PS A Data-Driven Method for Processing and Analysis of Gas Chromatography-Mass Spectrometry (GC-MS) Signals in Differentiation of Oil Samples*

Le Lu[1], Andrew Bishop[1], Tianxing Wang[1], Rosemarie Bisquera[1], and Yongchun Tang[1]

Search and Discovery Article #42401 (2019)**
Posted July 29, 2019

*Adapted from poster presentation given at 2019 AAPG Annual Convention and Exhibition, San Antonio, Texas, May 19-22, 2019
**Datapages © 2019 Serial rights given by author. For all other rights contact author directly. DOI:10.1306/42401Lu2019

[1] PEER Institute, Covina, California (le.lu@peeri.org)

Abstract

Due to the large number of organic compounds existing in crude oil samples and the variations in compositions, differentiation of oil and extract data by gas chromatography-mass spectrometry (GC-MS) analysis is time consuming and laborious. The process relies heavily on the skills of a limited pool of experienced analysts, and as a consequence delivers subjective outcomes. Comparison typically requires alignment in the time domain as a data pre-processing step, which can be challenging as elution orders can vary depending on test procedures, conditions and data vintages. Machine learning methods are known for their unparalleled ability to efficiently handle large volumes of data, intelligently extracting diagnostic features, and establishing complicated non-linear relationship between data and interpretations. In this study, a data-driven method is proposed to assist the differentiation of geochemistry samples, based on a database of oil and source rock extract GC-MS measurements and machine learning techniques.

Chromatogram peaks can be consistently located by a recurrent neural network classifier with the application of a continuous wavelet transform to the total signal. Compound assignment is performed via supervised classification of the mass spectra. The machine learning models are trained with a database of interpreted oil and known source rock extract samples, including various data vintages and instrument types. By comparing automatically assigned, comprehensive compound assignment between individual samples, major differences in the abundance of common compounds and the identification of missing species enables quantitative discrimination between critical samples of interest. Diagnostic compounds identified by this process can be used as a basis for robust production allocation schemes and higher confidence oil-source correlations.

Selected References

Christensen, J.H., G. Tomasi, and A.B. Hansen, 2005, Chemical fingerprinting of petroleum biomarkers using time warping and PCA: Environmental Science & Technology, v. 39/1, p. 255-260.

Fang, G., J.Y. Goh, M. Tay, H.F. Lau, and S.F.Y. Li, 2013, Characterization of oils and fats by 1H NMR and GC/MS fingerprinting: Classification, prediction and detection of adulteration: Food Chemistry, v. 138/2-3, p.1461-1469.

Lu, Y., I. Cohen, X.S. Zhou, and Q. Tian, 2007, Feature selection using principal feature analysis: Proceedings of the 15th ACM international conference on Multimedia, p. 301-304.

Simoneit, B.R., 2005, A review of current applications of mass spectrometry for biomarker/molecular tracer elucidations: Mass Spectrometry Reviews, v. 24/5, p. 719-765.

Wang, L., Y. Lei, Y. Zeng, L. Tong, and B. Yan, 2013, Principal feature analysis: A multivariate feature selection method for FMRI data: Computational and Mathematical Methods in Medicine.

A Data-Driven Method for Processing and Analysis of Gas Chromatography-Mass Spectrometry (GC-MS) Signals in Differentiation of Oil Samples

Le Lu, Andrew Bishop, Tianxing Wang, Rose Bisquera and Yongchun Tang (tang@peeri.org)

INTRODUCTION

In geochemistry production allocation, it is necessary to identify the essential differences between the molecular signatures of representative end-member samples. Such differences need to be reproducible for replicate analyses of the same sample, but sufficiently different to consistently distinguish relative proportions of the oil end-members of interest. Finding such 'needles in the haystack' is a laborious task for human analysts. Can machine learning approaches help?

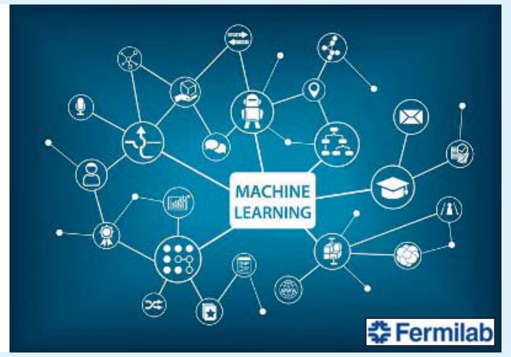
Machine Learning (ML)

ML algorithms are known for their unparalleled ability to efficiently handle large volumes of data, intelligently extracting diagnostic features, and establishing complicated non-linear relationships between data and interpretations. Two of the most commonly applied approaches are *Supervised* and *Unsupervised Learning* (see inset boxes for further details).

Definition

Machine learning is the scientific study of algorithms and models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.

https://en.wikipedia.org/wiki/Machine_learning



Objectives

In this study, a data driven method is proposed to assist the differentiation of geochemistry samples, based on a synthesis of oil and source rock extract GC-MS analysis datafiles and machine learning techniques.

Supervised Learning

Algorithms are designed to build a mathematical model on the basis of a training data set, with both known inputs and outputs, i.e. data and associated interpretations. Through multiple iterations, the algorithms learn the requisite functions to arrive at the desired outcome.

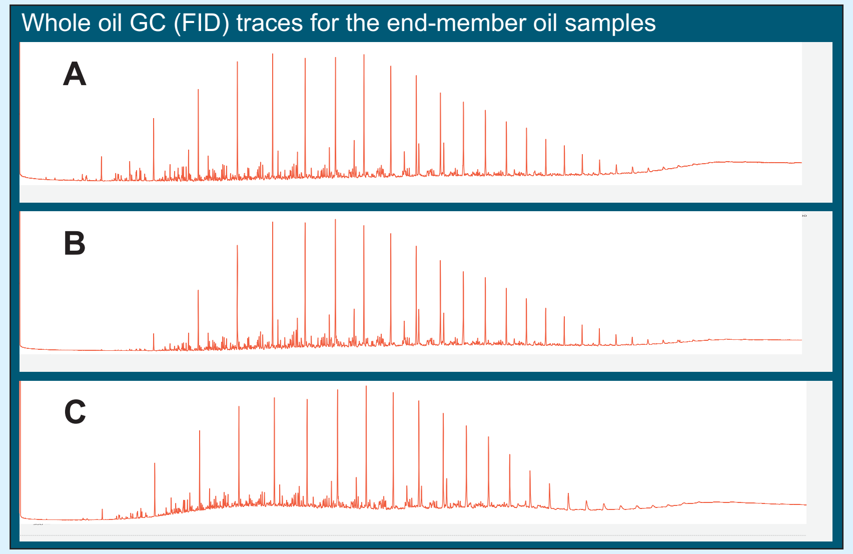
Unsupervised learning

These algorithms are designed to handle situations where only input data is available. Thus, they seeking structure in the data using methods such as cluster analysis. Groupings can then be identified, and the algorithms will highlight the occurrence or absence of such groupings in future datasets.

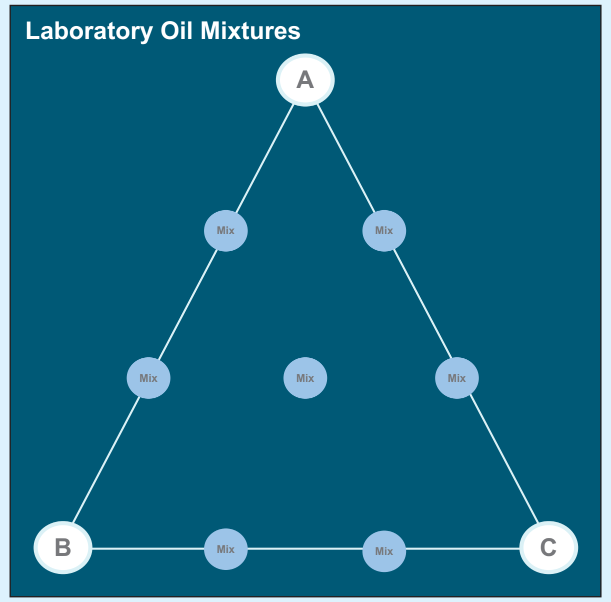
SAMPLES

We have developed our initial data driven method on the basis of three oil samples, selected from the PEER Institute's collection of oil and source rock extract samples. Samples were chosen from the same geological province, with a view to them being similar enough to provide a realistic test of a typical production allocation scenario, but with some evident differences on the basis of fluid properties to afford a reasonable probability of success. GC traces for these samples is shown in the figure to the right.

Whole oil GC (FID) traces for the end-member oil samples



Laboratory Oil Mixtures



To test how effectively this approach is with regard to production allocation studies, a series of oil mixtures were prepared in the laboratory. A simplex mixture design was employed for this purpose (see left). Three sets of binary end-member mixtures were prepared in the proportion of ⅓ and ⅔. A remaining mix was prepared with ⅓ of each oil.

Analysis

A clean hydrocarbon fraction was prepared for each sample, analyzed with an aromatic hydrocarbon SIM method, and run in duplicate. The m/z 57 ion was included to permit identification of the n-alkanes, which are used for automatic trace alignment. Major peaks were quantified in all traces, and identified simply by a code representing the ion and the retention time.

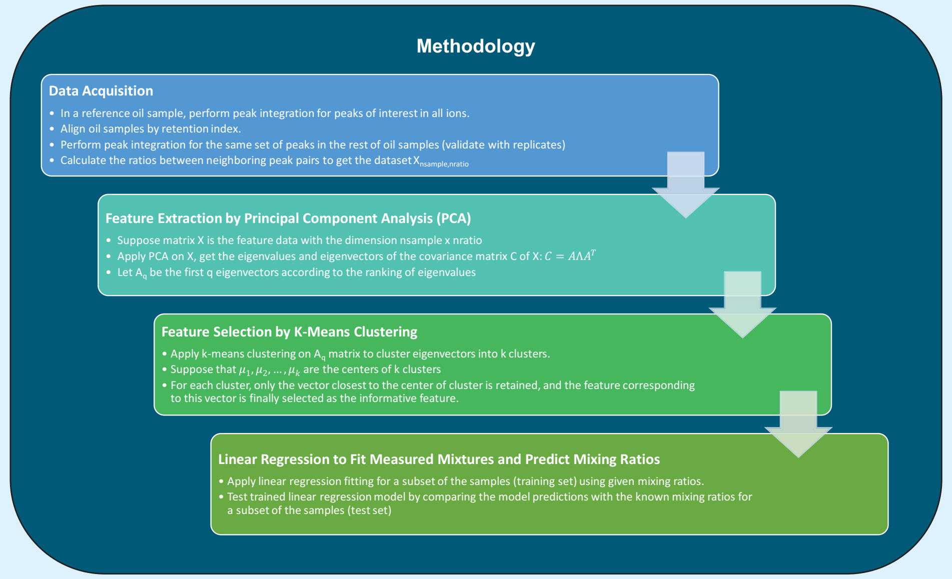
APPLICATION OF MACHINE LEARNING

Motivation

Feature selection is essential in dimension reduction in analysis of oil GC-MS measurements and in production allocation practice, due to the large number of organic compounds existing in crude oil samples and the variations in compositions.

Compared with feature extraction methods such as principal component analysis (PCA) and independent component analysis (ICA), which finds a mapping from the original feature space to a lower dimensional feature space, feature selection picks a subset of the original features. The selected subset is helpful in understanding and discriminating the samples.

Multivariate principal feature analysis (PFA) considers the dependencies between the features when calculating the ranking scores for them, retain more useful information with fewer features.



Methodology

Data Acquisition

- In a reference oil sample, perform peak integration for peaks of interest in all ions.
- Align oil samples by retention index.
- Perform peak integration for the same set of peaks in the rest of oil samples (validate with replicates)
- Calculate the ratios between neighboring peak pairs to get the dataset $X_{nsample,nratio}$

Feature Extraction by Principal Component Analysis (PCA)

- Suppose matrix X is the feature data with the dimension nsample x nratio
- Apply PCA on X, get the eigenvalues and eigenvectors of the covariance matrix C of X: $C = A\Lambda A^T$
- Let A_q be the first q eigenvectors according to the ranking of eigenvalues

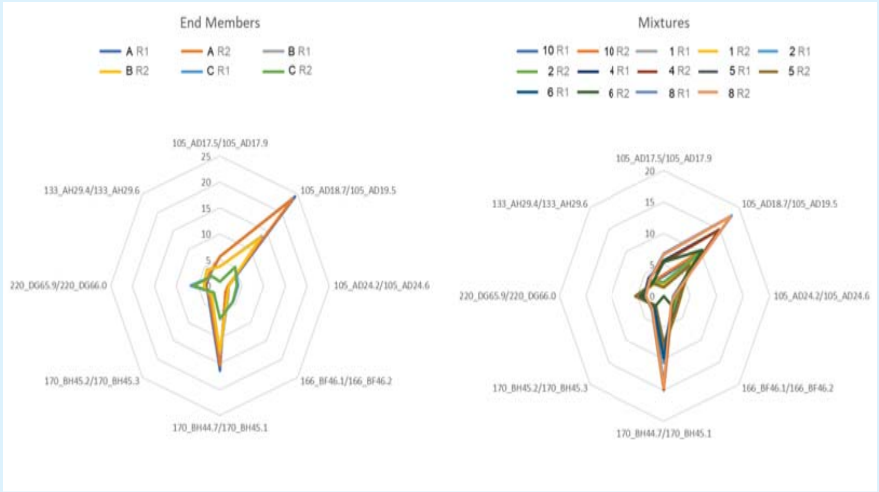
Feature Selection by K-Means Clustering

- Apply k-means clustering on A_q matrix to cluster eigenvectors into k clusters.
- Suppose that $\mu_1, \mu_2, \ldots, \mu_k$ are the centers of k clusters
- For each cluster, only the vector closest to the center of cluster is retained, and the feature corresponding to this vector is finally selected as the informative feature.

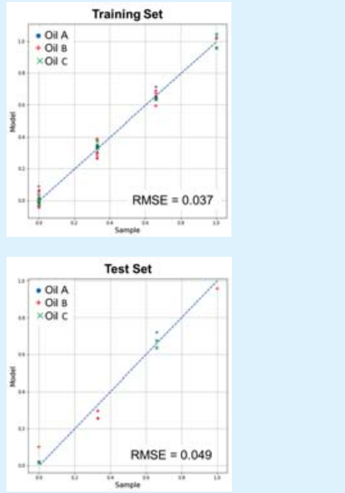
Linear Regression to Fit Measured Mixtures and Predict Mixing Ratios

- Apply linear regression fitting for a subset of the samples (training set) using given mixing ratios.
- Test trained linear regression model by comparing the model predictions with the known mixing ratios for a subset of the samples (test set)

Star Diagrams of Oil Samples with selected PFA features



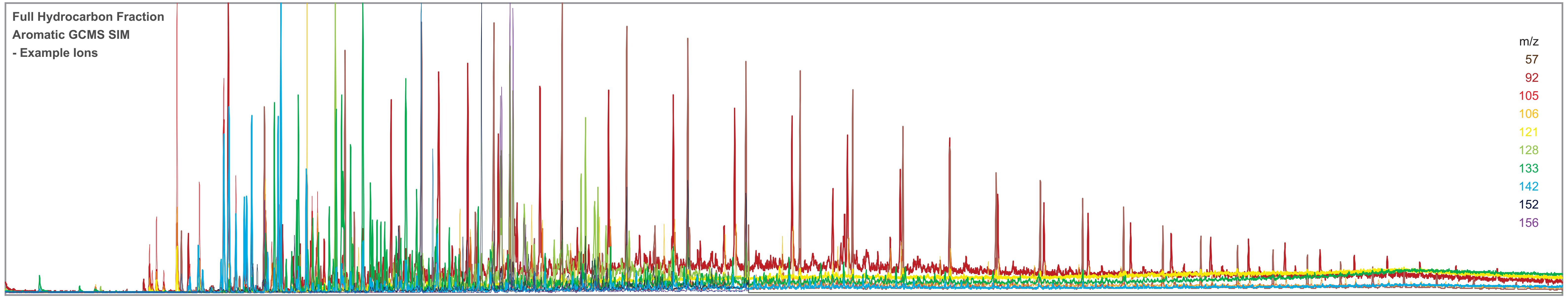
Mixture Estimation Results

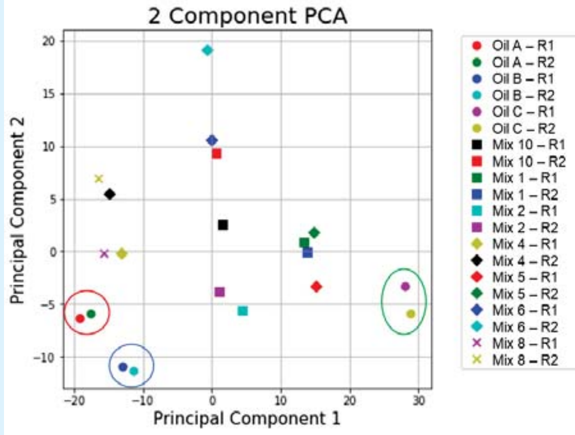


Full Hydrocarbon Fraction

Aromatic GCMS SIM

- Example Ions





Principal component analysis (PCA) for oil samples using peak ratios in GC-MS. GC-MS runs for the same sample are generally well clustered, whilst those from different samples are separated.

DISCUSSION AND CONCLUSIONS

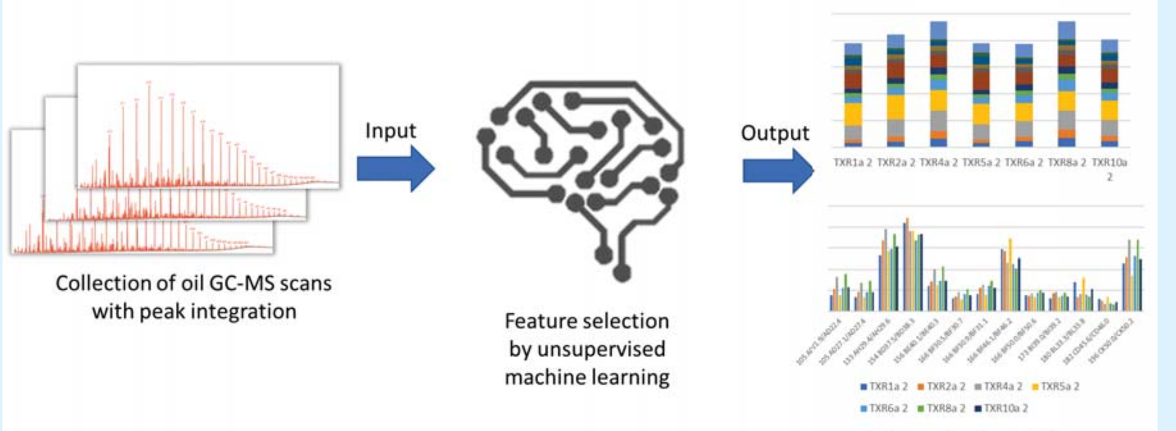
This study is the first to use PFA for feature selection in processing and analysis of oil sample GC-MS signals. The concept of PFA was proposed by Lu et al.[1] in 2007, and its applications have been explored in face tracking, content-based image retrieval, functional magnetic resonance imaging analysis[2].

Spatiotemporal characteristics of up to thousands of peak ratios from multiple ion scans in GC-MS is exploited using the principal components criteria, in order to retain the most of the information.

The PFA method adopts the k-means clustering to identify the specific subset of the original features, while most feature extraction methods such as PCA and k-means solely can only produce a projection to a new reduced component space.

The extracted features are used to train a linear regression model for the mixing ratios of the end members in this study, due to the small training and test sets. The predictions for the test set show promising ability of this method in terms of optimal feature selection and prediction by the fingerprinting. It is viable to use more complicated supervised machine learning algorithms, such as deep neural network and decision tree, for the regression and prediction when more sample data is available.

[1] Lu, Y., Cohen, I., Zhou, X.S. and Tian, Q., 2007, September. Feature selection using principal feature analysis. In Proceedings of the 15th ACM international conference on Multimedia (pp. 301-304). ACM.
[2] Wang, L., Lei, Y., Zeng, Y., Tong, L. and Yan, B., 2013. Principal feature analysis: A multivariate feature selection method for FMRI data. Computational and mathematical methods in medicine, 2013.



Collection of oil GC-MS scans with peak integration

Feature selection by unsupervised machine learning

Selected subset of features for fingerprinting

GRAPHICAL SUMMARY