Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 82 (2006) 50 – 58
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A comparative study of point-to-point algorithms for matching spectra

Jianfeng Li [a], D. Brynn Hibbert [a,*], Stephen Fuller [b], Gary Vaughn [b]

[a] School of Chemistry, The University of New South Wales, Sydney, Australia
[b] Environmental Forensic and Analytical Science, Department of Environment and Conservation (NSW), Lidcombe, Australia

Received 11 January 2005; received in revised form 19 April 2005; accepted 10 May 2005
Available online 19 October 2005

Abstract

Matching spectra is necessary for database searches, assessing the source of an unknown sample, structure elucidation, and classification of spectra. A direct method of matching is to compare, point by point, two digitized spectra, the outcome being a parameter that quantifies the degree of similarity or dissimilarity between the spectra. Examples studied here are correlation coefficient squared and Euclidean cosine squared, both applied to the raw spectra and first-difference values of absorbance. It is shown that spectra do not fulfill the requirements for a normal statistical interpretation of the correlation coefficient; in particular, they are not normally distributed variables. It is therefore not correct to use a Student's *t*-test to calculate the probability of the null hypothesis that two spectra are not correlated on the basis of a correlation coefficient between them. We have investigated the effect on the similarity indices of systematically changing the mean and standard deviation of a single Gaussian peak relative to a reference Gaussian peak, of changing one peak, and of changing many peaks, in a simulated 10-peak spectrum. Squared Euclidean cosine is least sensitive to changes and the first-difference methods are most sensitive to changes in mean and standard deviation of peaks. A shift of the center of a peak has a greater effect on the indices than increases in peak width, but a decrease in peak width does lead to significant changes in the indices. We recommend that if these indices are to be used to match spectra, appropriate windows should be chosen to avoid dilution by regions with no significant change.

Keywords: Matching spectra; Correlation coefficient; Euclidean cosine; Similarity index

1. Introduction

The comparison of two spectra is necessary for classification of a spectrum [1], searching a database of spectra to identify an unknown sample [2], to decide if two materials come from a common source[3], in process control against the target spectrum of an acceptable product [4], or to elucidate the structure of a compound [5]. It is realized using a measure of the similarity between the spectra, or conversely, the distance of one spectrum from the other in some measurement space. If the queried spectrum is in the database, a perfect match can be achieved, but if only part of spectrum can be found, the result might be a number of partially matched hits. In environmental analysis, when material spilled in the environment is exposed to weathering, chemical, physical and biological processes will happen [6]. If so, the spectrum of a spill will not always make an exact match with the spectrum of its source, and so, to correctly identify a spill for forensic applications requires some allowed tolerance to be applied. In quality control of herbal medicines, due to the changes of season, place of harvest, pre-processing and the conditions of analyses, chromatographic fingerprints of the same herbal medicine are not always the same [7]. Therefore, any method of matching spectra will need to distinguish between the same material that has been changed and different materials with similar spectra.

Methods for comparing spectra can be divided into direct and indirect methods. Direct matching methods use the spectral data directly, and indirect matching methods use derived information from spectra. The latter relies on identification of selected peaks and the extraction of information from them and has been used by human experts employing visual comparison [8], old computer spectral databases or comparison by simple mathematical calculations such as the measurement of ratios [9]. Multivariate data analysis techniques [10], artificial neural networks [11] and distance/angle [12] methods are direct methods which treat digitized spectra directly without any prior

* Corresponding author.
E-mail address: b.hibbert@unsw.edu.au (D.B. Hibbert).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.05.015

identification of peaks. (Note that it is also possible to use multivariate methods on peak area, or ratio data).

Vibrational and electronic spectra of mixtures can rarely be deconvoluted and assigned to individual components in contrast to the output of other methods such as nuclear magnetic resonance (NMR), chromatography or mass spectrometry, in that individual molecules do not give a single, or a small number of, identifiable peaks. Small informative peaks and overlapped peaks in Fourier transform infrared (FTIR) spectra are not easily identified by computer software and the shape of a peak, which is important for comparison, is difficult to describe accurately. These difficulties can be partially avoided by using point-to-point matching methods because all the data points in a spectrum are used. Similarity/distance methods based on point-to-point matching also have the distinct advantage, compared to pattern recognition techniques, that they only require two spectra and not a set of spectra belonging to different classes. Point-to-point matching is a direct method in which equal-length vectors describing two spectra (intensities, absorbances or detector response) are compared point by point, and a single statistic calculated. The Pearson correlation coefficient is an example of such a similarity index.

In our previous work [13] on matching spectra of petroleum oils, we have found that although different oils can exhibit very different spectra, they can also be very similar. A spectrum of a slightly weathered oil is almost identical to the spectrum of a fresh sample, but it is possible that the difference between the spectrum of a fresh oil and its weathered derivative is greater than the difference between this spectrum and the spectrum of another, highly similar, fresh oil. If we draw the distributions of a similarity measure of such a situation, we see a broader distribution of the spectral similarity of different oils, a narrower distribution for spectra of the same oil but there is often an overlap region leading to false positive or false negative assignments. The success, or otherwise, of a matching method, therefore, rests on its ability to discriminate subtle differences in samples that are inherently similar. The task becomes harder with real samples from the environment because of weathering and introduction of interfering species such as water.

Measures of similarity usually have a defined range, for example, the Pearson's correlation coefficient lies between -1 and 1 or its square between 0 and 1. The minimum or maximum similarity is not always met in the real world, nor is the distribution of values normal. The meaning of the actual value of a similarity index depends on the situation in which it is applied. A correlation coefficient of 0.99 does not mean a match in all situations. It is the analyst's responsibility to decide whether a pair of spectra matches according to the actual situation. This cannot be done without a knowledge (explicit or from experience) of the distribution of the index, against which a particular result is judged.

An IR spectrum not only depends on the particular functional groups, it also reflects the arrangement of these functional groups within a molecule. An IR spectrum is thus, in contrast to NMR or mass spectra, predominantly a property of the whole molecule and not just the sum of the properties of its constituents. The characteristic band of a functional group and the shifts when it connects to different neighboring structures have been described [14,15]. Not only is there not a complete spectral library of the form of bands arising from a particular group in all chemical environments, but also the simple summation of the contributions of all the bands of functional groups in a molecule does not give the real spectrum. It is therefore not possible to predict the spectrum of a complex environmental sample, even if the constituent compounds are known. The only thing we can do is to investigate the effect of the change of peaks of a spectrum itself. To deconvolute an IR spectrum into Gaussian peaks is more difficult than to fit, for example, an X-ray photoelectron spectroscopy (XPS) spectrum. It is impossible to start from a real spectrum and decompose it into small Gaussian peaks. We have therefore conducted the study reported here by simulating increasingly complex spectra, which have been compared pairwise to yield distributions of similarity indices. Starting from a two simulated Gaussian peaks, we investigate the effect, on a number of similarity measures, of differences in the position and width. The study is extended to changes in a single peak among a simulated spectrum of 10 random Gaussian peaks, then to changing more peaks. Finally, we report the distribution of similarity indices for real spectra, augmented by simulated spectra derived from the variance of fast Fourier transform (FFT) coefficients.

2. Theory

2.1. Similarity indices

A number of measures of similarity have been proposed that can be classed as a Minkowski distance

$$D_{1,2} = \left(\sum_i |x_{1,i} - x_{2,i}|^m \right)^{1/m}. \quad (1)$$

The spectra are described by vectors of equal length with individual elements $x_{1,i}$ and $x_{2,i}$. The Euclidean distance is given by $m=2$, and Manhattan (city block) distance is when $m=1$. Statistical measures include the correlation coefficient, and for approaches based on binary variables, the best known is the Tanimoto index, which counts the proportion of points that are mutually above or below a threshold [16]. Similarity indices for use with infrared are discussed by Varmuza et al. [20].

Four point-to-point similarity indices are studied here: squared correlation coefficient (Cor), squared first-difference correlation coefficient (DCor), squared Euclidean cosine (Euc) and squared first-difference Euclidean cosine (DEuc). Their definitions can be found in Table 1.

It is seen that the difference between correlation coefficient and Euclidean cosine is that the data is mean centered in the calculation of correlation coefficient. For a symmetrical peak, on taking the first difference, the mean of the spectrum is zero and so DCor=DEuc. The first derivative of a spectrum is often taken to remove the effect of a sloping baseline.