



Introducing Principal Coordinate Analysis (PCoA) Assisted EEMF Spectroscopic Based Novel Analytical Approach for the Discrimination of Commercial Gasoline Fuels

Riham El Kurdi¹ · Keshav Kumar² · Digambara Patra¹

Received: 16 June 2020 / Accepted: 1 September 2020 / Published online: 7 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In the present work, a novel analytical procedure by integrating principal coordinate analysis (PcoA) with excitation-emission matrix fluorescence (EEMF) spectroscopy was introduced for discriminating the commercial gasoline fuels. The PcoA technique involved analysis of the distance matrices containing the dissimilarity information and it can serve as an efficient tool for capturing the major as well as subtle compositional differences among the analyzed commercial gasoline samples. The utility of the proposed PcoA assisted EEMF analytical procedure was successfully tested by discriminating gasoline fuel samples belonging to five different industrial brands. The obtained results clearly showed that combination of PcoA and EEMF could provide a simple, sensitive and economical analytical procedure to carry out the rapid analyses of the gasoline samples belonging to different brands.

Keywords Principal coordinate analysis · Excitation-emission matrix fluorescence spectroscopy · Gasoline · Discrimination · Principal component analysis

Introduction

Gasoline is a petroleum-derived product and mainly used as fuel for the internal combustion (IC) engines [1–3]. The composition of the gasoline samples apart from the refining processes are also heavily influenced by the geographical and environmental factors [1–4]. The usage of low grade or unspecified gasoline as a fuel can significantly reduce the performance as well as longevity of the IC engines [1, 2, 4]. It can also increase the emission of greenhouse gases causing the global warming [1, 2, 4, 5]. It is important that a simple, sensitive and cost effective analytical procedure must be available that could routinely be used for classifying different types of gasoline samples.

Gasoline is essentially a mixture of alkanes (C_nH_{n+4} , where n is usually 5–12) [6]. It also contain a number of polycyclic aromatic hydrocarbons (PAHs) in varying concentrations [4, 7]. The PAHs have the rigid molecular framework making them highly fluorescent in nature [4, 8, 9]. The presence of PAHs allows fluorescence based analysis for the gasoline samples. Excitation-emission matrix fluorescence (EEMF) spectroscopy is multidimensional fluorescence technique that allows the simultaneous visualization of the fluorescence response of the mixture of fluorophores in a single three-dimensional plot [3, 10–12]. EEMF spectroscopy describes the variation of the emission spectral profile acquired as an increasing function of the excitation wavelength and vice versa. EEMF can serve as a useful technique to *fingerprint* the samples [3, 10–12]. EEMF spectroscopy has been successfully used as an analytical tool in different fields including pharmaceutical, food, agriculture and petroleum [3, 10–24]. The successful usage of EEMF technique could mainly be attributed to the availability of user-friendly software that enable acquisition of the data in an automatic and swift manner [25]. The modern software also enable the removal of Rayleigh scattering signals at the data acquisition stage without asking for user inputs

✉ Keshav Kumar
keshavkumar29@gmail.com

✉ Digambara Patra
dp03@aub.edu.lb

¹ Department of Chemistry, American University of Beirut, Beirut, Lebanon

² Present address: Hochschule Geisenheim University, Geisenheim, Germany

[25]. The EEMF data sets have the trilinear structure that also allowed its assimilation with various chemometric techniques [3, 13, 25].

Principal coordinate analysis (PcoA) belongs to the class of multidimensional scaling technique that involves projection of the data set in an abstract space spanned by a set of Cartesian axes [26–31]. The projection of the data in orthogonal space simplifies the data set and subsequently makes the data interpretation easier [26–31]. In PcoA, the distance among the samples in the orthogonal space directly reflect their relationship among each other [26–31]. All the samples with similar characteristics would appear in close proximity of each other and can easily be grouped together. Theoretically, a great resemblance could be seen between PcoA and principal component analysis (PCA) [32–34]. Both the approaches involve Eigen analysis i.e. computation of Eigenvector and Eigen values [26–34]. However, the difference mainly arises from the fact that PCA involve Eigen analysis of covariance matrix whereas PcoA involve the Eigen analysis of the user-specified distance matrix containing the dissimilarity information [26–34]. As PCA is based on covariance matrix, thus it could be considered as a method of choice for analyzing the data sets of the samples belonging to specific groups. Whereas, PcoA is based on the distance matrix, thus, one can easily argue that it could be a method of choice to capture the major as well as subtle compositional differences among the analyzed samples.

PcoA assisted EEMF spectral analysis can serve as a useful and cost-effective analytical technique for the analysis of the real life samples such as gasoline. However, for unknown reasons the analytical utility of PcoA-EEMF has not been explored hitherto. The present work explore the advantages associated with PcoA and EEMF technique towards the analysis of gasoline samples belonging to different brands having minor to significant compositional differences. To the best of our knowledge, it is the first report that describe the application of PcoA assisted EEMF analyses for the discrimination of gasoline samples.

Material and Methods

Gasoline Samples

Two type of gasoline samples G98* and G95* for each of the following four brands *apeck*, *hypco*, *total*, *wardiyeh* and *medco* were procured from the local vendors in Lebanon. These two type of gasoline mainly differ in the ethanol percentage. The G98 contain 2% ethanol whereas G95 contain 5% ethanol. The G98 sample was further diluted using the ethanol to create another set of 11 samples labelled as G97, G95, G90, G88, G85, G83, G80, G77, G75 G72 and G70 for each brand type. The G95 and G95* differ in the sense that G95 is created in the lab and the G95* come from the refinery.

These two kind of samples for each brands were also included to see if the proposed approach can also discriminate G95* and G95 containing same amount of ethanol.

Data Acquisitions

The 3D emission and excitation measurements were using Jobin-Yvon-Horiba Fluorolog III fluorimeter and the FluorEssence program. The 100 W Xenon lamp was used as excitation source. The R-928 operating at a voltage of 950 V was used as the detector. The EEMF spectra were collected over the excitation-wavelength range of 250–550 nm (in a step size of 10 nm) and emission-wavelength range of 300–650 nm (in a step size of 2 nm).

Computational Platform

All the analyses and EEMF data arrangement were carried out on MATLAB 2017a platform.

Results and Discussion

EEMF Characteristics of *apeck*, *hypco*, *total*, *wardiyeh* and *medco* Brands

The EEMF spectra of G98* sample for each of the five brands *apeck*, *hypco*, *total*, *wardiyeh* and *medco* are shown in Fig. 1. All the five gasoline types were mainly found to show the fluorescence in the excitation wavelength range of 300–450 nm and emission-wavelength range of 320–500 nm. The EEMF spectra clearly suggest that samples of each of the five brands are highly fluorescent in nature.

Working Scheme for Carrying out PcoA on EEMF Data Sets

In the first step, the Rayleigh scattering eliminated EEMF data sets of the samples must be arranged in a three-way array \underline{X} of dimension emission-wavelength \times excitation-wavelength \times sample. The data should further be unfolded along the sample mode (i.e. the third mode) to generate a two-way array X of dimension sample \times (emission-wavelength \times excitation-wavelength). As discussed above, input for PcoA is a distance matrix containing the dissimilarity values. Thus, in the second step, a distance matrix D of dimension sample \times sample must be created using a suitable method, some of the commonly used approaches are Euclidean, Spearman, Chebychev, Manhattan etc. [35]. In the third step, a squared proximity matrix A of dimension, sample \times sample, should be obtained by squaring each element of matrix D . It helps in magnifying the significant dissimilarity values between the samples. In the fourth step, a matrix Y of dimension, sample \times sample, must

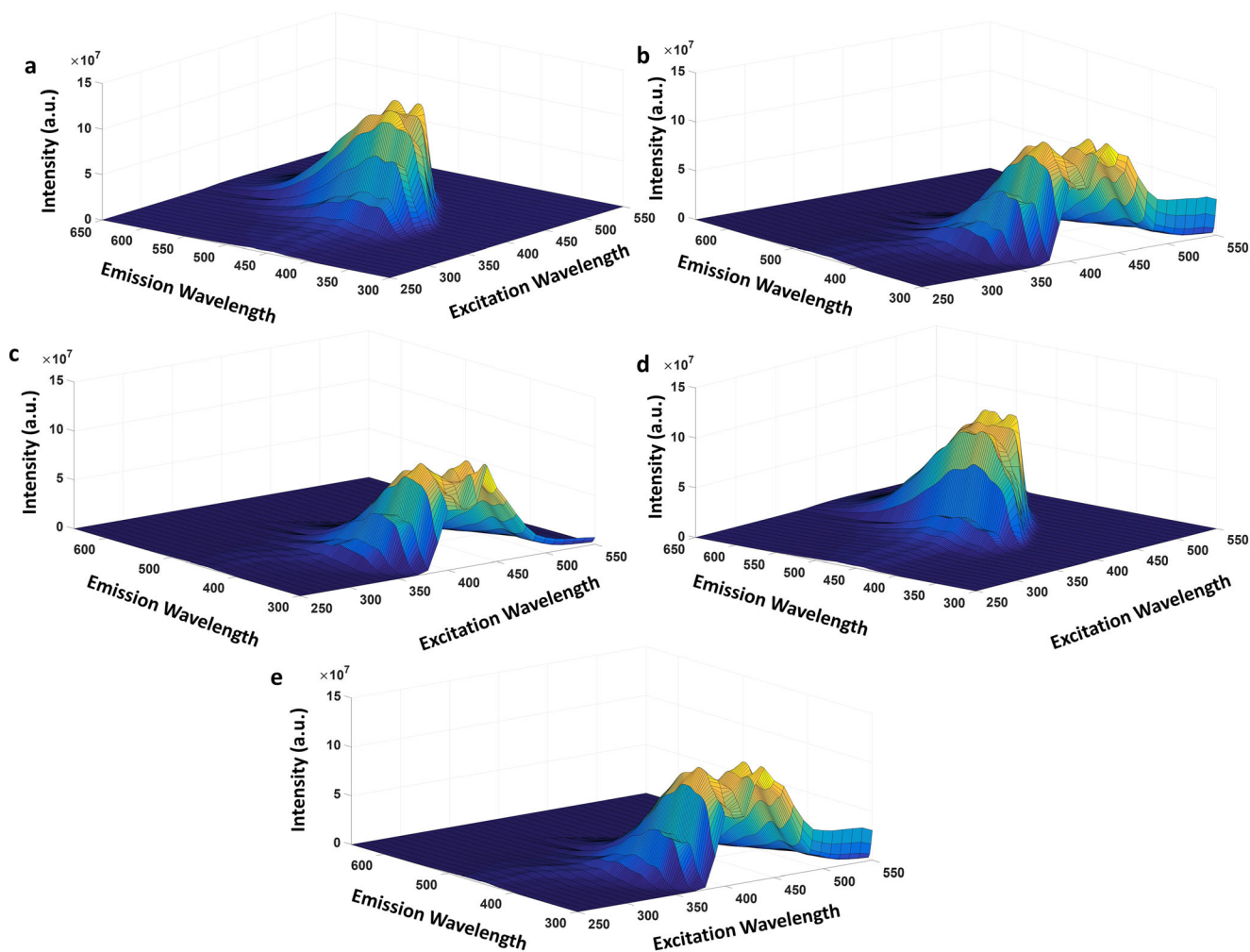


Fig. 1 EEMF spectra of (a) *apeck*, b *hypco*, c *total*, d *wardiyeh* and e *medco* type gasoline samples (G98*)

be obtained by subjecting the matrix *A* to Gower centering [27]. It essentially helps in positioning the origin of a new set of axes at the centroid of scattered samples and simultaneously preserve distances among the samples. In the fifth step, the matrix *Y* should be subjected to the singular value

decomposition (SVD) [33, 34] and the column of the right singular matrix that contain all the relevant information regarding the sample composition could be plotted to create PcoA plots. As discussed earlier, all the samples with similar composition in PcoA plot will appear in close vicinity of each

Fig. 2 Unfolded-EEMF spectral profiles of the gasoline samples

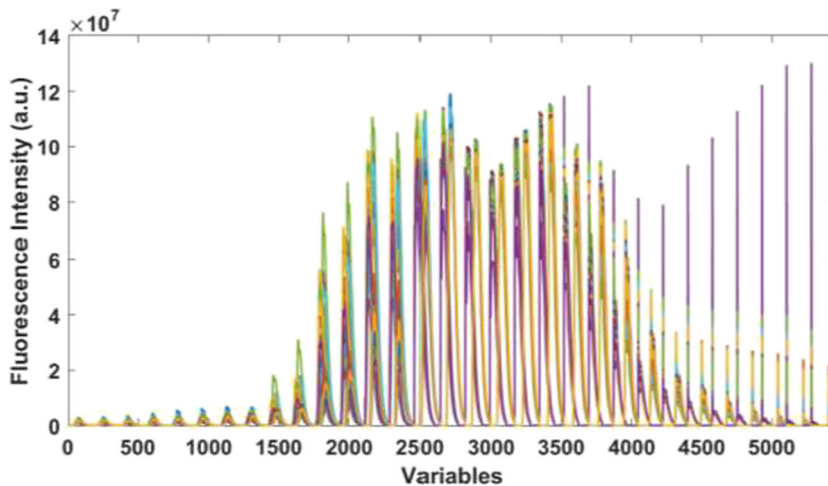
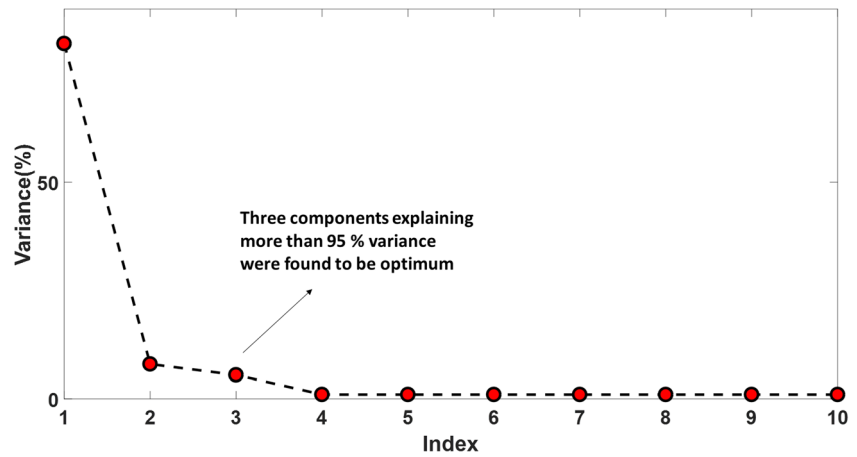


Fig. 3 The amount of variance explained by each component against index plot. It suggest that PcoA model developed with first three components must be sufficient to capture the differences in the analyzed gasoline samples



other. The steps 2–5 dealing with PcoA can be easily summarized using Eqs. 1–4.

$$D = \begin{bmatrix} d_{11} & \cdots & d_{1j} \\ \vdots & \ddots & \vdots \\ d_{i1} & \cdots & d_{ij} \end{bmatrix} \quad (1)$$

where d_{ij} is the element of matrix D, describing the dissimilarity value between the i^{th} and j^{th} samples or in other words the i^{th} and j^{th} row of the matrix X. The d_{ij} values are calculated using a suitable distance matrix approach.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{bmatrix} \quad (2)$$

where $a_{ij} (= -0.5 d_{ij}^2)$ is the element of matrix A, describing the square proximity between the i^{th} and j^{th} samples

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1j} \\ \vdots & \ddots & \vdots \\ y_{i1} & \cdots & y_{ij} \end{bmatrix} \quad (3)$$

where $y_{ij} (= a_{ij} - a_c - a_r - a_o)$ is the element of the matrix Gower centered matrix Y and a_c , a_r , and a_o are the column wise, row wise and overall mean of the matrix A. In Eqs. 1–3, $i \in [1, I]$

and $j \in [1, J]$.

$$Y = USV^T \quad (4)$$

where U, S, and V are the left singular matrix, diagonal matrix containing the singular values and right singular matrix, respectively.

Application of PcoA on EEMF Data Sets of Gasoline Samples

EEMF data sets for all the 65 samples belonging to four brands were arranged in a three-way array of dimension $176 \times 31 \times 65$ (emission wavelength \times excitation wavelength \times samples). In the next step, three way array was unfolded along the sample to generate two array of dimension 65×5456 (sample \times (emission wavelength \times excitation wavelength)). The unfolded-EEMF spectral profiles for each of the 65 gasoline samples are shown in Fig. 2.

In the next step, the distance matrix D of dimension 65×65 (sample \times sample) containing the dissimilarity values d_{ij} between the i^{th} and j^{th} sample was obtained using Eq. 1. Each element d_{ij} of matrix D was calculated using Chebyshev distance matrix approach [35] summarized using Eq. 5

Fig. 4 PcoA plot comprised over PcoA1, PcoA2 and PcoA3 was successfully found to classify the gasoline samples in four groups (Group 1, Group 2, Group 3, Group 4 and Group 5). The Group 1, Group 2, Group 3, Group 4 and Group 5 were found to contain the gasoline samples of *apecck*, *hypco*, *total*, *wardiyeh* and *medco* brands, respectively

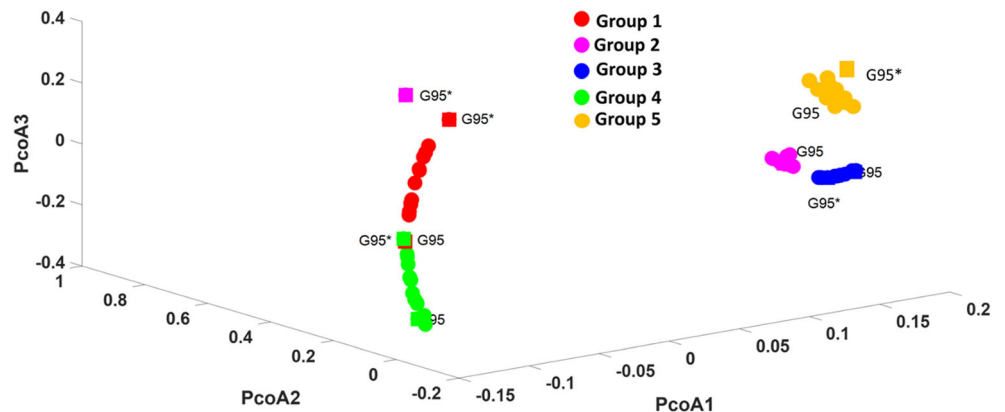
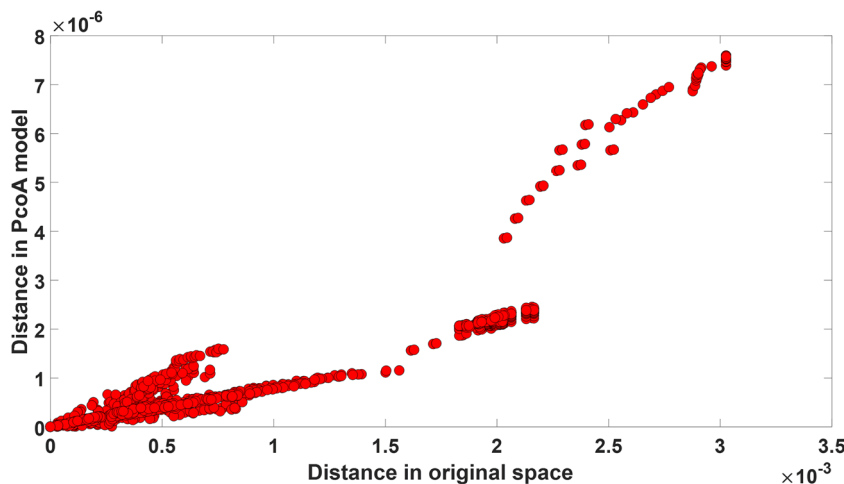


Fig. 5 The Shepard analysis clearly demonstrates the diagonal (i.e. linear) relationship between the distances among the samples in the original space and the distances among the samples in the PcoA model. It clearly shows that despite the compositional complexities of gasoline samples PcoA has successfully preserved the all the dissimilarity information of the original space



$$d_{ij} = \max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \dots, |x_{ik} - x_{jk}|) \quad (5)$$

where x_{ik} and x_{jk} are the k^{th} variable of the i^{th} and j^{th} sample (i.e. i^{th} and j^{th} row of matrix X). Chebyshev distance is also known as the maximum value distance; it essentially examines the absolute magnitude differences between all the variables for a given pair of samples. In the next step, the square proximity matrix A of dimension 65×65 was obtained using the Eq. 2. The square proximity matrix was further subjected to Gower centering to obtain double centered Y matrix of dimension 65×65 . The Gower centered matrix Y was further subjected to SVD analysis and the left singular matrix U , diagonal matrix D containing the singular values in the descending order and the right singular matrix V was calculated using Eq. 4. The optimization of the number of right singular vectors required to capture all the major variance of the data set is another important step. In the present work, it was achieved by plotting the variances associated with each singular vector against the indices. In the plot, the index beyond which the addition of new singular vector do not increase the variance appreciably is identified and it could further be taken as optimum for PcoA. The variance associated with a given singular vector can be calculated using the Eq. 6.

variance (%) associated with i^{th} singular vector

$$: \frac{S_{i,i}}{\sum_{i=1}^I S_{i,i}} \times 100 \quad (6)$$

where $S_{i,i}$ is the i^{th} diagonal element of the diagonal matrix S . The variance versus index plot, shown in Fig. 3, clearly suggested that first three singular vectors describing more than 95% variance is optimum for capturing major as well as subtle compositional differences among the analyzed gasoline samples.

The next step comprised the visualization of the dissimilarities in PcoA plot that was generated by plotting columns first three columns of the right singular matrix V . The PcoA plot, shown in Fig. 4, clearly indicated that analyzed gasoline samples mainly belong to the five groups. Group1, Group2, Group3, Group4 and Group5 were found to contain the gasoline samples of *apec*, *hypco*, *total*, *wardiyeh* and *medco* brands, respectively. It could also be seen within group variations for Group 1 and Group 4 are much higher than those observed in the case of Group 2, Group 3 and Group 5. It suggested that fluorescence intensities of the *apec* and *wardiyeh* brands are quite sensitive to the ethanol dilution. Nevertheless, PcoA assisted EEMF analysis clearly

Fig. 6 The calculated and measured ethanol (%) in different samples of Group1 (*apec*), Group 2 (*hypco*), Group 3 (*total*), Group 4 (*wardiyeh*) and Group 5 (*medco*) brands

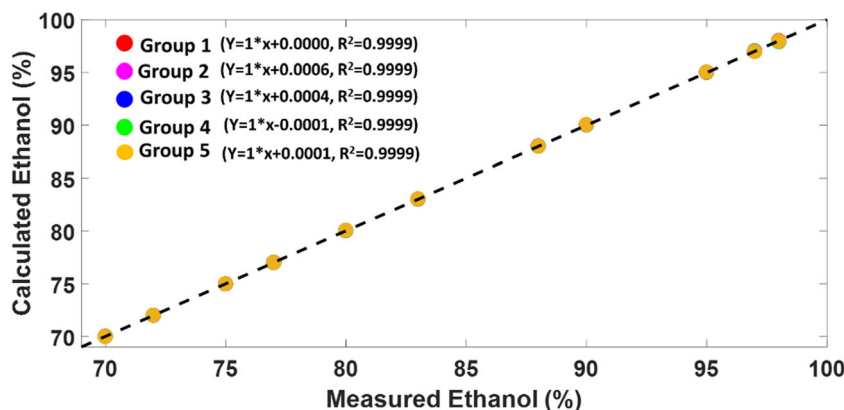
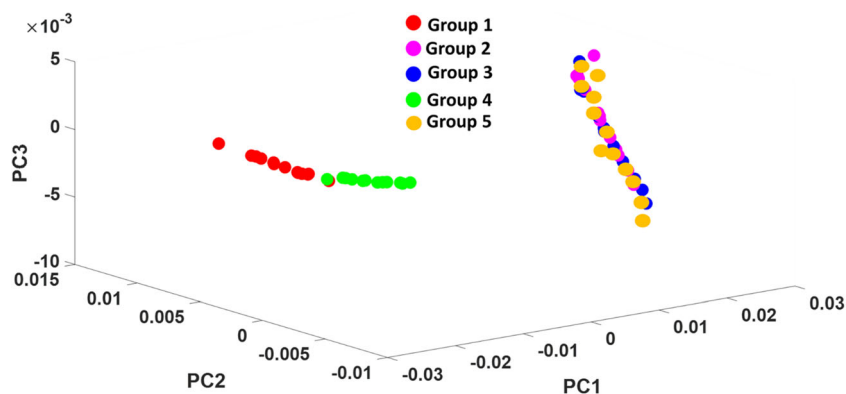


Fig. 7 PCA score plot comprised over first three components PC1, PC2 and PC3. PCA model clearly failed to discriminate the gasoline samples



discriminated the gasoline samples based on their brands. The G95 and G95* samples for each brand type were also clearly marked in PcoA plot. It showed that the proposed approach can also successfully discriminate if the samples (G95 and G95*) were prepared in the laboratory or comes directly from the refinery.

The next step involved the verification of the fact that whether PcoA has retained the original dissimilarities among the analyzed samples. It was verified by performing the Shepard analysis [27, 30] that involved plotting the original distance against the distances between the samples in PcoA model. The Shepard plot obtained for the present case, shown in Fig. 5, clearly indicated a diagonal (or linear) relationship between these two distances in different space despite the compositional complexities of the gasoline samples. The Shepard analysis verified that developed PcoA model has substantially simplified the data sets while preserving all relevant dissimilarity information.

Detection of Ethanol Content in Different Samples of *apeck*, *hypco*, *total*, *wardiyeh* and *medco* Brands

The unfolded-EEMF data sets for each of the five brands were arranged in five matrices of dimensions 12×5456 (sample \times (emission wavelength \times excitation wavelength)). The ethanol concentration (%) in different samples of each brands were arranged in a matrix of size 12×1 . It is to be noted that G95* sample was omitted from the analysis from each brands because it was found to have different fluorescence characteristics compared to those obtained from dilution of G98* sample. The unfolded-EEMF and concentration matrices for each brands were subjected to PLS analysis [33, 34] with three factors. The developed PLS models for each brand were found to explain more than 99.9% variances of spectral and concentration data matrices. The measured and calculated ethanol (%) were plotted against each other, shown in Fig. 6. The regression equations and square of the correlation coefficients (R^2) for each brands are also reported in the Fig. 6. The obtained results clearly showed that despite the differences in the dilution effect on fluorescence intensities as observed in each

gasoline brand, it is possible to detect and calibrate the ethanol concentration.

Comparing PCA with PcoA for the Analysis of EEMF Spectral Data of Gasoline Samples

To demsonstate the potential advanatges that one could achieve by caarying out PcoA, the obtained results were further compared with those obtained from unfolded-EEMF data sets of PCA. A three componenet PCA model was developed that was found to describe more than 95% variacnes of the spectral data set. The PCA score plot comprised of first three principal componenets PC1, PC2 and PC3 are shown in Fig. 7. It can be seen that developed PCA model successfully disriminated the gasoline samples of the *apeck* (Group 1) and *wardiyeh* (Group 4) brands whereas it completely failed to disriminate the gasoline samples of *hypco*, *total* and *medco* brands. One can clearly see that unlike PCA, the PcoA model can capture both major as well as subtle compositional differences among the analyzed gasoline samples belonging to different brands. The better discrimination achieved using PcoA can be attributed to the fact that it is based on the distance matrix containing the dissimilarity values and attempts to capture major as well as subtle compositional differences while retaining their true relationship in the original space. Whereas, PCA uses covariance matrix as the input and attempts to explain maximum variation of the data sets with as few components as possible.

Conclusions

In the present work, a simple, sensitive and cost-effective analytical method was proposed by combining PcoA with EEMF spectroscopy for discriminating the gasoline samples belonging to different industrial brands. PcoA approach involved the Eigen analysis of the distance matrix containing the dissimilarity information and therefore it can successfully capture even the subtle compositional differences. The utility of the proposed PcoA and EEMF based analytical approach

was successfully demonstrated by discriminating the gasoline samples belonging to five different brands. The PcoA was also found to perform better than PCA approach towards the analysis of gasoline samples.

Acknowledgments Financial support provided by Munib and Angela Masri Institute of Energy and Natural Resources Grant, American University of Beirut, Lebanon to carry out this work is greatly acknowledged.

References

- Lopez RR, Elizalde-Martinez I, Balderas-Tapia L (2010) Complete catalytic oxidation of methane over Pd/CeO₂-Al₂O₃: the influence of different ceria loading. *Catal Today* 150:358–336
- Scherzer J, Gruia AJ (1996) *Hydrocracking science and technology*. Marcel Dekker, Inc., New York
- Kumar K, Tarai M, Mishra AK (2017) Unconventional steady-state fluorescence spectroscopy as an analytical technique for analyses of complex-multifluorophoric mixtures. *Trends Anal Chem* 97:216–243
- IARC monographs Occupational exposures in petroleum refining; crude oil and major petroleum fuels, IARC Monogr Eval Carcinog Risks Hum. 45 (1989) 1–322
- Gooyal P (2003) Sidhartha, present scenario of air quality in Delhi: a case study of CNG implementation. *Atmos Environ* 37:5423–5431
- Ritter S (2005) Gasoline. *Sci Technol* 83:37
- Kumar K, Mishra AK (2012) Quantification of ethanol in ethanol-petrol and biodiesel in biodiesel-diesel blends using fluorescence spectroscopy and multivariate methods. *J Fluoresc* 22:339–347
- Lakowicz JR (2006) *Principles of fluorescence spectroscopy*. Springer, New York
- Valuer B (2001) *Molecular fluorescence: principles and applications*. Wiley-VCH Verlag GmbH, New York
- Freearde M, Hatchard CG, Parker CA (1971) Oil spilt at sea: its identification, determination, and ultimate fate. *Lab Pr* 20:35–40
- Warner IM, Callis JB, Davidson ER, Goutermann M, Christian GD (1975) Fluorescence analysis: a new approach. *Anal Lett* 8: 665–681
- Rho JH, Stuart JL (1978) Automated three-dimensional plotter for fluorescence measurements. *Anal Chem* 50:620–625
- Kumar K, Mishra AK (2013) Analysis of dilute aqueous multifluorophoric mixtures using excitation-emission matrix fluorescence (EEMF) and total synchronous fluorescence (TSF) spectroscopy: a comparative evaluation. *Talanta* 117:209–220
- Warner IM, Callis JB, Davidson ER, Christian GD (1976) Multicomponent analysis in clinical chemistry by use of rapid scanning fluorescence spectroscopy. *Clin Chem* 22:1483–1492
- Warner IM, Christian GD, Davidson ER, Callis JB (1977) Analysis of multicomponent fluorescence data. *Anal Chem* 49:564–573
- Warner IM, Davidson ER, Christian GD (1977) Quantitative analyses of multicomponent fluorescence data by the methods of least squares and non-negative least sum of errors. *Anal Chem* 49:2155–2159
- Coble PG (1996) Characterization of marine and terrestrial DOM in seawater using excitation emission matrix spectroscopy. *Mar Chem* 51:325–346
- Sierra MM, Giovanela M, Parlanti E, Soriano-Sierra EJ (2005) Fluorescence fingerprint of fulvic and humic acids from varied origins as viewed by single-scan and excitation/emission matrix techniques. *Chemosphere* 58:715–733
- Sheng GP, Yu HQ (2006) Characterization of extracellular polymeric substances of aerobic and anaerobic sludge using three-dimensional excitation and emission matrix fluorescence spectroscopy. *Water Res* 40:1233–1239
- Richards-Kortum R, Rava RP, Petras RE, Fitzmaurice M, Sivak M, Feld MS (1991) Spectroscopic diagnosis of colonic dysplasia. *Photochem Photobiol* 53:777–786
- M. Brewer, U. Utzinger, E. Silva, D. Gershenson, R.C. Bast, M. Follen, R. Richards-Kortum, Fluorescence spectroscopy for in vivo characterization of ovarian tissue., *Lasers Surg. Med.* 29 (2001)128–135
- Johnson DW, Callis JB, Christian GD (1977) Rapid scanning fluorescence spectroscopy. *Anal Chem* 49:747–757
- Wolfbeis OS, Leiner M (1985) Mapping of the total fluorescence of human blood serum as a new method for its characterization. *Anal Chim Acta* 167:203–215
- E. Sikorska, A. Romaniuk, I. V. Khmelinskii, R. Herance, J.L. Bourdelande, M. Sikorski, J.Kozioł (2004) Characterization of edible oils using total luminescence spectroscopy in: *J. Fluoresc*, pp. 25–35
- Kurdi RE, Kumar K, Patra D (2018) Random initialisation of the excitation-emission matrix fluorescence spectral variables in constraint fashion for subsequent multivariate curve resolution alternating least square analysis on a peculiarly designed calibration set: simultaneous sensing of nine polycyclic aromatic hydrocarbons in water samples. *Spectrochim Acta A* 204:354–361
- Young G, Householder AS (1938) Discussion of a set of points in terms of their mutual distances. *Psychometrika* 3:19–22
- Gower JC (1996) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53:325–338
- Krzanowski WJ, Marriott FHC (1994) *Multivariate analysis, part I: distributions*. Halstead Press London, Ordination and Inference
- Borg I, Groenen P (1997) *Modern multidimensional scaling: theory and applications*. Springer, New York
- Legendre P, Legendre L (1998) *Numerical ecology*, second edn. Elsevier, Amsterdam
- Kumar K, Cava F (2018) Principal coordinate analysis assisted chromatographic analysis of bacterial cell wall collection: a robust classification approach. *Anal Biochem* 550:8–14
- Kumar K (2017) Principal component analysis (PCA) the most favourite tool in chemometrics. *Reson* 8:747–759
- Brereton GR (2009) *Chemometrics for pattern recognition*. John Wiley & Sons, Chichester
- Kramer R (1998) *Chemometric techniques for quantitative analysis*. Marcel Dekker, New York
- Zadora G, Martyna A, Ramos D, Aitken C (2014) Statistical analysis in forensic science: evidential value of multivariate physico-chemical data, John Wiley & Sons, Chichester

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.