Statistical approach to determining the cut-off value in the algorithms of substance identification in Fourier transform infrared spectroscopy

A A Oranov¹,², S V Bashkin¹ and I L Fufurin¹,³

¹ Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5/1, Moscow, Russian Federation, 105005
² Email: alexandr.oranov@gmail.com
³ Email: igfil@mail.ru

Abstract. The purpose of the paper was to find the approach to determining the cut-off value in the spectral library matching identification algorithm. We used Pearson correlation coefficient as a similarity measure between the spectra, and according to its value between the reference spectra we divided the spectral library into classes. We considered a situation when one of the investigated spectra had an additive narrowband white noise component with a Gaussian distribution. We examined the impact of the cut-off value on the possibility of calculating the probability of right identification in a single identification experiment by means of probability distribution function, when the signal-to-noise ratio is known a priori. The signal-to-noise ratio was taken from the value area where analytical expressions were unknown. In the research we applied numerical methods. The algorithm of determining the cut-off value based on the fundamental principles of the ROC-curves analysis was proposed. The cut-off value was obtained where possible, and the situation for every class was analysed. Maximal probabilities of right identification were obtained.

1. Introduction
The problem of remote control of air pollution is relevant nowadays. Besides studying the atmosphere, determining the presence of toxic and hazardous substances in industrial emissions and in places of technological disasters is of great interest. Due to the fact that the presence of research staff in such conditions may be harmful to health, non-sampling methods of chemical analysis are used.

One of the characteristic properties, individual for each substance, is the optical spectrum, which allows using it as an identifier. Currently, there is a large number of optical devices based on various technologies. One of the most popular technologies is Fourier transform infrared spectroscopy (FTIR) [1]. The infrared (IR) spectrum of a substance corresponds to the oscillatory motion of the specific functional groups presented in the molecule as well as the overall molecular structure. Therefore, the IR spectrum of each substance is unique and FTIR plays a crucial role in identification of organic and inorganic substances. One of the main benefits of FTIR is the ability to work with lower intensities [1]. This feature makes it possible to detect even the self-emission spectrum of the substance. Some features and applications of FTIR is considered in [2 – 7].

In portable chemical express analysis devices and express monitoring complexes, an identification method based on the spectral library matching is widely used. The spectral library consists of the spectra obtained under laboratory conditions with a high signal-to-noise ratio (SNR). The main work
on searching and comparing the spectra of substances in the spectral library has been done by Clerk [8] and Luinge [9]. This method seems to be more reliable and accurate than such methods as PCA (principal component analysis) [10] or multilayer neural networks [11]. In order to compare the two spectra, a certain quantitative characteristic should be introduced. This characteristic is usually called a similarity measure. Different similarity measures were observed in [12, 13].

In every spectral library matching identification algorithm with any similarity measure between spectra, the cut-off value plays an important role. The choice of the cut-off value affects the ability of the identification algorithm to distinguish between noise and the spectrum of substances. Moreover, if the investigated spectrum has noise components, the cut-off value determines the probability of correct identification of the substance. Hence, determination of the cut-off value is an important objective.

2. Pearson correlation coefficient
One of the possible similarity measures is Pearson correlation coefficient [14]. In some sources it is also called SCM (spectrum correlation measure). Pearson correlation coefficient is commonly used in identification tasks [15–17].

\[
\rho_{ij} = \frac{\mathbf{\bar{r}_i}(\nu) - \mathbf{\bar{r}_i}(\nu)^T \cdot \mathbf{\bar{r}_j}(\nu) - \mathbf{\bar{r}_j}(\nu)^T \cdot \mathbf{\bar{r}_j}(\nu)}{\| \mathbf{\bar{r}_i}(\nu) - \mathbf{\bar{r}_i}(\nu) \| \cdot \| \mathbf{\bar{r}_j}(\nu) - \mathbf{\bar{r}_j}(\nu) \|},
\]

where \( \mathbf{\bar{r}_i}(\nu) \) and \( \mathbf{\bar{r}_j}(\nu) \) are vectors representing comparable IR spectra (here and further spectra are presented as algebraic vectors of finite length).

From expression (1) it follows that \( \rho_{ij} \in [-1,1] \). Equality to unity takes place when the spectra compared are identical. As \( \mathbf{\bar{r}_i}(\nu) \) and \( \mathbf{\bar{r}_j}(\nu) \) are represented by the set of points specified by fragmentation \( G \) in the frequency range \( [v_{\text{beg}}, v_{\text{end}}] \): \( G = \{ v_i \mid i = 1..N, v_{\text{beg}} < v_1 < ... < v_N < v_{\text{end}} \} \), there exists a measurement error in every point. When the random error is much smaller than the error due to the different nature of the substances, Pearson correlation coefficient can be used as an identifier [17].

In [18, 19] the identification algorithm based on Pearson correlation coefficient calculation was described as follows. The correlation coefficient of the investigated spectrum with each spectrum from the spectral library is calculated according to formula (1), then the maximum value of the correlation coefficient should be chosen, and if it exceeds the predetermined cut-off value, the corresponding substance is considered to have been identified.

3. Numerical identification experiment set up
For all further simulations, computations and graphics, Python 3.6 was used.

3.1. Model of compared spectra
Let us consider a situation when one of the measured spectra has an additive narrowband white noise component with a Gaussian distribution. Spectral library consists of \( M \) spectra: \( \{ \mathbf{r}_1, \mathbf{r}_2,...,\mathbf{r}_M \} \). In numerical identification experiment two spectra are compared. Reference spectrum of \( k \)-th substance from spectral library is represented by vector \( \mathbf{\bar{r}}_k \). Investigation of spectrum of some \( k' \)-th substance, which is thought to be obtained with some noise, is represented by vector \( \mathbf{\bar{r}}_{k'} = \mathbf{\bar{r}}_k + \mathbf{\bar{e}} \). Each component of noisy vector \( \mathbf{\bar{e}} \) matches with normal distribution \( N(0,\sigma^2_e) \) with determined parameters, and is uncorrelated with others \( \langle \mathbf{\bar{e}}_i, \mathbf{\bar{e}}_j \rangle = \sigma^2_e \delta_{ij} \).

3.2. Spectral library
For numerical experiment a spectral library was formed, including \( M = 9 \) reference IR spectra of the following substances: 1 – Pentanol \( (C_5H_{11}OH) \); 2 – Propanol \( (C_3H_8O) \); 3 – Butanol \( (C_4H_{10}O) \); 4 –
Heptane (C\textsubscript{7}H\textsubscript{16}); 5 – Methanol (CH\textsubscript{3}OH); 6 – Ethanol (C\textsubscript{2}H\textsubscript{5}OH); 7 – Ethylbenzene (C\textsubscript{8}H\textsubscript{10}); 8 – Benzene (C\textsubscript{6}H\textsubscript{6}); 9 – Toluene (C\textsubscript{7}H\textsubscript{8}). The spectra were taken in the spectral range [500 sm\textsuperscript{-1}, 4000 sm\textsuperscript{-1}] with the spectral resolution R = 4 sm\textsuperscript{-1}. The source was the open IR spectral library of National Institute of Standards and Technology (NIST), United States of America (USA).

The spectra in the spectral library were divided into classes according to Pearson cross-correlation matrix, presented in figure 1. For class 1 there are values of correlation coefficients with other substances from the spectral library higher than 0.90, for class 2 – the values are not higher than 0.84, for class 3 – the values are not higher than 0.67.

4. Numerical identification experiment

Let us express the signal-to-noise ratio as \( SNR = \sigma^2_{x_k} \cdot \sigma^{-2}_{\xi} \). In [10] analytic expression of Pearson correlation coefficient, we obtained a function of SNR for the described model. According to it, we calculated all characteristics of Pearson correlation coefficient as a random value and defined the cut-off values for the desired probability of right identification. Significant restriction of the results obtained was the limitation: \( SNR \gg 0.5 \). So, for the\( SNR \leq 0.5 \), we can apply only numerical methods. We considered this situation, and all identification experiments were carried out at \( SNR = 0.5 \).

4.1. Probability of right identification

If signal-to-noise ratio is known a priori, and single identification experiment showed that the identified substance belongs to the spectral library, then the probability of right identification can be calculated by means of density of probability distribution function \( \rho(\xi) \):

\[
P = \frac{1}{r^*} \rho(\xi),
\]

where \( P \) is the probability of right identification; \( r^* \) is the cut-off value; \( \rho(\xi) \) is the correlation coefficient of the reference spectrum with the same noisy reference spectrum. Function \( \rho(\xi) \) can be obtained as a result of series of numerical experiments of calculating (1) many times with fixed SNR. The typical view of this function is presented in figure 2. The probability of right identification calculated by (2) is shown in figure 3.
Figure 2. Probability distribution function. Every point was obtained as a result of $10^3$ numerical experiments of calculating Pearson correlation coefficient for pentanol and pentanol’s noisy spectrum with different SNR values.

Figure 3. Dependence of probability of right identification from the cut-off value, calculated by means of probability distribution function of pentanol with $SNR=0.5$.

Significant limitation of this approach is that the investigated spectrum must contain spectral lines of the reference spectrum and the spectral library consists of a single substance. So, the aim is to find such cut-off values for every substance, which will make (2) appropriate to be uses with the spectral library consisting of several substances without depending on the investigated spectrum.

4.2. Determination of the cut-off values

To find the cut-off values, we will use the ROC-curves analysis. The fundamental principles of this approach are described in table 1. The following characteristics of the identification algorithm are very important: True Positive Rate – $TPR = TP \cdot (TP + FN)^{-1}$; False Positive Rate – $FPR = FP \cdot (FP + TN)^{-1}$. The curves TPR and FPR depend on the cut-off values.

Table 1. Confusion matrix for the identification algorithm. Abbreviations: $TP$ – true positives; $FP$ – false positives, $FN$ – false negatives, $TN$ – true negatives. The presented concepts are always used to check the correctness of the work algorithm.

| Predicted Lines of reference spectrum are present | Actual Lines of reference spectrum are present | Actual Lines of reference spectrum are not present |
|-------------------------------------------------|---------------------------------------------|-----------------------------------------------|
| TP                                              | FN                                           | TP                                             |
| FP                                              | TN                                           | FN                                             |

In our case, equality of $FPR$ to zero means that from some cut-off value $r^*$, the algorithm doesn’t identify noise as a spectrum of some substance from the spectral library. If the cut-off value $r^* > r^*_1$, two situations are possible. In the first case, if the algorithm identifies some substance, then it is truly this substance and $TPR$ curve starts to coincide with the probability of the right identification curve. In
the second case, which can take place when the reference spectra are similar enough, the algorithm can confuse substances with each other due to the overlapping of probability distributions. To study this situation, we should introduce some new characteristic \(-\ \text{TPR}' = \text{TP} \cdot \left( \text{TP} + \text{FN} + \text{WP} \right)^{-1}\). The addend, which we will call \(\text{WP}\) (Wrong Positives), represents the cases, when the algorithm identifies noisy spectrum of \(k\)-th substance from the spectral library as a noisy spectrum of \(k'\)-th substance and gives the wrong output (\(k \neq k'\)). If it takes place, the cut-off value should be higher than value \(r^*_2\), i.e. determined by the equality of \(\text{WP}\) to zero, so, \(\text{TPR}' = \text{TPR} \) for every \(r^* > r^*_2\). The example of this situation for one of the substances from the spectral library is illustrated in figure 4.

![Figure 4. Determination of the cut-off value for heptane (class 2). Every point was obtained as a result of \(10^4\) numerical experiments of substance identification (\(\text{SNR}=0.5\)).](image)

Findings of research show that the algorithm of determination of the cut-off value consists in the following steps:

1. It is necessary to carry out a number of identification experiments for the chosen substance from the spectral library for different \(r^*\). In every series of the experiments, in the input of the identification algorithm the noisy spectrum of the chosen substance and pure noise should be entered alternately.
2. Based on the results obtained, it is necessary to calculate \(\text{FPR}\) and \(\text{TPR}'\) for every series, and then plot \(\text{FPR}\), \(\text{TPR}\) and \(\text{TPR}'\) curves for the whole range of \(r^*\) values.
3. Then, the value \(r^*_1\) is determined. It should match the first equality \(\text{FPR}\) to zero. It is important to check if there are any cases of wrong identification (\(\text{WP} \neq 0\)) for \(r^* > r^*_1\) in \(\text{TPR}'\). If not, then the cut-off value \(r^* = r^*_1\) is found, and the value \(r^*_2\), that matches the first equality \(\text{WP}\) to zero in \(\text{TPR}'\) is determined, and \(r^* = r^*_2\) is chosen.

5. Results
After applying the described algorithm of determining the cut-off values to every substance from the spectral library, we generalized the classes. The general cut-off value for the class was chosen as the highest of the cut-off values for each substance of this class, as they were very close to each other. So, for identifying the substance we had to choose the general cut-off value for its class, and this substance was identified with the corresponding probability. All results are presented in table 2.

Table 2. The results of determination of general cut-off values and corresponding probabilities of right identification in a single identification experiment
| General cut-off value | Probability of right identification |
|-----------------------|-------------------------------------|
| Class 1               | Undefined                           |
| Class 2               | $r^* = 0.565$                       |
|                      | $P = 0.52$                          |
| Class 3               | $r^* = 0.150$                       |
|                      | $P = 1.00$                          |

As it is seen from table 2, the cut-off values for the substances from class 1 and, as follows, the general cut-off value, and the probability of right identification in a single identification experiment could not be calculated in the proposed way. Despite this fact, we were able to determine that the substance under study definitely belongs to class 1 without claiming what concrete substance it is. It will be enough indeed for identifying hazardous substances similar to each other. For class 2, we also obtained the results, but the observable frequency of right identification with the lower cut-off values is higher, so other methods should be used for better results.

6. Conclusion
In this paper, we explored the problem of determining the cut-off value. The criterion for the determination was to make appropriate the ability to calculate the probability of right identification in a single identification experiment by means of probability distribution function of Pearson correlation coefficient of reference spectrum with same noisy reference spectrum in case of a very low signal-to-noise ratio. We found out that for the spectral library, divided on classes due to the similarity of the reference spectra, the general cut-off values can be separately determined for every class. We obtained the general cut-off values for the classes and calculated the corresponding probabilities of right identification. Dividing the spectral library into classes makes it possible to identify whether the investigated substance belongs to the definite class even if the probability of right identification of concrete substance can not be calculated.

It should be noted, that the proposed way to determine the cut-off value could be used for any criteria similarity measures and measures based on calculating the quantity difference between the compared spectra.

References
[1] Morozov A N and Svetlichniy S I 2016 Osnovi furie-spectroradiometrii [Fundamentals of Fourier spectroradiometry] (Moscow: Nauka)
[2] Kochikov I V, Morozov A N and Fufurin I L 2012 Numerical procedures for substances identification and concentration calculation in the open atmostphere by processing a single ftir measurement Comp. Optics 36(4) pp 554–61
[3] Morozov A N, Tabalin S E, Fufurin I L and Svetlichnyi S I 2013 Physical principles of the design of an interferometer with a rotating plate, Journal of Optical Technology A Translation of Opticheskii Zhurnal 80(8) pp 495–98
[4] Bashkin S V, Boiko A Yu, Kornienko V N, Kochikov I V, Le'l'kov M V, Mironov A I, Morozov A N, Pozvonkov A A, Samorodov A S, Svetlichnyi S I, Tabalin S Ye, Fufurin I L and Shlygin P E 2016 Experimental results of investigating panoramic fourier transform infrared spectrometer Herald of the Bauman Moscow State Technical University, Series Natural Sciences 2(65)
[5] Samsonov D A, Tabalina A S and Fufurin I L 2017 QCL spectroscopy combined with the least squares method for substance analysis Journal of Physics: Conf. Series 918
[6] Morozov A N, Svetlichnyi S I and Fufurin I L 2015 Correction of wavenumbers in a FTIR imaging spectrometer Doklady Physics 60(9) pp 388–91
[7] Bashkin S V, Vagin V A, Vintaikin I B, Dvoruk S K, Kornienko V N, Kochikov I V, Mironov A I, Morozov A N, Svetlichnyic S I, Tabalin S E and Fufurin I L 2017 Calculation of instrumental functions as a method for analyzing the alignment of the IR channel of a scanning Fourier-transform IR spectroradiometer *Optics and Spectroscopy* **123**(4) pp 658–63

[8] Clerc J T, Pretsch E and Zurcher M 1986 Performance Analysis of Infrared Library Search Systems *J. Chem. Inf. Comput. Sci.* **25**(3) pp 241–4

[9] Luinge H J 1990 Automated interpretation of vibrational spectra *Vib. Spectrosc.* **1**(1) pp 1–8

[10] Schoonjans V and Questier F 2001 Assessing molecular similarity/diversity of chemical structures by FT-IR spectroscopy *J. of Pharmac. and Biomed. Anal.* **24**(4) pp 613-27

[11] Hemmer M C and Gasteiger J 2000 Prediction of three-dimensional molecular structures using information from infrared spectra *Anal. Chim. Acta* **420**(2) pp 145–54

[12] Zurcher M, Clerc J T, Farkas M and Pretsch E 1988 General theory of similarity measures for library search *Anal. Chim. Acta* **206**(0) pp 161–172

[13] Lebedev K S 1993 Data base usage in IR and mass spectroscopy for the structural detection of organic substances *J. of Anal. Chem.* **48** pp 851-63

[14] Pearson K 1895 Notes on regression and inheritance in the case of two parents *Proc. Royal Society of London* **58** pp 347–52

[15] Beil A, Daum R, Matz G and Harig R 1998 Remote sensing of atmospheric pollution by passive FTIR spectrometry in Spectroscopic Atmospheric Environmental Monitoring *Proc. SPIE* **3493** pp 32–43

[16] Morozov A N, Kochikov I V, Novgorodskaya A V, Sologub A A and Fufurin I L 2015 Statistical estimation of the probability of the correct substance detection in FTIR spectroscopy *Comp. Optics* **39**(4) pp 614-21

[17] Vasiljev N S and Morozov A N 2014 Subsncie identification by error deformed spectra *Comp. Optics* **38**(4) pp 856–64

[18] Clerbaux C, Hadji-Lazaro J, Turquety S, Mège G and Coheur P-F 2003 Trace gas measurements from infrared satellite for chemistry and climate applications *Atm. Chem. and Phys.* **3**(5) pp 1495–508

[19] Harig R and Matz G 2001 Toxic cloud imaging by infrared spectrometry: A scanning FTIR system for identification and visualization *Field Anal. Chem. & Techn*. **5**(1-2) pp 75–90