useful resolution method for THz spectra data. Several chemical multicomponent systems were built to test the performances of the three approaches in the THz domain. The results show that MCR-ALS produces unique and accurate results in contrast to NMF or SMMA. The resolved spectra of pure components from MCR-ALS demonstrate close agreement with actual ones. In addition, the characteristic absorption peaks can also be precisely extracted. The numbers of components of various systems were predicted by SVD. The combination of THz spectroscopy and chemometrics could be successfully used to characterize unknown mixtures, which is of great practical significance in areas such as detecting counterfeit drug products and analyzing biochemical systems.

Acknowledgments The authors are grateful to Professor Jianming Dai, Yiwen E, and Qi Jin at the University of Rochester for the valuable discussions.

Funding This work was supported by National Natural Science Foundation of China (No. 61473255 and No. 61873234), the National Key R&D Program of China (No. 2017YFC1403801), and the Fundamental Research Funds for the Central Universities (No. 2019FZA5007).

References

1. R. C. Martinez, E. R. Gonzalo, M. J. A. Moran, J. H. Mendez, Sensitive method for the determination of organophosphorus pesticides in fruits and surface waters by high-performance liquid chromatography with ultraviolet detection, J. Chromatograph. A, vol. 607(1), 37–45 (1992).

2. S. Armenta, S. Garrigues, M. D. L. Guardia, Determination of iprodione in agrochemicals by infrared and Raman spectrometry, Anal. Bioanal. Chem., vol. 387(8), 2887–2894 (2007).
3. P. Y. Han, M. Tani, M. Usami, S. Kono, R. Kersting, X.-C. Zhang, A direct comparison between terahertz time-domain spectroscopy and far-infrared Fourier transform spectroscopy, J. Appl. Phys., vol. 89(4), 2357–2359 (2001).
4. Y. C. Shen, Terahertz pulsed spectroscopy and imaging for pharmaceutical applications: A review, Int. J. Pharm., vol. 417(1-2), 48–60 (2011).
5. H. J. Shin, S. J. Oh, S. I. Kim, H. W. Kim, J.-H. Son, Conformational characteristics of β-glucan in laminarin probed by terahertz spectroscopy, Appl. Phys. Lett., vol. 94(11), 111911 (2009).
6. J. Sibik, J. A. Zeitler, Direct measurement of molecular mobility and crystallisation of amorphous pharmaceuticals using terahertz spectroscopy, Adv. Drug Deliver. Rev., vol. 100, 147–157 (2016).
7. S. Q. Du, H. Li, L. Xie, L. Chen, Y. Peng, Y. M. Zhu, H. Li, P. Dong, J. T. Wang, Vibrational frequencies of anti-diabetic drug studied by terahertz time-domain spectroscopy, Appl. Phys. Lett., vol. 100(14), 143702 (2012).
8. M. Takahashi, Y. Ishikawa, Terahertz vibrations of crystalline α-D-glucose and the spectralchange in mutual transitions between the anhydride and monohydrate, Chem. Phys. Lett., vol. 642, 29–34 (2015).
9. J. Y. Qin, L. Xie J, Y. B. Ying, Feasibility of terahertz time-domain spectroscopy to detect tetracyclines hydrochloride in infant milk powder, Anal. Chem., vol. 86(23), 11750–11757 (2014).
10. U. Puc, A. Abina, M. Rutar, A. Zidanšek, A. Jeglič, G. Valušis, Terahertz spectroscopic identification of explosive and drug simulants concealed by various hiding techniques, Appl. Opt., vol. 54(14), 4495–4502 (2015).
11. F. Alsmeyer, H. J. Koss, W. Marquardt, Indirect spectral hard modeling for the analysis of reactive and interacting mixtures, Appl. Spectrosc., vol. 58(8), 975–985 (2004).
12. D. M. Mittleman, R. H. Jacobsen, R. Neelamani, R. G. Baraniuk, M. C. Nuss, Gas sensing using terahertz time-domain spectroscopy, Appl. Phys. B, vol. 67(3), 379–390 (1998).
13. Y. Chen, Y. Ma, Z. Lu, L. Qiu, J. He, Terahertz spectroscopic uncertainty analysis for explosive mixture components determination using multi-objective micro-genetic algorithm, Adv. Eng. Softw., vol. 42(9), 649–659 (2011).
14. P. F. X. Neumaier, K. Schmalz, J. Borngräber, R. Wylde, H.-W. Hübers, Terahertz gas-phase spectroscopy: chemometrics for security and medical applications, Analyst, vol. 140(1), 213–222 (2015).
15. X. Li, D. B. Hou, P. J. Huang, J. H. Cai, G. X. Zhang, Component spectra extraction from terahertz measurements of unknown mixtures, Appl. Opt., vol. 54(30), 8925–8934 (2015).
16. L. Duvillaret, F. Great, J. L. Coutaz, A reliable method for extraction of material parameters in terahertz time-domain spectroscopy, IEEE J. Sel. Topics Quantum Electron., vol. 2(3), 739–746 (1996).
17. Y. H. Ma, X. Li, P. J. Huang, D. B. Hou, Q. Wang, G. X. Zhang, THz spectral data analysis and components unmixing based on non-negative matrix factorization methods, Spectrochim. Acta A Mol. Biomol. Spectrosc., vol. 177, 49–57 (2017).
18. L. Blanchet, C. Ruckebusch, J. P. Huvenne, A. D. Juan, Hybrid hard-and soft-modeling applied to difference spectra, Chemometr. Intell. Lab. Syst., vol. 89(1), 26-35 (2007).
19. C. Gendrin, Y. Roggo, C. Collet, Self-modeling curve resolution of near infrared imaging data, J. Near Infrared Spectrosc., vol. 16(3), 151–157 (2008).
20. D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. D. Jong, C. K. Mann, Handbook of chemometrics and qualimetrics: Part b, Technometrics, vol. 52(2), 302 (1998).
21. W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem., vol. 63(14), 1425–1432 (1991).
22. K. Awa, T. Okumura, H. Shinzawa, M. Otsuka, Y. Ozaki, Self-modeling curve resolution (SMCR) analysis of near-infrared (NIR) imaging data of tablets, Anal. Chim. Acta., vol. 619(1), 81–86 (2008).
23. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nat., vol. 401(6755), 788–791 (1999).
24. I. Kopriva, I. Jerić, Blind separation of analytes in nuclear magnetic resonance spectroscopy: improved model for nonnegative matrix factorization, Chemometr. Intell. Lab. Syst., vol. 137, 47–56 (2014).
25. G. Livanos, M. Zervakis, N. Pasadakis, M. Karelioti, G. Giakos, Deconvolution of petroleum mixtures using mid-FTIR analysis and non-negative matrix factorization, Meas. Sci. Technol., vol. 27(11), 114005 (2016).

26. M. Maeder, Evolving factor-analysis for the resolution of overlapping chromatographic peaks, Anal. Chem., vol. 59(3), 527–530 (1987).
27. E. J. Karjlainen, The spectrum reconstruction problem: use of alternating regression for unexpected spectral components in two-dimensional spectroscopies, Chemom. Intell. Lab. Syst., vol. 7(1-2), 31–38 (1989).
28. J. Jaumot, R. Gargallo, A. D. Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemometr. Intell. Lab. Syst., vol. 76(1), 101–110 (2005).
29. Y. Ma, D. B. Hou, J. H. Cai, P. J. Huang, G. X. Zhang, The spectral resolution of unknown mixture based on THz spectroscopy with self-modeling technique, Chemometr. Intell. Lab. Syst., vol. 150, 65–73 (2016).
30. X. B. Huang, P. J. Huang, X. Li, Y. H. Ma, D. B. Hou, G. X. Zhang, Analysis of Terahertz Time Domain Spectroscopy of Mixtures Based on Indirect Hard Modeling Method, Spectrosc. Spect. Anal., vol. 37(10), 3021–3026 (2017).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.