# Éditorial

Fabio Rocca, Laurent Duval




Dossier

*This paper is a part of the hereunder thematic dossier published in OGST Journal, Vol. 69, No. 2, pp. 195-372 and available online here*

Cet article fait partie du dossier thématique ci-dessous publié dans la revue OGST, Vol. 69, n°2, pp. 195-372 et téléchargeable ici

DOSSIER Edited by/Sous la direction de : **L. Duval**

# *Advances in Signal Processing and Image Analysis for Physico-Chemical, Analytical Chemistry and Chemical Sensing*

## Progrès en traitement des signaux et analyse des images pour les analyses physico-chimiques et la détection chimique

*Oil & Gas Science and Technology – Rev. IFP Energies nouvelles*, Vol. 69 (2014), No. 2, pp. 195-372

*Oil & Gas Science and Technology – Rev. IFP Energies nouvelles*, Vol. 69 (2014), No. 2, pp. 195-206

DOI: 10.2516/ogst/2014004

Dossier

*Advances in Signal Processing and Image Analysis for Physico-Chemical, Analytical Chemistry and Chemical Sensing*
Progrès en traitement des signaux et analyse des images pour les analyses physico-chimiques et la détection chimique

# Éditorial

## PROGRÈS EN TRAITEMENT DES SIGNAUX ET ANALYSE DES IMAGES POUR LES ANALYSES PHYSICO-CHIMIQUES ET LA DÉTECTION CHIMIQUE

Les animaux et les humains traitent les informations sensorielles (vision, ouïe, toucher, odorat, goût) depuis la nuit des temps. Dans le cadre des connaissances actuelles, on estime qu'un être humain peut assimiler un taux d'information maximal de quelques dizaines d'éléments binaires (bits) par seconde, soit une poignée de lettres d'une dictée aléatoire. Par contraste, des millions de bits par seconde seraient nécessaires pour décrire la quantité phénoménale de signaux qui nous atteignent à chaque instant, essentiellement sous les aspects visuels, auditifs et haptiques (du toucher). Dennis Gabor, qui a reçu le prix Nobel de physique pour l'invention de l'holographie, et contribué de manière significative au traitement du signal, remarquait en 1959 dans l'éditorial d'un volume des *IRE Transactions on Information Theory* :

> "Il existe un écart d'échelle de l'ordre du million entre les 20 bits par seconde que l'œil humain peut absorber, d'après les psychologues, et ce que nous propose l'image télévisuelle."

Les chiffres peuvent avoir changé. Néanmoins, ces nombres révèlent la capacité des humains (et des animaux également pour ce cas) à traiter les signaux perçus pour en extraire la poignée de bits d'information requis, qui seront mémorisés et employés pour mener les tâches de la vie courante. Par conséquent, une description économe, ou parcimonieuse, de ces signaux est fondamentale à leur compréhension et à leur assimilation. Ce processus, qui se déroule en continu, est plus aisé à remarquer dans certains exploits des animaux, que la technologie ne sait pas encore égaler.

Considérons par exemple une nuée de chauves-souris s'échappant d'une grotte, chacune utilisant son propre radar, malgré l'interférence issue de ses congénères. Et pourtant, les chauves-souris ne se heurtent pas plus entre elles qu'aux parois de la grotte. Dans un volume aussi réduit et léger qu'une tête de chauve-souris, il n'existe pas aujourd'hui de système radar pleinement fonctionnel, aussi performant en temps réel. De la même manière, aucun drone n'a la capacité d'un oiseau de proie à identifier une cible non coopérative et à fondre sur elle sans dommage. De tels exemples sont innombrables, mais cette simple observation peut servir à présenter le dossier de ce journal consacré aux applications à la chimie du traitement du signal. Elle indique que tout traitement de signal a pour objectif d'extraire les informations utiles de mesures physiques bruitées, où les perturbations peuvent provenir d'interférences liées aux limites du système d'acquisition, à l'instar de l'altitude de vol limitant le champ visuel de l'aigle, ou d'interférences dues à la présence de signaux non désirés, dont le mélange peut être identifié à un « bruit » dans le cas des chauves-souris. De plus, une fois le signal séparé du bruit, il faut le comprendre, en d'autres termes estimer les principaux paramètres des observables avec la meilleure précision possible. Et seulement alors le lapin peut être capturé.

La révolution numérique a impacté le traitement du signal au même titre que les autres disciplines, et a produit un corpus de méthodes pouvant être appliquées dans des situations très diverses, et cependant partageant quelques principes. Il existe un espace de données (celui des observables) et un espace de modèles, dont on extrait des paramètres. La toute première opération est le stockage des signaux générés, encore dans l'espace des données « réelles » dans l'espace informatique, par la création de tableaux multidimensionnels de nombres ou de vecteurs. L'échantillonnage et la quantification, discrétisant à la fois l'espace des données et celui des modèles pour s'adapter à la quantité finie de mémoire disponible dans les ordinateurs, ont longtemps été des freins aux capacités de traitement des signaux. Aujourd'hui, cette difficulté s'est évanouie dans le « big data », des ordinateurs pétaflop, réalisant $10^{15}$ opérations à virgule flottante par seconde et dans les systèmes de stockage pétaoctets.

L'espace des données peut être de dimension 1 pour de l'audio monocanal, de dimension 2 pour les images, trois pour la vidéo ou les volumes de données, quatre pour les informations de vol (trois pour l'espace et une pour le temps), cinq pour les données sismiques (deux pour la position spatiale des sources en surface, deux pour celle des capteurs, une pour la dimension temporelle au long de laquelle se succèdent les échos réfléchis dans le sous-sol), etc. L'espace des modèles peut également être de dimension élevée : quatre pour la localisation d'une cible se déplaçant dans l'espace (pour une chauve-souris chassant une mouche), six pour des données élastiques (trois de position et trois autres pour les paramètres élastiques locaux). La dimension du vecteur d'observation (dans l'espace données) est indépendante de celle de l'espace de support : l'audio en stéréo reste monodimensionnel en temps ; même si le vecteur mesuré possède deux composantes (gauche et droite). Cela est important également dans les applications chimiques où l'on utilise les mesures couplées ou multimodales (à l'instar de la stéréo : deux oreilles pour la même symphonie). Différentes techniques peuvent alors être utilisées conjointement afin de mieux estimer les mêmes paramètres : la dimension du vecteur observé est augmentée, pas celle de l'espace des modèles.

Les outils du traitement de signal sont de plus bien définis et réutilisables pour différents objectifs et différentes applications. L'outil de base est le débruitage, cette opération pouvant aller du plus simple au plus complexe. La régression linéaire est un cas simple. Elle consiste en l'identification de relations linéaires entre deux variables en présence de perturbations. L'ajustement d'une droite au travers d'un graphe bivarié, typiquement par minimisation de l'erreur quadratique, en est l'avatar le plus évident. Plus complexe est le lissage d'un signal en prenant comme données des $N$-uplets de points consécutifs, et en les ajustant localement par un polynôme de faible degré $M$ ($M < N$ pour les filtres de Savitzky-Golay).

La suppression des bruits peut de surcroît être obtenue par des moyens plus fins, en considérant l'effet « soirée ». Il est bien connu que lors d'une soirée, il est possible d'écouter à la dérobée la conservation menée dans un autre groupe de participants, en supprimant la perception des discussions interférentes, avec un bon degré d'efficacité. Ce mécanisme correspond approximativement à la méthodologie de la séparation de sources : le signal reçu est formé d'une combinaison de sources diverses (les échanges), et il est possible de démêler, de manière univoque à certains instants, les processus indépendants qui ont ainsi été combinés. En d'autres termes, la somme algébrique de deux discussions n'est pas une discussion en elle-même. Cela implique que chaque « discussion » doit être caractérisée par des propriétés morphologiques (son contenu fréquentiel) ou statistiques (absence de corrélation) spécifiques, la rendant discernable : si l'une était un bruit purement gaussien, on ne pourrait la séparer d'un autre bruit gaussien. Cependant, on peut facilement séparer une salve d'impulsions d'une sinusoïde.

Ce dernier exemple nous mène à une autre méthodologie pour comprimer un ensemble parcimonieux. Une salve d'impulsions, signal potentiellement compliqué si échantillonné précisément mais sans intelligence, peut être paramétrée de manière très efficace en mesurant ses amplitudes et temps d'arrivée et en identifiant la forme commune des impulsions.

En pratique, être capable de s'apercevoir qu'un signal peut être paramétré de façon simple, par un ensemble parcimonieux d'événements, est un problème différent et bien plus complexe. Il s'agit, en substance, de compression. Il est alors essentiel d'identifier les différents composants comprimés, réalisant une décomposition parcimonieuse, par minimisation de la norme d'une erreur judicieusement choisie. Remarquons que le choix d'une moindre norme quadratique nous rendrait trop tolérants aux faibles erreurs, empêchant la convergence. Rappelons-nous que Sherlock Holmes accordait une grande attention aux moindres détails pour résoudre une affaire : les moindres carrés ne sont guère fructueux en criminalistique. Bien qu'encore loin de l'objectif final, nous approchons ainsi du problème de la « compréhension ». Pratiquement, nous recherchons la simplification d'un ensemble de données très complexes par transformation en une combinaison d'événements simples pouvant être aisément détectés et estimés. Nous sommes conduits à obéir à la loi d'économie ou rasoir d'Ockham (souvent attribué à Guillaume d'Ockham) : *Entia non sunt multiplicanda praeter necessitatem*. Bien sûr, cette approche n'est utile que si elle approche la réalité : c'est-à-dire si les paramètres ainsi identifiés sont suffisamment proches des paramètres physiques dont nous avons réellement besoin. Une chauve-souris qui n'attrape pas de mouche reste affamée.

Le traitement du signal en chimie analytique sert à l'analyse qualitative (détection : quel composé est présent ?) et quantitative (estimation : en quelle quantité ?), afin d'étudier les propriétés physiques et chimiques de composés et de mélanges de substances naturelles ou chimiques. Il doit s'appuyer sur de nombreuses interactions physico-chimiques d'atomes et des molécules. Sa spécificité, vis-à-vis de l'analyse chimique de routine, réside dans l'amélioration continue des méthodes analytiques, de la conception d'expérience et de la chimiométrie, « l'art d'extraire de l'information pertinente d'expériences chimiques ». Cette technique emprunte essentiellement aux champs de l'analyse multivariée et des statistiques.

Un signal chimique monodimensionnel standard, souvent nommé spectre, est caractérisé en chaque point par une amplitude liée à la proportion d'un certain composé. La variable ordinale n'est pas restreinte au temps ou à l'espace. Elle représente une propriété physico-chimique pouvant réaliser la séparation entre composés élémentaires, par exemple le point d'ébullition (température), la migration (masse moléculaire), la sensibilité aux champs électromagnétiques (rapport masse/charge), etc. Le signal chimique résultant est composé approximativement d'une combinaison linéaire d'une série de pics de différents signaux et de bruit. Le modèle le plus simple est donc un mélange linéaire.

Les spectres élémentaires sont souvent non-négatifs, et reflètent les constantes stœchiométriques des équations chimiques à l'équilibre (conservation de la masse, de la charge et des atomes). De nouvelles contraintes de parcimonie sur les espèces chimiques en présence sont apparues récemment. Le besoin de séparation impliquant plusieurs propriétés (point d'ébullition, structure électronique) a émergé. Ce besoin a engendré des techniques couplées, combinant des techniques élémentaires en paires, triplets, etc. Par exemple, la chromatographie bidimensionnelle produit un signal à deux dimensions. Le couplage peut être étendu en dimension supérieure, au prix de problèmes sévères de gestion de données.

Les articles dans ce dossier d'OGST décrivent un état de l'art en traitement du signal pour la chimie.

Le premier, « *Traitement de signaux par analyse multivariée* », par **J.R. Beattie**, traite de l'Analyse Multivariée (AM) et de son potentiel à transformer l'analyse de signaux. Les méthodes multivariées permettent des prétraitements de signaux pouvant en éliminer des variations non pertinentes, en prélude à l'analyse des signaux d'intérêt. Du débruitage haute-fidélité (suppression de « non-signaux » non reproductibles) devient ainsi possible, de manière cohérente et reproductible. Des « non-signaux » reproductibles peuvent être supprimés, des signaux interférents éliminés, et des fluctuations d'amplitude normalisées. La validité de ces approches requiert des propriétés particulières. Une certaine stationnarité est notamment nécessaire. Une rapide description de la méthodologie en AM est d'abord présentée, autour de l'analyse en composantes principales, dans laquelle la matrice de

données est représentée de façon parcimonieuse comme somme de dyades correspondant aux vecteurs et valeurs propres principaux.

Le premier exemple s'intéresse à la composition de cellules souches et différentiées, cartographiée en utilisant l'intensité de bandes spectrales spécifiques de marqueurs biochimiques comme l'ARN. Le bruit présent résulte d'une acquisition rapide. L'application d'un débruitage basé sur la décomposition en valeurs singulières a permis d'obtenir des images fortement contrastées à partir de données bruitées, et de distinguer cellules souches et cellules différentiées.

Dans un deuxième exemple, l'auteur discute de l'amélioration introduite par Martens d'une méthode de prétraitement spectral existante : la correction de dispersion multiplicative, dans laquelle les erreurs sont corrigées par normalisation de décalage et de moyenne. Martens rend cette méthode adaptable par la construction de deux bases de données, l'une de signaux interférents, l'autre de « signaux cibles ». Ces bases de données sont combinées et soumises à une réduction de données multivariées. L'application de régressions linéaires multiples aux données originales, nommée correction de dispersion multiplicative étendue, permet de soustraire les contributions des signaux non souhaités, et de conserver celles relatives aux signaux souhaités et aux résidus. Cette technique a été appliquée à toute une gamme de signaux, incluant proche-infrarouge, spectroscopie Raman et spectroscopie infrarouge à transformée de Fourier.

Le troisième exemple présente les travaux de Wold *et al*., introduisant la correction orthogonale de signaux pour éliminer des signaux interférents dominants. Les signaux multivariés non liés à la cible sont identifiés, et leur contribution soustraite, ne préservant que les signaux corrélés à la cible. Dans un quatrième exemple, les moindres carrés partiels sont utilisés pour calculer les facteurs d'échelle requis pour combiner des mélanges complexes de signaux différents, ne possédant pas de caractéristique commune permettant de normaliser leurs intensités. La méthodologie basée sur les moindres carrés partiels permet de réduire de moitié la prédiction de variance, par comparaison avec des méthodes traditionnelles.

Un autre exemple est proposé : de nombreux mélanges pétroliers sont constitués d'une variété d'hydrocarbures aliphatiques et aromatiques ainsi que d'autres constituants, contaminants ou molécules actives avec des groupes fonctionnels. Les analyses standard peuvent impliquer le calcul de rapports analyte/hydrocarbure ou des paramètres analogues. En identifiant les composantes principales contenant des contributions aux différents hydrocarbures, il serait possible de calculer la contribution globale d'une classe de molécules par combinaison linéaire d'espèces individuelles présentes dans les données.

Un deuxième article traite d'analyse de données RMN : « *Analyse de données RMN : une approche paramétrique basée sur une décomposition en sous-bandes adaptative* », par **E.-H. Djermoune**, **M. Tomczak** et **D. Brie**. L'article s'intéresse à une méthode d'analyse de données de spectroscopie par Résonance Magnétique Nucléaire (RMN) mono- et bidimensionnelles par une décomposition spectrale adaptative. Elle est obtenue au travers de filtrages et de décimations successifs. Cette méthode propose une sélection automatique du degré de décimation et induit une décomposition adaptée aux signaux. Elle permet de plus de réduire le temps de traitement, et simplifie le choix des paramètres classiques par rapport à un traitement sur le signal complet. La performance de la méthode est évaluée sur des données de RMN du carbone 13 1-D et 2D. Les données sont essentiellement filtrées passe-bande, et sous-échantillonnées, de façon à ne conserver que les composantes utiles du spectre, réduisant ainsi l'impact de composantes spectrales non désirées, et contenant principalement du bruit.

Le troisième article, « *Mélange de Gaussiennes spatialisé et sélection de modèle pour la segmentation non-supervisée d'images spectrales* », traite de la segmentation non supervisée d'images hyperspectrales par modèles de mélanges spatialisés de gaussiennes, écrit par **S.X. Cohen** et **E. Le Pennec**. Il décrit un nouvel algorithme non supervisé la segmentation d'images hyperspectrales. Cet algorithme étend le modèle classique de mélange de gaussiennes en incorporant également les propriétés spatiales des signaux analysés. Un mélange de

$K$ classes modélise le spectre, chacune de distribution gaussienne, dont les proportions de mélange dépendent de la position. Le nombre de classes, ainsi que d'autres paramètres, sont estimés, une garantie théorique accompagnant cette estimation. Une implémentation efficace est également décrite. Enfin, des tests de segmentation non supervisée sont menés sur un jeu de données réelles.

Le quatrième article, « *Analyse en composantes morphologiques pour les retouches d'images de diffraction des rayons X en incidence rasante utilisée pour la caractérisation structurale des couches minces* », traite de l'analyse en composantes morphologiques pour la désocclusion d'images de diffraction des rayons X en incidence rasante écrit par **G. Tzagkarakis**, **E. Pavlopoulou**, **J. Fadili**, **G. Hadziioannou** et **J.-L. Starck**. La diffraction des rayons X en incidence rasante (*GIXD, Grazing Incidence X-ray Diffraction*) est une méthode de caractérisation très employée, appliquée ici à l'étude de la structure de couches minces. Pour ce qui concerne les couches organiques, le confinement de la couche au substrat génère des structures GIXD bidimensionnelles anisotropes, comme celles observées sur des films de polythiophene, utilisés comme couches actives pour les applications photovoltaïques. Les défauts potentiels des détecteurs employés peuvent altérer la qualité des images acquises, avec un impact sur le processus d'analyse et sur l'information structurelle extraite. Le succès de l'analyse en composantes morphologiques en traitement d'images a motivé l'abord du problème de la récupération d'informations manquantes dans les images GIXD, du fait de défauts potentiels du détecteur. Les structures géométriques présentes dans les images de GIXD peuvent être représentées de manière parcimonieuse par une combinaison de transformées surabondantes, précisément en curvelets et transformées en ondelettes non-décimées, offrant une description simple et compacte de leur contenu informationnel. L'information manquante est alors restaurée par analyse en composantes morphologiques appliquée dans un cadre de désocclusion, exploitant la représentation parcimonieuse ainsi obtenue. L'évaluation expérimentale indique que cette méthode est très efficace pour restaurer l'information manquante sous la forme de pixels brûlés aléatoirement ou de rangées entières brûlées, jusqu'à un taux de 50% des pixels. L'approche proposée peut ainsi s'appliquer à la résolution des problèmes liés à la performance du détecteur à l'acquisition. Ceci est important pour les expériences de type synchrotron, car le temps de faisceau alloué aux usagers est très réduit, tandis que toute défaillance technique peut nuire au déroulement d'un projet expérimental. De surcroît, en permettant d'éviter des temps d'acquisition longs et la répétition de mesure, l'approche proposée présente un intérêt supplémentaire.

Le cinquième article, « *Approche problème inverse pour l'alignement de séries en tomographie électronique* », traite d'une approche problème inverse pour l'alignement de séries d'images en tomographie électronique, par **V.-D. Tran**, **M. Moreaud**, **É. Thiébaut**, **L. Denis** et **J.M. Becker**. La tomographie électronique (ou nanotomographie) est l'une des principales techniques d'acquisition récentes employées en raffinage. Elle donne accès à la mesure de propriétés morphologiques de particules, un ingrédient essentiel de la caractérisation de supports catalytiques. Des volumes 3D sont reconstruits à partir d'ensembles de projections à différents angles réalisées par un Microscope Électronique à Transmission (MET). Cette technique donne une information réellement tridimensionnelle à l'échelle nanométrique. L'enjeu principal réside dans le corecalage des projections qui contribuent à la reconstruction. Les techniques d'alignement actuelles emploient couramment des marqueurs fiduciaires, comme des particules d'or, permettant un alignement cohérent des images. Quand cette pratique n'est pas possible, l'alignement des images est obtenu par corrélation de projections adjacentes. Cependant, cette méthode échoue parfois. Les auteurs proposent une nouvelle méthode basée sur une approche en problème inverse du corecalage. Elle est constituée de deux étapes. La première consiste en un processus d'alignement initial, s'appuyant sur une fonction de coût basée sur des statistiques robustes mesurant la similarité entre une projection et les projections précédentes. Elle réduit des décalages importants liés à l'acquisition. Dans la deuxième étape, ces projections grossièrement recalées initialisent un processus itératif d'alignement/raffinement qui alterne entre (i) reconstruction volumique et (ii) recalage des projections acquises sur des projections calculées à partir du volume

reconstruit en (i). Quand ce processus est terminé, on obtient une reconstruction correcte du volume. La méthode est testée sur des données simulées, et démontre une estimation fiable des paramètres de translation, rotation et changement d'échelle. Elle est évaluée avec succès sur différents supports de catalyseurs.

Enfin, le dernier article, «*Développement de réseaux de capteurs chimiques intelligents par des méthodes de séparation source fondée sur l'analyse de composantes indépendantes non linéaire* », traite de la conception de réseaux intelligents d'électrodes sélectives d'ions, basés sur de la séparation de sources par analyse non-linéaire en composantes indépendantes, par **L.T. Duarte** et **C. Jutten**. Le développement de réseaux de capteurs chimiques basés sur la Séparation de Sources Aveugle (SSA) est une solution prometteuse aux problèmes d'interférence liés aux électrodes sélectives d'ions. La motivation principale de cette approche est la simplification de l'étape consommatrice en temps qu'est la calibration. Tandis que les premiers travaux sur ce problème ne prenaient en compte que le cas où les ions avaient la même valence, ce travail vise à proposer une méthode de SSA fonctionnelle pour des ions de charges différentes. Dans cette situation, le modèle de mélange résultant appartient à une classe particulière de systèmes non linéaires qui n'a pas été encore étudiée dans la littérature. Pour aborder ce processus de mélange, les auteurs utilisent un réseau récurrent comme système de séparation. De plus, pour ce qui est de la stratégie d'apprentissage de SSA, une approche de minimisation de l'information mutuelle est développée, à partir de la notion de différentielle de l'information mutuelle. La méthode opère par traitement en lots et peut être utilisée en analyse hors ligne. La validation de la méthode proposée est étayée par des expériences dans lesquelles les paramètres de mélanges sont extraits de données réelles.

L'objectif principal de ce dossier est de porter à l'attention de la communauté scientifique quelques problèmes pertinents de traitement de signaux chimiques. Bien entendu, les interactions possibles entre chimie et traitement du signal sont nombreuses et les articles de ce dossier ne peuvent être exhaustifs. Nous citerons l'inférence statistique pour la simulation de dynamique moléculaire à large échelle ou l'estimation de concentration en bioréacteurs comme sujets additionnels. À l'inverse, dans d'autres articles, des auteurs suggèrent l'usage de systèmes chimiques pour effectuer des tâches de traitement de signal.
Le lecteur à la recherche de plus d'informations est invité à lire les références [1, 2].

## REFERENCES

1 Duval L., Duarte L.T., Jutten C. (2013) An overview of signal processing issues in chemical sensing, in 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada.

2 Wentzell P.D., Brown C.D. (2000) Signal Processing in Analytical Chemistry, in *Encyclopedia of Analytical Chemistry*, R.A. Meyers (ed.), John Wiley & Sons Ltd, Chichester, pp. 9764–9800.

Fabio Rocca
*Politecnico di Milano,*
*Membre du Comité éditorial d'OGST*

Laurent Duval
*IFP Energies nouvelles*

# Editorial

## ADVANCES IN SIGNAL PROCESSING AND IMAGE ANALYSIS FOR PHYSICO-CHEMICAL, ANALYTICAL CHEMISTRY AND CHEMICAL SENSING

Animals and humans have been processing sensory information (traditionally sight, hearing, touch, smell and taste) for ages. The maximal information rate that can be assimilated by humans is estimated around tens of binary units (bits) per second, say the few letters correspondent to a random dictation, to the best of our present knowledge. On the other side, millions and millions of bit/s are necessary to describe the immense quantity of signals that hit us at any time, that is the visual, audio, and haptic (touch) information. Dennis Gabor, who was awarded to the Nobel Prize in Physics for the invention of holography, and made several outstanding contributions to signal processing, remarked in a Guest editorial to a 1959 volume of the IRE Transactions on Information Theory:

> *"There is a gap of a million or so between the 20 bits per second which, the psychologists assure us, the human eye is capable of taking in, and what we offer it in a television picture."*

The figures may have changed. Yet, these numbers exemplify the capability of humans (and animals, at that) to process the signals that are perceived in order to extract the few needed bits of information that are then memorized and used to proceed with the normal life. A parsimonious or sparse description of the signals is therefore mandatory for their understanding and assimilation. This happens all the time, and it is even more visible in feats of animals that are not as yet reachable by the available technology.

Let us consider a swarm of bats flying out of a cave, each of them using its own acoustical radar, notwithstanding the interference of the others. They neither hit each other, nor hit the walls of the cave. No radar system of today would be capable of achieving such interference free behaviour in real time, with full functionality, in such a small weight and space as that of the head of a bat. Again, no drone has the capability of a bird of prey of identifying the non-cooperative target and then flying down to grab it, without harming itself. The examples are so numerous that it would be impossible to count. This observation, however, is useful to introduce an issue of this journal dedicated to signal processing for chemical applications.

The point is that any sort of signal processing has the aim of extracting the useful information from a noisy measurement, where the noise could be the interference due to the limited capabilities of the measurement system, such as the flying height limits the visual capabilities of an eagle, or the interference due to other unwanted signals, that we collectively identify as "noise" in the case of the bats. Moreover, once the signal is separated from the noise, then it has to be understood, namely the main parameters of the observables have to be estimated with the highest possible precision and only then can the rabbit be captured.

The digital revolution impacted signal processing too, and generated a wide set of methodologies that are applied in extremely different situations, but still share the same principles. There is a data space (that of the observables), and a model space, that of the

parameters to be extracted. The very first operation is the storage of the incoming signals, the real data space, into the computer data space, namely creating a multidimensional array of numbers or vectors. Sampling and quantization, that yield the discretization of both the data space and the model space to accommodate for the finite extent of the memory available on the computers, have been for long a cause for limitation of the signal processing capabilities. Nowadays, big data that allow petaflops computers, that are computers that can process $10^{15}$ floating point operations per second, and petabyte bit stores make this problem irrelevant.

The data space can be one dimensional for mono audio, two dimensional for images, three for images in motion or data volumes, four for flight information (three for space and one for time), five for seismic data (two for the location of the sources on the earth surface, two for the location of the receivers, and one for the time axis along which the echoes of the sounding experiment are positioned) etc. But the dimensionality of the model space can also be high: four dimensional for the location of targets moving in space (a bat chasing a fly), six dimensions for the elastic data (three for the location and three for the local elastic parameters). The dimensionality of the observation vector (the data space) is independent from that of the space of support (the model space): stereo audio is one dimensional in time, even if the measured vector has two components (left and right). This is important also for chemical applications where hyphenated or multimodal measurements are used (like stereo: two ears for the same symphony). Multiple techniques can be used to better estimate the same parameters: the dimensionality of the observation vector is augmented, but that of the model space is left unchanged.

The tools for signal processing are also well defined and reused for different aims and applications. The first tool is noise removal, this operation being possibly simple but also complex. The simple case consists in linear regression, that is the identification of a linear relation between two variables in presence of disturbances. The fit of a line to a cross plot, typically carried out by minimizing the square error is the obvious case. But then, it can be made more complex: we can smooth a noisy signal using as data sets $N$-tuples of successive data points, and locally fitting to them a lower order polynomial of degree $M$ ($M < N$, for Savitzky-Golay filters).

However, noise removal can be achieved in a more clever way: let us consider the cocktail party effect. It is well known that, in a cocktail party, we can eavesdrop a conversation of people talking within another group, suppressing the perception of all the interfering talks, with a good degree of success. This mechanism corresponds approximately to the source separation methodology: the received signal is the combination of many different sources (the talks), and we can achieve their separation finding, at times in an unique way, the independent processes that have been combined. In other words, the algebraic sum of two talks is not another talk. This implies that the "talk" has to be characterized by some morphological (frequency content) or statistical (decorrelation) properties that stand out making it recognizable: for one, was it pure random Gaussian noise, it would be inseparable from any other pure Gaussian noise. On the other hand, we can easily separate a train of spikes from a sinusoid, if we wish so.

This last example leads us to another quite complex methodology that is the compression of a sparse set. A train of spikes, a very complex signal if sampled accurately but without intelligence, can easily be parameterized in a very efficient way by measuring their amplitudes and arrival times, and their common wave shape. In fact, being able to appreciate that a signal can be parameterized in a simple way, with a sparse set of events, is a different and much more complex problem. This, in substance, is indicated as compression. It is essential to identify the compressed components, achieving the parsimonious decomposition, minimizing some appropriate norm of the error. Notice that if we used the usual least squares norm, we would be too lenient with the small errors, and we would not achieve convergence. But we remember that Sherlock Holmes always looks very carefully at the details in order to unravel the mystery: least squares would not lead forensic sciences to very useful result.

This is to say that, while still very far from the ultimate goal, we are approaching the problem of "understanding". Namely, we are attempting the simplification of a very complex set of data, transforming it into the combination of simple events that can thus be well detected and estimated. We tend to obey the law of parsimony, or Ockham's razor — *Entia non sunt multiplicanda praeter necessitatem* — often attributed to William of Ockham. Obviously, all this is useful if it is true, that is if the parameters that are identified are close enough to the physical ones that we need. If the bat does not reach the fly, it stays hungry.

Signal processing in analytical chemistry is intended for the qualitative (detection: what compound is present?) and quantitative (estimation: how much of it?) analyses, to study physical and chemical properties of compounds and mixtures of natural or artificial materials. It relies on many chemical-physical interactions of atoms and molecules. Its specificity, with respect to routine chemical analysis, resides in the continuous improvement of analytical methods, experimental designs and chemometrics, *i.e.* "the art of extracting chemically relevant information from data produced in chemical experiments". The latter essentially borrows methods from multivariate analysis and statistics.

A typical one-dimensional chemical signal (the spectrum, as often called) is characterized by amplitude at each point related to the proportion of a certain component. The ordinal variable is not restricted to time or space. It represents a physical-chemical property, which performs the separation between elementary components, *e.g.* boiling point (temperature), migration (molecular mass), sensitivity to electro-magnetic fields (mass-to-charge ratio), etc. The resulting chemical signal is approximately composed of a linear combination of a sum of peaks of different signals and noise. Hence, the simplest model is a linear mixture.

Elementary spectra are typically non-negative and take into account the stoichiometric constants of balanced chemical equations (conservation of mass, charge or atoms). Recently, sparsity constraints on chemical species have come into play. The need for a separation based on two or more chemical properties (*e.g.* boiling point and electronic structure) has emerged. This has given birth to hyphenated techniques, combining techniques in pair, triple, etc. For instance, two-dimensional or comprehensive chromatography generates a two-dimensional signal. Hyphenation may be extended to higher dimensions, providing an enhancement of resolution at the costs of more drastic data management problems.

The papers of the OGST dossier that follows describe the state of the art in signal processing for chemistry.

The first, *"Multivariate Analysis for the Processing of Signals"*, authored by **J.R. Beattie** discusses Multivariate Analysis (MA) and its capabilities of transforming the signal analysis. Multivariate methods allow pre-processing of signals with the aim of eliminating irrelevant variations prior to analysis of the signal of interest. High-fidelity denoising (removal of irreproducible non-signals) is made possible, consistent and reproducible. Reproducible non-signals are removed, interfering signals are eliminated and signal amplitude fluctuations can be standardized. However, signal properties have to be suitable for MA (a form of stationarity is mandatory). After a short description of the MA methodology, namely the principal components analysis, where the data matrix (more data than variables, indeed) is parsimoniously represented as a sum of dyads corresponding to its principal eigenvectors and eigenvalues.

In a first example, an application is given where the intensities of bands that were unique to specific biochemicals such as RNA were used to map the differing compositions of stem cells and differentiated cells. The rapid acquisition times made the data noisy. Application of the SVD-based (Singular Value Decomposition, a breed of Principal Components Analysis) denoising allowed chemical images of high contrast to be generated from the noisy data, enabling identification of the cellular regions that could be used to differentiate between stem cells and differentiated cells.

In a second example, the author discusses a multivariate-based improvement, introduced by Martens, to an existing spectral pre-processing method: Multiplicative Scatter Correction (MSC, in which multiplicative scattering errors were corrected through

standardising the offset and mean value of the spectra). Martens introduced the concept of an adaptable MSC based on construction of a database on interfering signals and a database of desired ('target') signals. These databases were then combined and subjected to a multivariate data reduction and multiple linear regressions onto the original data. The technique was termed Extended Multiplicative Scatter Correction (EMSC). The contribution from undesired signals was removed from the original data while the contributions from the desired signals and the residual were retained. The technique has been applied to a range of signal types, including NIR, Raman Spectroscopy and FTIR.

In a third example the author presents the work of Wold *et al.* who introduced Orthogonal Signal Correction (OSC) as a means to eliminate interfering dominant signals. Multivariate signals that are unrelated to the target are identified and their contribution removed, leaving only signals that are correlated with the target. In a fourth example, Partial Least Squares (PLS) are used to calculate the scaling factors necessary to combine complex mixtures of different signals that do not share a common feature suitable for standardising the intensity of the signal. The PLS methodology resulted in a halving of the prediction variance compared with using traditional methods. As another example, in many oil systems, a range of different aliphatic and aromatic hydrocarbons and other constituents/contaminants/active molecules with functional groups may be present. Standard analyses may involve calculating the ratio of an analyte to the total hydrocarbon or some similar parameter. By identifying which principal component contains contributions from different species of hydrocarbons, it would be possible to calculate the overall contribution of the class of molecules based on the linear sum of the individual species present in the dataset.

A second paper deals with *"NMR Data Analysis*: *A Time-Domain Parametric Approach Using Adaptive Subband Decomposition"*, authored by **E.-H. Djermoune**, **M. Tomczak** and **D. Brie**. The paper discusses a data analysis method for one- and two-dimensional Nuclear Magnetic Resonance (NMR) spectroscopy, based on an adaptive spectral decomposition. This is achieved through successive filtering and decimation steps. The method leads to an automated selection of the decimation level and consequently to a signal-adaptive decomposition. Moreover, it enables one to reduce the processing time and makes the choice of usual free parameters easier, comparatively to the case where the whole signal is processed at once. The efficiency of the method is demonstrated using 1-D and 2-D 13C NMR signals. In essence, the data are band-pass filtered and sub-sampled, so that only the meaningful parts of the spectrum are retained, reducing the impact of the unwanted spectral components, considered containing mainly noise.

The third paper, "*Unsupervised Segmentation of Spectral Images with a Spatialized Gaussian Mixture Model and Model Selection*", deals with unsupervised segmentation of hyper spectral images with spatialized Gaussian mixture model and model selection, authored by **S.X. Cohen**, and **E. Le Pennec**. A novel unsupervised hyper spectral image segmentation algorithm is described. This algorithm extends the classical Gaussian mixture model based unsupervised classification technique by also incorporating to the model the spatial properties of the signal to be analyzed. A mixture of $K$ classes modelizes the spectrum, each having a Gaussian distribution, whose mixing proportions depend on the position. The number of classes as well as all the other parameters are estimated and a theoretical guaranty for this estimation is provided. An efficient implementation is also described. Finally, some numerical experiments of unsupervised segmentation on a real dataset are carried out.

The fourth paper deals with a *"Morphological Component Analysis for the Inpainting of Grazing Incidence X-Ray Diffraction Images Used for the Structural Characterization of Thin Films"*, authored by **G. Tzagkarakis, E. Pavlopoulou**, **J. Fadili**, **G. Hadziioannou** and **J.-L. Starck**. Grazing Incidence X-ray diffraction (GIXD) is a widely used characterization technique, applied for the investigation of the structure of thin films. As far as organic films are concerned, the confinement of the film to the substrate results in anisotropic 2-dimensional GIXD patterns, such those observed for polythiophene-based films, which are used as active layers in photovoltaic applications. Potential malfunctions of the detectors utilized

may distort the quality of the acquired images, affecting thus the analysis process and the structural information derived. Motivated by the success of Morphological Component Analysis (MCA) in image processing, in this study the authors tackle the problem of recovering the missing information in GIXD images due to potential detector's malfunction. The geometrical structures, which are present in the GIXD images can be represented sparsely by means of a combination of over-complete transforms, namely, the curvelet and the un-decimated wavelet transform, resulting in a simple and compact description of their inherent information content. Then, the missing information is recovered by applying MCA in an inpainting framework, by exploiting the sparse representation of GIXD data in these two over-complete transformed domains. The experimental evaluation shows that the proposed approach is highly efficient in recovering the missing information in the form of either randomly burned pixels, or whole burned rows, even at the order of 50% of the total number of pixels. Thus, their approach can be applied for healing any potential problems related to detector performance during acquisition, which is of high importance in synchrotron-based experiments, since the beam time allocated to users is extremely limited and any technical malfunction could be detrimental for the course of the experimental project. Moreover, the non-necessity of long acquisition times or repeating measurements, which stems from these results adds extra value to the proposed approach.

The fifth paper deals with the *"Inverse Problem Approach for the Alignment of Electron Tomographic series"* authored by **V.-D. Tran**, **M. Moreaud**, **É. Thiébaut**, **L. Denis** and **J.-M. Becker**. Electron tomography (or nanotomography) is one of the main novel acquisition techniques in the refining industry, yielding the morphological measurements of particles, which have become an essential part in the characterization of catalyst supports. 3D volumes are reconstructed from sets of projections from different angles made by a Transmission Electron Microscope (TEM). This technique provides real three-dimensional information at the nanometric scale. A major issue in this method is the requirement of co-registration of the projections that contribute to the reconstruction. The current alignment techniques usually employ fiducial markers such as gold particles for a correct alignment of the images. When the use of markers is not possible, the correlation between adjacent projections is used to align them. However, this method sometimes fails. A new method based on the inverse problem approach is proposed for the co-registration. The proposed approach is composed of two steps. The first step consists of an initial alignment process, which relies on the minimization of a cost function based on robust statistics measuring the similarity of a projection to its previous projections in the series. It reduces strong shifts resulting from the acquisition between successive projections. In the second step, the pre-registered projections are used to initialize an iterative alignment-refinement process which alternates between (i) volume reconstructions and (ii) registrations of measured projections onto simulated projections computed from the volume reconstructed in (i). At the end of this process, a correct reconstruction of the volume is available. The method is tested on simulated data and shown to estimate accurately the translation, rotation and scale of arbitrary transforms. It also has been successfully tested with real projections of different catalyst supports.

Finally, the last paper deals with the *"Design of Smart Ion-Selective Electrode Arrays Based on Source Separation through Nonlinear Independent Component Analysis"*, and is authored by **L.T. Duarte** and **C. Jutten**. The development of chemical sensor arrays based on Blind Source Separation (BSS) provides a promising solution to overcome the interference problem associated with Ion-Selective Electrodes (ISE). The main motivation behind this new approach is to ease the time-demanding calibration stage. While the first works on this problem only considered the case in which the ions under analysis have equal valences, the present work aims at developing a BSS technique that works when the ions have different charges. In this situation, the resulting mixing model belongs to a particular class of nonlinear systems that have never been studied in the BSS literature. In order to tackle this sort of mixing process, the authors adopted a recurrent network as separating

system. Moreover, concerning the BSS learning strategy, a mutual information minimization approach is developed based on the notion of the differential of the mutual information. The method work requires a batch operation, and, thus, can be used to perform off-line analysis. The validity of the proposed approach is supported by experiments where the mixing model parameters were extracted from actual data.

The major goal of this dossier is to bring some relevant problems of chemical signal processing to the scientific community at large. Of course, the possible interactions between the chemistry and signal processing are very wide, and the papers contained in this issue cannot be exhaustive. Additional topics include *e.g.* statistical inference for simulating large scale molecular dynamics, or estimation of chemical concentration in bioreactors, etc. Conversely, in other papers, authors suggest using chemical systems for doing signal processing tasks.

The reader interested for more information is invited to look at the papers [1, 2].

## REFERENCES

1 Duval L., Duarte L.T., Jutten C. (2013) An overview of signal processing issues in chemical sensing, in 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada.

2 Wentzell P.D., Brown C.D. (2000) Signal Processing in Analytical Chemistry, in Encyclopedia of Analytical Chemistry, R.A. Meyers (ed.), John Wiley & Sons Ltd, Chichester, pp. 9764–9800.

Fabio Rocca
*Politecnico di Milano,*
*Member of the Editorial Board of OGST*

Laurent Duval
*IFP Energies nouvelles*