Anal. Chem. **2010,** *82,* 106–114

Information Extraction from a Complex Multicomponent System by Target Factor Analysis

Limin Shao

Department of Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

Peter R. Griffiths*

Department of Chemistry, University of Idaho, Moscow, Idaho 83844-2343

A theoretical investigation into the mechanism of information extraction by target factor analysis (TFA) is presented from experimental data in the form of a matrix, and the results were validated using composite spectra obtained by open-path Fourier transform-infrared (FT-IR) spectrometry. The composite spectra were generated by adding the spectral information of a target molecule with known path-integrated concentrations to the raw open-path FT-IR spectra obtained in a pristine atmosphere. Target molecules are deemed to be detected when the weighted correlation coefficient between the calculated spectrum of the analyte and its reference spectrum exceeds 0.90. The effective detection by TFA is shown to depend on the variation of their concentrations over the period of the measurement and not necessarily on the magnitude of concentration. When TFA fails to detect an analyte at high, but relatively constant, concentration that varies so little as to have low variance, blank spectra, i.e., spectra in which the analyte is known to be absent, are included in the data matrix. This procedure effectively increases the variance of the concentrations in the whole data set, and TFA detects the analyte.

Target factor analysis (TFA)[1] is a self-modeling technique that rotates purely mathematical results obtained by principal component analysis (PCA) into vectors of physical significance, such as spectra or concentration profiles of pure components. Applications of TFA in chemistry are found widely in chromatography,[2−4] reaction mechanics and kinetics,[5−7] and spectroscopic analysis.[8,9] Besides in those conventional fields, TFA is also applied in medical and pharmaceutical research.[10,11] Results of those applications demonstrate the efficiency and effectiveness of TFA to handle large and complex data sets.

In a previous report,[12] we demonstrated that TFA could be used to identify the presence of trace compounds in air from open-path Fourier transform-infrared (OP/FT-IR) spectra measured in a continuous monitoring session. The results indicated that TFA could extract the spectrum of a given target compound even when its spectral features are obscured in the raw spectrum either because of its low concentration or because of serious spectral interferences. TFA shows the potential to decrease the limit of detection (LOD) of OP/FT-IR spectrometry and to allow a warning to be given when the target molecule is present at a concentration above its LOD. In this article, we examine some of the quantitative aspects of TFA, especially the conditions needed to detect the presence of molecules at very low concentration.

THEORY

Throughout this article, boldface lower- and upper-case letters denote vectors and matrices, respectively. All vectors are column vectors, the transpose of which are row vectors, indicated with superscript t. The subscript is the matrix size.

In analytical chemistry, the acquisition of data in a series of measurements during a temporal process such as chromatography or continuous process monitoring is fairly common. In such cases, a bilinear matrix may be constructed by arranging these data in a row-wise manner. Let us consider the case of a series of spectral measurements where the spectra are converted to absorbance (in a format such that the ordinate scale varies approximately linearly with the concentration of each component.) Suppose $\mathbf{D}_{m\times n}$ is such a matrix that comprises m spectra; with each spectrum having n data points. The mathematical bilinear model for $\mathbf{D}$ is

$$\mathbf{D}_{m\times n} = \mathbf{C}_{m\times p}(\mathbf{S}_{n\times p})^{\mathrm{t}} = \sum_{i=1}^{p} \mathbf{c}_i(\mathbf{s}_i)^{\mathrm{t}} \qquad (1)$$

* To whom correspondence should be addressed.

(1) Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; John Wiley and Sons: New York, 2002.
(2) McCue, M.; Malinowski, E. R. *Appl. Spectrosc.* **1983**, *37*, 463–469.
(3) Gemperline, P. J. *Anal. Chem.* **1986**, *58*, 2656–2663.
(4) van Zomeren, P. V.; Metting, H. J.; Coenegracht, P. M. J.; de Jong, G. J. *J. Chromatogr., A* **2005**, *1096*, 165–176.
(5) Tam, K. Y.; Chau, F. T. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 25–42.
(6) Carvalho, A. R.; Wattoom, J.; Zhu, L.; Brereton, R. G. *Analyst* **2006**, *131*, 90–97.
(7) Day, J. P. R.; Campbell, R. A.; Russell, O. P.; Bain, C. D. *J. Phys. Chem. C* **2007**, *111*, 8757–8774.
(8) Liang, X.; Andrews, J. E.; de Haseth, J. A. *Anal. Chem.* **1996**, *68*, 378–385.
(9) Kuzmanovski, I.; Trpkovska, M.; Soptrajanov, B.; Stefov, V. *Vib. Spectrosc.* **1999**, *19*, 249–253.
(10) Tetteh, J.; Mader, K. T.; Andanson, J.-M.; McAuley, W. J.; Lane, M. E.; Hadgraft, J.; Kazarian, S. G.; Mitchell, J. C. *Anal. Chim. Acta* **2009**, *642*, 246–256.
(11) Kauffmana, J. F.; Dellibovi, M.; Cunninghamb, C. R. *J. Pharm. Biomed. Anal.* **2007**, *43*, 39–48.
(12) Shao, L.; Griffiths, P. R. *Anal. Chem.* **2007**, *79*, 2118–2124.

10.1021/ac901246x CCC: $40.75
Published on Web 12/04/2009

where p is the number of compounds and $\mathbf{C}$ and $\mathbf{S}$ are the concentration and spectral matrices of all compounds, respectively. Vectors $\mathbf{c}_i$ and $\mathbf{s}_i$ represent the concentration array and the spectrum of the ith compound, respectively; matrix $\mathbf{c}_i(\mathbf{s}_i)^t$ is the contribution of this compound to the entire data matrix, and is referred to as the concentration-spectral data. Additionally, $\mathbf{c}_i$ and $\mathbf{s}_i$ are the ith column vectors of matrices $\mathbf{C}$ and $\mathbf{S}$, respectively. The first step of TFA is principal component analysis (PCA), which is well-known in analyzing matrix data. Therefore, we tried to find the connection between the purely mathematical results of PCA on $\mathbf{D}_{m\times n}$ and some physical properties of the experimental data.

PCA on Bilinear Matrix of Experimental Data. Let us assume that the covariance between any two concentration vectors or any two spectral vectors is relatively low and is negligible compared to the variance of either of the two vectors. This assumption is reasonable for real experimental data because in practice it is rare that two different compounds have such similar spectra or concentration profiles that a high covariance occurs.

From a mathematical perspective, the goal of PCA on matrix $\mathbf{D}$ is to find a way to combine all column vectors so that the combination has maximum variance. The combination is called the first principal component score vector, and the vector to give the combination is called the first principal component loading vector. Therefore, to perform PCA is to obtain a vector, $\mathbf{a}$, that maximizes the variance of the product, $\mathbf{Da}$, subject to the constraint $\mathbf{a}^t\mathbf{a} = 1$. Let V denote the variance of $\mathbf{Da}$, i.e., the first principal component score vector, then we have

$$V = \mathrm{var}(\mathbf{Da}) \tag{2}$$

where "var" denotes the calculation of variance.

Inserting eq 1 into eq 2 yields

$$V = \mathrm{var}\left(\sum_{i=1}^{p} \mathbf{c}_i(\mathbf{s}_i)^t\mathbf{a}\right) \tag{3}$$

$(\mathbf{s}_i)^t\mathbf{a}$ is a scalar, so eq 3 can be rewritten as

$$V = \mathrm{var}\left(\sum_{i=1}^{p} ((\mathbf{s}_i)^t\mathbf{a})\mathbf{c}_i\right) \tag{4}$$

If the covariance of any two of the concentration arrays can be neglected, the variance V approximately equals

$$V = \sum_{i=1}^{p} \left((\mathbf{s}_i)^t\mathbf{a}\right)^2\mathrm{var}(\mathbf{c}_i) = \sum_{i=1}^{p} \left((\mathbf{s}_i)^t\mathbf{a}\right)^2 v_i^c \tag{5}$$

where v_i^c is the variance of concentration array $\mathbf{c}_i$. With the use of the Lagrange multiplier method to maximize V, a new objective function, L, is found

$$L = V - \lambda(\mathbf{a}^t\mathbf{a} - 1) = \sum_{i=1}^{p} \left((\mathbf{s}_i)^t\mathbf{a}\right)^2 v_i^c - \lambda(\mathbf{a}^t\mathbf{a} - 1) \tag{6}$$

where λ is the Lagrange multiplier. Differentiating L with respect to $\mathbf{a}$ and setting the derivative equal to zero, we have

$$\frac{\partial L}{\partial \mathbf{a}} = 2\sum_{i=1}^{p} (\mathbf{s}_i)^t\mathbf{a}v_i^c\mathbf{s}_i - 2\lambda\mathbf{a} = 0 \tag{7}$$

Rearranging eq 7 gives

$$\sum_{i=1}^{p} (\mathbf{s}_i)^t\mathbf{a}v_i^c\mathbf{s}_i = \lambda\mathbf{a} \tag{8}$$

In eq 8, $(\mathbf{s}_i)^t\mathbf{a}$ is actually the dot product of vectors $\mathbf{s}_i$ and $\mathbf{a}$. The dot product can be written as $||\mathbf{s}_i||\cdot||\mathbf{a}||\cos\theta_i$, where $||\mathbf{s}_i||$ and $||\mathbf{a}||$ are the lengths of $\mathbf{s}_i$ and $\mathbf{a}$, and equal $((\mathbf{s}_i)^t\mathbf{s}_i)^{1/2}$ and 1, respectively; θ_i is the angle between the two vectors. Thus eq 8 can be rewritten as

$$\sum_{i=1}^{p} \sqrt{(\mathbf{s}_i)^t\mathbf{s}_i}v_i^c\cos\theta_i\mathbf{s}_i = \lambda\mathbf{a} \tag{9}$$

Equation 9 indicates that the first principal component loading vector in the PCA of the bilinear matrix of experimental data, $\mathbf{a}$, is a weighted sum of the spectra of all compounds, $\mathbf{s}_i$, or can be regarded as the weighted average of those spectra. The weight for the jth compound is determined by two parts, $((\mathbf{s}_j)^t\mathbf{s}_j)^{1/2}v_j^c$ and the angle between vectors $\mathbf{s}_j$ and $\mathbf{a}$. The higher is $((\mathbf{s}_j)^t\mathbf{s}_j)^{1/2}v_j^c$, the more is spectrum $\mathbf{s}_j$ included in $\mathbf{a}$, the closer to zero is angle θ_j, and the larger is the weight. For the jth compound, the variance of the concentration-spectral data, i.e., the matrix $\mathbf{c}_j(\mathbf{s}_j)^t$, is $((\mathbf{s}_j)^t\mathbf{s}_j)v_j^c$. Therefore, the variances of the concentration-spectral data of the compounds are closely related to the constitution of the first principal component loading vector. If the variance of the concentration-spectral data of the jth compound is higher than those of the remaining compounds, the first principal component loading vector is composed mainly of the spectrum of this compound, i.e., $\mathbf{s}_j$.

Next we investigate the constitution of the first principal component score vector. Let $\mathbf{u}$ denote the first principal component score vector of $\mathbf{D}$; it is obtained by

$$\mathbf{u} = \mathbf{Da} \tag{10}$$

where $\mathbf{a}$ is the first principal component loading vector discussed above. Inserting eq 1 into eq 10 gives

$$\mathbf{u} = \sum_{i=1}^{p} \mathbf{c}_i(\mathbf{s}_i)^t\mathbf{a} \tag{11}$$

$(\mathbf{s}_i)^t\mathbf{a}$ is a scalar, so eq 11 can be rewritten as

$$\mathbf{u} = \sum_{i=1}^{p} \left((\mathbf{s}_i)^t\mathbf{a}\right)\mathbf{c}_i \tag{12}$$

Equation 12 shows that the first principal component score vector, $\mathbf{u}$, is the weighted sum of the concentration profiles of all compounds. The weight for the jth compound is determined by the dot product of $\mathbf{s}_j$ and $\mathbf{a}$. As discussed above, if the variance of the concentration-spectral data of the jth compound is higher than those of the remaining compounds, $\mathbf{a}$ is composed mainly of spectrum $\mathbf{s}_j$, and the weight for $\mathbf{c}_j$, $(\mathbf{s}_j)^t\mathbf{a}$, is higher than the

other weights. As a result, the first principal component score vector, **u**, is composed mainly of the concentration profile of the *j*th compound, $\mathbf{c}_j$. The above investigation can be generalized to the case of more than one principal component score and loading vector, since the principal component score vectors are mutually orthogonal and so are the loading vectors.

Thus the result of performing PCA on a bilinear matrix of experimental data is to redistribute the concentration and the spectral information of all compounds into a series of principal component score and loading vectors, respectively. These principal component score vectors are sorted in descending order of eigenvalues, as are the loading vectors. The important (or principal) information tends to be largely dispersed into the first few principal component score and loading vectors, while the unimportant (or nonprincipal) information is largely retained in the later principal component score and loading vectors and could be lost when they are neglected as residual. In other words, whether or not certain information is principal in PCA is determined by the variance of corresponding concentration-spectral data, not necessarily by the magnitude of concentration or spectral value. By performing PCA on a bilinear matrix, we were able to explore the mechanism of information extraction through TFA.

Mechanism of Information Extraction through TFA from the Data Matrix. As introduced above, the first step of TFA is PCA. The result of PCA on the bilinear matrix **D** in eq 1 can be expressed as follows

$$\mathbf{D}_{m\times n} = \mathbf{U}_{m\times q}(\mathbf{V}_{n\times q})^{\mathrm{t}} + \mathbf{R}_{m\times n} = \mathbf{D}^{\#}_{m\times n} + \mathbf{R}_{m\times n} \qquad (13)$$

where q is the number of principal component score (or loading) vectors and equals p in eq 1 if Beer's Law is obeyed exactly. **U** and **V** are the principal component score and loading matrices that are correlated with the concentration and spectral information in matrices **C** and **S**, respectively. $\mathbf{D}^{\#}$ is referred to as the principal matrix that contains the principal information in raw data matrix **D**; **R** is the residual matrix and contains the nonprincipal information that is neglected.

Malinowski[1] showed that TFA is performed on the loading matrix **V** rather than the raw data matrix **D**. Thus for the target compound, only when the original spectral information measured in **D** is sufficiently retained in **V** after PCA can this information be extracted through TFA. As shown in eq 13, PCA redistributes the raw information in **D** into $\mathbf{D}^{\#}$ and **R**; the redistribution is based on the variance of the concentration-spectral data. For example, for the *j*th compound, the concentration-spectral data are represented by matrix $\mathbf{c}_j(\mathbf{s}_j)^{\mathrm{t}}$, where $\mathbf{c}_j$ is the concentration array and $\mathbf{s}_j$ is the spectrum of the compound of unit concentration. The variance of matrix $\mathbf{c}_j(\mathbf{s}_j)^{\mathrm{t}}$, V_j, is given by

$$V_j = \left((\mathbf{s}_j)^{\mathrm{t}}\mathbf{s}_j\right)v_j^{\mathrm{c}} \qquad (14)$$

where v_j^{c} is the variance of $\mathbf{c}_j$. In our previous report, the concentrations of the analyte, ammonia, in air was usually low during the monitoring period. However, it fluctuated significantly due to meteorological reasons, which resulted in a reasonably large concentration variance, i.e., v_j^{c} in eq 14. As a result, the spectral information was retained in the principal matrix and retrieved by TFA.

From the above discussion, we conclude that the capability of extracting analytical information through TFA is primarily determined by the variance; thus TFA can detect analytes at low concentration as long as the variance of the concentration-spectral data is large enough or, as we show later, can be increased artificially. Although detection by TFA is not directly related to the magnitude of the concentration of the target compound, in many practical cases the results of TFA can still be arbitrarily correlated with the magnitude of its concentration provided that its concentration varies.

Using TFA to Obtain the Upper Limit of Standard Deviation of the Concentrations. Suppose the measured spectral information of the target compound was redistributed by PCA into the first n principal component loading vectors in order to ensure an effective TFA result. However, when the number of loading vectors in TFA is intentionally decreased from n to m, the aforementioned spectral information might be excluded if the target compound is a trace one, and TFA fails to extract the target spectrum. In this case, the measured spectral information of the trace compound is in the residual matrix, **R** in eq 13, and discarded, which means that the variance of the concentration-spectral data of the trace compound is smaller than the variance of **R**. In other words, the variance of **R** in this case is the upper limit of the variance of the concentration-spectral data of the target compound, i.e., V_j in eq 14; then we can calculate the upper limit of the variance of the concentrations of the target compound, i.e., v_j^{c} in this equation and eventually obtain the upper limit of the standard deviation.

We have implemented this operation by gradually increasing the number of loading vectors in TFA and calculating the variance of the residual matrix at the same time. During this process, we inspect the extracted spectrum and use the weighted correlation coefficient[13] as an auxiliary criterion to determine when the main features of the target spectrum are about to appear in the extracted spectrum. The weighted correlation coefficient is similar to the conventional correlation coefficient except that larger weights are assigned to those wavelengths that have relatively high absorbance in the reference spectrum. It should be noted that if the target compound is such a major component that its spectral information already appears in the first loading vector, the upper limit of standard deviation through TFA is no longer available.

EXPERIMENTAL SECTION

A total of 92 OP/FT-IR spectra were measured at a resolution of 1 cm^{-1} in pristine air where the only infrared-active compounds present above the detection limit were H_2O, CO_2, CH_4, and N_2O. Of these molecules, only water vapor has significant absorption in the spectral region used in this study (1250−880 cm^{-1}.) The experimental conditions are the same as those in our previous paper.[12] Background spectra were acquired over a path-length of 372 m at intervals of approximately 1 min. The spectral and concentration data of the target compound, which is known not to be present in the background spectra, are added to the raw measurements in the following way

$$\mathbf{D} = \mathbf{D}^{*} + \mathbf{cs}^{\mathrm{t}} \qquad (15)$$

(13) Griffiths, P. R.; Shao, L. *Appl. Spectrosc.* **2009**, *63*, 916–919.

where **D** is the composite matrix; **D*** is the raw data matrix; **c** is the concentration vector, and **s** is the spectrum of unit concentration. By using composite data matrices, we were able to control the magnitude and the profile of the concentrations of the target compound.

The target compounds we selected were diethyl ether and ammonia. At 1 cm^{-1} spectral resolution, diethyl ether shows a single relatively broad absorption band in the region from 1250 to 880 cm^{-1}, while ammonia shows a number of sharp, well-resolved lines in its vibration−rotation spectrum. The two spectra have such different features that we could objectively investigate the information extraction through TFA over a wide scope. We recognize that by adding scaled reference spectra to measured background spectra, we are ignoring the fact that vapor-phase spectra vary slightly with temperature and pressure. However, these effects are very small for molecules for which the rotational fine structure is not resolvable, i.e., the spacing between the lines in the rotation−vibration spectrum is less than the width of these lines. For condensed-phase samples for which the effect of intermolecular interactions on the spectrum may be relatively large, TFA becomes difficult, although not insurmountably so because the effect of intermolecular interactions can often be taken care of by adding more eigenvectors.

The concentration profiles were chosen to be Gaussian curves with various peak heights, peak positions, and full widths at half height (fwhh). Since the fwhh is equal to $2(2 \ln 2)^{1/2}\sigma$, where σ is the standard deviation of the Gaussian function, widths are given subsequently in terms of σ rather than fwhh. Other profiles, such as rectangular and triangular, were also investigated and led to similar results. All data matrices in this investigation were mean centered as a preprocessing procedure.

RESULTS AND DISCUSSION

Validating the Theory. Initially, we performed TFA on matrix **D*** (see eq 15); the results confirmed that no spectral information due to diethyl ether or ammonia was present in the raw data. Therefore, the results of TFA on the composite matrix **D** are exclusively related to the concentration-spectral data that we added. Several composite data matrices were constructed to validate the theory that the detection of TFA is primarily related to variance in the concentration of the target compound. Since the spectrum in eq 15 is of unit concentration, the peak height of the profile is the maximum concentration of diethyl ether or ammonia in the corresponding data set.

The first concentration profile has a peak height of 5 ppm-m and $\sigma = 5$ (fwhh = 11.8), see Figure 1a. The maximum absorbance of the strongest ether band, corresponding to a maximum path-integrated concentration of 5 ppm-m, was 0.007 au. The noise level of the measured spectra is about 0.0008 au estimated by the standard deviation of absorbance within 1008 and 988 cm^{-1}.[14] The spectrum that was extracted by TFA is shown in Figure 1b. As mentioned previously, the weighted correlation coefficient, wcc, was used for the unequivocal identification of target compound. When the value of the wcc exceeds 0.90, we showed that there is a high probability that the target compound is present. In the case of the data shown in Figure 1, the weighted correlation coefficient between the extracted spectrum and the reference spectrum is 0.9988. Thus both visual inspection and the high value of wcc indicate that diethyl ether was identified unequivocally. The variance of the added concentration-spectral data of diethyl ether was calculated to be 3.54×10^{-4}. When the peak position of the concentration profile was shifted without changing the peak height or fwhh, the values of the wcc and variance of the concentration-spectral data were essentially unchanged.

When the width was increased so that $\sigma = 150$, as shown in Figure 2a, all the concentrations were close to the maximum of 5 ppm-m, only varying slightly during the measurement period. Figure 2b shows the spectrum extracted by TFA. The wcc between the extracted spectrum and the reference spectrum is 0.7410; thus TFA did not detect diethyl ether for this profile. The variance of the concentration-spectral data is only 4.87×10^{-6}, i.e., about 70 times less than the case for the profile shown in Figure 1. Shifting the peak position of the concentration profile made a negligible change to the values of wcc and the variance. It is clear, therefore, that the information extracted through TFA is not related to the magnitude of concentration of the target compound but rather to the variance of the concentration.

For the second concentration profile, the peak height was 1 ppm-m and $\sigma = 0.375$, i.e., the analyte effectively only appeared at one point, see Figure 3a. Therefore, the peak height and width of this profile were reduced significantly from the cases shown in Figures 1 and 2. The extracted spectrum is shown in Figure

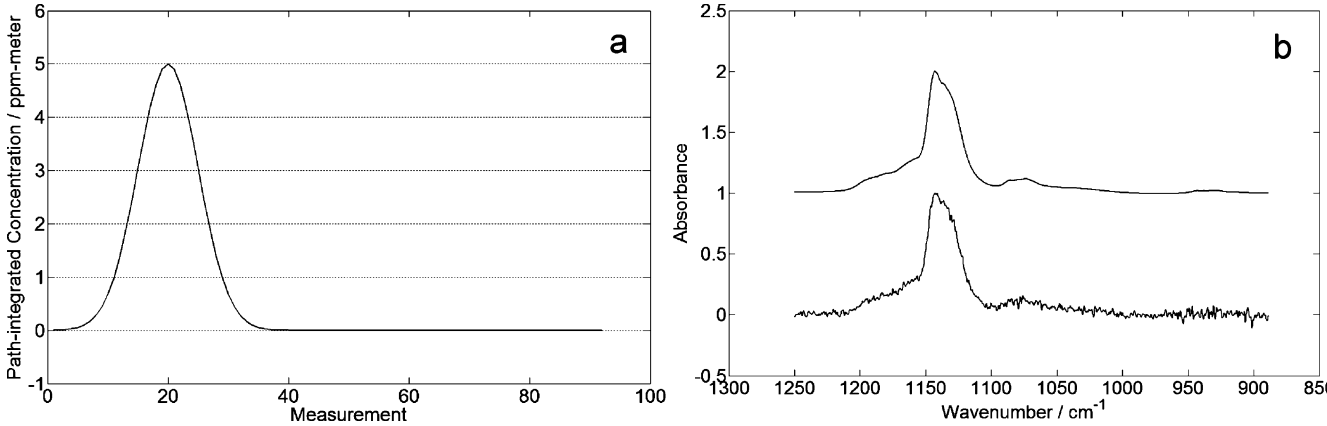

Figure 1. (a) The Gaussian concentration profile and (b) the reference spectrum of diethyl ether (above) and the spectrum extracted by TFA (below).

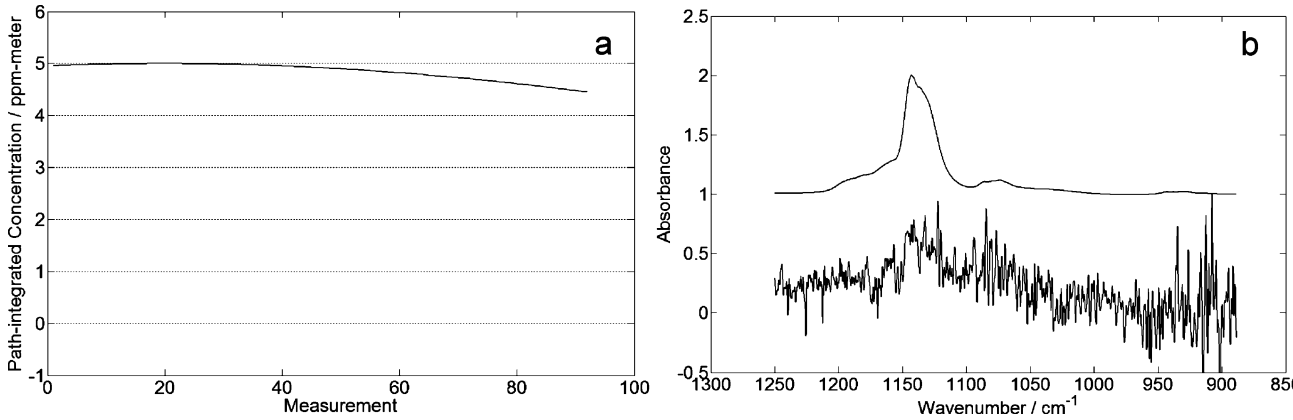

Figure 2. (a) The Gaussian concentration profile after increasing the fwhh and (b) the reference spectrum of diethyl ether (above) and the spectrum extracted by TFA (below).

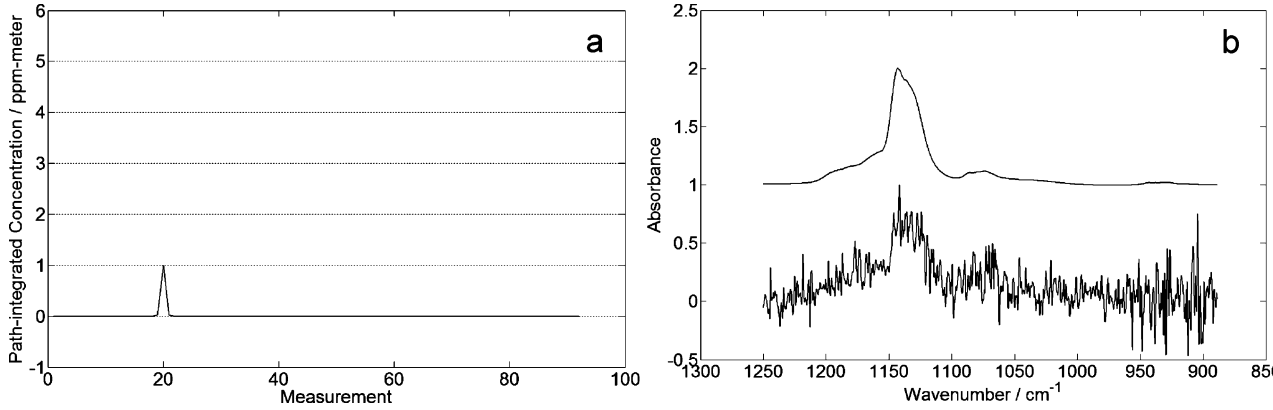

Figure 3. (a) The Gaussian concentration profile and (b) the reference spectrum of diethyl ether (above) and the spectrum extracted by TFA (below).

3b. For this example, the wcc between the extracted spectrum and the reference spectrum was 0.8728, which is still not high enough for the presence of diethyl ether in the infrared beam to be confirmed. The variance of the concentration-spectral data was 1.96×10^{-6} (i.e., almost 200 times less than for the data shown in Figure 1). When the peak position of the concentration profile was changed and TFA was performed on the corresponding composite data matrix, all the extracted spectra were quite similar and the wcc values were always around 0.87. Thus for this data set, the ineffective detection was caused by low concentration leading to a low variance.

To investigate the effect of variance on TFA further, we constructed a more complex concentration profile by replicating the small Gaussian peak shown in Figure 3a 18 times at different positions; the resulting profile is shown in Figure 4a. In this case, the maximum concentration is still the same as the second profile for which TFA is ineffective, but the variance is increased from 1.96×10^{-6} to 2.79×10^{-5}. The extracted spectrum is shown in Figure 4b; since the wcc is 0.9926, the presence of diethyl ether in the beam is confirmed.

The results obtained using ammonia as the target were very similar, despite the difference between the width of the spectral features. The results obtained with 1 and 18 narrow peaks in the concentration profile (analogous to the results shown in Figures 3 and 4) are shown in Figure 5. We also studied the effect of rectangular and triangular concentration profiles and found that

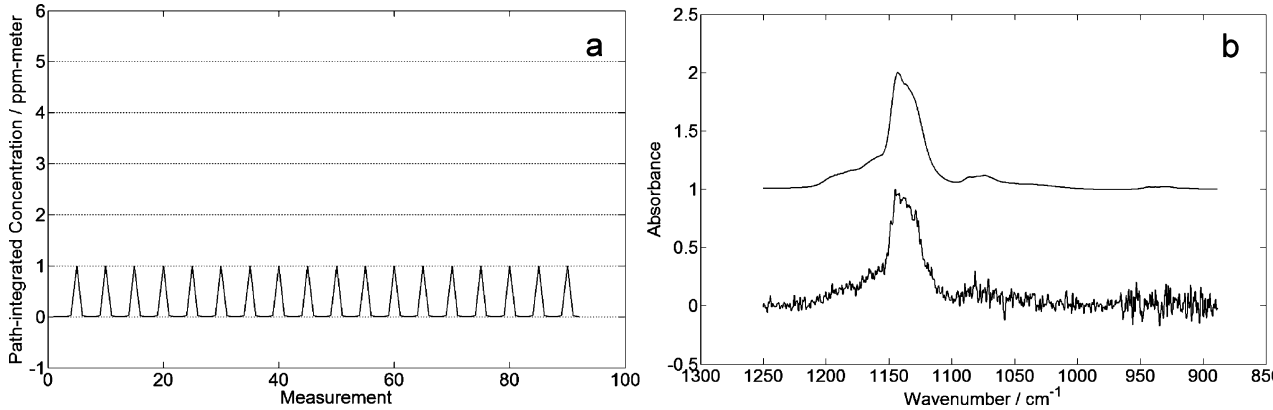

Figure 4. (a) The concentration profile of 18 equally spaced replicates of the Gaussian peak in Figure 3a and (b) the reference spectrum of diethyl ether (above) and the spectrum extracted by TFA (below).

detection by TFA was hardly affected by the distribution of concentration values. Thus the theory that information extraction through TFA is primarily related to the variance of the concentration-spectral data has been validated computationally.

Increasing the Variance of a Data Array by Adding Zero Elements. In an attempt to increase the effectiveness of TFA when the analyte is present at high concentration but with a low variance, we investigated the feasibility of increasing the variance by adding spectra in which the analyte was known to be absent to the data matrix, which is equivalent to adding zero values to the concentration array. Let **C** denote the original concentration array with n elements, i.e., $\mathbf{C} = [c_1, c_2, \ldots, c_n]$. For sake of convenience, we use S and SS to denote the sum and the sum of the squares of the elements in **C**, respectively.

$$S = \sum_{i=1}^{n} c_i \tag{16}$$

$$\mathrm{SS} = \sum_{i=1}^{n} c_i^2 \tag{17}$$

The variance of **C**, V, can be calculated in the following form (here we used population variance)

$$V = \frac{n \cdot \mathrm{SS} - S^2}{n^2} \tag{18}$$

Now we construct another data array, $\tilde{\mathbf{C}}$, by padding the original array **C** with m zero elements, so the new array is

$$\tilde{\mathbf{C}} = [\underbrace{c_1, c_2, \ldots, c_n}_{n}, \underbrace{0, 0, \ldots, 0}_{m}]$$

Let $\tilde{V}$ denote the variance of $\tilde{\mathbf{C}}$; thus, we can calculate $\tilde{V}$ by the definition of population variance,

$$\tilde{V} = \frac{\sum_{i=1}^{n+m} (\tilde{c}_i - \bar{\tilde{C}})^2}{n + m} \tag{19}$$

where $\bar{\tilde{C}}$ is the average of $\tilde{\mathbf{C}}$ and can be obtained by

$$\bar{\tilde{C}} = \frac{\sum_{i=1}^{n+m} \tilde{c}_i}{n + m} = \frac{\sum_{i=1}^{n} c_i}{n + m} \tag{20}$$

With eqs 16, 17, and 20, eq 19 can be simplified into

$$\tilde{V} = \frac{(n + m) \cdot \mathrm{SS} - S^2}{(n + m)^2} \tag{21}$$

In order to understand how m (the number of zero elements added) affects the variance, we calculate the first derivative of $\tilde{V}$ with respect to m.

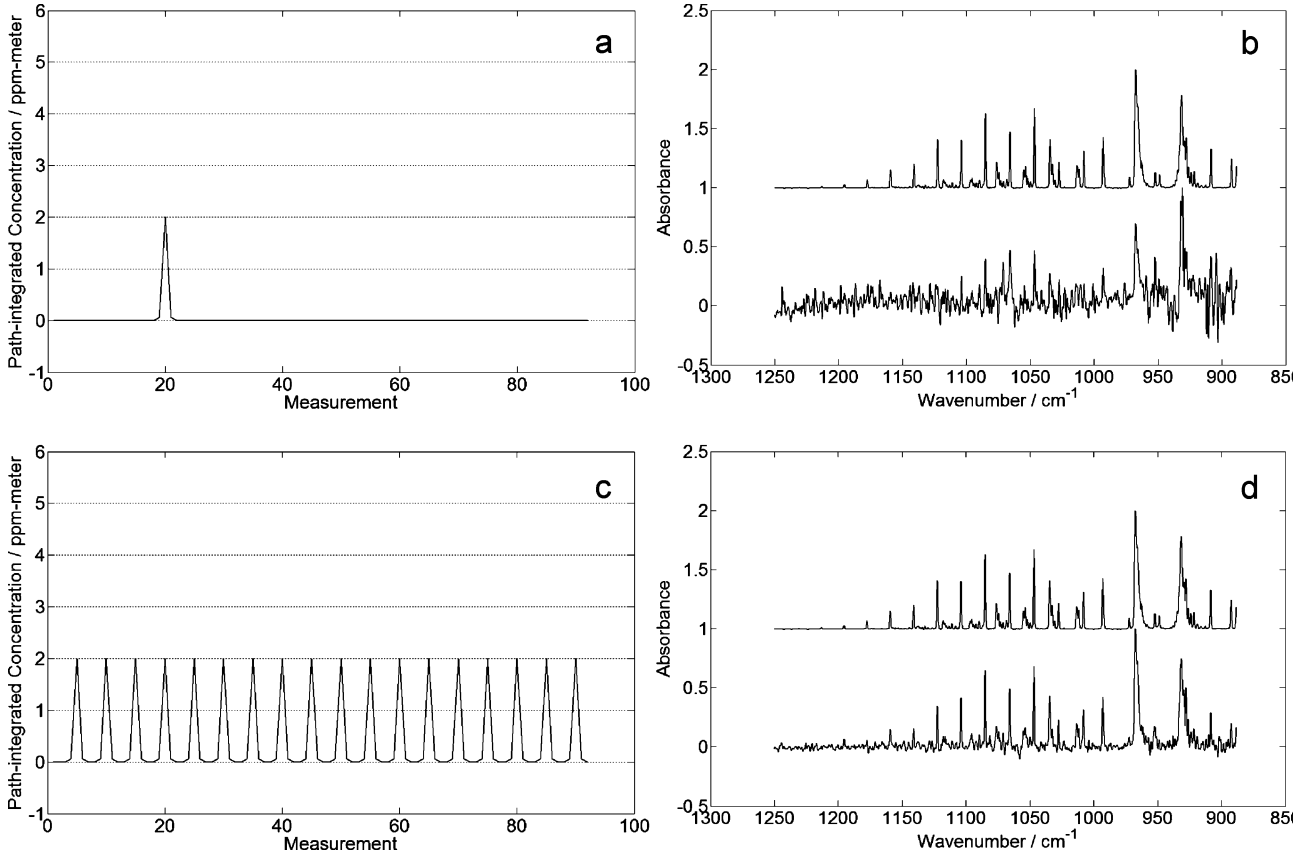

Figure 5. (a) The Gaussian concentration profile and (b) the spectrum extracted by TFA (below). (c) The concentration profile of 18 equally spaced replicates of the Gaussian peak in Figure 5a and (d) the spectrum extracted by TFA (below). The reference spectrum of ammonia is shown as the upper trace in parts b and d, respectively.

$$\frac{d\tilde{V}}{dm} = \frac{2S^2 - (n+m)\cdot SS}{(n+m)^3} \quad (22)$$

If the variance could be increased by adding the m zero elements to the data array, the first derivative is positive. Therefore,

$$2S^2 > (n+m)\cdot SS \quad (23)$$

The inequality can be rewritten as

$$m < \frac{2S^2}{SS} - n \quad (24)$$

In order to get a valid value of m as an integer, it requires

$$\frac{2S^2}{SS} - n > 1 \quad (25)$$

By inserting eq 18 into the above equation, we obtain the following result,

$$V < \frac{n-1}{n+1}\left(\frac{S}{n}\right)^2 = \frac{n-1}{n+1}\bar{C}^2 \quad (26)$$

where $\bar{C}$ is the average of data array **C**.

Inequality 26 is the condition that the variance of a data array can be increased by adding some zero values. If the condition is not satisfied, adding zero values will decrease the variance. For large n, $(n-1) \approx (n+1)$, so the inequality can be simplified as

$$V < \bar{C}^2 \quad (27)$$

Therefore, adding spectra without the analyte, so that the corresponding concentrations are zero, is an effective way to increase the variance in the case that elements of the data array are all high; i.e., $\bar{C}$ is large but only varies slightly, so that V is small.

The inclusion of too many spectra with zero analyte concentration will eventually decrease the variance because, as shown in eq 22, a large value of m will result in a negative first derivative. When applying TFA to identify certain compounds from experimental data, therefore, it is useful to include some "blank" measurements into the data matrix to increase the concentration variance. However, the addition of too many "blank" measurements will ultimately decrease this variance. The addition of too many spectra to the data matrix also means that increased computation resources will be required.

Thus TFA should work more effectively in a "moving fixed window" mode, so that a fixed number of measurements are arranged into the matrix for TFA. In this way, the number of "blank" measurements, i.e., m in eq 22 is restricted. The "moving window" mode not only reduces the computation time for TFA but increases the time resolution of the result, compared to the analysis of the entire data matrix. This is probably one reason for the so-called "window target-testing factor analysis" that has been found to be effective in qualitative analysis of complex mixtures by other researchers.[15]

To test this theory, i.e., to increase the variance in the concentration profile, we added together in a row-wise manner the raw data matrix (i.e., where no diethyl ether is present in the beam) and the composite used for Figure 2, where the spectrum of ether has been added with a high path-integrated concentration; thus, the large matrix is

$$\begin{bmatrix} \mathbf{D}^* \\ \mathbf{D} \end{bmatrix}$$

Since the raw data matrix does not contain spectral information due to diethyl ether, the concentration profile of diethyl ether for the large matrix is the combination of the same number zero concentration values as the profile of Figure 2a and the profile itself, as shown in Figure 6a. The variance of the concentration-spectral data is thus greatly increased, from 4.87×10^{-6} to 1.07×10^{-3}. We performed TFA on the large matrix and found a perfect match between the extracted spectrum and the reference, as shown in Figure 6b. Clearly the effectiveness of TFA this time is due to the increase in variance that was achieved by adding spectra where the analyte concentration was zero. Similar results were found for ammonia, as shown in Figure 7. From the concentration profile given in Figure 6a, we can see that the concentration difference from zero is roughly 10 times the variation in concentrations of the analyte. According to analysis of variance considerations, the difference from zero contributes significantly more to the variance than the variance of concentrations does, so this result should not be surprising.

Obtaining the Limit of Standard Deviation of Concentrations. In this section, we investigate the limit of the standard deviation for the concentrations of the target compound. With the

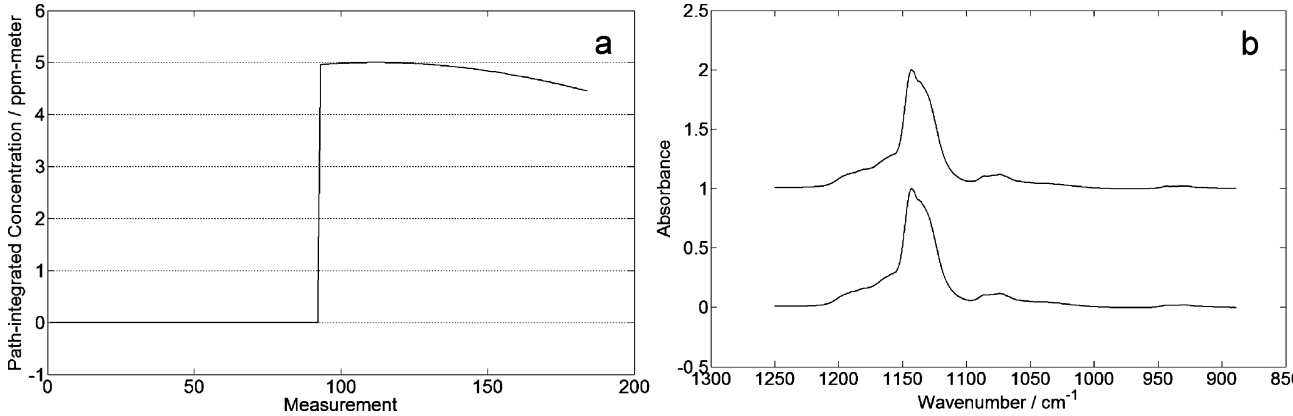

Figure 6. (a) The concentration profile obtained by combining the raw data with no ether present with the data set used for Figure 2. (b) The reference spectrum of diethyl ether (above) and the spectrum extracted from this data set by TFA (below).

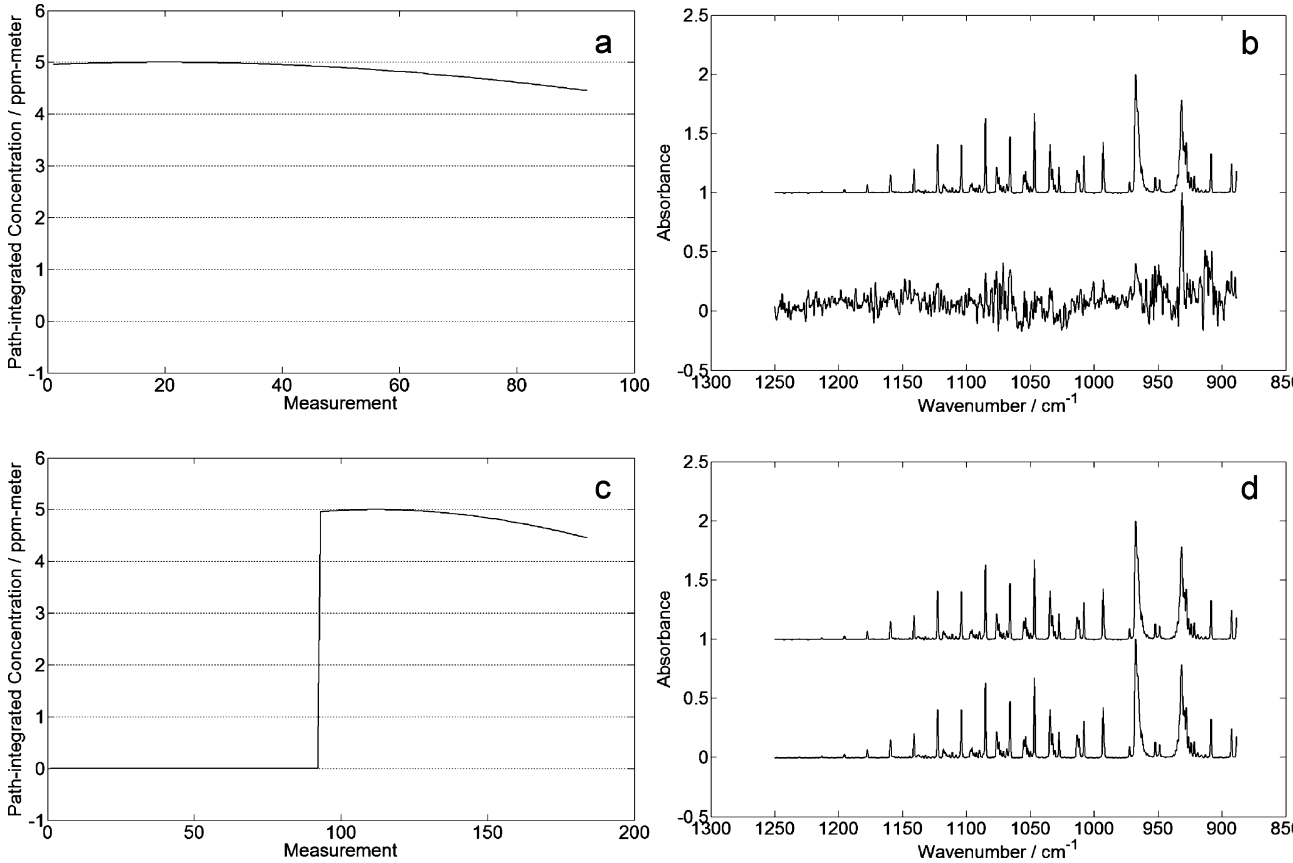

Figure 7. (a) The Gaussian concentration profile for ammonia with $\sigma = 150$; (b) the reference spectrum of ammonia (above) and the spectrum extracted by TFA (below); (c) the concentration profile obtained by combining the raw data with no ammonia present with a data set used for part a. (d) The reference spectrum of ammonia (above) and the spectrum extracted by TFA from the data set shown in part c (below).

Table 1. Limits of Standard Deviation Obtained through TFA and the Actual Standard Deviations of the Concentration Distributions, and the Weighted Correlation Coefficients Obtained for $n = (n_{crit} - 1)$ and n_{crit}[a]

C_{max}	$n_{crit} - 1$	LOSD	SD	wcc for $n_{crit} - 1$	wcc for n_{crit}
5	1	1.43	1.40	−0.83	1.00
4	1	1.30	1.12	−0.77	0.99
3	1	1.19	0.84	−0.67	0.98
2	2	0.94	0.56	−0.63	0.94
1	3	0.80	0.28	0.42	0.91
0.8	4	0.76	0.22	0.79	0.94
0.7	5	0.75	0.20	0.88	0.92
0.6	5	0.74	0.17	0.62	0.89

[a] Table headings are, from left to right, maximum concentration in units of ppm m; maximum number of loading vectors used in the TFA that gives no spectral features of the target compound; limit of standard deviation obtained from TFA; standard deviation calculated from the concentration profile; weighted correlation coefficient, wcc, between the extracted and the reference spectrum of diethyl ether calculated with the number of loading vectors shown in column 2; wcc calculated with n_{crit} loading vectors.

use of diethyl ether as the target, a total of eight different Gaussian concentration profiles were generated with different peak heights, C_{max}, keeping the width ($\sigma = 5$ min) and peak position (20) constant. The values of C_{max} are shown in the first column of Table 1; because the spectrum is of unit concentration (i.e., the path integrated concentration was 1 ppm-m), these numbers are equal to the maximum concentrations of diethyl ether in corresponding composite data matrices. The number of loading vectors used in the TFA, n, was gradually increased to a maximum value of 20 (It should be noted that using an excessive number of loading vectors could lead to a false result). Each time that n was increased, we tried to determine a certain value of n, n_{crit}, when the main spectral features of diethyl ether just appeared in the extracted spectrum, and corresponding wcc exceeds 0.90; we then calculated the limit of standard deviation according to our theoretical analysis. The results are summarized in Table 1. As can be seen in this table, all standard deviations are smaller than corresponding limits obtained through TFA, which is consistent with our theoretical analysis. When we decreased the maximum of the concentration profile to 0.5 ppm m, TFA could not extract the spectrum of diethyl ether even with 20 loading vectors and the wcc value was below 0.90.

An example of how n_{crit} is determined is shown in Figure 8, for which $C_{max} = 0.8$ ppm-m. Traces a and b are the spectra extracted by TFA using 5 and 4 loading vectors; the reference spectrum of diethyl ether is c. The weighted correlation coefficients between a and c was 0.94, which is greater than 0.90 and is therefore deemed to be a good match. The value of wcc for spectrum b was 0.79, which is too low for a good match. Therefore the value of n_{crit} is 5. Both by visual inspection of the spectra and from the values of wcc, we concluded that there is insufficient spectral evidence of diethyl ether in extracted spectrum b, while there is adequate evidence in a. When only four loading vectors were used in TFA, most of the spectral information from diethyl ether is contained in the residual matrix. When the variance of the residual was

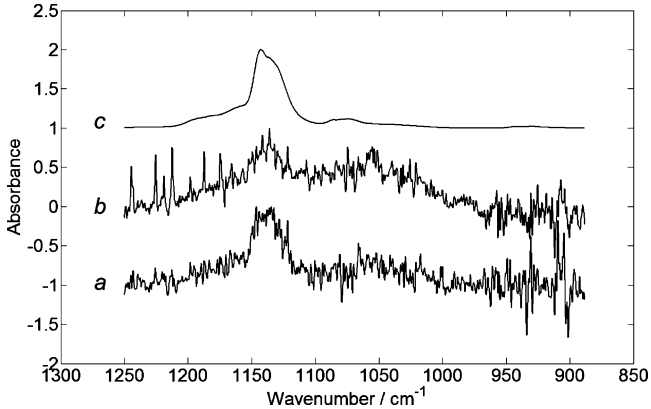

Figure 8. The spectra extracted through TFA with (a) 5 and (b) 4 loading vectors and (c) the reference spectrum of diethyl ether.

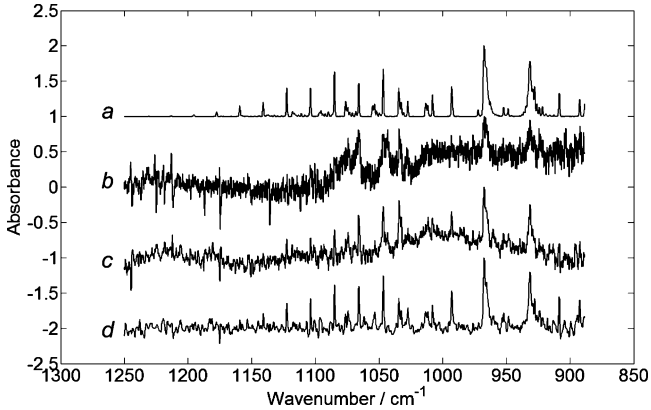

Figure 9. (a) Reference spectrum of ammonia and (below) the spectra extracted by TFA using (b) 8, (c) 9, and (d) 10 loading vectors.

calculated, the limit of standard deviation was shown to be 0.76. Thus, according to the theory given above, the standard deviation of the concentrations of diethyl ether in this data matrix should be smaller than 0.76. This conclusion is correct, as shown by the data in Table 1.

To further investigate the reliability of the limit of standard deviation obtained from TFA, we processed some real OP/FT-IR data.[12,16] OP/FT-IR spectra were acquired near a dairy pen in southern Idaho. The instrument was mounted on an open field about 230 m to the west side of the dairy pen. Although the air over this field should be pristine, wind could carry ammonia from the dairy pen into the IR beam. We performed TFA on the data matrix with ammonia as the target, and the results are shown in Figure 9. In Figure 9, spectra b, c, and d were extracted using 8, 9, and 10 loading vectors. The values of wcc between the extracted spectra and the reference were 0.673, 0.849, and 0.971. Apparently, no spectral features of ammonia can be found with high confidence in spectrum b, which means most information of ammonia is retained in the residual matrix. Therefore, we calculated the limit of standard deviation according to the residual matrix when 8 loading vectors were used, and the value was 0.0035 ppm. To validate this upper

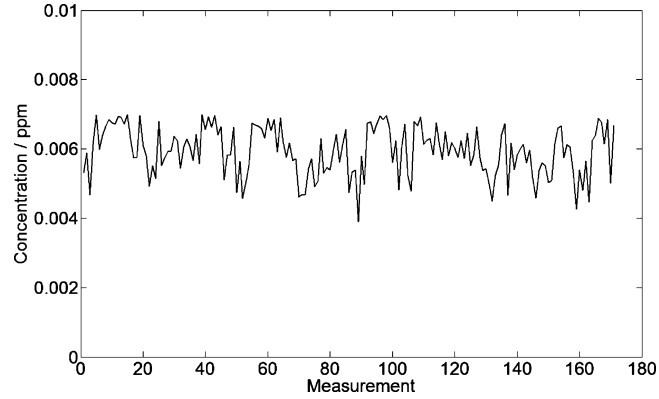

Figure 10. Variation of the concentration of ammonia obtained by PLS regression on the measured OP/FT-IR spectra.

limit, the concentrations of ammonia in all measurements were calculated by partial least-squares (PLS) regression[16] and the result is shown in Figure 10. For all spectra, the maximum concentration was calculated to be ~0.0070 ppm and the average was 0.0059 ppm; the standard deviation is 0.0007 ppm. Thus we could find the actual standard deviation is below the limit we have obtained from TFA, which again confirms our theoretical analysis. Additionally, spectrum d in Figure 9 clearly demonstrates the presence of ammonia. This evidence for the presence of ammonia is much more convincing than any of the values of the concentration of ammonia given by PLS regression because the strongest absorption line of ammonia in any of the of the measured OP/FT-IR spectra is of the same order as the noise level.

CONCLUSIONS

The results shown in this article have demonstrated that the successful extraction of spectra from an experimental data matrix by TFA is primarily determined by the variance of the concentration-spectral data of the target compound and not necessarily by the magnitude of the concentrations; it is only slightly affected by how the concentrations are distributed. This conclusion holds for other PCA-related techniques. TFA can detect the presence of trace compounds as long as the variance of their concentrations is sufficiently high. TFA fails when the analyte is present at high concentration with a small variance. In practice, however, it is rare that the concentration of a target compound remains at a high level and remains constant. If this is the case, the concentration variance may be increased by including some blank measurements. Once a target compound has been identified, TFA can possibly provide the limit of standard deviation of the concentrations of the target compound. In summary, TFA is a powerful, yet reliable, method to identify trace components in a complex multicomponent system. A very low rate of false detections is attainable, especially when visual inspections are also carried out.

ACKNOWLEDGMENT

This work was funded by Contract W91ZLK08P0739 from the Edgewood Chemical Biological Center (ECBC), Edgewood Arsenal, U.S. Army, and by the National Natural Science Foundation in China (Grant No. 20705032).

Received for review June 8, 2009. Accepted August 21, 2009.

AC901246X

(14) Russwurm G. M. Childers, J. W. *FTIR Open-Path Monitoring Guidance Document*, Document TR-4423-99-03, 3rd ed.; ManTech Environmental Technology: Research Triangle Park, NC, 1999,

(15) Lohnes, M. T.; Guy, R. D.; Wentzell, P. D. *Anal. Chim. Acta* **1999**, *389*, 95–113.

(16) Griffiths, P. R.; Shao, L; Leytem, A. B. *Anal. Bioanal. Chem.* **2009**, *393*, 45–50.