IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 21, Issue 2, Ser. 2 (Mar. – Apr. 2025), PP 01-10
www.iosrjournals.org

Overall and Local Optimization: A Multi-Substance Identification Algorithm Based on Fourier Transform Infrared Spectroscopy

PENG Yajin[1], YU Bo[1*], SUI Feng[2], YAO Zhiwen[2], ZHOU Zhiming[2]
[1](College of Science, China Three Gorges University, China)
[2](China Shipbuilding Industry Group Co. Ltd Alphapec Instrument Hubei， China)

Abstract: *This paper proposes a multi-substance identification and analysis algorithm based on Fourier Transform Infrared (FTIR) spectroscopy. The algorithm begins with baseline correction and data length normalization of the FTIR spectra. From the perspective of overall optimization, the mixture data is analyzed using the non-negative least squares algorithm to identify potential substances in the database. Subsequently, key features such as the maximum values of spectral peaks, peak widths, and the left and right minimum values are extracted to construct a feature parameter vector. Finally, from the perspective of local optimization, a similarity function is employed to match and identify the most similar substance in the database. By iterating this process, multi-substance identification is achieved. Experimental results demonstrate the method's high accuracy and reliability in identifying mixture components, providing an efficient and precise approach for qualitative analysis. This method holds significant potential for future applications in chemical analysis.*

.***KeyWords****: Fourier Transform Infrared Spectroscopy; Multi-Substance Mixture; Qualitative Analysis; Overall Optimization; Local Optimization*

Date of Submission: 14-03-2025 Date of Acceptance: 26-03-2025

I. Introduction

In modern industrial production and chemical management, the importance of multi-substance identification has grown significantly. With the continuous increase in chemical products, accurately identifying toxic and hazardous substances is crucial for ensuring safety and environmental protection. Accurate substance identification technology not only prevents safety risks but also enhances production efficiency. Establishing a fast and reliable identification method is essential for addressing potential risks and improving emergency response capabilities. Fourier Transform Infrared Spectroscopy (FTIR), as an efficient and environmentally friendly analytical technique, has been widely applied in fields such as toxic gas detection[1]and drug component analysis[2]. It has become a core tool in multi-substance identification and analysis. Compared to traditional methods, FTIR detection often faces challenges such as insufficient sensitivity and long response times, which may lead to severe consequences due to delayed results. In contrast, FTIR technology can monitor trace harmful gases in real-time, making it an ideal solution. Additionally, FTIR can identify unknown chemical substances, ensuring rapid responses to toxic gas leaks. Another key application of FTIR is drug component analysis, where it effectively identifies main components and impurities, supporting drug quality control and regulation. This ensures drug safety and enhances therapeutic efficacy. As technology advances, the potential applications of FTIR will continue to expand, offering more comprehensive safeguards for safety and health.

Given the broad applications of FTIR-based multi-substance identification in toxic gas detection and drug property analysis, researchers have proposed various methods for analyzing FTIR data. In 1994, Curk et al. [3] described how to use FTIR to monitor the cultivation process of lactic acid bacteria in beer production and identify different species of lactic acid bacteria. They attempted to improve traditional identification methods using FTIR and other analytical techniques, providing rapid, reliable, and routine analysis for many samples. However, there is still room for improvement, such as exploring additional spectral windows (single or combined) and increasing the number of species and strains analyzed to create a more reliable database. In 2012, Jiang An[4] proposed several improved algorithms for the qualitative analysis of complex mixtures using FTIR, including an asymmetric least squares baseline correction algorithm, a continuous wavelet transform-based FTIR fitting algorithm, and a support vector machine-based qualitative analysis algorithm using prior knowledge. However, when dealing with more complex spectral data, these algorithms may face challenges in achieving local optimization, affecting the correction results. In 2019, Liu Caizheng et al. [5] proposed a method for identifying mixture components using Raman spectroscopy. This method involves background correction and denoising of the Raman spectra, followed by fitting the Raman peaks using the Voigt function to construct feature vectors for standard library substances and the mixture to be identified. Finally, the mixture components are identified by correlating the feature vectors with the database. However, this method struggles to effectively identify spectral data with overlapping peaks. In 2023, Chen Bin et al. [6] addressed the qualitative analysis of mixtures by using a multi-feature fusion Backpropagation (BP) neural network model based on infrared spectroscopy and compared it with a logistic regression model. The multi-feature fusion logistic regression model and BP neural network model, which incorporate both spectral features and full-spectrum information, exhibit stronger discriminative capabilities compared to single-matching algorithms. However, these methods still face challenges such as high model training complexity and strong dependence on data quality. In summary, existing methods either achieve local optimization with poor overall performance or achieve overall optimization but fail to accurately identify features, leaving room for improvement in accuracy and flexibility.

In FTIR-based multi-substance identification, the principle of overall optimization quickly narrows down a subset of the database that likely contains the target substance, ensuring rapid response when dealing with large databases. On the other hand, the principle of local optimization focuses on precisely extracting the spectral features of the target substance from the subset obtained through overall optimization, emphasizing detailed accuracy, which is crucial for improving identification accuracy. Therefore, the core task of this paper is to find a balance between overall and local optimization to enhance the flexibility and accuracy of identification. To achieve this, we propose a new FTIR-based multi-substance identification and analysis algorithm that integrates non-negative least squares, spectral peak identification, similarity function matching, and iterative processes. First, when dealing with complex data, the non-negative least squares algorithm is used to quickly determine a subset from the database that likely contains the target substance, providing a solid foundation for subsequent in-depth analysis and identification of multiple substances, ensuring the accuracy and reliability of the results. Second, spectral peak identification is used to extract key features from the spectra, and a similarity function is defined based on these features. Finally, the similarity function matching method is used to match the spectral peaks of the mixture with the subset, and the most similar target substance is selected. By iterating this process, the goal of multi-substance identification can be achieved. Theoretically, this algorithm can iterate indefinitely until all substance components in the test data are identified.

The structure of this paper is as follows: Section 1 describes the algorithm and presents the FTIR-based multi-substance identification and analysis algorithm. Section 2 discusses the numerical results and analysis. The final section concludes the paper.

II. Algorithm Description

This section describes the main methods used in the algorithm. First, the Fourier Transform Infrared (FTIR) spectral data is preprocessed, including baseline correction and data length normalization. Next, the non-negative least squares algorithm is used to analyze the multi-substance mixture from an overall perspective. Finally, feature vectors for the substances in the database and the multi-substance mixture are constructed, and the similarity function is used to analyze the feature vectors of the multi-substance mixture and the database data one by one.

Preprocessing of FTIR Spectral Data

Preprocessing of FTIR spectral data [7-9] is a crucial step to ensure data quality and analytical accuracy. It eliminates noise and baseline drift generated during data acquisition while enhancing model adaptability through standardization and regularization. The preprocessing includes multiple steps aimed at improving data quality, such as baseline correction to remove drift and data length normalization to align test and database data for subsequent analysis.

First, this paper crops the data to ensure that the horizontal coordinate ranges of the test data and the database data are consistent. Assume there are K spectral datasets, each containing NK sample points, where the k-th dataset has a horizontal coordinate set $V_k = \{v_{k1}, v_{k2}, \ldots, v_{kN_k}\}$ and a vertical coordinate set $I_k = \{I_{k1}, I_{k2}, \ldots, I_{kN_k}\}$ .The horizontal coordinate range of each dataset is calculated, i.e., the minimum value $v_{min,k} = \min(V_k)$ and the maximum value $v_{\max,k} = \max(V_k)$ of the horizontal coordinates. Then, the overall horizontal coordinate range for all datasets is calculated as follows:

$$x_{min} = \max\{v_{min,1}, v_{min,2}, \ldots, v_{min,K}\}, \quad x_{max} = \min\{v_{max,1}, v_{max,2}, \ldots, v_{max,K}\} \tag{1}$$

For each dataset, the horizontal coordinate range is cropped to the overall horizontal coordinate range $[x_{min}, x_{max}]$. That is, for the k-th dataset, data points that satisfy $x_{min} \le x_k \le x_{max}$ are selected, resulting in the cropped dataset $\{V_k^{'}\}$:

$$\{V_k^{'}\} = \{V_k\} \quad when \quad x_{min} \le x_k \le x_{max} \tag{2}$$

This ensures that the horizontal coordinate ranges of all datasets are unified to a common interval. Next, spectral alignment is performed to ensure that all datasets have consistent horizontal coordinate intervals and the same number of sampling points. This paper uses spline interpolation to map the horizontal coordinate values of each dataset to a new, uniformly spaced set of sampling points. For the k-th dataset, the cropped horizontal coordinate range is $\left[x_{min}, x_{max}\right]$, and a new, uniformly spaced horizontal coordinate set with M elements is defined, evenly distributed within the interval $\left[x_{min}, x_{max}\right]$, as follows:

$$x_i = v_{min} + (i-1)\cdot\Delta v \quad where \quad \Delta v = \frac{v_{max} - v_{min}}{M-1}, \;\; i = 1, 2, \ldots, M. \tag{3}$$

Then, for the k-th dataset, spline interpolation is used to interpolate the original vertical coordinates Ik onto the new horizontal coordinate set $\left\{V_k^{''}\right\}$, resulting in a new vertical coordinate set $\left\{I_k^{'}\right\}$.Through this process, each dataset has the same number of elements M within the unified horizontal coordinate range, and the horizontal coordinate values are uniformly spaced.

After aligning all datasets, this paper performs baseline correction [10-15] on the test data to eliminate or reduce baseline shifts caused by factors such as instrument drift and sample background. This $\left\{V_k^{''}\right\}$ involves correcting the vertical coordinates of each dataset so that the baseline is close to zero. Common baseline correction methods include polynomial baseline correction, adaptive baseline correction, and piecewise polynomial fitting. Common baseline correction methods include polynomial baseline correction, adaptive baseline correction, and piecewise polynomial fitting. This paper employs the effective extremum correction method, whose basic idea is based on effective extrema to better capture the main trends and effective details of the signal. Based on this, the effective extremum baseline correction method is derived. This method not only retains the effective information of the test data and the database but also makes their baselines consistent, facilitating subsequent analysis.

Through the above steps, the preprocessed dataset can be compared within the same horizontal coordinate range, eliminating differences caused by varying numbers of sampling points, horizontal coordinate ranges, and baseline shifts. This provides consistent spectral data for subsequent analysis, effectively improving the accuracy and reliability of FTIR spectral analysis and laying a solid foundation for further data analysis.

Non-Negative Least Squares Algorithm

After preprocessing the test data and the database data, this paper obtains *n+1* vectors of length *m*, where *n* vectors correspond to the vertical coordinates of the substances in the database, i.e., $\left\{I_k^{'} : k = 1, 2, \ldots, n\right\}$. The remaining vector is the test data, denoted as *y*. By combining $I_k^{'}$ as column vectors, a matrix $A \in □^{m\times n}$ is obtained. The goal now is to find a limited number of column vectors from the n column vectors of A such that the error between their linear combination and y is minimized. This is exactly the problem that traditional least squares methods are well-suited to solve, i.e., finding $x \in □^{n}$ such that $\left\|Ax - y\right\|_2$ is minimized in the 2-norm sense. However, some components of x obtained by this method may be negative, which contradicts the physical meaning of the problem. Therefore, this paper adds a constraint that every component of x must be greater than or equal to zero. This is the basic idea of Non-negative Least Squares (NNLS). For more details on Non-negative Least Squares, refer to references [16,17]. Specifically, Non-negative Least Squares (NNLS) solves for the coefficient vector $x* \in □^{n}$ such that

$$\left\|Ax*\right\|_2^2 = \min_{x\geq 0}\left\|Ax - y\right\|_2^2 \tag{4}$$

where $\left\|□\right\|_2$ denotes the 2-norm. To solve the NNLS problem, researchers have developed various effective algorithms, such as the steepest descent method[18], Newton's method[19], and the conjugate gradient method[20]Some researchers have also transformed this problem into a linear complementarity problem, resulting in a feasible interior-point algorithm, as described in reference[21]. Due to space limitations, this paper will not elaborate on these algorithms in detail. Interested readers can refer to[18].

Using the NNLS algorithm, a subset of the database that likely contains the target substance can be quickly determined. In fact, by sorting all components of $x*$ in descending order and selecting the top ten largest components, the corresponding database data and substances can be identified. The next task is to find the substance most likely contained in the target mixture from these ten substances. In other words, after determining an overall optimal range using the NNLS algorithm, the principle of local optimization is applied within this range to identify the target substance. To achieve this, this paper studies the local features of the database data and test data and proposes a definition for the similarity function.

Similarity Function

To define the similarity function, it is first necessary to study the local information of the FTIR spectral data, such as spectral peaks and peak widths. In FTIR spectra, spectral peaks are regions where the absorption intensity of the sample changes significantly at specific wavelengths or wavenumbers, typically representing the characteristics of the components in the sample. In mixtures, spectral peaks often shift or deform due to the overlapping of adjacent peaks or interference from certain components. This phenomenon not only affects the clarity of the peaks but may also lead to misjudgment of the mixture components, posing challenges to the accuracy of peak identification. To better illustrate the characteristics of spectral peaks, Figure 1 shows an example of a multi-substance mixture's FTIR spectrum. The shape characteristics of each peak are clearly displayed. Figure 2 shows the spectral peak characteristics of the mixture, including the relative positions and shapes of the peaks. By observing this figure, the importance of spectral peaks in mixture analysis can be more intuitively understood. Peak width typically refers to the width of the spectral peak, reflecting the extent of the peak's spread in the frequency domain. Under constant data collection conditions, the peak width is relatively stable. By measuring the peak width in the spectrum and comparing it with the peak width of standardized substances, the corresponding substances can be further identified in the database.

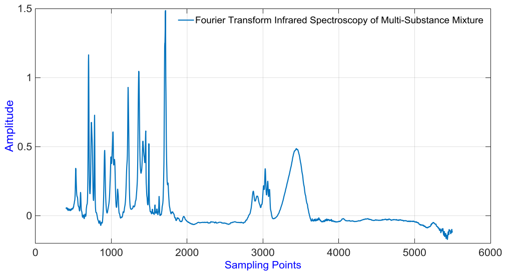

Figure 1: Fourier infrared spectroscopy of mixtures

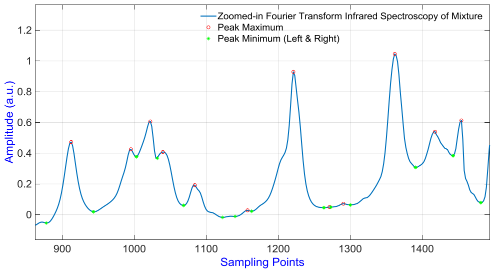

Figure 2: Spectral peak characterization in mixtures

By observing the spectra of mixtures and 342 common chemical substances based on FTIR, spectral peak characteristics are typically represented as a series of peaks and troughs. To clearly describe these characteristics, this paper extracts the maximum values of the peaks, the width of the peaks, and the left and right minimum values to form the spectral peak feature parameter vector of the data. Let the *i*-th spectral peak feature be represented as a vector $(\alpha_i^{\max}, \lambda_i^{\max}, S_i, \alpha_i^{\min_l}, \alpha_i^{\min_r})$, $i = 1, 2, \ldots, N$, where N is the number of peaks. Here, $\alpha_i^{\max}$、$\lambda_i^{\max}$、S_i、$\alpha_i^{\min_l}$、$\alpha_i^{\min_r}$ represent the horizontal and vertical coordinates of the peak maximum, the peak width $S_i = \alpha_i^{\min_r} - \alpha_i^{\min_l}$, and the horizontal coordinates of the left and right minima, respectively. These elements together form the overall feature vector of the spectral peak, facilitating the identification and differentiation of different components in multi-substance analysis. The maximum point is located at the highest point of each main peak, indicating the concentration of the main components in the sample. The peak width describes the width of the spectral peak, and through the mathematical model of the peak, structural information of complex mixtures can be revealed. The left and right minimum points are located on either side of each peak, indicating the shape and symmetry of the spectral peak.

By capturing the spectral peaks and their width information from the database and test data using the horizontal coordinates of the local maxima and their left and right minima, a similarity function is defined to determine their similarity. This enables the identification of the substance in the database that is most similar to the test data. Specifically, let the *i*-th spectral peak feature vector of a substance in the database be $(\alpha_i^{\max}, \lambda_i^{\max}, S_i, \alpha_i^{\min_l}, \alpha_i^{\min_r})$, $i = 1, 2, \ldots, N$, where N is the number of peaks, and let the *k*-th spectral peak feature vector of the test data be $(\alpha_k^{\max}, \lambda_k^{\max}, S_k, \alpha_k^{\min_l}, \alpha_k^{\min_r})$, $k = 1, 2, \ldots, M$, where M is the number of peaks. First, the horizontal distance between the test data and the *i*-th peak in the database is calculated. Based on the principle of minimizing the distance between the test data peak and the *i*-th peak in the database, a peak from the test data that is closest to the *i*-th peak in the database is selected. Let its feature vector be $(\alpha_*^{\max}, \lambda_*^{\max}, S_*, \alpha_*^{\min_l}, \alpha_*^{\min_r})$, and let the distance be denoted as $l = \left|\alpha_*^{max} - \alpha_i^{max}\right|$ .Next, the similarity between the selected test data peak and the *i*-th peak in the database is defined. The test data peak and the database peak are considered similar if they satisfy the following three conditions:

1. The minimum distance l is less than or equal to half the width of the test data peak, i.e., $l \le \frac{S}{2}$;

2. The horizontal coordinate of the maximum of the i-th peak in the database lies within the peak of the test data, i.e., $\alpha_i^{max} \in \left[\alpha_*^{min_l}, \alpha_*^{min_r}\right]$;
3. The vertical coordinate of the maximum of the i-th peak in the database is close to that of the test data peak, i.e., $\left|\lambda_i^{max} - \lambda_*^{max}\right| \le \varepsilon$, where ε is a predefined threshold (set to 0.02 in this paper);

For two similar peaks, the *i*-th peak in the database can be assigned a positive value.

$$R^+(\lambda_i^{max}, \lambda_*^{max}) = \begin{cases} 1.5 \quad if \quad \lambda_i^{max} \ge \dfrac{\max \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}; \\ 1 \quad if \quad \lambda_i^{max} < \dfrac{\max \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}, \quad and \quad \lambda_i^{max} \ge \overline{\lambda_k^{max}}; \\ 1 \quad if \quad \lambda_i^{max} < \overline{\lambda_k^{max}}, \quad and \quad \lambda_i^{max} \ge \dfrac{\min \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}; \\ 0.5 \quad if \quad \lambda_i^{max} < \dfrac{\min \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}. \end{cases} \tag{5}$$

Otherwise, the i-th peak in the database is assigned a negative value

$$R^+(\lambda_i^{max}, \lambda_*^{max}) = \begin{cases} -1.5 \quad if \quad \lambda_i^{max} \ge \dfrac{\max \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}; \\ -1 \quad if \quad \lambda_i^{max} < \dfrac{\max \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}, \quad and \quad \lambda_i^{max} \ge \overline{\lambda_k^{max}}; \\ -1 \quad if \quad \lambda_i^{max} < \overline{\lambda_k^{max}}, \quad and \quad \lambda_i^{max} \ge \dfrac{\min \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}; \\ -0.5 \quad if \quad \lambda_i^{max} < \dfrac{\min \lambda_k^{max} + \overline{\lambda_k^{max}}}{2}. \end{cases} \tag{6}$$

Here, $\max \lambda_k^{max}$ is the maximum vertical coordinate of the peaks in the test data, $\overline{\lambda_k^{max}}$ is the mean vertical coordinate of all peaks in the test data, and $\min \lambda_k^{max}$ is the minimum vertical coordinate of the peaks in the test data. Finally, each feature in the database data is assigned a value, and the similarity between the database data and the test data is calculated as follows:

$$R = \frac{\sum R^+(\lambda_i^{max}, \lambda_*^{max}) + \sum R^-(\lambda_i^{max}, \lambda_*^{max})}{N}. \tag{7}$$

Multi-Substance Identification Analysis Algorithm

After preprocessing the test data, the NNLS algorithm is used to obtain a subset of the database that likely contains the target substance from an overall optimization perspective. Then, the similarity function is used to identify the most likely target substance from a local optimization perspective. In this way, the first substance in the mixture corresponding to the test data is identified. By subtracting the product of the database data corresponding to this substance and the weighting coefficient obtained from the NNLS algorithm from the test data, a new set of data is obtained. This new test data is then analyzed again using the NNLS algorithm to obtain a new subset from the database. The spectral peak features of the first target substance are removed from this new subset, and the similarity function is used to identify the second target substance. This process is repeated until all possible substances in the mixture are identified. Specifically, the algorithm is as follows:

Algorithm: Multi-Substance Identification and Analysis Algorithm Based on FTIR Spectroscopy:

Input: Test data, i.e., the mixture to be analyzed;

Output: The names of two or three common chemical substances that match the mixture, along with their similarity values.

1. Import the data of 342 common chemical substances based on FTIR spectroscopy to form the database. Preprocess both the database data and the test data;
2. Calculate the spectral peak feature vectors for both the test data and the database;
3. Use the NNLS algorithm to identify the top ten candidate substances from the database that are similar to the test data for identifying the first target substance;
4. Calculate the similarity values between the ten candidate substances and the test data using the similarity function, sort them, and identify the first target substance with the highest similarity;

5. Calculate the similarity values between the ten candidate substances and the test data using the similarity function, sort them, and identify the first target substance with the highest similarity data;
6. Use the NNLS algorithm again on this new test data to obtain a new subset from the database. Remove the spectral peak features of the first target substance from this new subset, and use the similarity function to identify the second target substance;
7. Repeat steps (5) and (6) to identify the third target substance;
8. Combine the three target substances in 1:1 pairwise combinations and a 1:1:1 combination to form four new simulated mixture datasets;
9. Calculate the spectral peak feature vectors for the four simulated mixtures obtained in step (8);
10. Use the similarity function to calculate the similarity values between the four simulated mixtures and the test data, sort them, and output the names of all substances in the simulated mixture with the highest similarity value, along with its similarity value.

III. Experiments and Results

In the previous section, this paper introduced the steps of the multi-substance identification and analysis algorithm based on Fourier Transform Infrared (FTIR) spectroscopy. In this section, numerical experiments will be conducted to verify the high accuracy and effectiveness of the proposed multi-substance identification and analysis algorithm. This paper uses the Alphapce ALPHAPEC5010 portable FTIR spectrometer (spectral range: 500cm^{-1}~5000cm^{-1}, maximum resolution: 1cm^{-1}, signal-to-noise ratio: 45000:1, transmittance repeatability better than 0.5%T) to sample 342 common chemical substances and 40 mixture samples, which are used as the database data and test data for spectral analysis. The following two multi-substance mixtures will be selected for numerical experiments. These mixtures are chosen because they are representative and can more effectively test the algorithm's performance. The numerical experiments in this paper were conducted on a personal computer equipped with an Intel i5-8250U processor (1.60GHz), 12GB RAM, and the Microsoft Windows 10 operating system, using MATLAB (R2017b). This paper presents the data obtained using the proposed non-negative least squares method and similarity function method. Additionally, to demonstrate the differences in data between different substances, the multi-substance mixtures are divided into three-substance mixtures and two-substance mixtures for data analysis. The data for each stage are listed step-by-step in Tables 1 to 14, and the comparison between the multi-substance mixtures and the fitted spectra of the target substances is displayed in the same window.

Table 1: Three-substance mixtures and the use of non-negative least squares to obtain the first ten alternative substances and non-negative least squares coefficients for the identification of the first target substance

	Multi-substance mixture: Acetone + Styrene + Benzyl Alcohol	
Ranking	Top 10 Candidate Substance Names	Non-negative Least Squares Coefficients
1	Acetone	1.3363
2	Styrene	0.8498
3	Benzaldehyde	0.5398
4	4-Hydroxy-4-methyl-2-pentanone	0.3138
5	Benzyl alcohol	0.2483
6	Biphenyl	0.2136
7	Amphetamine hydrochloride	0.1192
8	Trimethyl borate	0.1090
9	Dimethyl sulfide	0.0952
10	Phenylacetic acid	0.0945

Table 2: Similarity values of the three-substance mixture and the first ten alternative substances used to identify the first target substance compared with each other using the similarity function

	Multi-substance mixture: Acetone + Styrene + Benzyl Alcohol	
Ranking	Top 10 Substances	Similarity Value
1	Acetone	0.8750
2	Styrene	0.5000
3	Biphenyl	0.2222
4	Phenylacetic acid	0.0833
5	4-Hydroxy-4-methyl-2-pentanone	0.0455
6	Benzaldehyde	-0.0667
7	Dimethyl sulfide	-0.2857
8	Amphetamine hydrochloride	-0.3235
9	Trimethyl borate	-0.4167
10	Benzyl alcohol	-0.7308

Table 3: Three-substance mixtures and the use of non-negative least squares to obtain the first ten alternative substances and non-negative least squares coefficients for the identification of the second target substance

Multi-substance mixture: Acetone + Styrene + Benzyl alcohol		
Ranking	Top 10 Substances	Non-negative Least Squares Coefficient
1	Benzyl alcohol	10.1488
2	Diphenylmethanol	11.4775
3	Tetraphenyl-1,2-ethanediol	12.2202
4	Styrene	12.2557
5	Dibenzylamine	12.5690
6	Furfuryl alcohol	12.6376
7	1-Phenylethanol	12.6629
8	Diphenylmethane	12.7440
9	Dibenzyl ether	12.7696
10	Triphenylmethane	12.8199

Table 4: Similarity values of the three-substance mixture and the first ten alternative substances used to identify the second target substance compared using the similarity function

Multi-substance mixture: Acetone + Styrene + Benzyl alcohol		
Ranking	Top 10 Substances	Similarity Value
1	Benzyl alcohol	0.8636
2	Styrene	0.5417
3	Diphenylmethane	0.5000
4	Dibenzylamine	0.3889
5	Diphenylmethanol	0.1923
6	1-Phenylethanol	0.1000
7	Dibenzyl ether	0.0833
8	Tetraphenyl-1,2-ethanediol	0.0000
9	Triphenylmethane	-0.0357
10	Furfuryl alcohol	-0.0909

Table 5: Three-substance mixtures and the use of non-negative least squares to obtain the first ten alternative substances and non-negative least squares coefficients for the identification of the third target substance

Multi-substance mixture: Acetone + Styrene + Benzyl alcohol		
Ranking	Top Ten Substances	Non-negative Least Squares Coefficient
1	4-Hydroxy-4-methyl-2-pentanone	8.7613
2	Benzaldehyde	8.8477
3	Styrene	8.9432
4	4-Phenylcyclohexanone	9.1081
5	Phenylacetic acid	9.1611
6	2-Pyrrolidone	9.3309
7	cis-Stilbene	9.4156
8	Furfural	9.5995
9	Acetophenone	9.6006
10	4-Phenylurazole	9.6161

Table 6: Similarity values of the three-substance mixture and the first ten alternative substances used to identify the third target substance compared using the similarity function

Multi-substance mixture: Acetone + Styrene + Benzyl alcohol		
Ranking	Top Ten Substances	Similarity Value
1	Styrene	0.5000
2	cis-Stilbene	0.0417
3	4-Phenylurazole	-0.1818
4	4-Phenylcyclohexanone	-0.2105
5	Phenylacetic acid	-0.3333
6	Benzaldehyde	-0.3667
7	Acetophenone	-0.4231
8	2-Pyrrolidone	-0.6000
9	4-Hydroxy-4-methyl-2-pentanone	-0.6500
10	Furfural	-0.6667

Table 7: 1:1 combination of a three-substance mixture with the three target substances obtained and 1:1:1 combination and similarity values calculated using the similarity function

Multi-substance mixture: Acetone + Styrene + Benzyl alcohol		
Ranking	Target Substance	Similarity Value
1	Acetone + Styrene + Benzyl alcohol	0.9750
2	Styrene + Benzyl alcohol	0.9444
3	Acetone + Benzyl alcohol	0.9375

4	Acetone + Styrene	0.6786

Table 8: Two-substance mixtures and the use of non-negative least squares to obtain the first ten alternative substances and non-negative least squares coefficients for the identification of the first target substance

Multi-substance mixture: Ethanol + Ethylene glycol		
Ranking	Top 10 Substances	Non-negative Least Squares Coefficient
1	Ethylene glycol	1.2711
2	Ethanol	1.0071
3	N-Methyl-2,2-iminodiethanol	0.6200
4	Benzene	0.6093
5	Anthracene	0.5952
6	cis-1,5-Cyclooctadiene	0.4963
7	Tetrahydrofuran (THF)	0.4316
8	9,10-Diphenylanthracene	0.4213
9	Diiodomethane	0.3919
10	Cyclohexene	0.3761

Table 9: Similarity values of the two-substance mixtures and the first ten alternative substances used to identify the first target substance compared using the similarity function

Multi-substance mixture: Ethanol + Ethylene glycol		
Ranking	Top 10 Substances	Similarity Value
1	Ethanol	0.7500
2	Ethylene glycol	0.6667
3	Diiodomethane	0.2500
4	Tetrahydrofuran (THF)	0.2000
5	cis-1,5-Cyclooctadiene	0.1000
6	Cyclohexene	0.0909
7	N-Methyl-2,2-iminodiethanol	0.0500
8	Benzene	-0.1250
9	Anthracene	-0.1667
10	9,10-Diphenylanthracene	-0.4286

Table 10: Two-substance mixtures and the use of non-negative least squares to obtain the first ten alternative substances and non-negative least squares coefficients for the identification of the second target substance

Multi-substance mixture: Ethanol + Ethylene glycol		
Ranking	Top 10 Substances	Non-negative Least Squares Coefficient
1	Ethylene glycol	13.8772
2	1,5-Pentanediol	16.4148
3	N-Methyl-2,2-iminodiethanol	16.9025
4	Polyvinyl alcohol (PVA)	17.1447
5	Glycerol (or Glycerin)	17.2766
6	2-Propyn-1-ol (Propargyl alcohol)	18.3000
7	Diphenhydramine	18.6301
8	Atropine	18.8749
9	Benzyl alcohol	18.9556
10	Morphine	19.2505

Table 11: Similarity values of the two substance mixtures and the first ten alternative substances used to identify the second target substance compared using the similarity function

Multi-substance mixture: Ethanol + Ethylene glycol		
Ranking	Top 10 Substances	Similarity Value
1	Ethylene glycol	0.7500
2	Glycerol (or Glycerin)	0.1250
3	Benzyl alcohol	0.0625
4	1,5-Pentanediol	0.0000
5	Polyvinyl alcohol (PVA)	-0.0833
6	2-Propyn-1-ol (Propargyl alcohol)	-0.2500
7	Diphenhydramine	-0.2500
8	Atropine	-0.2679
9	N-Methyl-2,2-iminodiethanol	-0.3571
10	Morphine	-0.5455

Table 12: Two-substance mixtures and the use of non-negative least squares to obtain the first ten alternative substances and non-negative least squares coefficients for the identification of the third target substance

	Multi-substance mixture: Ethanol + Ethylene glycol	
Ranking	Top 10 Substances	Non-negative Least Squares Coefficient
1	4-Methyl-2-pentanol	11.1180
2	1-Propanol	11.5467
3	Mixed xylenes	11.5648
4	1-Phenylethanol	11.5981
5	2-Butanol	11.7024
6	Polyvinyl alcohol (PVA)	11.7622
7	Diphenylmethanol	11.8157
8	1,3-Butanediol	11.8292
9	Styrene-butadiene rubber (SBR 1500)	11.8383
10	2-Aminophenol	11.8897

Table 13: Similarity values of the two substance mixtures and the first ten alternative substances used to identify the third target substance compared using the similarity function

	Multi-substance mixture: Ethanol + Ethylene glycol	
Ranking	Top 10 Substances	Similarity Value
1	Diphenylmethanol	-0.2857
2	Polyvinyl alcohol (PVA)	-0.3000
3	1,3-Butanediol	-0.4583
4	Styrene-butadiene rubber (SBR 1500)	-0.5000
5	4-Methyl-2-pentanol	-0.5625
6	Mixed xylenes	-0.5625
7	2-Butanol	-0.5833
8	1-Phenylethanol	-0.6000
9	2-Aminophenol	-0.6750
10	1-Propanol	-0.7727

Table 14: 1:1 combinations of two-substance mixtures with the three target substances obtained and 1:1:1 combinations and similarity values calculated using the similarity function

	Multi-substance mixture: Ethanol + Ethylene glycol	
Ranking	Target Substance Data Simulation	Similarity Value
1	Ethanol + Ethylene glycol	0.8750
2	Ethanol + Ethylene glycol + Diphenylmethanol	0.4375
3	Ethanol + Diphenylmethanol	0.4167
4	Ethylene glycol + Diphenylmethanol	0.2222

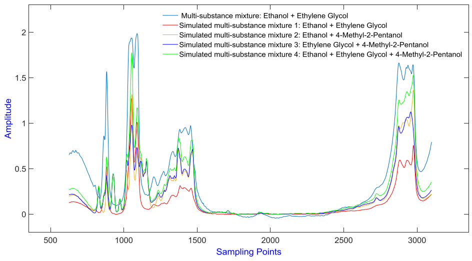

Figure 3: Comparison of Fourier infrared spectra of three-substance mixtures and Fourier infrared spectra of simulated multi-substance mixtures

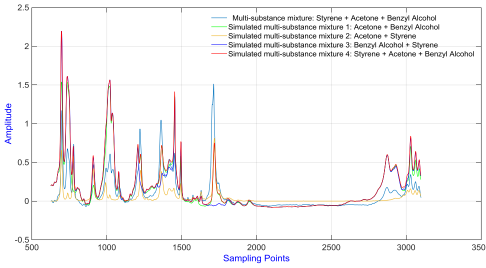

Figure 4: Comparison of Fourier infrared spectra of two-substance mixtures and Fourier infrared spectra of simulated multi-substance mixtures

By examining Tables 1 to 14, the method combining overall and local optimization demonstrates high accuracy and reliability in identifying mixture components. The non-negative least squares method identifies a subset of the database containing target substances, but the NNLS coefficients are often very close, making it difficult to determine the most similar substance. Therefore, local optimization is applied within this subset to identify the target substance. Analysis of Tables 2, 4, 6, 9, 11, and 13 shows that similarity values are positively correlated with the match between substances and test data: higher values indicate better matches, while values below 0 indicate poor fits. Table 7 shows that the similarity values of the three target substances with the test data exceed 0.5, and subsequent verification confirms a high degree of fit (close to 1). In contrast, Table 14 shows that only two target substances have similarity values above 0.5, while the third has a value below 0, indicating a poor fit. Figures 3 and 4 further illustrate that higher similarity values correspond to better matches between simulated mixtures and test data.

In summary, the combination of the non-negative least squares method and the similarity function enables accurate identification of true substances in multi-substance identification analysis. The non-negative least squares method ensures that a subset containing the true substances in the test data is obtained from the entire database, while the similarity function can accurately and quickly identify the true substances in the test data within the optimal range. It can be concluded that the method proposed in this paper not only effectively and quickly identifies the components of multi-substance mixtures but also demonstrates high practicality and reliability in real-world applications.

IV. Conclusion

This paper proposes a multi-substance identification and analysis algorithm based on Fourier Transform Infrared spectroscopy. By integrating overall and local optimization strategies, the method significantly improves the accuracy of substance identification. It utilizes the non-negative least squares algorithm, a similarity function matching algorithm, and an iterative process to achieve multi-substance identification. Theoretically, the algorithm can iterate indefinitely until all substance components in the test data are identified. Experimental results demonstrate that the method effectively leverages both overall and local optimization, enabling accurate qualitative analysis and identification of multi-substance mixtures.

References

[1]. Pan Dongning, Zhao Leihong, Xie Wenjun. Detection method of precursor mixed gases based on neural network and Fourier spectral analysis[J]. Infrared, 2024, 45(4): 46-52.

[2]. Shen Yunxia, Zhao Yanli, Zhang Ji, et al. Identification of different processed Gentiana rigescens by Fourier transform infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2016, 36(5): 1369-1373.

[3]. Curk M C, Peledan F, Hubert J C. Fourier transform infrared (FTIR) spectroscopy for identifying Lactobacillus species[J]. FEMS microbiology letters, 1994,123, (3):241-248.

[4]. Jiang An. Research on qualitative analysis methods of complex mixtures based on infrared spectroscopy[D]. Graduate University of Chinese Academy of Sciences, 2012.

[5]. Liu Caizheng, Zhu Qibing, Huang Min, et al. Component identification method of mixtures based on Raman spectroscopy[J]. Laser & Optoelectronics Progress, 2019, 56(8): 253-259.

[6]. Chen Bin, Zheng Xiaohuan, Geng Dechun, et al. Research on component identification of mixtures based on multi-parameter fusion infrared spectroscopy[J]. Journal of Instrumental Analysis, 2023, 42(3): 323-329.

[7]. Fang Yonghua, Kong Chao, Lan Tiange, et al. Noise removal and baseline correction of spectra using wavelet transform[J]. Optics and Precision Engineering, 2006, 14(6): 1088-1092.

[8]. Ching, PC,So, HC,Wu, SQ.On wavelet denoising and its applications to time delay estimation[J].IEEE TRANSACTIONS ON SIGNAL PROCESSING,1999,47,(10):2879-2882.

[9]. Wang Xin, Lü Shilong, Li Yan, et al. Automatic baseline correction of gas spectra based on baseline drift model[J]. Spectroscopy and Spectral Analysis, 2018, 38(12): 3946-3951.

[10]. Sun Yi, Du Zhenhui, Yin Xin, et al. Research on baseline correction methods for online gas analysis using near-infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2008, 28(10): 2282-2284.

[11]. Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Recent advances in research and application of near-infrared spectroscopy analysis technology in China[J]. Analytical Instruments, 2006, (2): 1-10.

[12]. Ning Zhiqiang, Liu Jiaxiang, Wu Yue, et al. Infrared spectrum baseline correction method based on improved iterative polynomial fitting[J]. Laser & Optoelectronics Progress, 2020, 57(3): 247-253.

[13]. Gao Rongqiang, Fan Shifu, Yan Yanlu, et al. Research on data preprocessing of near-infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2004, 24(12): 1563-1565.

[14]. Chen, Yanyan,Zou, Caineng,Mastalerz, Mariaet al.Applications of Micro-Fourier Transform Infrared Spectroscopy (FTIR) in the Geological Sciences-A Review[J].INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES,2015,16,(12):30223-30250.

[15]. Peng, Jiangtao,Peng, Silong,Xie, Qiong et al.Baseline correction combined partial least squares algorithm and its application in on-line Fourier transform infrared quantitative analysis[J].ANALYTICA CHIMICA ACTA,2011,690,(2):162-168.

[16]. Sun Yanbo. Equivalence among non-negative linear least squares problem, linear complementarity problem, and fixed-point problem[J]. Journal of Anhui Normal University (Natural Science Edition), 2015, 38(6): 537-540.

[17]. Zhou Shaomin, Liu Jianxin, Sun Huanle. Research on one-dimensional MT regularization inversion based on non-negative least squares method[J]. Chinese Journal of Engineering Geophysics, 2017, 14(3): 253-261.

[18]. Lawson C L, Hanson R J. Solving least squares problems[M]. Society for Industrial and Applied Mathematics, 1995.

[19]. He Yongbin, Fan Xiaotao, An Hongyan, et al. Theoretical analysis of solutions for linear least squares problems[J]. Journal of Chengdu University of Technology (Natural Science Edition), 2003, 30(5): 529-533.

[20]. Gander W. Least squares with a quadratic constraint[J]. Numerische Mathematik, 1980, 36(3): 291-307.

[21]. Yong Longquan. A strictly feasible interior-point algorithm for non-negative linear least squares problems[J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2010, 26(4): 84-89, F0003.