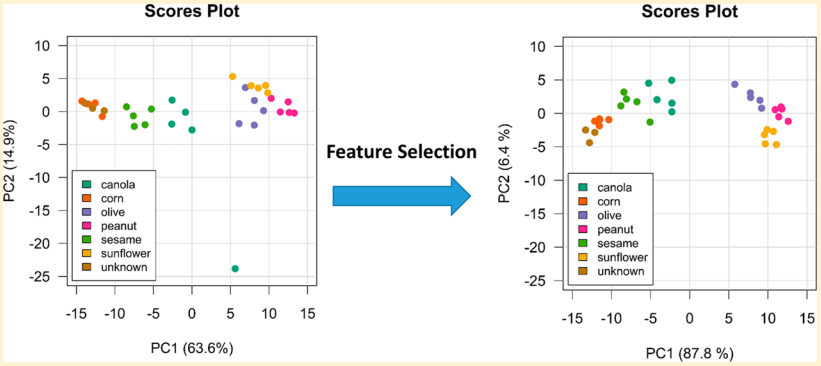
JOURNAL OF CHEMICAL EDUCATION
Cite This: J. Chem. Educ. 2019, 96, 1790−1792
Letter
pubs.acs.org/jchemeduc

Comment on "Identification of Edible Oils by Principal Component Analysis of ^{1}H NMR Spectra"

Tai-Sheng Yeh*

Department of Food Science and Nutrition, Meiho University, Pingtung County 91202, Taiwan

Supporting Information



ABSTRACT: Recently, a useful article by Anderson et al. (*J. Chem. Educ.* **2017**, *94*, 1377−1382) was published in this *Journal* demonstrating how NMR spectroscopy in conjugation with principal component analysis can be applied for identification of edible oils. An update is provided that stresses the importance of feature selection for the purpose of principal component analysis.

KEYWORDS: *Upper-Division Undergraduate, Analytical Chemistry, NMR Spectroscopy*

NMR spectroscopy is one of the most powerful and versatile analytical techniques for food science studies[1] and metabolomics.[2,3] Due to the high prevalence of olive oil adulteration with cheap oil,[4] GC and NMR spectroscopy have been frequently employed in edible oil authentication.[5−7] Recently, a useful article by Anderson et al. was published in this *Journal* demonstrating how NMR spectroscopy in conjugation with chemometrics can be applied for identification of edible oils.[8]

Although Anderson et al. mentioned that their principal component analysis (PCA) approach was similar to that of Rusak et al.,[9] there were subtle differences in terms of feature variable selection for PCA analysis. In Figure 2 of Rusak's report, the absorption signals from seven peak positions were input for PCA analysis. The seven peaks had these wavenumbers in the IR spectra: 3007.23, 2924.97, 2854.23, 1464.41, 1377.46, 1162.73, and 722.90. Even with signals from just these seven peaks, the cluster of olive oil and sunflower was well-separated in Figure 3 of Rusak et al.[9] The file uploaded by Anderson et al. to MetaboAnalyst[10] consisted of whole NMR spectra. In previous NMR work by Vigli et al.[11] and Popescu et al.,[12] and the more recent work by Zhang et al.,[6] only selected feature peaks were used for PCA analysis. The cluster of olive oil and sunflower oil in score plots was well-separated in all these related works. This letter uses data provided by Anderson et al. with feature peak selection to improve the PCA analysis.

FEATURE SELECTION IS IMPORTANT

PCA is a dimension reduction technique by linear combination of the original variables into different principal components (PCs). Using a large or small number of NMR feature peaks usually will not cause a great difference in performance as long as the variables are effective. However, the problem in the current case could be due to the peak grouping step in the data pretreatment process of MetaboAnalyst. In MetaboAnalyst, a moving window of 0.03 ppm was used to group peaks together.[13] Because MetaboAnalyst sums close peaks together, information contained in the NMR data such as multiplets will be removed. This change in feature selection will impede the performance of PCA analysis. The ^{1}H NMR chemical shifts of the linoleyl group triplet and the linolenyl group triplet are 2.79−2.70 and 2.84−2.79 ppm.[14] The 0.03 ppm moving window will have a great chance to mix the linoleyl and linolenyl group signals together. The 1.02−0.92 ppm chemical shift of the linolenyl group triplet and 0.92−0.80 ppm chemical shift of all of the acyl groups except the linolenyl group triplet could also face the same situation. Using a few selected feature peaks as in a previous publication[11] could be easier for the first-time encounter with PCA analysis. Because the effective feature peaks are well-separated and will not sum together in

Received: February 15, 2019
Revised: June 24, 2019
Published: July 16, 2019

DOI: 10.1021/acs.jchemed.9b00133
J. Chem. Educ. 2019, 96, 1790−1792

MetaboAnalyst, the PCA performance by the selected feature peaks is comparable to that in the previous publication. After the initial success with the selected feature peaks, students could expand the list of feature peak assignments from the more recent publication[14] in their next PCA analysis. Better performance could also be obtained by an NMR data processing R package such as speaq 2.0, which uses wavelets to summarize the peaks.[13] R programing might be a bit challenging for students with no prior experience. Using MetaboAnalyst will be easier, but one must be aware of the details of data preprocessing procedures for better PCA performance.

Improved PCA Results for Score Plot

By using the peaks in Table 1 of Vigli et al.[11] as input variables for PCA analysis by MetaboAnalyst, the resulting score plot is shown in Figure 1. The input file in csv format and

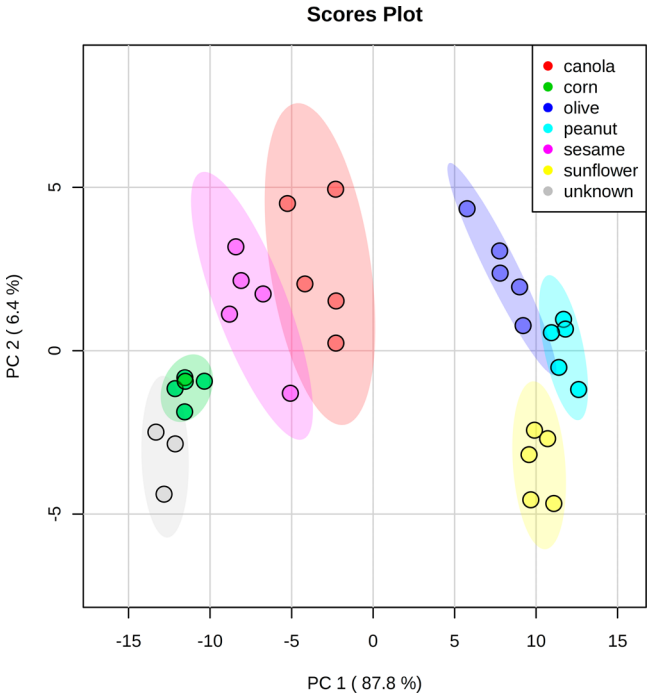


Figure 1. Score plot for PC1 and PC2 using selected peak intensities as input.

MetaboAnalyst execution result as a zip file are included in the Supporting Information (all_raw_select.csv and Download_0213_2019.zip). The color shaded area in each cluster is the 95% confidence region. The cluster of sunflower oil with the 95% confidence region is well-separated from olive oil. The explained variance for PC1 87.8% plus PC2 6.4% is higher than the previous report of PC1 78.7% plus PC2 7.7% in Figure 3 (top row) of Anderson et al.[8] Improved PCA performance for the cluster of six vegetable oils can be seen in Figure 1.

Easier Result Visualization with a Smaller Number of Feature Variables

As MetaboAnalyst sums close peaks together, correct interpretation of the resulting loading plot will be difficult. The original input file consisted of 1100 peaks; after data preprocessing by MetaboAnalyst, the peak list reduced to 163 peaks (data_normalized.csv in the Supporting Information Download_0214_2019.zip). By using just 10 variables (Figure 2), the influence of variables on PC1 and PC2 becomes easier

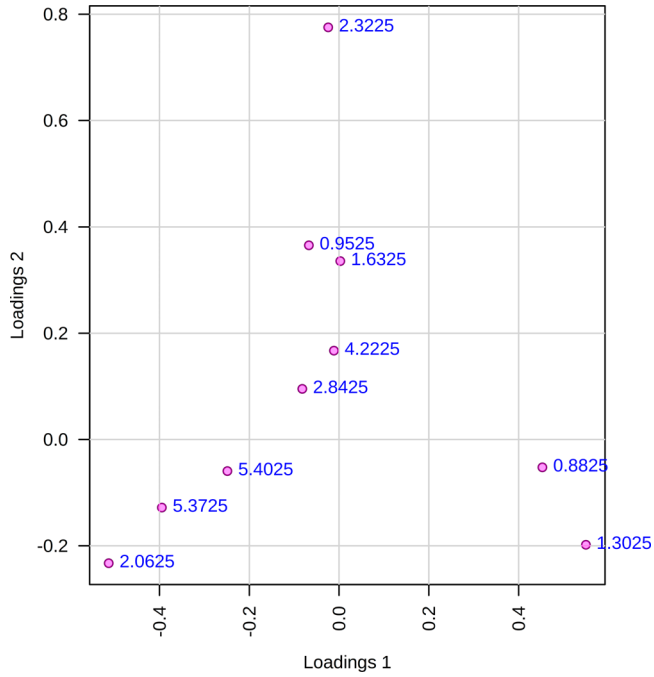


Figure 2. Loading plot for PC1 and PC2 using selected peak intensities as input.

to visualize. The opposite influence of the 0.88 and 1.30 ppm peak pair versus the 2.06 and 5.37 ppm peak pair shows the same trend as the work by Popescu et al.[12] This can be shown, for example, for the olive, peanut, and sunflower sample on the right-hand side of PC1, and for the canola, sesame, and corn oil sample on the left side of PC1.

Too many variables will be a bit too complex to visualize and understand the biplot for PCA analysis with ease. The biplot for the PCA analysis (Figure 3) is easier to understand. Biplots can link the score plot of different oils and the loading plot of

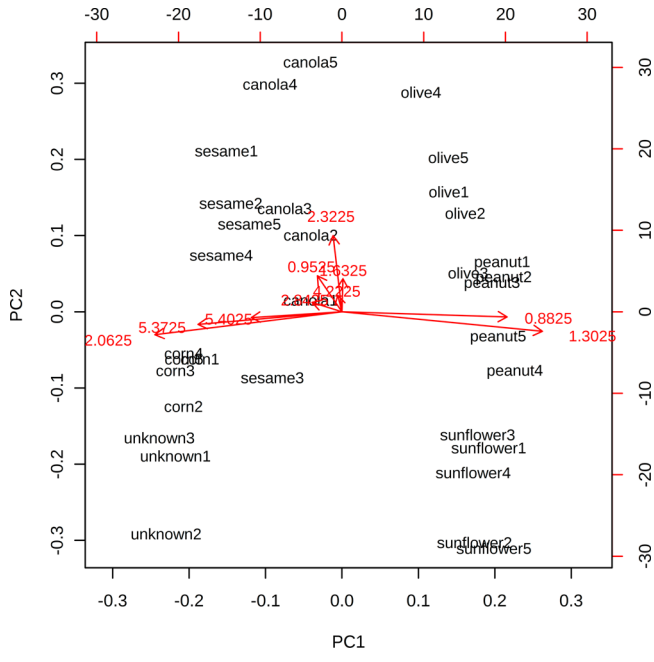


Figure 3. Biplot for PC1 and PC2 using selected peak intensities as input.

DOI: 10.1021/acs.jchemed.9b00133
J. Chem. Educ. 2019, 96, 1790−1792

NMR peaks. The relatively small influence of the linolenyl group peak at 2.84 ppm is due to the relatively small amount of linolenic acid in most vegetable oils. The linolenic acid in corn, olive, peanut, sesame, and sunflower oils is 1.26, 0.54, 0.05, 0.77, and 0.28, respectively, in Table 3 of the work of Zhang et al.[6] Because the amount of linolenic acid in vegetable oils did not vary greatly, the position in the score plot was quite close to zero.

The olive oil, peanut oil, and sunflower oil samples have a high positive score in PC1 due to 0.88 and 1.30 ppm. The large negative score in PC1 for corn oil is due to 2.06 and 5.37 ppm. From the biplot of Popescu et al.,[12] sunflower oil is associated with more linoleic acid, and both peanut oil and olive oil are associated with more oleic acid. The linoleic acid in corn, olive, peanut, sesame, and sunflower oils is 53.6, 9.58, 32.8, 43.5, and 61.4, respectively, in Table 3 of the work of Zhang et al.[6] Therefore, peanut oil is closer to olive oil than sunflower oil in both the score plot and the biplot.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.9b00133.

Note for upload NMR spectra bins with selected feature peaks for MetaboAnalyst (PDF, DOCX)

Example of selected NMR feature peaks input csv files for PCA analysis in zipped file (ZIP)

Copy of the execution result with feature peaks from MetaboAnalyst Download_0213_2019 (ZIP)

Copy of the execution result with whole spectra from MetaboAnalyst Download_0214_2019 (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: x00010091@meiho.edu.tw.

ORCID

Tai-Sheng Yeh: 0000-0002-7798-2669

Notes

The author declares no competing financial interest.

■ REFERENCES

(1) Hatzakis, E. Nuclear Magnetic Resonance (NMR) Spectroscopy in Food Science: A Comprehensive Review. *Compr. Rev. Food Sci. Food Saf.* **2019**, *18* (1), 189−220.

(2) Nagana Gowda, G. A.; Raftery, D. Recent Advances in NMR-Based Metabolomics. *Anal. Chem. (Washington, DC, U. S.)* **2017**, *89* (1), 490−510.

(3) Markley, J. L.; Brüschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34−40.

(4) Moore, J. C.; Spink, J.; Lipp, M. Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *J. Food Sci.* **2012**, *77* (4), R118−R126.

(5) Fang, G.; Goh, J. Y.; Tay, M.; Lau, H. F.; Li, S. F. Y. Characterization of oils and fats by ^{1}H NMR and GC/MS fingerprinting: Classification, prediction and detection of adulteration. *Food Chem.* **2013**, *138* (2), 1461−1469.

(6) Zhang, Y.; Zhao, Y.; Shen, G.; Zhong, S.; Feng, J. NMR spectroscopy in conjugation with multivariate statistical analysis for distinguishing plant origin of edible oils. *J. Food Compos. Anal.* **2018**, *69*, 140−148.

(7) Ravaglia, L. M.; Pizzotti, A. B. C.; Alcantara, G. B. NMR-based and chemometric approaches applicable to adulteration studies for assessment of the botanical origin of edible oils. *J. Food Sci. Technol.* **2019**, *56* (1), 507−511.

(8) Anderson, S. L.; Rovnyak, D.; Strein, T. G. Identification of Edible Oils by Principal Component Analysis of ^{1}H NMR Spectra. *J. Chem. Educ.* **2017**, *94* (9), 1377−1382.

(9) Rusak, D. A.; Brown, L. M.; Martin, S. D. Classification of Vegetable Oils by Principal Component Analysis of FTIR Spectra. *J. Chem. Educ.* **2003**, *80* (5), 541−543.

(10) Chong, J.; Soufan, O.; Caraus, I.; Xia, J.; Li, C.; Wishart, D. S.; Bourque, G.; Li, S. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **2018**, *46* (W1), W486−W494.

(11) Vigli, G.; Philippidis, A.; Spyros, A.; Dais, P. Classification of Edible Oils by Employing ^{31}P and ^{1}H NMR Spectroscopy in Combination with Multivariate Statistical Analysis. A Proposal for the Detection of Seed Oil Adulteration in Virgin Olive Oils. *J. Agric. Food Chem.* **2003**, *51* (19), 5715−5722.

(12) Popescu, R.; Costinel, D.; Dinca, O. R.; Marinescu, A.; Stefanescu, I.; Ionete, R. E. Discrimination of vegetable oils using NMR spectroscopy and chemometrics. *Food Control* **2015**, *48*, 84−90.

(13) Beirnaert, C.; Meysman, P.; Vu, T. N.; Hermans, N.; Apers, S.; Pieters, L.; Covaci, A.; Laukens, K. speaq 2.0: A complete workflow for high-throughput 1D NMR spectra processing and quantification. *PLoS Comput. Biol.* **2018**, *14* (3), No. e1006018.

(14) Zhu, M.; Shi, T.; Chen, Y.; Luo, S.; Leng, T.; Wang, Y.; Guo, C.; Xie, M. Prediction of fatty acid composition in camellia oil by ^{1}H NMR combined with PLS regression. *Food Chem.* **2019**, *279*, 339−346.

DOI: 10.1021/acs.jchemed.9b00133
J. Chem. Educ. 2019, 96, 1790−1792