Self-Weighted Correlation Coefficients and Their Application to Measure Spectral Similarity

PETER R. GRIFFITHS* and LIMIN SHAO

Department of Chemistry, University of Idaho, Moscow, Idaho 83844-2343 (P.R.G.); and Department of Chemistry, University of Science and Technology of China, Hefei, Anhui, 230026, P.R. China (L.S.)

A technique for spectral searching with noisy data is described that improves the performance over contemporary approaches. Instead of simply calculating the correlation coefficient between the spectrum of an unknown and a series of reference spectra, greater weight is given to the more intense features in the reference spectra. The weight array, w, is given by $|\mathbf{r}|/\{1 + \mathbf{d}\}$, where the vector r represents the reference spectrum and the difference vector, d, contains the difference between the sample and reference data points, equal to $|\mathbf{s} - k\mathbf{r}|$, where k is a scaling factor that eliminates the effect of signal strength. By this approach, a large weight is only given to those points that have relatively high absorbance and are close to their counterparts in the reference spectrum. This technique was shown to give significantly improved performance when applied to noisy spectra of trace atmospheric components obtained by target factor analysis.

Index Headings: **Weighted correlation coefficient; Spectral searching; Target factor analysis; TFA.**

INTRODUCTION

Comparing the spectrum of an unknown sample to a large database of reference spectra is a common task for the identification of chemicals and materials. While visual examination by a trained spectroscopist often yields reliable results, time constraints and the lack of expertise of many chemists in the interpretation of spectra may well mean that this approach is quite limited and reliable automated methods are called for. Several similarity metrics have been used to measure the similarity of infrared spectra;[1] these include such approaches as the calculation of Euclidean distance, the sum of the absolute differences, and the correlation coefficient.

The correlation coefficient has been applied to quantify spectral similarity for many years[2] and good performance was often found.[1,3,4] It has been shown that the correlation coefficient is closely related to the dot product and the cosine of the angle between vectors; the latter two are other common measures of spectral similarity.[5,6] Even though the correlation coefficient is one of the more popular metrics for spectral similarity, in our previous investigation that involved the identification of unknowns during target factor analysis (TFA),[7] we found that a high value of the correlation coefficient did not always indicate the correct match. This deficiency has also been noticed by other researchers.[8,9]

There are several reasons for this problem, including the presence of interferences in the sample spectrum. These interferences are either neglected when spectra are inspected manually or taken care of by using reverse searching approaches in automated searching approaches. However, the presence of spectral interferences always decreases the value of the correlation coefficient and results in inconsistent behavior in practice.

In the case of visual inspection of the similarity, most spectroscopists initially focus on the more intense bands in the spectrum, trying to confirm whether some key group frequencies are present or absent.[10] Thus, they obtain a general idea of the chemical class of the compound before delving deeper and studying weaker spectral features. Indeed they often neglect other parts that are less likely to contain useful spectral features until they obtain a good idea of the general chemical class of the unknown. Only then do they investigate the finer details of the spectrum.

Obviously the approach to visual comparison by observing the stronger bands in the spectrum is a *local* means of classification, whereas calculation of the correlation coefficient is a global one that takes into consideration all points, whether rich in spectral features or not. Incorporating the localization approach to visual comparison into the correlation coefficient provides a possible way to improve its performance in measuring similarity of spectra. The way that we propose to accomplish this end is to weight the correlation coefficient based on the intensity of the reference spectral data. With this approach, the weighted correlation coefficient should represent the visual inspection more effectively than the conventional unweighted correlation coefficient.

The concept of weighting the correlation coefficient was introduced by Bland and Altman in processing clinical data[11] and was found to be better than the conventional approach. The problem with respect to spectroscopic interpretation is how to design a functional, yet less arbitrary, weighting scheme. Several weighting schemes have been proposed, each of which is effective in certain cases.[9,11–13] In this paper, we use a weighting scheme that simulates the first step in visual inspection, and we apply it to measure the similarity between experimental infrared spectra and reference spectra in a database. The results show the effectiveness of this weighted correlation coefficient when interferences, such as noise and spectral information from other species, are present.

THEORY

Throughout this paper, a spectrum is represented with a boldface lower-case letter; the same letter without boldface and with subscript i, denotes the ith data point of the spectrum, i.e., the ith element of the vector. All vectors are column vectors, the transpose of which are row vectors, indicated with superscript t.

Let three vectors, $\mathbf{s} = [s_i]$, $\mathbf{r} = [r_i]$ and $\mathbf{w} = [w_i]$ ($i \in [1, n]$), be the sample spectrum, the reference spectrum, and the weight array with n data points. According to Bland and Altman,[11] the weighted correlation coefficient, *wcc* (based on Pearson's correlation coefficient), between **s** and **r** is

Received 2 March 2009; accepted 8 May 2009.
* Author to whom correspondence should be sent. E-mail: pgriff@uidaho.edu.

0003-7028/09/6308-0916$2.00/0

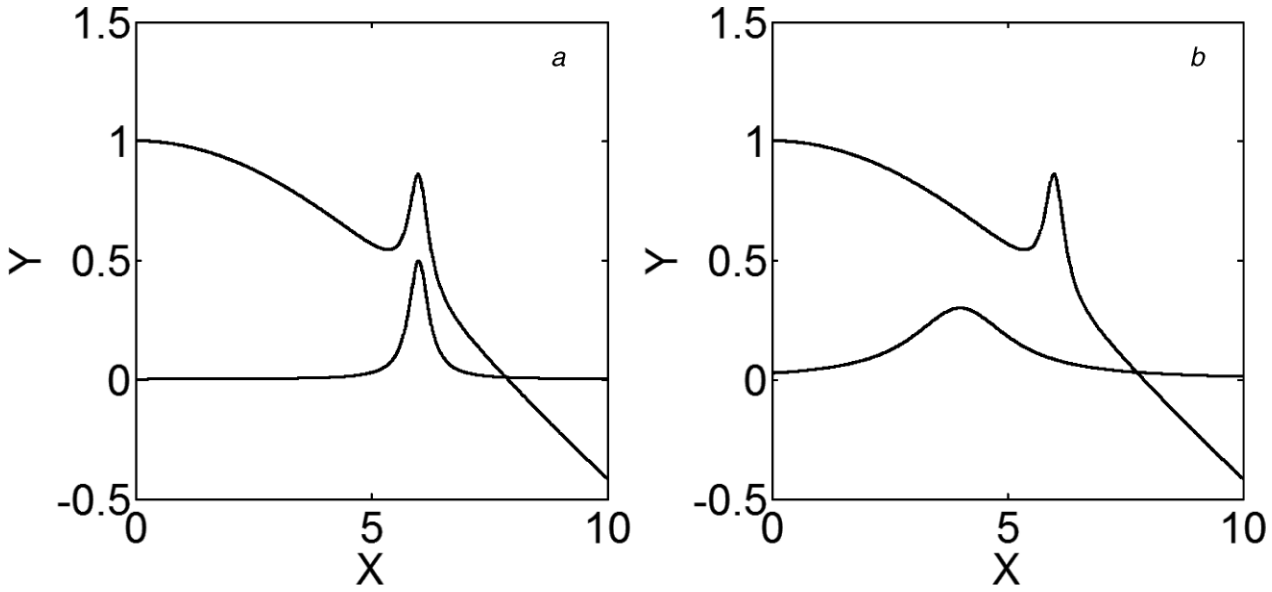

FIG. 1. Simulated sample spectrum with a nonzero baseline, shown as the upper trace in both (**a**) and (**b**). The two lower traces in (**a**) and (**b**) are the two simulated reference spectra.

$$wcc = \left(\sum w_i \cdot s_i \cdot r_i - \sum w_i \cdot s_i \sum w_i \cdot r_i \Big/ \sum w_i\right) \div \left\{\left[\sum w_i \cdot s_i^2 - \left(\sum w_i \cdot s_i\right)^2 \Big/ \sum w_i\right] \times \left[\sum w_i \cdot r_i^2 - \left(\sum w_i \cdot r_i\right)^2 \Big/ \sum w_i\right]\right\}^{1/2} \quad (1)$$

where all summations are from $i = 1$ to n. If we use Fisher's definition of the correlation coefficient, the corresponding *wcc* is as follows:

$$wcc = \frac{\sum w_i \cdot (s_i - \bar{s}) \cdot (r_i - \bar{r})}{\sqrt{\left[\sum w_i \cdot (s_i - \bar{s})^2\right]\left[\sum w_i \cdot (r_i - \bar{r})^2\right]}} \quad (2)$$

where $\bar{s} = (\Sigma\ w_i \cdot s_i)/\Sigma\ w_i$ and $\bar{r} = (\Sigma\ w_i \cdot r_i)/\Sigma\ w_i$ are weighted means of the sample spectrum and the reference. The relationship between Eqs. 1 and 2 is analogous to the relationship between Pearson's and Fisher's definition of the correlation coefficient.[14] When the weight array, **w**, has unit elements, Eqs. 1 and 2 change into the ordinary Pearson's and Fisher's correlation coefficient, respectively. The key to the use of the weighted correlation coefficient is obtaining an appropriate weight array, **w**; for this work, we attempted to construct the weight array by simulating visual inspection.

In general, visual inspection comprises two steps, first, finding some features from the reference spectrum that appear to be similar to the spectrum of the unknown, and then determining whether these features fit the sample spectrum at the same locations. In this discussion, the term *features* implies large values in the reference spectrum; if one is attempting to fit an eigenvector, both positive and negative values should be considered. The term *fit* means that the differences between the sample spectrum and the reference are sufficiently small. Incorporating this strategy into the weighting scheme design, we define the weight array, **w**, as

$$\mathbf{w} = \frac{|\mathbf{r}|}{\mathbf{1} + \mathbf{d}} \quad (3)$$

where the numerator is the absolute values of the reference spectrum and **d** contains the differences between the sample and reference spectral data points. The differences are calculated by

$$\mathbf{d} = |\mathbf{s} - k\mathbf{r}| \quad (4)$$

where k is a scaling factor that eliminates the effect of signal strength difference. The scaling factor k is obtained by using the following equation in a least-squares manner:

$$k = (\mathbf{r}^t\mathbf{s})/(\mathbf{r}^t\mathbf{r}) \quad (5)$$

It can be seen from Eq. 3 that large weights only occur for those points that have relatively high absolute values in the reference spectrum *and* are fairly close to the counterparts in the sample spectrum. This is consistent with the type of visual comparison used by many spectroscopists who are attempting to gauge whether the major spectral features of a particular reference spectrum fit the sample spectrum. In spectral regions where the values of the reference spectrum are low, or where there are large differences between the sample and the reference spectrum, the corresponding weights are small. This is also consistent with visual comparison, which simply means that less attention is paid to these non-feature points or the fit is poor.

RESULTS AND DISCUSSION

First we investigated the use of simulated spectra to evaluate the performance of the proposed weighted correlation coefficient. A narrow and a broad Lorentzian peak were employed as reference spectra 1 and 2, shown as the lower traces in Figs. 1a and 1b, respectively. The sample spectrum was prepared by adding a cosine signal as the nonzero baseline to reference spectrum 1, shown as the upper trace in Figs. 1a and 1b.

By visual inspection of Fig. 1, one would conclude that the

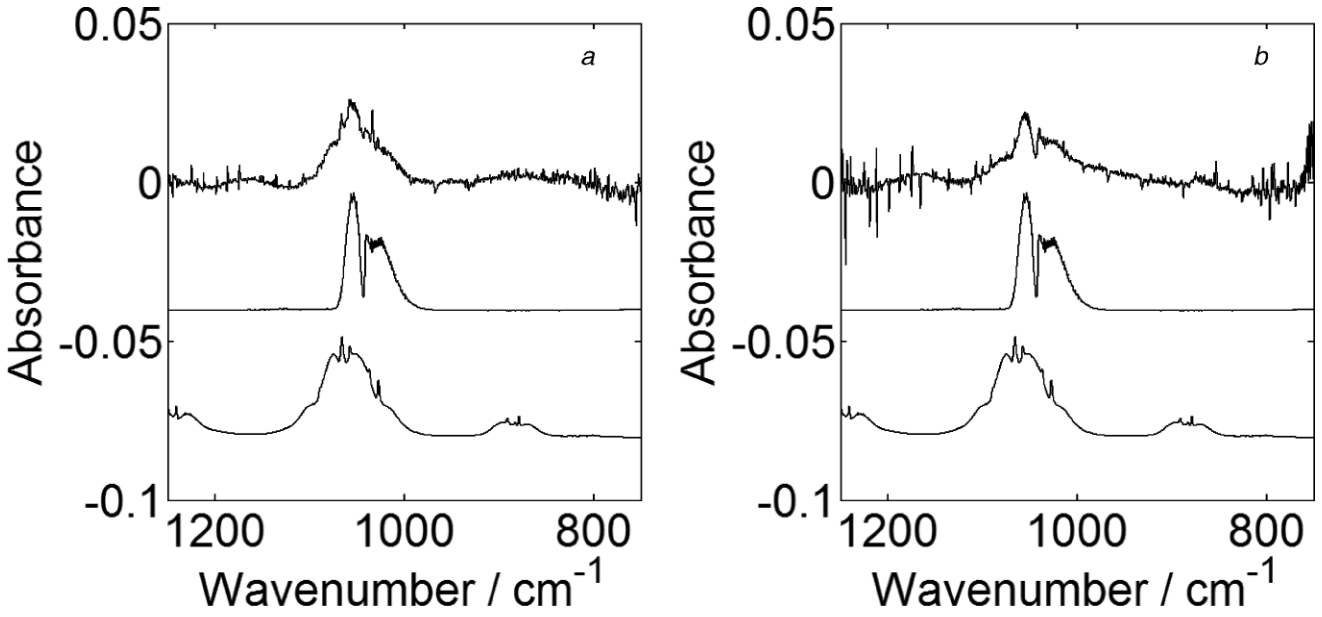

FIG. 2. Results of target factor analysis of data sets (**a**) 5 and (**b**) 7 in our previous investigation. The upper traces are the results of target factor analysis, and the middle and the lower traces are the reference FT-IR spectra of ozone and ethanol, respectively.

sample spectrum matched reference spectrum 1 better than reference spectrum 2, despite the nonzero baseline. However, the conventional correlation coefficients between the sample spectrum and references 1 and 2 were calculated to be 0.12 and 0.44, which indicates a poorer match between the sample and reference spectrum 1 than that between the sample and reference spectrum 2. The obvious reason for the apparently incorrect result found with the conventional correlation coefficient is the interference of the nonzero baseline. We calculated the weighted correlation coefficients of the two cases, and the values were 0.80 and 0.37, which is consistent with visual inspection.

We then studied the use of the weighted correlation coefficient on data obtained in our previous investigation into the application of target factor analysis for open-path Fourier transform infrared (FT-IR) spectrometry.[7] Several data sets were obtained in this project, of which we selected two. In the first, the presence of trace amounts of ethanol in the beam caused a weak band centered near 1050 cm^{-1}, while in the second a low concentration of ozone gave rise to a similar band in the same region but with a different shape. As shown in Table 2 of that paper, visual inspection indicated the absence of ozone in data set 5, and its presence in data set 7 (see Fig. 2).

The conventional correlation coefficients when the spectrum of ozone was used as the reference for data sets 5 and 7 are 0.83 and 0.79, respectively, which not only did not imply a good fit in either case, but indicated that the spectrum shown in Fig. 2a was a slightly better match to the reference spectrum of ozone than the spectrum in Fig. 2b. Apparently, the strong interferences in the spectrum shown in the upper trace of Fig. 2b caused the conventional correlation coefficient to yield a low value for the correlation coefficient. When the weighted correlation coefficient was applied to the two cases, the values of *wcc* were calculated to be 0.87 and 0.94, respectively, correctly indicating that the result of TFA in Fig. 2b was more likely to show the presence of ozone.

We calculated the weighted correlation coefficients for all data sets in our previous paper[7] and have listed them in Table I. One can see that all the values of weighted correlation coefficient conform to the results of visual inspections. In practice, we found that when $wcc > 0.90$, the analyte could be detected reliably.

CONCLUSION

A weighted correlation coefficient has been shown to be a more reliable measure of the similarity between experimental and reference infrared spectra than the standard unweighted correlation coefficient. The application of a weighted correlation coefficient simulated the process of visual inspection by assigning large weights to the data points with obvious spectral features. The effectiveness of this approach was demonstrated and shown to be useful, especially when the sample spectrum is degraded with some interferences. This weighted correlation coefficient can also be used for spectral library searching. The hit list thereby generated is more reliable and led to the correct conclusion. We will examine the results of this study in a subsequent paper. This weighted correlation coefficient should be applied to the case in which a benchmark, such as a reference spectrum, is available, so that the rational weighting array can be constructed.

TABLE I. Qualitative confirmation of the presence of ethanol and ozone using conventional and weighted (a/b) correlation coefficient and visual inspection (+/−).

Data set	Ethanol[a]	Ozone[a]
3	0.95/0.99 (+)	0.71/0.78 (−)
4	0.87/0.95 (+)	0.66/0.73 (−)
5	0.95/0.94 (+)	0.83/0.87 (−)
6	0.84/0.84 (−)	0.91/0.97 (+)
7	0.81/0.89 (−)	0.79/0.94 (+)

[a] A plus or minus symbol in parentheses indicates confirmation of the presence or absence, respectively, of detection by visual inspection of the appropriately rotated eigenvectors.

ACKNOWLEDGMENTS

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).

1. S. R. Lowry, "Automated Spectral Searching in Infrared, Raman and Near-infrared Spectroscopy", in *Handbook of Vibrational Spectroscopy*, J. M. Chalmers and P. R. Griffiths, Eds. (John Wiley and Sons, Chichester, UK, 2002), vol. 3, pp. 1948–1961.
2. K. Tanabe and S. Saeki, Anal. Chem. **47,** 118 (1975).
3. K. Baumann and J. T. Clerc, Anal. Chim. Acta **348,** 327 (1997).
4. K. Varmuza, M. Karlovits, and W. Demuth, Anal. Chim. Acta **490,** 313 (2003).
5. S. E. Lappi and S. Franzen, Spectrochim. Acta, Part A **60,** 357 (2004).
6. J.-J. P. Sievert and A. C. J. H. Drouen, "Spectral matching and peak purity", in *Diode Array Detection in HPLC,* L. Huber and S. A. George, Eds. (Marcel Dekker, New York, 1993), pp. 51–126.
7. L. Shao and P. R. Griffiths, Anal. Chem. **79,** 2118 (2007).
8. L. Teng and L. W. Chan, J. VLSI Sign. Process. Syst. **50(3),** 267 (2008).
9. Y. S. Liu, Q. H. Meng, R. Chen, J. S. Wang, S. M. Jiang, and Y. Z. Hu, J. Chromatogr. Sci. **42,** 545 (2004).
10. F. A. Miller, in *Course Notes on the Interpretation of Infrared and Raman Spectra*, D. W. Mayo, F. A. Miller, and R. W. Hannah, Eds. (Wiley-Interscience, Hoboken, NJ, 2004), Chap. 1.
11. J. M. Bland and D. G. Altman, BMJ **310,** 633 (1995).
12. H. R. Karfunkel, B. Rohde, F. J. J. Leusen, R. J. Gdanitz, and G. J. Rihs, Comput. Chem. **14,** 1125 (1993).
13. R. de Gelder, R. Wehrens, and J. A. Hageman, J. Comput. Chem. **22,** 273 (2001).
14. S. Plata, Appl. Math. Lett. **19,** 499 (2006).