Classification of Vegetable Oils by Principal Component Analysis of FTIR Spectra

David A. Rusak,* Leah M. Brown, and Scott D. Martin
Department of Chemistry, University of Scranton, Scranton, PA 18510; **rusakd2@scranton.edu*

Downloaded via NORTH CAROLINA STATE UNIV on February 14, 2019 at 06:47:38 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

An IR spectrum usually contains more information than is actually useful. Often the entire spectrum is acquired in the interest of obtaining the position or absorbance values of a few peaks. In this experiment, the useful information is not easy to identify. The variation in the IR spectra of different vegetable oils is subtle, and it is hard to classify different oils by visual comparison of these spectra. However, by using peak positions as variables in a principal component analysis (PCA), the vegetable oils can be correctly classified.

PCA is described as a variable-reduction technique because it reduces redundant information in a set of data. In this experiment, PCA eliminates data that are not useful in making a distinction among the IR spectra of vegetable oils. However, the information that is most useful in making this distinction is retained. PCA has been used to analyze data sets to determine the origin of different wines, juices, and oils *(1–3).* In particular, Dahlberg et al. have used the technique to distinguish 27 different cooking oils and margarines *(3).*

An experiment that uses principal component regression of UV–vis spectra to quantify the active ingredient in a medicinal syrup has appeared recently in this *Journal (4).* In contrast to that experiment, our experiment is not quantitative and hence does not involve regression of the principal component scores. We identify an unknown vegetable oil based on the similarity of the principal component scores of its IR spectrum to those of known vegetable oils. In order to assess the similarity, we generate plots of principal component scores and observe the proximity of the unknown oil scores to the clusters of known scores.

The IR spectra of peanut, canola, sunflower, olive, and soybean oil are similar in appearance. It is difficult to identify the oils by the variation within the IR specta. In this laboratory assignment, the PCA is "trained" by acquiring spectra of peanut, canola, sunflower, olive, and soybean oil and determining the peak position of several of the more intense peaks in each spectrum. These data are entered in a Minitab worksheet. Next, spectra of the unknowns are acquired and the peak positions for each spectrum are also entered into the worksheet. Each row of the worksheet consists of the peak positions for a single spectrum.

PCA treats these peak positions as vectors ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$) and forms linear combinations of the vectors by assigning a weight (a_1, a_2, ..., a_n) to each vector. The weights are chosen to maximize the variation in the linear combinations formed from each set of peak positions. This maximization of variation is subject to the constraint that the sum of the squared weights is equal to one. The linear combinations created are called principal components and can be expressed in the form $\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + ... + a_n\mathbf{x}_n$. Detailed descriptions of principal components analysis can be found in texts on multivariate data analysis *(5–7).*

When two principal components are used as the axes on which to plot the peak position data, the result is that most of the variation present in the original *n*-dimensional data set can be expressed in a two-dimensional scatter plot. By assessing the proximity of the unknowns to the knowns in this principal component space, unknowns can be classified.

PCA can be used, in general, to make distinctions between data sets that are highly correlated. In chemistry it has been used extensively to help make distinctions between data sets acquired for different samples within a similar group (i.e., wines, oils, inks, etc.). The analysis is also commonly used in the field of economics where it treats correlated variables such as interest rates, unemployment, and stock prices. Introducing the method in the instrumental analysis lab is advantageous because it gives students the opportunity to apply it to specific problems. Students are given a broadly applicable chemometric tool and must use it to treat data that they acquire.

This experiment also gives students a perspective on the uncertainty associated with data analysis. In a typical quantitative analysis experiment, uncertainty manifests itself as an error bar. In this experiment, uncertainty leads to poorly defined clusters that can ultimately lead to misidentifications. Students can observe graphically the effect that uncertainty in the data has on their ability to identify an unknown.

Experimental

FTIR Data Collection

Students are given small, labeled vials of the five known oils to be analyzed. We used peanut, sunflower, canola, olive, and soybean oil. Other oils or different types of olive oils can also be used. At least four replicate IR spectra of each oil should be acquired in order to establish the training data set. Salt plates or a NaCl cell with a very small, fixed path length can be used to acquire spectra. The most important consideration, in either case, is the magnitude of the absorbance values obtained. Vegetable oils absorb strongly in several regions of the IR. It can be difficult to determine the precise peak position of a very strong absorption band. These peaks are noisy near their maxima because very little light reaches the detector. Ideally, the maximum absorbance values for the strongest bands should be less than three. The cell or salt plates should be cleaned with dichloromethane between samples. The spectrum and a table of absorbance wavenumber values should be saved for each sample.

Our data were acquired in a 0.025-mm fixed path-length cell with a Nicolet Avatar FTIR (Thermo Nicolet, Madison, WI) using 0.5-cm^{-1} resolution. The peak finder reported peak positions with two digits to the right of the decimal. The actual precision with which the peak positions can be determined is obviously not as good as the data imply. Peak positions that differ by less than 0.25 cm^{-1} (the distance between the data points) are not statistically different. Nevertheless,

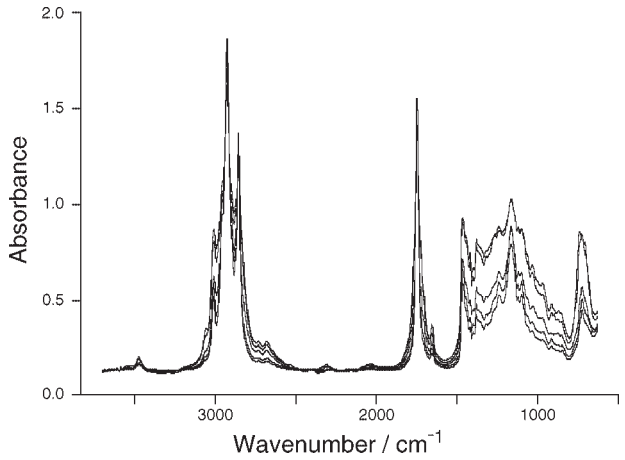

Figure 1. The IR spectra of 5 different types of vegetable oil: canola, peanut, sunflower, soy, and olive.

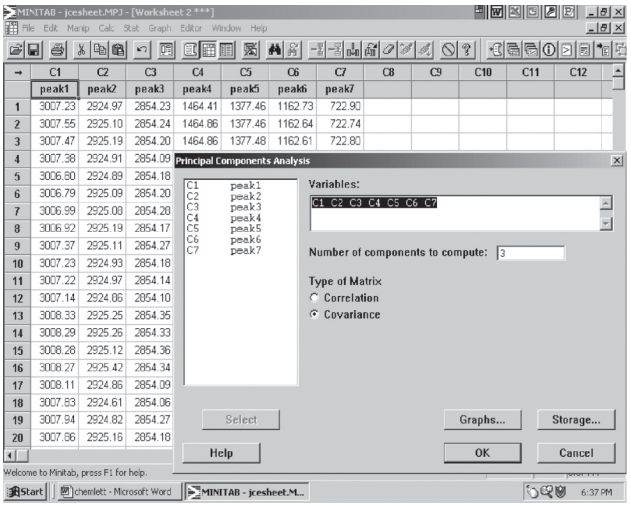

Figure 2. The Minitab worksheet. Each row corresponds to a single spectrum. Each column consists of the wavenumbers at which a specific peak appears in the different spectra.

the peak positions were accurate enough to distinguish among the five oils. The experiment has also been repeated successfully with a Mattson Galaxy 5000 FTIR (Thermo Mattson, Madison, WI) using 1-cm^{-1} resolution.

Students acquire one spectrum of each of the unknown oils. Cells or plates need to be rinsed thoroughly with dichloromethane and allowed to dry for several minutes between samples. Students must be careful not to get the sample oil on the outside surfaces of the cell or plates. Absorbance values in the C−H stretch region should be similar in all the spectra acquired (RSD ~20%). A plot superimposing the spectra from five different oils is shown Figure 1.

Students should try to identify the unknown oils by comparing these spectra to the spectra of the knowns. While it may be possible to identify one or two of the unknowns, it is very challenging to identify all of them. In the process of comparing the spectra, students should note differences in peak positions. This information will ultimately be used in the PCA. Seven peak positions were used in this experiment.[1]

Students extract the information to be used in the PCA from the spectra; they determine the peak positions for peaks of interest in each spectrum. The position of the C−H stretching vibrations, C−H bending vibrations, and a few peaks in the fingerprint region were successfully used as variables in our PCA. Most of the differences in the spectra appear in these regions *(3)*. Many IR software packages can determine peak positions simply by clicking on the peaks or setting a threshold value. Alternatively, the absorbance and wavenumber tables can be used to determine the peak positions. The number of peak positions for each spectrum is not critical. Students can investigate the effect of the data set size on the accuracy of the classification. Once the positions of the chosen peaks are determined for each spectrum, the PCA can be performed.

We used Minitab software for our PCA analysis.[2] Statistica can also perform PCA, and a PCA routine is relatively simple to construct using Mathcad. Our worksheet consists initially of 7 columns and 25 rows. Each column consists of the peak positions (cm^{-1}) that a specific peak appears. Each row corresponds to an individual spectrum. The worksheet is shown in Figure 2.

PCA is accomplished using the "multivariate" option under "stat" in the toolbar. The variables to be transformed (columns) are entered separated by a space. The number of principal components to calculate is chosen, a correlation or covariance matrix is selected, and the coefficients (i.e., the values a_1, a_2, ..., a_7) and scores (the position of each spectrum in this newly defined coordinate system) are stored in the worksheet. The coefficients of each principal component appear as a column (i.e., 3 principal components from 7 peaks gives 3 columns and 7 rows). The scores of each spectrum on a particular principal component also appear as a column (i.e., 3 principal components and 25 spectra gives 3 columns and 25 rows). If different symbols are assigned to each type of oil and each unknown, the plots of scores will facilitate the identification of the unknowns. Inspection of the coefficients allows students to determine which peaks are most useful in distinguishing the oils.

NMC Technique

As an alternative to PCA, it has been suggested that these oils can be classified using their IR spectra and nearest means classification (NMC). The NMC technique can be applied using any spreadsheet as follows:

1. Calculate the mean of each wavenumber for each class of oil.
2. Calculate the absolute difference between the unknown wavenumber and the means for each class and each band.
3. Sum the differences for each class across all seven bands.
4. Assign the unknown to the class with the smallest sum of differences.

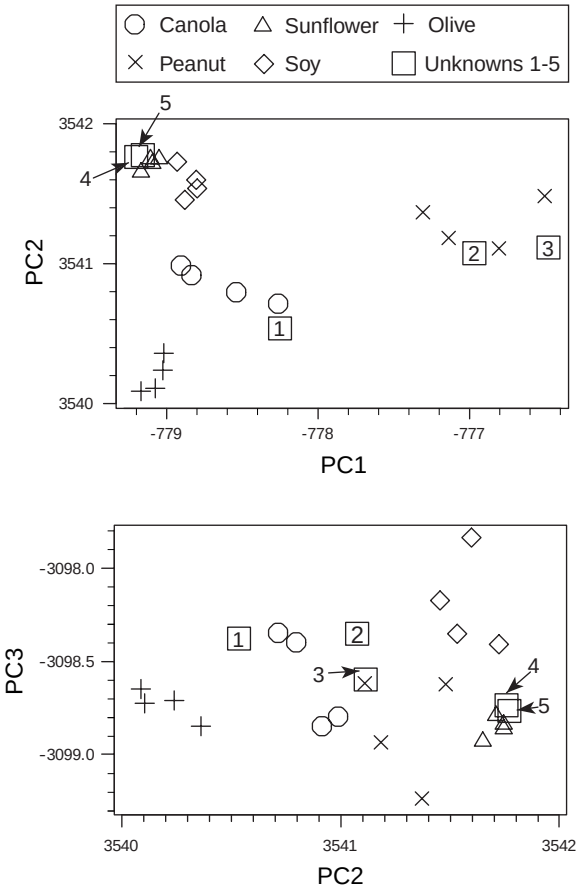

Figure 3. Plots of the principal component scores. The PC2 versus PC1 plot (top) was used to assign unknown 1 to the canola group and unknowns 2 and 3 to the peanut group. The PC3 versus PC2 plot (bottom) was used to assign unknowns 4 and 5 to the sunflower group.

Hazards

Dichloromethane (methylene chloride) is a carcinogen. It can irritate eyes and skin. Standard eye and skin protection should be used. Excessive inhalation can cause nasal and respiratory irritation. Swallowing can cause gastrointestinal irritation.

Results

In order to determine the identity of all the unknowns, two plots were used. Each plot used the principal component scores for each spectrum. Different symbols were used to represent each type of known oil and each unknown oil. The plots are shown in Figure 3. The PC2 versus PC1 plot shows that several of the oils (olive, canola, and peanut) form distinct clusters. From this plot it was determined that unknown 1 was canola oil and that unknowns 2 and 3 were peanut oil. In the PC3 versus PC2 plot sunflower oil and soy oil, which were previously indistinguishable, are well separated. Unknowns 4 and 5 were determined to be sunflower oil because they were found in close proximity to the sunflower oil knowns in both plots.

Use of a larger number of peaks or spectra in the PCA may contribute to production of better-defined clusters in the plots of principal component scores. In our case, the coefficients for the principal components indicate that peak 1 and peak 7 were the most useful in distinguishing the spectra. Other analyses such as refractive index or viscosity can be used in conjunction with the IR spectra in a PCA analysis of vegetable oils. If other analyses are performed, the data acquired can be entered into the worksheet as another column.

Supplemental Material

The experimental procedure and notes for the students are available in this issue of *JCE Online*.

Notes

1. It has been suggested that normalized absorbance values could also be used in the PCA, but we have not done this experiment.

2. A 30-day free trial version of Minitab can be downloaded at *http://www.minitab.com/products/13/demo/index.htm* (accessed Feb 2003). A one-semester (5-month) rental version is also available for a small price.

Literature Cited

1. Kwan, W.–O.; Kowalski, B. R. *J. Food Sci.* **1978,** *43,* 1320.
2. Bayer, S.; McHard, J. A.; Winefordner, J. D. *J. Agric. Food Chem.* **1980,** *28,* 1306.
3. Dahlberg, D. B.; Lee, S. M.; Wenger, S. J.; Vargo, J. A. *Appl. Spec.* **1997,** *51,* 1118.
4. Ribone, M. E.; Pagani, A. P.; Olivieri, A. C.; Goicoechea, H. C. *J. Chem. Educ.* **2000,** *77,* 1330.
5. Dunteman, G. H. Principal Components Analysis. In *Sage University Series on Quantitative Applications in the Social Sciences,* 07-069; Sage Publications: Newbury Park, CA, 1989.
6. Jackson, B. B. *Multivariate Data Analysis;* Richard D. Irwin, Inc.: Homewood, IL, 1983.
7. Gordon, A. D. *Classification;* Chapman & Hall: New York 1981.