Evaluating the benefits of data fusion and PARAFAC for the chemometric analysis of FT-ICR MS datasets from gas oil samples

Julie Guillemant, Marion Lacoue-Nègre, Alexandra Berlioz-Barbier, Luis P de Oliveira, Jean-François Joly, Ludovic Duponchel

Evaluating the benefits of data fusion and PARAFAC for the chemometric analysis of FT-ICR MS datasets from gas oil samples

Julie Guillemant†, Marion Lacoue-Nègre†, Alexandra Berlioz-Barbier†, Luis P. de Oliveira†, Florian Albrieux†, Jean-François Joly†, and Ludovic Duponchel*‡*

† IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize, France

‡ Univ. Lille, CNRS, UMR 8516 - LASIRE – Laboratoire avancé de spectroscopie pour les interactions, la réactivité et l'environnement, F-59000 Lille, France

Gas oils; hydrotreatment; data fusion; PARAFAC; multivariate analysis; mass spectrometry; chemometric; FT-ICR MS.

Abstract

Advanced characterization of the products of gas oils hydrotreatment is of high interest for the refiners and can be achieved by using ultra-high resolution mass spectrometry (FT-ICR MS). However, the analysis of gas oil samples by FT-ICR MS generates complex datasets with numerous variables whose exhaustive analysis requires the use of multivariate methods. Relevant information about nitrogen and sulfur compounds contained in several industrial gas oils are obtained by using three different ionization modes that are electrospray ionization (ESI) used in positive and negative polarities and atmospheric pressure photo-ionization (APPI) used in positive polarity. For datasets generated for a single ionization

mode, classical multivariate methods such as Principal Component Analysis (PCA) are commonly used. When the key information is spread into several ionization modes and thus into several datasets, a data fusion approach is highly interesting to simultaneously explore these datasets and can be followed by Parallel Factor analysis (PARAFAC). Nevertheless, many more variables are simultaneously considered when data fusion is performed and the sensitivity of PARAFAC and its ability to extract the most relevant variables compared to classical multivariate methods has not been assessed yet in the framework of FT-ICR MS. In this paper, the comparison of the classical data analysis (PCA) approach and the data fusion combined with the PARAFAC analysis approach is presented. The results have shown that applying PARAFAC on fused datasets is highly sensitive and able to put forward features and variables that individually identified through the classical data analysis with greater ease of implementation and interpretation of results. As an example, dibenzothiophenes and carbazole families (DBE 9) have explained most of the variance between samples and remain the most refractory compounds in hydrotreated samples. A significant difference in alkylation between the different types of gas oils has also been spotted. This paper validates the power and efficiency of this approach to explore complex datasets simultaneously without any loss of significant information.

Introduction

The improvement of the hydrotreatment (HDT) process is of major importance as the environmental specifications for sulfur content in commercial on-road diesels are getting more and more severe[1]. This process allows reducing the sulfur content to a very low concentration (below 10 ppm) to limit sulfur emissions and thus respect legal specifications. In gas oils, sulfur is found in different organic structures: thiophenic compounds, alkyl-benzothiophenes, and alkyl-dibenzothiophenes. The repartition of the sulfur structures in gas oils can be very different depending on their origins. Indeed, gas oil cuts found in refineries can be obtained from distillation of the crude oils (Straight Run gas oils called

SRGO) or can be produced from several conversion processes such as fluid catalytic cracking (FCC, producing gas oil cuts called LCO), coking (producing gas oil cuts called GOCK) and catalytic hydroconversion (producing gas oil cuts called FBGO with fixed-bed technology or EBGO samples with ebullating-bed technology).

The efficiency of the HDT process is reliant on the molecular composition because the kinetic of the hydrodesulfurization (HDS) reactions depends strongly on the structure of the sulfur compounds in the gas oil cuts[2]. The efficiency of hydrodesulfurization is also dependent on nitrogen compounds. Indeed, basic nitrogen compounds are deactivating the acidic catalysts used while the neutral nitrogen compounds are refractory and compete with sulfur compounds reducing the overall sulfur removal[3–6]. Thus, hydrodenitrogenation (HDN) is also targeted during hydrotreatment.

Molecular-based kinetic modeling approaches are increasingly used[7] to predict the reactivity of the different gas oil cuts and to optimize the operating conditions of the HDT process. There is then a real need for molecular characterization of these different gas oils[2]. The detailed characterization of sulfur compounds in gas oils can be achieved through atmospheric pressure photo-ionization (APPI) coupled to ultra-high resolution mass spectrometry (Fourier Transform Ion Cyclotron Resonance Mass Spectrometry known as FT-ICR MS) analysis[8–11] while nitrogen compounds are described using the electrospray ionization (ESI) source in two different polarities (negative polarity for neutral nitrogen compounds and positive polarity for basic nitrogen compounds)[12,13]. As FT-ICR MS generates big datasets, multivariate analysis is more and more used to fully explore the resulting datasets[14–16]. Generally, principal component analysis (PCA) is applied on single datasets obtained for a given ionization mode or on several datasets with variables being the molecular formula of the compounds and their corresponding abundances or relative intensities[17,18]. This methodology does not allow to directly assess the contribution from the aromaticity and the alkylation degrees throughout the loadings

analysis and thus the direct identification of the sulfur and basic/neutral nitrogen compounds explaining the observed variance between samples. Recently, a new approach has been introduced by merging several datasets obtained from different ionization modes and applying the PARAFAC method on this multi-dimensional hypercube[19]. This new approach seems promising to extract the key features (DBE, nC) of the different samples[20–22]. However, the exhaustivity and sensibility of this method have not been compared to classical multivariate methods when considering a large number of variables. In this paper, two different approaches have been studied to evaluate the benefits of the data fusion and the PARAFAC approach over the classical chemometric analysis (PCA) using the same variables that are the DBE (Double Bond Equivalent) and the number of carbon atoms. The classical data analysis has been first performed on single datasets to provide an exhaustive chemometric analysis of these datasets. Then, datasets have been merged and the PARAFAC method was applied to compare information extracted from the single PCA analysis and the fused PARAFAC analysis. Besides, the evolution of the nitrogen and sulfur species over hydrotreatment has been followed using these approaches and put forward some refractory species such as carbazoles or dibenzothiophenes.

Material and Methods

Gas oil samples. 23 different gas oils with various industrial origins were analyzed in this study: 5 SRGO, 3 LCO, 4 GOCK, 1 EBGO, 1 FBGO, 5 HDT and 4 blends (MIX). The macroscopic properties of these samples are shown in Table 1. The preparation and ionization conditions were optimized with a Design of Experiments (DoE) approach and detailed in a previous work[13].

FT-ICR MS analysis. Each gas oil sample was analyzed using ESI(+), ESI(-) and APPI(+)-FT-ICR MS considering 6 technical replicates. Mass spectrometry (MS) analyses were performed using a LTQ FT Ultra Fourier Transform Ion Cyclotron Resonance Mass Spectrometer (FT-ICR MS) (ThermoFisher Scientific, Bremen Germany) equipped with a 7T magnet (Oxford Instruments) and an ESI source (ThermoFisher Scientific) used in positive and negative modes and an APPI source (Syagen

Technology, Tustin CA, USA) used in positive mode. The mass range was set to *m/z* 98-1000. 70 scans with 4 µ-scans were recorded with an initial resolution set to 200,000 (transient length of 1.6s) at *m/z* 300 (center of average gas oil mass distribution). The transient signal was recorded to enable further

Table 1. Macroscopic properties of gas oil samples used in this study. The ASTM standard methods used for analysis are mentioned for each property.

Sample	Type (*)	Total sulfur (ppm) *Ref. method: ASTM D2622*	Total nitrogen (ppm) *Ref. method: ASTM D4629*	Basic nitrogen (ppm) *Ref. method: ASTM D2896*	Density at 15°C (g/cm3) *Ref. method: ASTM D4052*	Boiling point range (°C) *Ref. method: ASTM D86*
GO 1	SRGO	13555	115	47	0.8541	219-386
GO 2	SRGO	7044	254	100	0.8665	258-396
GO 3	SRGO	10979	350	129	0.8878	244-396
GO 4	SRGO	8892	114	42	0.8484	221-381
GO 5	SRGO	4189	96	48	0.8491	186-392
GO 6	LCO	9496	928	91	0.9130	199-386
GO 7	LCO	11074	1170	49	0.9413	248-390
GO 8	LCO	2231	496	141	0.9035	166-304
GO 9	GOCK	14796	893	404	0.8501	148-358
GO 10	GOCK	12723	838	390	0.8581	163-371
GO 11	GOCK	15314	1200	449	0.8640	173-375
GO 12	GOCK	24270	1260	569	0.8813	188-401
GO 13	EBGO	1248	1719	855	0.8712	199-429
GO 14	FBGO	344	195	121	0.8522	180-359
GO 15	MIX (65% GO 5 + 35% GO 6)	6400	380	63	0.8708	189-391
GO 16	MIX (67% GO 9 + 33% LCO)	14004	988	436	0.8576	151-351
GO 17	HDT from GO 16	190	93	14	0.8585	184-383
GO 18	HDT from GO 16	261	140	23	0.8591	187-386
GO 19	MIX (55% GO 5 + 30% GO 7 + 15% GO 11)	14162	586	122	0.8828	218-390
GO 20	HDT 3 from GO 19	626	205	38	0.8617	209-387
GO 21	HDT 2 from GO 19	2813	464	107	0.8678	211-388
GO 22	HDT 1 from GO 19	3656	723	330	0.8691	210-389
GO 23	MIX (50% LCO+50% LCO)	9125	925	98	0.9310	206-368

(*): SRGO = Straight Run Gas Oil; LCO = Light Cycle Oil; GOCK = Coker Gas Oil; EBGO = Gas Oil from Ebullating Bed reactor; FBGO = Gas Oil from Fixed Bed reactor; MIX = blended Gas Oil, HDT = Hydrotreated.

data processing. The ionization and ion transfer conditions for each ionization mode are available in Table 2. External mass calibration was performed using a home-made sodium formiate clusters solution (sodium formiate from VWR, Fontenay-sous-Bois, France) from about 90 Da to 1000 Da.

Table 2. Ionization and ion transfer conditions for each ionization mode. Tol=Toluene, MeOH=Methanol, AA=Acetic Acid and AmHy = Ammonium Hydroxide. (-) indicates the parameter is not considered..

Parameter	ESI(+)	ESI(-)	APPI(+)
% dilution	1	0.5	1
% solvents mix	25%-75% Tol-MeOH	25%-75% Tol-MeOH	75%-25% Tol-MeOH
% additive	0.05% AA	0.15% AmHy	-
Spray voltage (kV)	3.7	3.5	-
Tube lens (V)	110	-140	70
Capillary voltage (V)	50	-50	30
Capillary temperature (°C)	275	275	275
Vaporization temperature (°C)	-	-	250
Sheath gas (a.u)	-	5	20
Auxiliary gas (a.u)	-	-	5
Flow rate (µL/min)	5	5	10

Spectral data processing. The full data processing has been described elsewhere[13]. Briefly, the 70 transients have been summed and the resulting summed transient has been used to perform phase absorption and phase correction in order to obtain the absorption mode spectra for enhanced resolution and mass accuracy. After noise thresholding and peak picking, the obtained mass spectrum has been processed by a home-made software to perform molecular formula assignment using the following conditions $C_{0-50}H_{0-100}O_{0-3}N_{0-3}S_{0-3}$ with maximum content of heteroatoms in one molecular formula set to 3 and 5 ppm of maximum mass error as a first round. For clearer identification of the families identified in APPI(+) mode, the radical ions are identified as X families and protonated or deprotonated ions are identified as X[H] families. The mass spectrum has been then recalibrated using iterative mass recalibration considering the most intense family in each ionization mode with a maximum mass error set to 1 ppm for the second round of assignment. To compare the samples between one another, the relative intensities of the N1[H] or S1 compounds have been calculated and are equal to the peak intensity divided by the sum of intensities from all N1[H] or S1 peaks. Finally, the nitrogen and sulfur families have been attributed according to DBE values.

Chemometric analysis. This study is focused on the statistical analysis of neutral nitrogen, basic nitrogen, and sulfur compounds. These three families of interest are respectively identified using ESI(-) mode in the N1[H] class, using ESI(+) mode in the N1[H] class and using APPI(+) mode in the S1 class. For each replicate and a given ionization mode, a DBE as a function of carbon number plot (DBE=f(#C)) has been generated considering relative intensities of the peaks. Then, all the DBE=f(#C) plots have been concatenated to obtain 3D arrays of size 50x25x138 for each ionization mode where 50 corresponds to the range of the number of carbon atoms, 25 to the DBE range and 138 to the number of acquired MS spectra (23 samples times 6 replicates). Two strategies have then been followed: 1. The generated 3D cubes have been unfolded into 2D arrays of size 138x1250 (1250 corresponding to given DBE/number of carbon atoms pairs) for each ionization mode. Then a Principal Component Analysis (PCA) has been applied on each array[17]. 2. A low-level data fusion method has been applied to concatenate the previously obtained 3D cubes along the DBE axis obtaining one hypercube of size 50x75x138 containing the information of the three ionization modes. In this specific case, the PARAFAC method has been applied using the Alternating Least Squares algorithm (ALS)[19]. These two chemometric strategies are presented in Figure 1. Further information about both statistical approaches and data fusion methods can be found elsewhere[18,20,23].

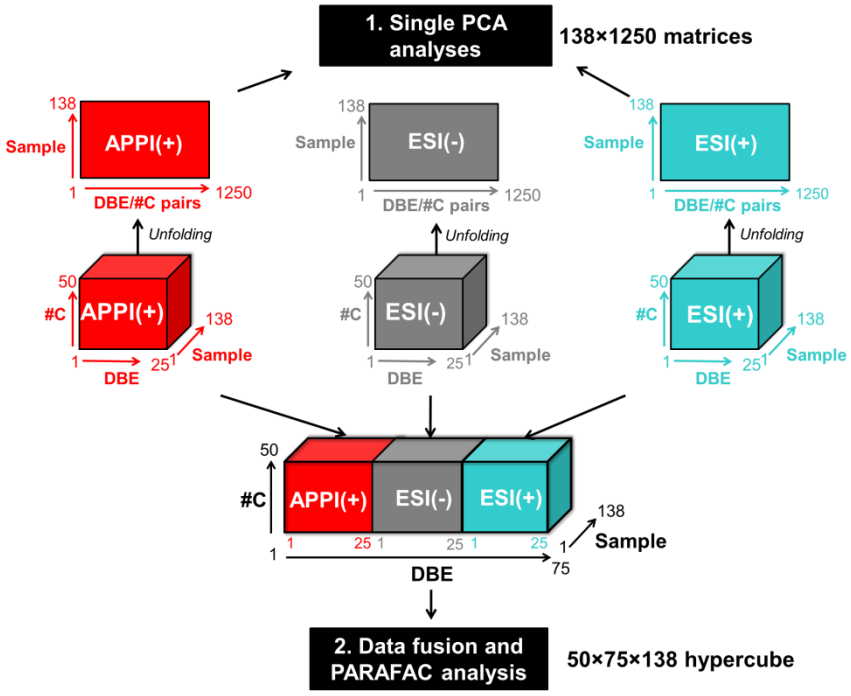

Figure 1. Chemometric strategies applied in the present work

All these models have been developed with the PLS_Toolbox version 8.6 (Eigenvector Research Inc, Wenatchee, WA, USA) for Matlab version R2018b (The Mathworks, Inc, Natick, MA, USA). For PCA analysis, the mixed blends samples (GO 15, 16, 19 and 23, see Table 1) were used for validation and all other samples were used for PCA calculation (calibration base). The matrices have been mean-centered before multivariate analysis. The mixed samples (GO 15, 16, 19 and 23) were also used for validation for the PARAFAC analysis which was performed with non-negativity constraints and the 3D matrices were block-scaled before data fusion.

Results and Discussion

1. PCA on individual datasets

Basic nitrogen compounds. PCA analysis was applied to the matrix obtained from the analysis of N1[H] compounds in ESI(+) mode. Four principal components were selected and studied expressing up to 93% of the total variance. The gas oil 8 (LCO type) was considered as an outlier (large Hotelling's t-

squared statistic T^2 and was discarded from the data set. The score plot obtained along the two first principal components (i.e. PC1-PC2) is shown in Figure 2A.

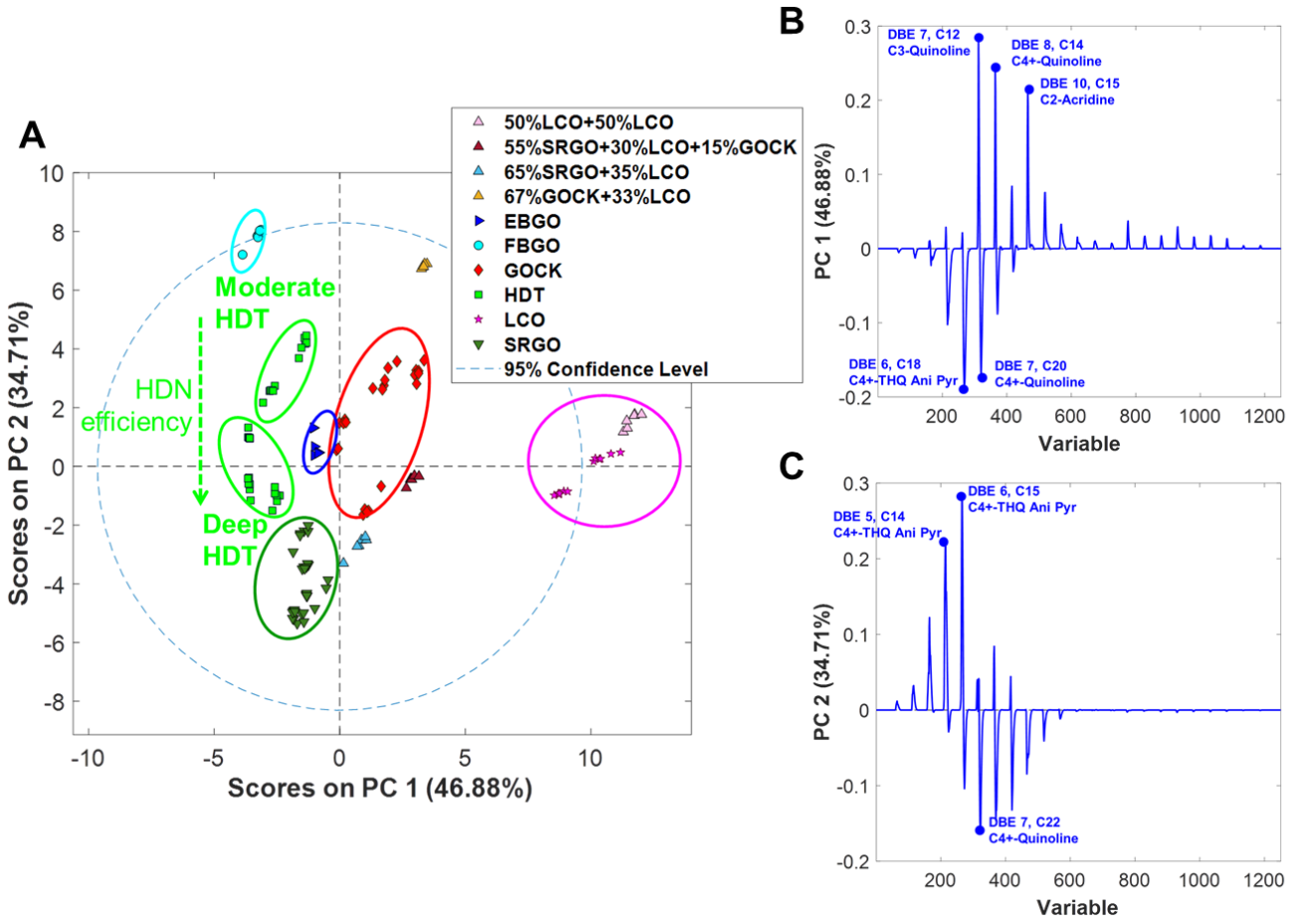

Figure 2. (A) Score plot along PC1 and PC2 obtained from ESI(+)-FT-ICR MS data for N1[H] class. (B) Loadings plots of PC1 and (C) PC2 for ESI(+)-FT-ICR MS data. THQ Ani Pyr = TetraHydroQuinolines Anilines Pyridines family

This projection allows assessing graphically the repeatability of the different analysis which is here satisfying as the replicates from each gas oil are projected close to one another. Moreover, some clusters are observed according to the type of gas oil considered which is related to the specificities of each process. From a general point of view, most samples are negatively or poorly projected over PC1 while the LCO samples are positively projected. As a matter of fact, the relative intensities of the different basic nitrogen families presented in Figure 3A are quite similar for each type of gas oil except for the sample GO 7 (LCO) as it has a very low content in poorly aromatic compounds (DBE < 6,

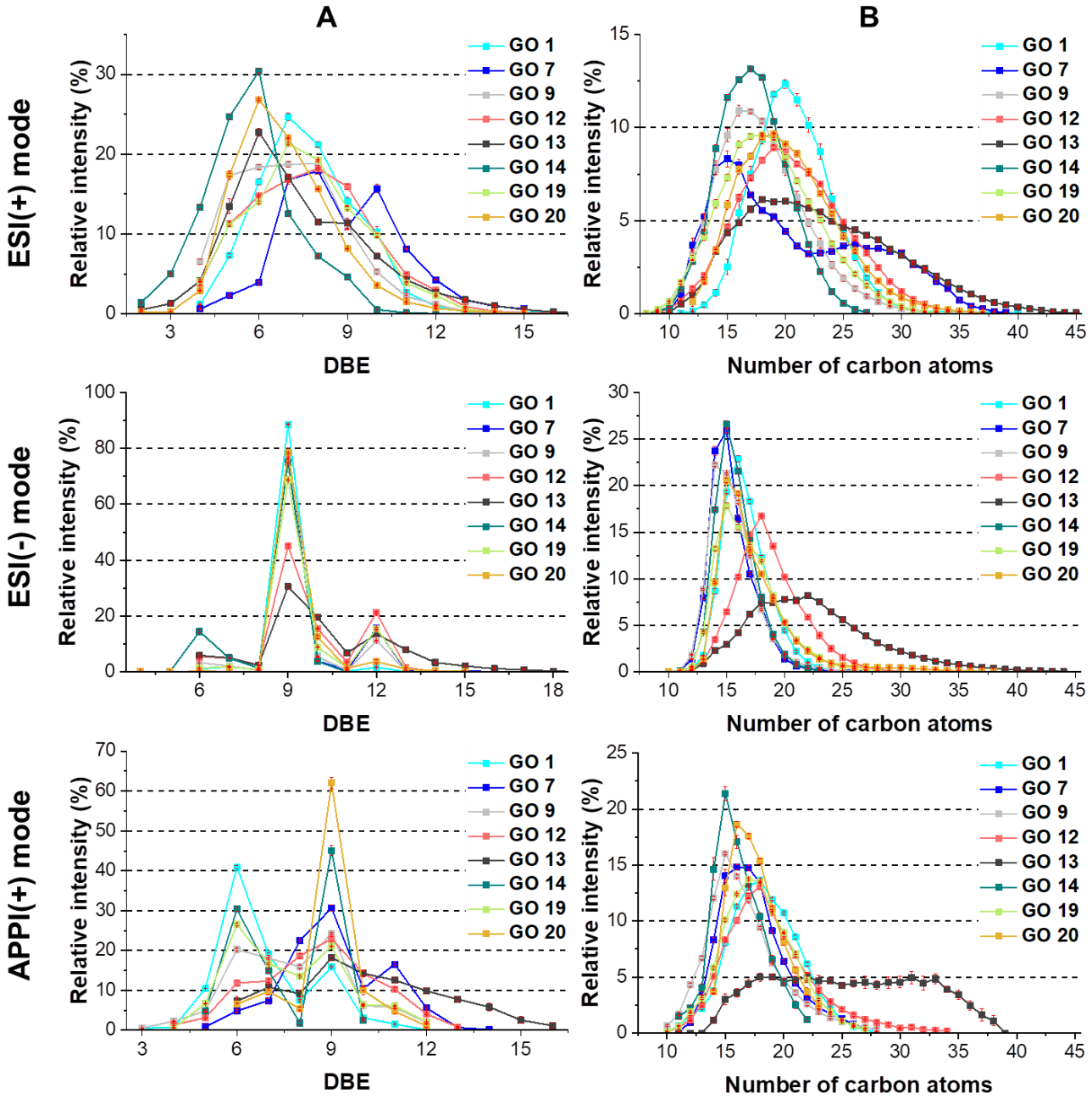

Figure 3. (A) Evolution of the relative intensities as a function of the DBE for different gas oil samples for the three ionization modes. (B) Evolution of the relative intensities as a function of the number of carbon atoms for the three ionization modes. The standard deviation bars calculated from the 6 replicates are indicated in red. The graphics obtained as a function of the global identified families are available in Figure S2 in Supporting Information. THQ Ani Pyr = THQ Anilines Pyridines, BT = Benzothiophenes, DBT = Dibenzothiophenes and NBT = Naphtobenzothiophenes. GO 1 = SRGO, GO 7 = LCO, GO 9 = GOCK, GO 12 = GOCK, GO 13 = EBGO, GO 14 = FBGO, GO 19 = MIX, and GO 20 = HDT.

TetraHydroQuinolines [THQ] Anilines Pyridines) and high content in very aromatic compounds (DBE > 10, Acridines). Besides, the global carbon atoms distribution of the sample GO 7 (LCO) shown in Figure 3B exhibits a bimodal distribution with maximum intensities centered over C15 and C28 while

only Gaussian distributions are observed for the other samples with an average alkylation level lower than C28. In Figure 2A, the variables having positive contributions along PC1 are poorly alkylated aromatic quinolines (DBE 7, C12 and DBE 8, C14) and poorly alkylated acridines (DBE 10, C15) while the variables having negative contributions correspond to poorly aromatic compounds but rather alkylated (DBE 6, C18 and DBE 7, C20). Thus, the projection of the samples over PC1 is consistent regarding the projection of the LCO samples that contain more aromatic compounds that are poorly alkylated.

PC2 mainly discriminates the moderate HDT samples including the sample GO 14 (FBGO) having positive scores and the SRGO samples having negatives ones. The variables having positive contributions in the loadings (Figure 2C) are poorly aromatic compounds (DBE 5, C14 and DBE 6, C15). The variables having negative contributions are more aromatic and more alkylated (DBE 7, C22). Thus, the observed projection of the samples is again consistent with the variables identified. The sample GO 14 (FBGO) is obtained through a relatively severe hydroconversion process (Fixed-bed hydroconversion) leading to quite light and poorly aromatic gas oils[24] and the ions from the THQ Anilines Pyridines (DBE 4-5-6) family represents up to 80% of all ions identified (see Figure 3A). On the other side, SRGO samples contain compounds that are a little bit more aromatic and much more alkylated such as quinolines family up to 50% with a maximum intensity over C20 (see Figures 3A and 3B).

Another interesting point from the PC1-PC2 score plot is the projection of the hydrotreated samples obtained from the sample GO 19. Indeed, two clusters are observed depending on the hydrotreatment level of the samples so-called Moderate HDT (> 460 ppm of total nitrogen) or Deep HDT (< 210 ppm of total nitrogen). Moreover, the increasing efficiency of hydrodenitrogenation (HDN) is translated into negative scores along PC2. Thus, the moderately hydrotreated samples would contain more poorly aromatic compounds such as THQ Anilines Pyridines whereas the deeply hydrotreated samples would

contain more aromatic compounds such as Quinolines. These compounds would be less efficiently removed throughout hydrotreatment than THQ Anilines Pyridines. This is observed in Figure S1 (Supporting Information) for the hydrotreated samples obtained from the sample GO 19. The THQ Anilines Pyridines family is more intense for the moderate HDT samples (samples GO 21 and 22 that are HDT 1 and HDT 2) than for the deep HDT sample (sample GO 20 that is HDT 3) while the relative intensities of the quinolines family are less intense for the moderate HDT samples than for the deep HDT sample. Moreover, these relative intensities evolutions seem to be related to the amount of total nitrogen in the sample and thus to the overall HDN efficiency. On the other side, the relative intensities of the acridines family are steady for every hydrotreated sample whichever operating conditions considered meaning that these compounds might be very refractory.

The gas oils obtained from mixed blends are used for validation purpose and their projection allow us to conclude on the additivity of the analysis as well as the efficiency of the chemometric model for the exploration of complex matrices. As an example, the mixed gas oils SRGO/LCO, SRGO/LCO/GOCK and LCO/LCO are correctly projected with respect to their original compositions. However, the mixed blend GOCK/LCO projection is quite surprising. The LCO used to produce the blend was not available for FT-ICR MS analysis so it might be atypical and the inconsistent projection of the GOCK/LCO mixed sample could be related unexpectedly to a high amount of THQ Anilines Pyridines family as it is largely positively projected over PC2.

The information contained in the PC3-PC4 score plot is quite similar to those extracted from the analysis of the PC1-PC2 score plot so it is not exhaustively discussed in this paper (see Figure S3 in Supporting Information). Briefly, the sample GO 13 (EBGO) is put forward over both PC3 and PC4 and the loadings are related to very alkylated compounds which are consistent with its composition (Figure 3B).

Neutral nitrogen compounds. PCA was applied to the matrix containing the neutral nitrogen compounds identified in the N1[H] class. Four principal components were again selected representing up to 97% of the total variance. The score plot obtained along PC1-PC2 is shown in Figure 4A. For the same reasons as before, the sample GO 8 (LCO) was excluded from the data set.

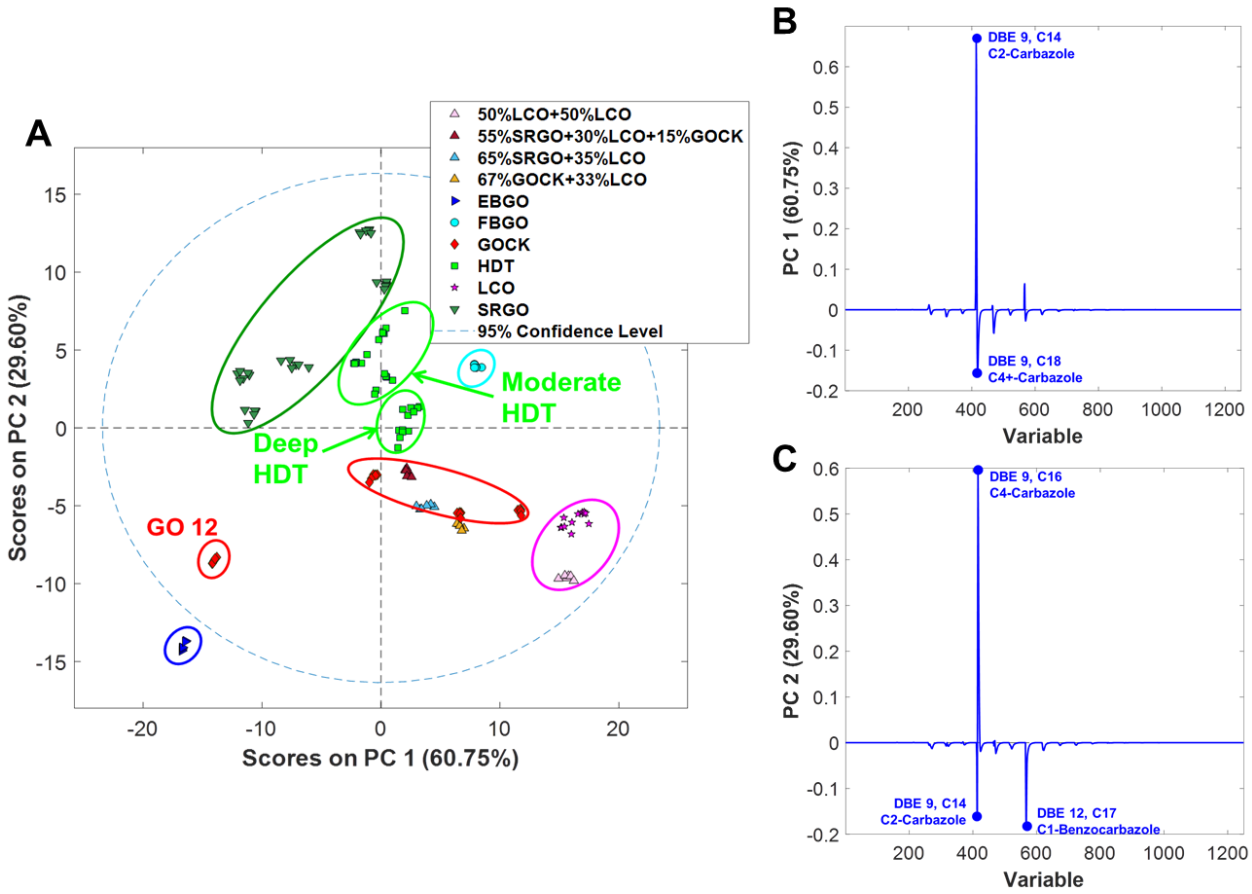

Figure 3. (A) Score plot along PC1 and PC2 obtained from ESI(-)-FT-ICR MS data for N1[H] class. (B) Loadings plots from PC1 and (C) PC2 from ESI(-)-FT-ICR MS data.

Again, good repeatability is observed regarding the projection of replicates for a given gas oil sample. Thus it indicates that ESI(-) mode also provides repeatable measurements.

Some clusters are also observed according to the type of gas oil considered. Globally, PC1 reflects the variance between the LCO samples that are positively projected over PC1 and the samples GO 12 (GOCK) and GO 13 (EBGO) that are negatively projected over PC2. It can be noted that the sample GO 12 (GOCK) is projected quite far away from the other GOCK samples over PC1 which could be

characteristic of unique features. The variables responsible for the projections over PC1 are shown in Figure 4B. It is worth noticing that only 2 types of molecules express more than 60 % of the total variance between the samples. These molecules correspond to compounds with DBE equal to 9 and with a number of carbon atoms respectively equal to 14 and 18 that we assumed to correspond to carbazoles family. In particular, the variable C2-Carbazole (classical carbazoles with two additional carbon atoms) expresses most of the PC1 variance and explains the positive projection of the LCO samples and most of GOCK ones and the negative projection of the EBGO sample and GO 12 (GOCK) over PC1. Thus, the main difference between these samples is based on the amount of C2-Carbazoles which is lower in the samples EBGO and GOCK 12 as they are much more alkylated and contain more C4+-Carbazoles (general denomination for classical carbazoles with at least four additional carbon atoms, here it is 6 carbon atoms) as plotted in Figure 3B – ESI(-) mode. The carbon atoms distribution of the sample GO 7 (LCO) is very centered around C14-15 which explains the strong positive contribution of the C2-Carbazole variable. The overall carbon atoms distribution of the sample GO 12 (GOCK) is shifted to a higher number of carbon atoms with maximum intensity for C18 compared to other GOCK samples such as sample GO 9 (GOCK) whose maximum intensity is observed at C14. This is also related to a higher amount of more aromatic compounds (i.e benzocarbazoles) that contain more carbon atoms and thus increases the overall carbon atoms distribution. Both higher alkylation and aromaticity degrees explain its atypical projection. As regards the sample GO 13 (EBGO), its distribution shown in Figure 3B is extremely shifted to a higher number of carbon atoms with a maximum intensity around C18-C23 due to its larger boiling point range (see Table 1) explaining its large negative projection along PC1.

Along PC2, the gas oils obtained from conversion processes with high content in neutral nitrogen (see Table 1) such as the EBGO, GOCK, and LCO samples are negatively projected while gas oils obtained from hydrotreament processes or crude distillation with lower nitrogen contents such as the SRGO, HDT, and FBGO samples are positively projected. Three molecules mainly explain PC2 variance: C2-

Carbazole, C4-Carbazole, and C1-Benzocarbazole as seen in Figure 4C. The distribution of the sample GO 1 (SRGO) is very centered over C16 (C4-Carbazole) which explains the strong positive contribution of this variable while the distributions of the samples GO 7 (LCO) and GO 9 (GOCK) are more centered over C14 (C2-Carbazole) explaining the negative contribution of this variable (see Figure 3B). A higher amount of benzocarbazoles in the samples GO 7 (LCO), GO 12 (GOCK) and GO 13 (EBGO) explains the contribution of the C1-Benzocarbazole as seen in Figure 3A.

Speciation according to the hydrotreatment level is observed in the score plot in Figure 4A. The variance between the hydrotreated samples is only expressed by PC2 through a negative translation and thus by a slightly higher contribution of C2-Carbazole compared to C4-Carbazole. Thus, the deep HDT samples are less alkylated than the moderate HDT samples indicating that the present deep hydrotreatment operating conditions might enhance the hydrogenation of heavy species such as C4-Carbazole rather than light species such as C2-Carbazole. The sample GO 14 (FBGO) is projected close to the hydrotreated samples due to the severe hydrotreatment occurring during the hydroconversion process.

This time, all the mixed blends are correctly projected including the GOCK/LCO mixed blend that was not correctly projected throughout the analysis of the ESI(+)-FT-ICR/MS dataset (see previous explanation).

The information obtained from the PC3-PC4 score plot and its corresponding loadings plots available in Supporting Information in Figure S4 are quite redundant and mostly express the unique features of the samples EBGO and FBGO due to their respective heavy and light character (regarding both aromaticity and alkylation).

Sulfur compounds. The results obtained considering APPI(+)-FT-ICR/MS data have been exhaustively discussed in a previous work[17]. 6 principal components have been considered as significant and explain

96.7% of the total variance. The score plot obtained over PC1-PC2 is available in Figure S5 in Supporting Information. Briefly, the samples are correctly clustered according to their process origins. The samples SRGO and HDT are separated over PC1 with the deep HDT samples being positively projected and the SRGO samples being negatively projected. The PC1 loading is shown in Figure 5B and reveals the strong positive contribution of the poorly alkylated dibenzothiophenes (DBE 9, C15) and the small contribution of alkylated benzothiophenes (DBE 6, C20). As a consequence, the negative projection of the SRGO samples over PC1 is explained by major content in alkylated benzothiophenes as seen in Figure 3A (in APPI(+) mode). On the opposite, the positive projection of the hydrotreated samples is related to their content in C3-dibenzothiophenes which is more intense in the deep HDT samples (S < 700 ppm) than moderate HDT samples (S > 2800 ppm). Dibenzothiophenes are identified in high contents in the sample HDT 20 as seen in Figure 3A which confirms its refractory character. On the opposite, moderate HDT samples contain more benzothiophenes compounds that are more easily converted than dibenzothiophenes and thus are poorly contributing to the variance of the deep HDT samples.

Over PC2, the HDT, SRGO or FBGO samples are positively projected while the EBGO, GOCK or LCO samples are negatively projected. The PC2 loading (Figure 5B) show a positive contribution for alkylated benzothiophenes (DBE 6, C18) explaining the SRGO samples projection while more aromatic and moderately alkylated molecules are negatively contributing and are related to the negative projection of the EBGO, LCO and GOCK samples over PC2. Then, the HDT, SRGO and FBGO samples contain poorly aromatic and very alkylated compounds while the EBGO, GOCK and LCO samples contain more aromatic compounds that are less alkylated. This is consistent with the compositions of the gas oils SRGO 1, LCO 7, GOCK 9 and 12, EBGO 13 and HDT 20 in terms of aromaticity and alkylation that are respectively visible in Figures 3A and 3B.

Finally, the samples used for validation (mixed blends) are also well clustered and well projected in the considered PCA model.

The score plots obtained over PC3 and PC4 and the corresponding loadings plots are available in Figure S5 in Supporting Information. They highlight the variance of the samples FBGO 14 and LCO 8 which is strongly influenced by their poorly aromatic (C4-BT) and poorly alkylated (C2-DBT) character.

Data fusion. Finally, the three datasets have been concatenated into a single hypercube using a methodology described elsewhere[19]. To explore this multi-dimensional matrix, the PARAFAC method has been used with three main modes: DBE, the number of carbon atoms and the samples. When optimizing the model, we tried several models with several number of components and we observed the evolution of the core consistency and % explained variable as a function of the number of components. The optimal model was the one corresponding to the highest core consistency and % explained variables observed corresponding to a given number of PARAFAC components. The models obtained with a higher number of components had core consistency values close to 0 which correspond to a very poor decomposition of the data by the PARAFAC algorithm. As a consequence, two PARAFAC components have been selected to decompose data with a core consistency equal to 94 (100 being the maximum value) explaining 82% of the total variance. The GO 8 (LCO) was excluded from both ESI(+) and ESI(-) data sets due to its large Hotelling's t-squared statistic T^2 but was considered in the APPI(+) one. When considering fused data, this gas oil still behaves as an outlier as nitrogen compounds are considered. This indicates that thcis approach is still able to retain the intrinsic characteristics of the samples even when more variables are considered.

The projections of the samples over these two components in two-dimensions and one-dimension representations are shown in Figure 6.

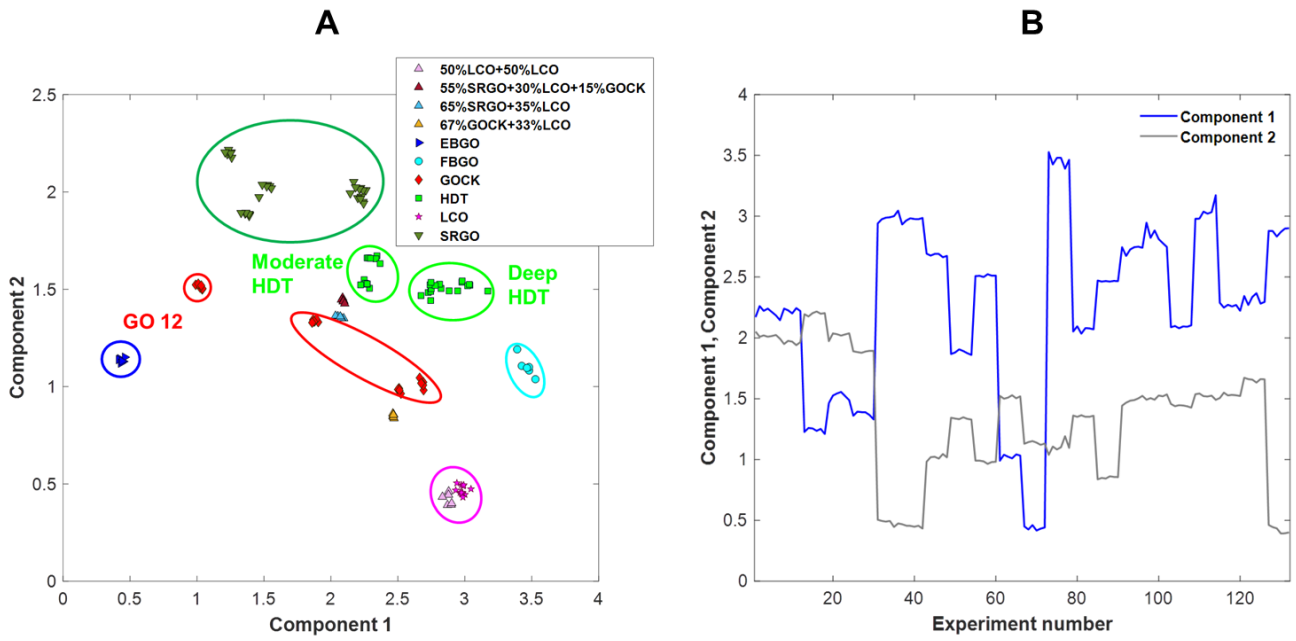

Figure 6. (A) Score plot along the two first components obtained after data fusion. (B) 1D representation of the scores of all samples over both components (Experiment numbers from 1 to 6 correspond to the 6 replicates of sample GO 1, from 7 to 12 to the 6 replicates of sample GO 2 and so on).

Data fusion does not induce any supplementary bias on the existent variance between samples as some well-defined clusters are observed within replicates of a given sample in Figure 6A. Besides, better-defined clusters as a function of gas oil type considered are observed compared to the previous single analyses. This indicates that the specificities of each gas oil type are logically better highlighted when a higher portion of the heteroatomic composition (N_{basic}, $N_{neutral}$ and Sulfur) of the gas oils is considered. Two clusters are also observed according to hydrotreatment level: a first cluster containing deeply hydrotreated samples (Deep HDT: N < 210 ppm, S < 700 ppm) and a second cluster containing moderately hydrotreated samples (Moderate HDT: N > 460 ppm, S > 2800 ppm). The sample GO 12 (GOCK) is excluded from the other GOCK samples cluster due to its atypical character which has only been mainly observed during the single analysis of ESI(-)-FT-ICR/MS dataset. Indeed, a higher alkylation shift has been identified for this sample compared to other GOCK samples and directly visible through its projection on the ESI(-)-FT-ICR MS score plot over PC1 and PC2 in Figure 4A. It is worth noticing that this alkylation shift is visible in every ionization mode (see Figure 3A) but has not been statistically significant among all other variables identified during the single analysis of the

datasets. The application of PARAFAC on fused data is then very efficient to highlight atypical samples and condense the most significant contributions from all samples. The loadings from the sample mode are presented in Figure 6B as a function of the experiment number. It is a 1D projection of the Figure 6A which confirms the high score over the first component of the LCO samples (experiments number from 31 to 42 and from 127 to 132), the FBGO sample (experiments number from 73 to 78) and the HDT samples (experiments number from 91 to 102 and from 109 to 126). Similarly, the samples that are best described by the second component have also high scores such as the SRGO samples (experiments number from 1 to 30), the HDT samples (experiments number from 91 to 102 and from 109 to 126) and the sample GOCK 12 (experiments number from 61 to 66).

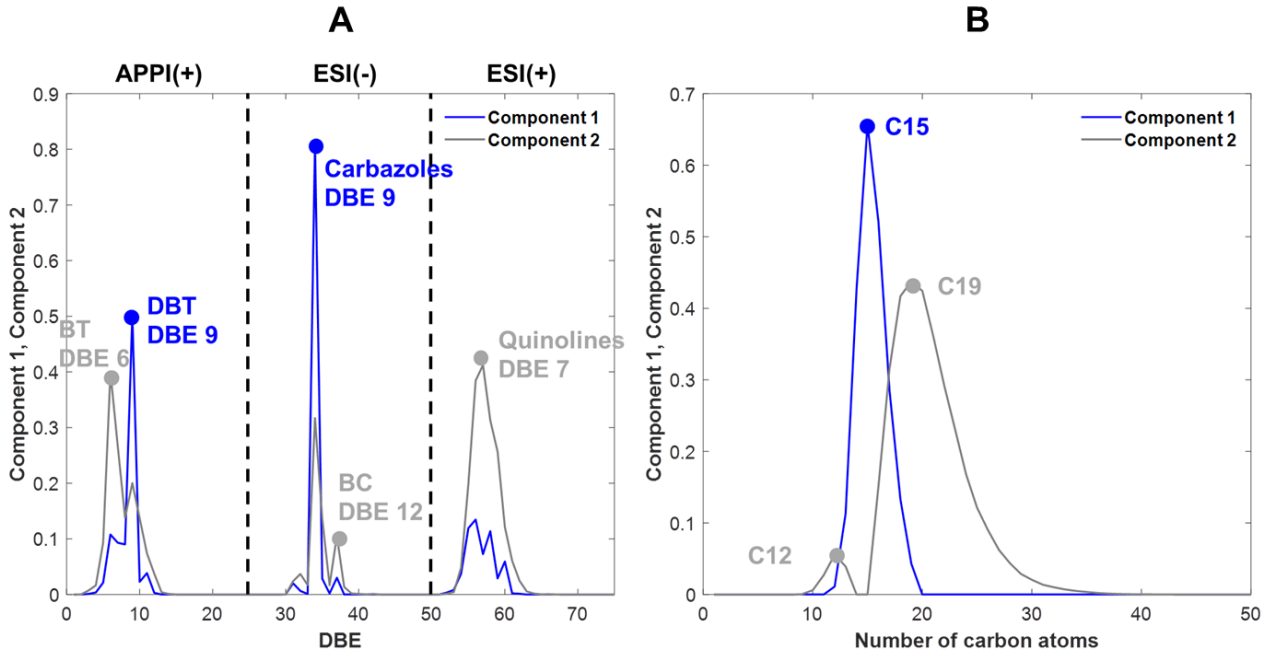

Figure 7. PARAFAC loadings. (A) Loadings from DBE mode. (B) Loadings from the number of carbon atoms mode.

The loadings for DBE mode are shown in Figure 7A. Globally, the first component is driven by the DBE 9 family within both APPI(+) and ESI(-) datasets which respectively correspond to dibenzothiophenes and carbazoles families. In particular, the carbazoles family represents the most intense contribution. The contributions from ESI(+) dataset are very low and spread over the whole DBE range. The projection of the hydrotreated samples and the FBGO sample over the first component

is then correlated to a very strong contribution of the DBT and carbazoles family that mainly remain in the sample as they are known to be refractory[2,25]. It should be noted that these observations are the same as the ones obtained throughout the single analysis of the ESI(-) and APPI(+) datasets hence proving the sensitivity and relevance of such data fusion approach to put forward significant information. Moreover, the score of the deep HDT samples over the first component is higher than those of the moderate HDT samples. This indicates that the DBT and carbazoles families are relatively more intense in the deep HDT samples whereas less refractory species such as benzothiophenes or benzocarbazoles have been mostly removed from the samples using deep hydrotreatment operating conditions. The same observation can be made for the FBGO sample that mostly contains DBT or carbazoles compounds. The separation of the sample GO 12 (GOCK) from the other GOCK samples is partly due to a smaller score on the first component as this sample contains less carbazoles and more benzocarbazoles (see Figure 3B). The projection of the LCO samples over the first component is mainly due to the fact that only neutral nitrogen compounds are mostly found in these samples (see Table 1) so they do not have significant contributions from basic nitrogen compounds and are best described by the first component. The very low contribution of the GO 13 (EBGO) sample to the explained variance of the first component is related to its amount of very aromatic compounds such as benzocarbazoles (DBE 12) that are not contributing to the variance of the first component.

The contributions of the second component are more equally spread over the three different ionization modes with a contribution from the DBE 6 family (BT) for APPI(+) data and to a lesser extent from the DBE 9 family (DBT). For ESI(-) mode, two contributions are also observed from the DBE 9 family (Carbazoles) and DBE 12 (Benzocarbazoles). Finally, the contributions from ESI(+) mode are spread over the whole DBE range with a maximum for the DBE 7 family (Quinolines). The contribution of the benzothiophenes family over the second component is correlated with the composition of the SRGO samples which have high benzothiophenes contents as they are not hydrotreated and are generally

poorly aromatic (see Figure 3B). The sample GO 12 (GOCK) also reveals higher content in benzocarbazoles (about 25%, see Figure 3B) and lower content in carbazoles (about 65%) compared to other GOCK samples which is consistent with an increased contribution of the second component. The basic nitrogen families are also contributing to the variance of the second component and explain the score of the HDT samples over this component. Indeed, most of the basic compounds are still found in HDT samples and there is no particularly intense family (see Figure 3B). The moderate score of the EBGO sample over the second component is mainly related to the small contribution of very aromatic neutral nitrogen compounds (Benzocarbazoles, DBE 12) which are intense in this very aromatic sample (see Figure 3B).

Figure 7B shows the loading corresponding to the number of carbon atoms mode. Two main distributions are observed: the distribution of the first component is focused over C15 while the distribution of the second component is spread over C12 and C19 with C19 being much more intense. Thus, compounds that are best described by the first component are less alkylated than those best described by the second component. This is observed as the SRGO samples are more alkylated than the LCO samples whichever dataset considered (see Figure 3A). The projections of most GOCK samples are intermediate between those observed for SRGO and LCO samples over both components reflecting their intermediate alkylation state compared to other feeds. The contribution of the FBGO sample over the first component is important as it is poorly alkylated due to severe hydroconversion conditions. The score of the deep HDT samples over the first component is a little bit higher than those observed for the moderate HDT samples reflecting a loss in alkylation when increasing the hydrotreatment level. It is also worth noticing that the distribution of the first component relies on the intense contribution of C15. C15 alkylation degree corresponds to C3-Carbazole or C3-DBT molecules that are refractory. As a consequence, these compounds are found in the deep HDT and FBGO samples which all have strong scores over the first component. On the opposite, the sample GO 13 (EBGO) shows a very low

contribution to the first component regarding its very alkylated character which is best described by the second component. Finally, the sample GO 12 (GOCK) is less contributing to the variance of the first component as it is globally more alkylated than the other GOCK samples and thus shows larger projection over the second component as already demonstrated before (see Figure 3B).

The mixed blends have been used as validation samples.. The obtained projections over both components according to their compositions were all consistent whereas it was not the case for the single analysis of the ESI(+)-FT-ICR MS dataset.

In summary, a single PARAFAC analysis allows extracting the main characteristics of each type of gas oil in terms of aromaticity as well as alkylation degrees. The projection of the LCO samples over both components reflects their very aromatic and poorly alkylated character. On the opposite, the projection of the SRGO samples is directly related to their poorly aromatic and very alkylated composition. Most GOCK samples have intermediate characteristics between the LCO and the SRGO samples whereas the unique character of the sample GO 12 is both due to higher alkylation and aromaticity degrees. The hydrotreated samples and the FBGO sample have very strong contributions over the first component which is consistent with their high contents in poorly alkylated refractory species such as C3-Dibenzothiophene and C3-Carbazole. As regards the PARAFAC efficiency gain observed, we estimated that this new approach is about 5 times faster than PCA approach in our case. Indeed, our usual PCA approach is repeated three times (one for each ionization mode) including data pre-processing, choice of the appropriate number of principal components, outliers detection... while these steps are only performed once using PARAFAC. Moreover, the interpretation of scores plots and loadings extracted from PCA is certainly the most time-consuming step because in our case we potentially consider four components for each of the three ionization modes that is at least 12 in total. In comparison, only two components from PARAFAC are interpreted simultaneously in parallel on the same graphics.

Conclusion

In this study, two different chemometric strategies have been assessed over a large gas oil database. As a first step, a classical chemometric approach has been followed applying PCA on selected single FT-ICR MS datasets corresponding to the main problematic compounds in the hydrotreatment processes that are the basic nitrogen, neutral nitrogen, and sulfur compounds. For a given dataset, some clusters have been observed according to gas oil type considered as well as speciation regarding the hydrotreatment level. The variables explaining these clusters have been identified through the analysis of the obtained loadings. Then, the evolution of the relative intensities of these given variables for the different samples has also been plotted to validate their relevance to explain the variance between samples and the efficiency of the chemometric model to extract significant variables. The most refractory compounds have been identified within each data set. As a second step, an innovative chemometric method has been assessed by merging the three datasets to obtain a single hypercube containing the information from the three different ionization modes. To explore this multi-dimensional matrix, the PARAFAC method has later been applied to this dataset studying simultaneously three different modes including DBE, number of carbon atoms and samples. The projection of the samples over the two principal components and the analysis of the obtained loadings have led to the same conclusions as those obtained throughout the single analysis of the datasets. This proves the efficiency of the PARAFAC method to explore very complex datasets and extract the most relevant variables to explain the variance between samples. Besides, it allows visualizing simultaneously the contribution of each ionization mode to the explained variance between samples which was not accessible through the single analysis of the datasets. The efficiency of such a method opens up perspectives for the analysis of complex datasets from different ionization modes as well as obtained with different sample introduction modes such as the comparison of direct infusion and gas or liquid chromatography coupled to FT-ICR MS analysis.

ASSOCIATED CONTENT

Supporting Information

Evolution of basic nitrogen families for the different hydrotreated samples, PC3-PC4 score plot and its corresponding loadings from ESI(+)-FT-ICR MS dataset and ESI(-)-FT-ICR MS dataset, PC1-PC2 and PC3-PC4 scores plots and its corresponding loadings from APPI(+)-FT-ICR MS dataset (PDF).

AUTHOR INFORMATION

Corresponding Author

Dr Marion Lacoue-Nègre, marion.lacoue-negre@ifpen.fr

Dr Ludovic Duponchel, ludovic.duponchel@univ-lille.fr

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript. All authors contributed equally.

REFERENCES

(1) Ma, X.; Sakanishi, K.; Mochida, I. Hydrodesulfurization Reactivities of Various Sulfur Compounds in Vacuum Gas Oil. Ind. Eng. Chem. Res. 1996, 35, 2487–2494.

(2) Valencia, D.; García-Cruz, I.; Uc, V. H.; Ramírez-Verduzco, L. F.; Aburto, J. Refractory Character of 4,6-Dialkyldibenzothiophenes: Structural and Electronic Instabilities Reign Deep Hydrodesulfurization. ChemistrySelect 2018, 3, 8849–8856.

(3) Rabarihoela-Rakotovao, V.; Diehl, F.; Brunet, S. Deep HDS of Diesel Fuel: Inhibiting Effect of Nitrogen Compounds on the Transformation of the Refractory 4,6-Dimethyldibenzothiophene Over a NiMoP/Al2O3 Catalyst. Catal Lett 2009, 129, 50–60.

(4) Sau, M.; Basak, K.; Manna, U.; Santra, M.; Verma, R. P. Effects of organic nitrogen compounds on hydrotreating and hydrocracking reactions. Catalysis Today 2005, 109, 112–119.

(5) Purcell, J. M.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. Speciation of nitrogen containing aromatics by atmospheric pressure photoionization or electrospray ionization fourier transform ion cyclotron resonance mass spectrometry. Journal of the American Society for Mass Spectrometry 2007, 18, 1265–1273.

(6) Nguyen, M.T.; Pirngruber, G.; Chainet, F.; Albrieux, F.; Tayakout-Fayolle, M.; Geantet, C. Molecular level insights into straight run/coker gas oil hydrodenitrogenation by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy Fuels 2019, 33, 3034–3046.

(7) Oliveira, L. P. de; Hudebine, D.; Guillaume, D.; Verstraete, J. J.; Joly, J. F. A Review of Kinetic Modeling Methodologies for Complex Processes. Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles 2016, 71, 45.

(8) Marshall, A. G.; Rodgers, R. P. Petroleomics: Chemistry of the underworld. PNAS, 105, 18090–18095.

(9) Marshall, A. G.; Rodgers, R. P. Petroleomics: The next grand challenge for chemical analysis. Accounts of chemical research 2004, 37, 53–59.

(10) Purcell, J. M.; Juyal, P.; Kim, D.-G.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. Sulfur Speciation in Petroleum: Atmospheric Pressure Photoionization or Chemical Derivatization and Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy Fuels 2007, 21, 2869–2874.

(11) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: A primer. Mass spectrometry reviews 1998, 17, 1–35.

(12) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. Elemental Composition Analysis of Processed and Unprocessed Diesel Fuel by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Energy Fuels 2001, 15, 1186–1193.

(13) Guillemant, J.; Albrieux, F.; Oliveira, L. P. de; Lacoue-Nègre, M.; Duponchel, L.; Joly, J.-F. Insights from Nitrogen Compounds in Gas Oils Highlighted by High-Resolution Fourier Transform Mass Spectrometry. Analytical chemistry 2019, 91, 12644–12652.

(14) Hur, M.; Yeo, I.; Park, E.; Kim, Y. H.; Yoo, J.; Kim, E.; No, M.-h.; Koh, J.; Kim, S. Combination of statistical methods and Fourier transform ion cyclotron resonance mass spectrometry for more comprehensive, molecular-level interpretations of petroleum samples. Analytical chemistry 2010, 82, 211–218.

(15) Hur, M.; Ware, R. L.; Park, J.; McKenna, A. M.; Rodgers, R. P.; Nikolau, B. J.; Wurtele, E. S.; Marshall, A. G. Statistically Significant Differences in Composition of Petroleum Crude Oils Revealed by Volcano Plots Generated from Ultrahigh Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectra. Energy Fuels 2018, 32, 1206–1212.

(16) Chiaberge, S.; Fiorani, T.; Savoini, A.; Bionda, A.; Ramello, S.; Pastori, M.; Cesti, P. Classification of crude oil samples through statistical analysis of APPI FTICR mass spectra. Fuel Processing Technology 2013, 106, 181–185.

(17) Guillemant, J.; Albrieux, F.; Lacoue-Nègre, M.; Pereira de Oliveira, L.; Joly, J.-F.; Duponchel, L. Chemometric Exploration of APPI(+)-FT-ICR MS Data Sets for a Comprehensive Study of Aromatic Sulfur Compounds in Gas Oils. Analytical chemistry 2019, 91, 11785–11793.

(18) Law, J.; Jolliffe, I. T. Principal Component Analysis. The Statistician 1987, 36, 432.

(19) Guillemant, J.; Berlioz-Barbier, A.; Albrieux, F.; Oliveira, L. P. de; Lacoue-Nègre, M.; Joly, J.-F.; Duponchel, L. Low-Level Fusion of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Data Sets for the Characterization of Nitrogen and Sulfur Compounds in Vacuum Gas Oils. Analytical chemistry 2020, 92, 2815–2823.

(20) Bro, R. PARAFAC. Tutorial and applications. Chemometrics and Intelligent Laboratory Systems 1997, 38, 149–171.

(21) Mirnaghi, F. S.; Soucy, N.; Hollebone, B. P.; Brown, C. E. Rapid fingerprinting of spilled petroleum products using fluorescence spectroscopy coupled with parallel factor and principal component analysis. Chemosphere 2018, 208, 185–195.

(22) Ohno, T.; Amirbahman, A.; Bro, R. Parallel factor analysis of excitation-emission matrix fluorescence spectra of water soluble soil organic matter as basis for the determination of conditional metal binding parameters. Environmental science & technology 2008, 42, 186–192.

(23) Data Fusion Methodology and Applications; Cocchi, M., Ed.; Elsevier: Cambridge, 2019.

(24) Billon, A.; Morel, F.; Morrison, M. E.; Peries, J. P. Les procédés IFP HYVAHL(r) et SOLVAHL(r) de conversion de résidus. Rev. Inst. Fr. Pét. 1994, 49, 495–507.

(25) Shin, S.; Sakanishi, K.; Mochida, I.; Grudoski, D. A.; Shinn, J. H. Identification and Reactivity of Nitrogen Molecular Species in Gas Oils. Energy Fuels 2000, 14, 539–544.