Article – Postprint Version

Please, cite this study as:

Domingo-Almenara X, Perera A, Ramírez N, Cañellas N, Correig X, Brezmes J. Compound identification in gas chromatography/mass spectrometry–based metabolomics by blind source separation. Journal of Chromatography A. Vol. 1409 (2015) 226–233. DOI: 10.1016/j.chroma.2015.07.044.

Compound identification in gas chromatography/mass spectrometry–based metabolomics by blind source separation

Xavier Domingo-Almenara [*] [†] **Alexandre Perera** [‡] **Noelia Ramirez** [*] [†] **Nicolau Cañellas** [*] [†] **Xavier Correig** [*] [†] **and Jesus Brezmes** [*] [†]

[*]Metabolomics Platform – IISPV, Department of Electrical and Automation Engineering (DEEEA), Universitat Rovira i Virgili, Tarragona, Catalonia, Spain,[†]Biomedical Research Networking Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain, and [‡]B2SLAB, Department d'Enginyeria de Sistemes, Automatica i Informatica Industrial, CIBER-BBN, Universitat Politecnica de Catalunya, Barcelona, Catalonia, Spain.

Corresponding author at: xavier.domingo@urv.cat.

Metabolomics GC–MS samples involve high complexity data that must be effectively resolved to produce chemically meaningful results. Multivariate curve resolution–alternating least squares (MCR–ALS) is the most frequently reported technique for that purpose. More recently, independent component analysis (ICA) has been reported as an alternative to MCR. Those algorithms attempt to infer a model describing the observed data and, therefore, the least squares regression used in MCR assumes that the data is a linear combination of that model. However, due to the high complexity of real data, the construction of a model to describe optimally the observed data is a critical step and these algorithms should prevent the influence from outlier data. This study proves independent component regression (ICR) as an alternative for GC–MS compound identification. Both ICR and MCR though require least squares regression to correctly resolve the mixtures. In this paper, a novel orthogonal signal deconvolution (OSD) approach is introduced, which uses principal component analysis to determine the compound spectra. The study includes a compound identification comparison between the results by ICA–OSD, MCR–OSD, ICR and MCR–ALS using pure standards and human serum samples. Results shows that ICR may be used as an alternative to multivariate curve methods, as ICR efficiency is comparable to MCR–ALS. Also, the study demonstrates that the proposed OSD approach achieves greater spectral resolution accuracy than the traditional least squares approach when compounds elute under undue interference of biological matrices.

Introduction

The analysis of samples from a metabolomics perspective allows the phenotyping of organisms at a molecular level [1]. At the same time, metabolomics provides a means of detecting early biochemical changes in organisms before the appearance of a disease and thus, a means of finding predictive biomarkers [2]. Among the analytical techniques used in metabolomics, gas chromatography-mass spectrometry (GC–MS) is a well stablished platform due to its robustness and its applicability to a wide range of matrices and metabolites through silylation of the polar groups.

Because of the high complexity of biological fluids, the complete chromatographic resolution of all the metabolites in a sample cannot be easily achieved as the co-elution of two or more of them usually occurs. The correct identification of co-eluted compounds depends mostly on the degree of the chromatographic separation and their spectral dissimilarity. Likewise, the metabolites in the samples usually occur at low concentrations and the background signal, inherent in the instrument and the sample biological matrix, interferes in their correct identification and quantification. The use of resolution algorithms, which can help extract the purest compound elution profile and spectra, is mandatory for GC–MS data processing.

One of the best–established algorithms for application to chromatographic data to resolve co-eluted compounds is multivariate curve resolution–alternating least squares (MCR–ALS) [3, 4]. MCR–ALS can resolve a mixture of compounds into a pure concentration profile matrix and a pure spectra matrix [5]. In recent years, a blind source separation (BSS) technique known as independent component analysis (ICA) [6], already widely applied for the resolution of spectroscopic mixtures [7, 8, 9, 10, 11], has also been applied for the resolution of GC–MS samples [12]. In a GC–MS chromatogram, the compounds elution profiles appear mixed with their respective spectra. In these cases, ICA-based approaches are able to recover the different independent sources contained in data and, eventually, resolve GC–MS data. MCR–ALS approaches this problem by minimizing the residual error between the data and the predicted model, whereas ICA focuses on estimating the original sources - or components - by maximizing their statistical independence. Actual ICA-based methods to resolve chromatographic data include mean-field ICA (MF–ICA) [13], post-modification based on chemical knowledge (PBCK) [14], window ICA (WICA) [15] and non-negative ICA [16]. Artificial immune system algorithms involving the use of ICA have also been proposed [17]. The first step of the resolution procedure in these methods is the use of ICA to resolve the mass spectrum for each compound in the mixture. The above-mentioned algorithms use different approaches to determine the elution profile of each compound, since the elution profiles determined by ICA tend to be inaccurate or affected by various ICA ambiguities such as negativity or variance (energy) indetermination [18]. Recently, these ICA-based methods were compared with MCR for the resolution of GC–MS data by Parastar and coworkers [19] who showed that the ICA-based resolutions methods show the same performance than MCR. A natural extension of ICA to recover co-eluted profiles might be independent component regression (ICR), which was first used to resolve mixtures in near infrared (NIR) spectra by Shao et al. [20], but whose efficiency on GC–MS data treatment has not yet been studied.

The use of least squares (LS) regression, common to most algorithms in GC–MS data resolution, has a major drawback,

induced by the inherent correlation between ions related to the same compound. This correlation yields an ion-redundancy which means that, for each compound, different ions, also called fragments or m/z, elute at the same retention time and with the same elution profile. When fitting the elution profiles to data, no correlation information between the ions is taken into account, so the LS regression does not distinguish between noise and the compound ions that are being regressed; this may introduce a bias into the LS regressors. This effect includes instrumental or experimental noise as baseline, peak-tailing, or compound co-elution. The performance of the resolution of mixtures with least squares may, therefore, depend on the correct estimation of the underlying model from the data.

This study proposes the use of ICR for GC–MS compound identification. In this approach, we integrate ICA and MCR with a novel orthogonal spectra deconvolution (OSD) as an alternative to least squares regression with a view to improve the determination of the compound spectra when compounds elute under the interference of a biological matrix.

Materials and methods

This section describes MCR–ALS, ICR and their variants integrated with the OSD algorithm (ICA–OSD and MCR–OSD). The proposed methods were evaluated by comparing the resolution of the spectra of 38 compounds in a pure standards sample and 25 compounds in a human serum sample. A match score between the resolved and the reference spectra was determined for each compound and method. The samples were processed by MCR, the proposed ICR, both ICA and MCR using the OSD approach (ICA-OSD and MCR-OSD). The goal was to use the different methods compared in this study to extract the most pure spectra for each compound. The spectra extracted were matched against a reference MS spectra database. For this study, the Golm Metabolome Database (GMD) [21] was used as a reference database.

Materials. A set of four pure standards samples - four sample repetitions - and a total of eight biological samples - four sample repetitions of a human serum sample, and two repetitions of two human urine samples from healthy volunteers - were used for evaluation. The standard mixture was composed of 26 metabolites (see Table S1 of the Supplementary Material) previously found in the human serum and urine metabolome [22]. First, all samples were characterized by a curated identification of the reference compounds (standards). The pure standards samples were taken as a reference to later identify the same compounds in the human serum and urine samples. Two compounds identified in the biological samples that are not included in the pure standards set were validated also analyzing their corresponding standard references.

The metabolites of the human serum and urine samples were extracted and derivatized following a standard protocol [23] with slight modifications to optimize the process. Extracts were analyzed using a 7890 gas chromatograph from Agilent (Palo Alto, CA. USA) coupled to a Pegasus IV TOF/MS from Leco (St. Joseph, MI, USA) using a DB5–MS capillary column (30 m × 0.25 mm × 0.25 μm, 5% diphenyl, 95% dimethylpolysiloxane) from Agilent. Analyses were performed by injecting 1 μL of the extracts into a split/splitless inlet at 250°C with a split flow of 5 mL min^{-1} and a helium constant flow of 1 mL min^{-1} (99.999%, Abelló Linde, Barcelona). The oven temperature of the GC was initially held at 50°C for 1 min, then raised to 285°C at a rate of 20°C min^{-1} and held at that temperature for 5 min. The GC–TOF/MS interface was set at 280°C and the ion source at 250°C. The mass spectrometer acquired m/z ratios between 35 and 600 amu at 10 Hz and an electron impact energy of 70 eV.

Data pre-processing and analysis. In order to analyze an entire dataset using the MCR or ICA-based approaches, each chromatogram was divided in chromatographic peak features (CPFs) using the same criteria as in [24]. The different CPFs contained several compounds, so the algorithm had to deconvolve them in case of co-elution. The number of factors or components used to initialize both MCR and ICA was determined by cross-validation (described in Section 2.6). A unimodality constraint [25] was applied to the resolved profiles and the same non-negative least squares algorithm was applied for both MCR and ICR. The simple mean spectra determined either by ICA–OSD, MCR–OSD, ICR or MCR in the different samples for each compound were compared using the dot product [26] against the GMD MS spectra database.

The masses 73, 74, 75, 147, 148, and 149 m/z were excluded before processing the sample, since they are ubiquitous mass fragments typically generated from compounds carrying a trimethylsilyl moiety [21]. They were also excluded in the identification. Only the fragments from m/z 70 to 600 were taken into account when comparing reference and empirical spectra, since this is the m/z range included in the downloadable GOLM database. Also, the human serum and urine samples signal was filtered using a Savitzky–Golay filter [27] and the baseline was removed using a semi-supervised spline interpolation to reduce the interaction of the biological matrix (described in Section 3.2). The ICA algorithm used was the joint approximate diagonalization of eigenvalues (JADE) [30].

Resolution of GC/MS mixtures by multivariate curve resolution–alternating least squares (MCR–ALS). The purpose of multivariate curve resolution – alternating least squares (MCR–ALS) is to decompose a data matrix containing a mixture of compounds into two matrices containing the resolved pure concentration profiles and pure spectra. MCR can mathematically be expressed as:

$$D = CS^T + E \qquad [1]$$

where D ($N \times M$) is the raw data matrix containing the mixture of compounds, C ($N \times k$) is the resolved concentration profile matrix, S ($M \times k$) is the resolved spectra matrix and E ($N \times M$) is the error matrix. In this notation, N is the number of chromatographic scans (retention time), M is the range of acquisition of the mass-charge ratio (m/z), and k is the number of components or compounds in the model. MCR–ALS uses an iterative least squares algorithm (ALS) to determine both C and S matrices by minimizing the error matrix E. A detailed explanation of MCR–ALS, together with pseudocode, is given elsewhere [29]. To optimize execution speed, we used our own implementation of the MCR–ALS algorithm. This was based on the R package *NNLS*, which uses the Lawson–Hanson non-negative least squares (NNLS) implementation. The package uses C routines to increase the computational speed.

Resolution of GC/MS mixtures by independent component regression (ICR). The proposed independent component regression (ICR) method consists of applying an independent component analysis (ICA), followed by a least squares regression (LS) using the ICA output as a regressor. In this manner, ICA is used to determine the elution profile of the different compounds in the mixture. Then, a least squares regression is used to determine the spectra of each compound by fitting the extracted elution profiles to the data. This implementation is the opposite of the extraction of the compound spectra to later

determine the elution profile, used in the above mention ICA-based implementations. Our ICA model can be expressed as:

$$D^T = AZ^T \quad [2]$$

Analogously to (Eq. 1), D ($N \times M$) is the original chromatographic raw data matrix, A ($M \times K$) is the mixing-matrix and Z ($N \times K$) is the independent components matrix. The Z matrix holds the elution response of the underlying components, but it presents two main ambiguities: (i) we cannot determine the energy or intensity of the resolved components and therefore they are not ordered by explained variance, and (ii) recovered sources do not fulfill non-negativity. Due to the first ambiguity, the recovered sources in Z are arbitrarily scaled and consequently they cannot be used for quantifying the concentration of compounds. Due to the second ambiguity, the extracted components can be negative or contain negative values - known to be caused by source signal overlapping, as explained in [16]-. According to [7, 12], the estimated sources in Z may appear negatively correlated with the data, i.e., the estimated elution profile may be a negative mirror image of the real one. Thus, the Z matrix contains only the qualitative shape—the elution profile model in the retention time dimension—of the underlying compounds. A natural strategy for avoiding such negativity ambiguity is the use of non-negative ICA (nnICA). This, however, adds a significant computational cost and does not solve the first ambiguity, which still has to be resolved by a least squares regression. Therefore, the following strategy is proposed to overcome both ICA ambiguities: all the profiles in Z that express more negative variance than positive variance are negatively rotated. After this step, a non-negative least squares regression (NNLS) is applied to resolve the variance ambiguity and to retrieve the spectrum for each compound. This is to determine a non-negative spectra matrix S that minimizes the error matrix E:

$$D = \hat{Z}S^T + E \quad [3]$$

where D ($N \times M$) is the raw data matrix, Z ($N \times k$) is the elution matrix and S ($M \times k$) the spectra matrix. The hat in $\hat{Z}$ denotes a normalized matrix, since real energies are not known a priori. The determined profiles are fitted in the different columns of the data matrix containing the different m/z values. For ICR, Z is the matrix analogous to the C matrix in MCR. The *JADE* R package is used for the implementation of the ICA-based algorithms.

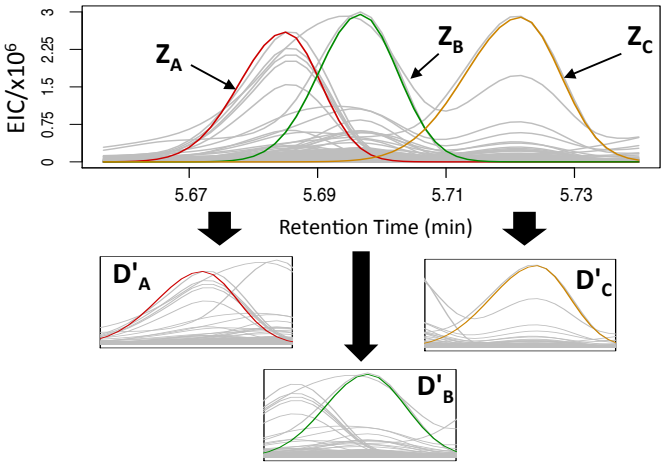

Fig. 1. Determination of D'_j for a given data matrix D, where three compounds appear co-eluted. The extracted ion chromatogram (EIC) of the original D matrix is shown (top). The grey lines represent the different m/z masses whereas the coloured lines represent the three resolved compounds for the case given. Each sub-data matrix D'_j is determined comprising the data for which each compound profile D_j is eluting. A cut-off of 5% is applied to all the profiles, so the D'_j sub-data matrix comprises the data in D for which the profile Z_j is non-zero.

Spectra extraction by orthogonal signal deconvolution (OSD). Orthogonal signal deconvolution (OSD) is a method to extract and deconvolve the spectra given only the compounds elution profile. In multivariate curve resolution or independent component regression, the spectra is determined by means of non-negative least squares, instead, in OSD principal component analysis (PCA) is used to determine the spectra of each compound as opposite to the use of least squares. For this study, a pre-process to determine the elution profiles is conducted by independent component analysis (ICA) or multivariate curve resolution (MCR), and are referred as ICA-OSD and MCR-OSD, respectively. In OSD, PCA is used to decorrelate the sub-data matrix and to determine which ions co-vary along the retention time, thus detecting the different ion-redundancies or ion-correlations related to each compound. However, PCA cannot be used directly to resolve an entire chromatographic mixture, since it is constrained to fulfill maximum variance and orthogonality [10]. PCA can be used to deconvolve spectra though if we force PCA to fulfill maximum variance and orthogonality just in the eluting space of the compound whose spectrum is to be extracted. Then, for each extracted compound profile j in Z (2), a D_j sub-data matrix is determined comprised only of the data of the retention time in which the compound Z_j is eluting (Figure). After that, a PCA is applied for each given window, i.e., each compound profile. Following the same notation, PCA can be mathematically described as:

$$D'_j = YW^T \quad [4]$$

where D'_j ($N \times M$) is the sub-data matrix to decompose, Y ($N \times M$) is the score matrix and W ($M \times M$) the loading or eigenvectors matrix. Matrix Y holds the retention time response of the different decomposed components and matrix W holds the spectra associated with each component, which includes the spectrum of the compound of interest and other unknown noise interferences. In both decomposed matrices, each component may have negative or positive variance. The component of interest associated with the compound whose spectrum is to be extracted is determined by comparing the different covariance responses in matrix Y with the reference profile in Z. This is to determine which component has the highest absolute correlation with the elution profile of the compound of interest. The spectra associated with the selected components are rotated according to the sign of the correlation coefficient with the compounds profile models. OSD algorithm can be summarized in the following steps:

1. Given a Z_j compound elution profile, determine a D_j sub-data matrix comprised only of the data of the retention time in which the compound is eluting.
2. Apply a PCA over D_j. The result is a score matrix Y and loading matrix W.
3. Determine the correlation coefficient between Z_j and each component in Y and select the component h with the highest absolute correlation value.
4. Select the component h in W, rotate W_h according to the sign of the previous determined correlation coefficient, and clip to zero all the negative values. W_h is now considered to be the spectrum of Z_j.

OSD uses a PCA-based approach in order to avoid the use of an LS regressor, which finds difficulties in discriminating noise and the compound ions that are being regressed, which itself

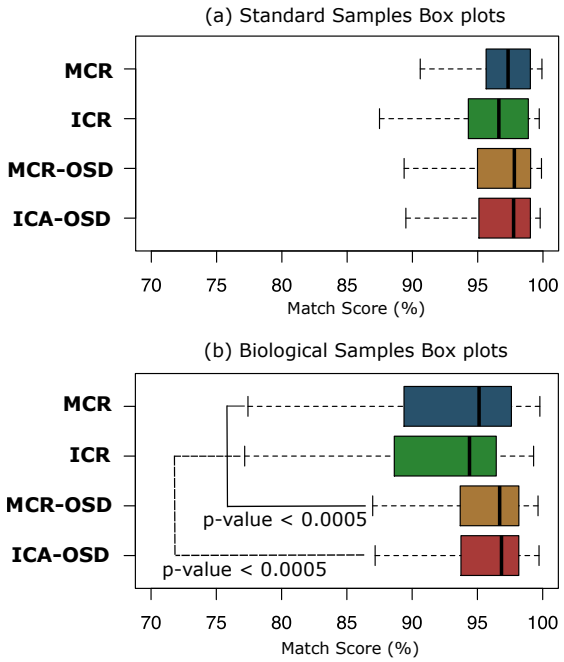

Fig. 2. Match score box plots. (a) The match score boxplot for the case of pure standards dataset and (b) for the case of biological samples dataset. Outliers in the boxplot are not shown. The ρ-values were determined with a paired wilcoxon test, with an alternative hypothesis that the OSD method performs better than LS. The sample size N was of N=152 for (a) and of N=80 for (b).

Table 1. Identification score results for the human serum and urine samples.

Name	ICA OSD	MCR OSD	ICR	MCR
Serum				
Leucine (1TMS)	99.61	99.26	86.55	89.26
Proline (1TMS)	99.34	99.49	95.42	95.60
Urea (2TMS)	98.45	96.78	95.77	95.24
Isoleucine (2TMS)	98.83	97.20	94.63	96.07
Proline (2TMS)	98.01	98.13	95.08	95.62
Glycine (3TMS)	99.25	99.26	98.56	98.60
Serine (3TMS)	98.18	98.24	96.77	96.81
Allo-threonine (3TMS)	97.35	94.67	87.52	96.21
Methionine (2TMS)	92.24	96.22	85.22	84.85
Aspartic acid (3TMS)	96.51	94.61	87.03	88.91
Phenylalanine (1TMS)	98.65	98.42	99.06	98.99
Cysteine (3TMS)	93.84	93.68	72.12	72.88
2-oxo-glutaric acid (2TMS)	87.48	87.43	73.48	74.40
Proline [+CO2] (2TMS)	98.78	98.74	98.28	98.53
Phenylalanine (2TMS)	97.35	97.01	95.11	95.09
Ornithine (3TMS)	98.22	98.19	97.47	97.91
Ornithine (4TMS)	98.21	98.19	98.92	98.99
Citric acid (4TMS)	96.81	96.78	95.30	95.27
Tyrosine (2TMS)	96.73	96.76	95.54	95.04
Myo-inositol (6TMS)	97.92	97.98	96.19	98.03
Cholesterol (1TMS)	92.58	92.23	92.84	92.23
Urine				
Urea (2TMS)	94.26	97.20	91.19	91.17
2-oxo-glutaric acid (2TMS)	80.26	77.99	73.12	73.26
Citric acid (4TMS)	94.41	90.67	88.26	88.83
Myo-inositol (6TMS)	90.74	91.10	91.97	95.35

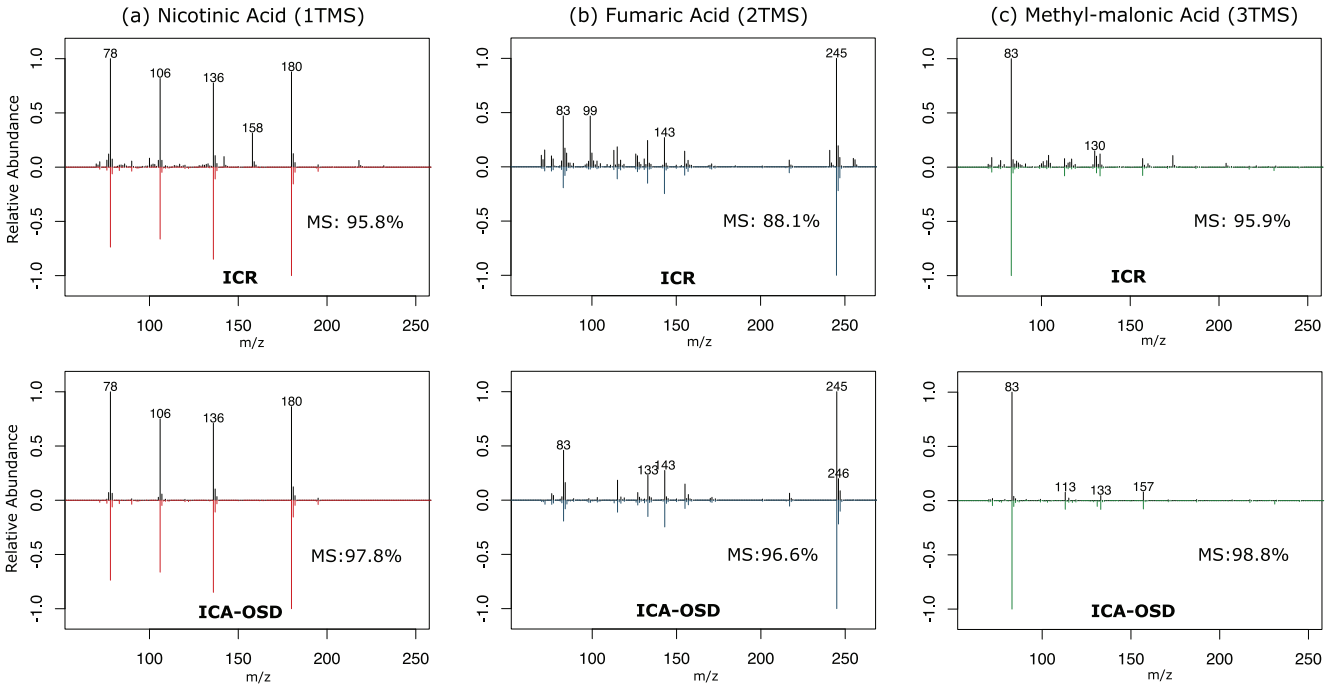

Fig. 3. Comparison of the standards dataset extracted spectra (black and positively displayed) and the reference GMD spectra (color and negatively displayed). Qualitative spectra differences can be seen between least squares (ICR) and OSD (ICA–OSD) approaches. The extracted spectra by ICR and ICA–OSD are shown in black for (a) nicotinic acid, (b) fumaric acid and (c) methyl-malonic acid. The reference spectra (color) are shown in the same axis, negatively rotated, for better visual appreciation. The match score (MS) is noted in each plot.

may introduce a bias to the LS regressors. This effect results in the extraction of the spectra with fragments that may not belong to the true compound spectrum or its intensity is over or underestimated. In OSD, principal component analysis is proposed to improve this limitation and to take advantage of the multivariate nature of GC/MS data. The difference in the application of PCA instead of NNLS resides in the fact that the PCA model takes into account the inherent noise always present in real data and which may have not been included in the ICA or MCR model.

Determination of number of components. Both MCR and ICA/ICR require a fixed number of components, also known as factors, to define their respective models. This parameter clearly affects the ICA or MCR outcome, as a correct estimation of components in the mixture leads to the construction of a model which better fits in data. In this study, a cross-validation approach was used to assure an appropriate determination of the number of components, which was implemented by the following steps: (i) Similarly to [31], divide the D matrix into D_{even} and D_{odd}. Each matrix contains every second row (scans) in D, and thus, all the columns (m/z channels) are preserved. (ii) Compute PCA over D_{even} and determine a L1 matrix containing the PCA loadings. (iii) For each column j in L_1 matrix, determine a matrix $T = [l_1, l_2, ..., l_j]$ containing all the L_1 columns from 1 to j, (iv) determine a rotation matrix T_p and compute the S_2 scores over D_{odd} (5), (v) project the variance explained by S_2 scores into D_{odd} by constructing the M_2 matrix (6). For each iteration j determine the residual sum of squares (RSS) error between D_{odd} and M_2.

$$T_p = (T^T T)^{-1} T \qquad \Rightarrow \qquad S_2 = D_{odd} T_p^T \qquad [5]$$

$$M_2 = S_2 T^T \qquad [6]$$

This method yields a decreasing RSS curve. The proper number of factors is determined when the addition of more components does not significantly decrease the explained variance, i.e., when the RSS error reaches a minimum.

Results and discussion

Pure standards dataset processing. The synthetic sample was processed by all the alternative strategies described. All the approaches led to the correct identification of all 26 metabolites in the original mixture design. Some of the compounds appeared in different trimethylsilyl (TMS) derivatives and therefore a total of 38 compounds was identified. Table S2 (see Supplementary Material) shows the complete list of metabolites identified, along with their match score by the different methods for a quantitative comparison reference. The identification match score is determined by the following steps: first, the normalized spectra for each compound in the four samples is averaged by a simple mean - the total sum of the spectra in each sample -. Then, the match score is determined by the dot product between the average resolved - extracted or empirical - and the reference spectra. The closer the score to one hundred, the more exact and pure the spectra extracted.

Overall identification performance for the studied methods is shown in the box plots of Figure (a). To increase the statistical power, these box plots were constructed by the match score for each metabolite and sample separately — each spectrum of the same compound in each sample was matched independently against the reference spectra —. It is clear that in the capability of the proposed ICR and OSD methods is comparable to the best extended MCR–ALS. Still, some qualitative differences between ICR and ICA-OSD extracted spectra can be appreciated when visually comparing the empirical (resolved) and reference spectra of certain compounds, specially in co-elution situations. Compounds showing important qualitative spectral differences between methods include nicotinic acid, fumaric acid and methyl-malonic acid (Figure 3), for which the OSD approach performs a better isolation of the compound-related ions from the other ions or fragments product of the co-elution with neighboring compounds. In Figure 3 (a), the OSD approach is able to discard the ion m/z number 158 for nicotinic acid as it is an interference due to the co-elution with isoleucine. The same observation can be seen in Figure 3 (b) where the ion number 99 for fumaric acid is detected as an outlier by the OSD approach and discarded from its resolved spectrum; this interference occurs as fumaric acid in co-elution with uracil (See Table S2 of Supplementary Material). Also, Figure 3 (c) shows the case of methyl-malonic acid, for which OSD extracted purer spectra, specially at low ion intensity levels.

Biological samples processing. In this case, the methods under study were tested in biological samples, where compounds appear in very low concentrations and with the interference of a biological matrix. Processing of the human serum and urine samples by the different methods led to the extraction of a total of an average of 230 compounds or components per sample by the OSD approaches, ICR and MCR. From all of them, 15 metabolites from the original pure standards experiment, and two that were not included in the standards dataset, were identified in different TMS derivatives, so a total of 25 compounds were found (Table) - 21 in human serum and 4 of them both in serum and urine -.

Raw data pre-processing included signal filtering using a Savitzky–Golay filter of third order with a 1.1 seconds window length, i.e., half the average peak width. Baseline was removed using a three-step spline interpolation. For each m/z channel, first, (i) a running minimum filter was used with window length 10 times the average peak width (k_{filter}) and from the resulting signal Υ_{min} the baseline standard deviation was determined (σ_b). After that, (ii) a same window length running medians filter was applied, and the resulting signal was Υ_{base}. The running medians filtered signal outcomes a good approximation of the underlying baseline, but to refine it and to avoid outliers each point in Υ_{base} was constrained not to have intensity above Υ_{min} plus σ_b. Finally, (iii) a spline interpolation was applied - with $k_{filter}/2$ degrees of freedom - to smooth Υ_{base}. The smoothed Υ_{base} was subtracted from the original raw data.

An overall identification capability for the studied methods is shown in the box plots of Figure (b). In the biological dataset, the OSD implementations display a more accurate identification of the metabolites in terms of match score and major qualitative differences between the regression (ICR/MCR–ALS) and OSD approaches can be observed. Compounds showing an important match score enhancement between least squares and OSD methods include proline (1TMS) and (2TMS), serine, methionine, aspartic acid, 2-oxoglutaric acid, cysteine, phenylalanine or urea.

Compounds showing important qualitative and quantitative differences between the least squares and OSD approaches include isoleucine, urea, aspartic acid and cysteine (Figure 6). Figure 6 shows that isoleucine low intense interfering ions are removed in the OSD approach. In the case of urea, the spectra is structurally the same between methods but in the OSD approach, the intensities of their ions are closer to the pure spectrum values, and this enhances the match score for the OSD case. Figure 6 also shows that m/z signals 128 and 176

for aspartic acid and 91 and 120 for cysteine are clearly interfering with the underlying pure spectrum of the compound, as the least squares approach is not able to diminish the signal disturbance. On the contrary, the OSD approach is able to deconvolve or discard those signals, and to correct their intensity so that they are closer to the pure spectrum value. This also reveals that OSD is not only an m/z classifier but also has a distinct multivariate deconvolution property. OSD is not only able to discard those m/z signals unrelated to the compound of interest, but is also able to correct, and therefore deconvolve, the intensity of the m/z response. This deconvolution property, a product of the benefits of the application of multivariate over univariate methods, is specially observable in the case of urea or cysteine (Figure 6) but also occurs in the remainder of the cases.

To compare the multivariate deconvolution capacity of the different approaches, the euclidean error distance was computed for all the normalized spectra and methods (Figure). For each compound, the euclidean distance was computed between the m/z of each reference spectra and the m/z of the different empirical spectra by each method, as described in the Supplementary Information. The figure shows that both ICA–OSD and MCR–OSD methods appear closer to the original spectra, since their distances are generally smaller. These results confirm that OSD acts as a multivariate method for spectra deconvolution.

In some cases, the use of OSD led to a decrease of the match score in comparison with LS, this occurs in the case of phenylalanine (1TMS) or myo-inositol (urine). This can be explained as the LS approaches are more conservative, since they do not make any presumption whether a certain fragment belongs or not to the compound spectrum being extracted. Therefore, the OSD approach may fail in detecting covariability between ions which may lead to an incorrect association of the true fragments of the compound. Both ICA-OSD and MCR-OSD exhibit similar performance as can be observed from Table S2 (Supplementary Materials) and Table , but there exist some differences in the match score for certain compounds between both OSD methods. This can be explained as the only input for OSD is the elution profile, previously determined in this study by MCR or ICA. Consequently, the elution profiles determined modify the amount of variance captured by PCA, including the amount of variance related to the spectra to be extracted, and the amount of outlier variance from neighboring compounds or noise, and this clearly determines the purity of the eventual extracted spectra. This thought, is the main advantage of OSD, as it is able to deconvolve the spectrum for a certain compound only with the shape of its elution profile, and therefore independent of the quality of the elution profiles extracted for the rest of the compounds.

Execution time comparison. Finally, the execution time differences between ICA–OSD, MCR–OSD, ICR and MCR are shown in Figure . Each method was tested by processing 2000 scans of raw data (3.3 min of sample) with a range of 566 m/z fragments. These bar plots show the mean speed of execution per scan. From this picture, it can be appreciated that both ICR and ICA–OSD offer the most rapid processing of the chromatogram. The total time difference between the ICA- and MCR-based methods becomes more important as the number of samples to process increases. Data were processed using a 2.4 GHz Intel Core 2 Duo processor with 4 GB of 1067 MHz DDR3 RAM.

Conclusion

This paper demonstrates the capability and suitability of independent component regression (ICR) for GC–MS compound identification as an alternative to multivariate curve resolution. The results given by ICR are comparable to the results given by MCR, but ICR is is superior in terms of execution time. This is of special interest in metabolomics due to the high amount of data that GC–MS currently generates and the quantity of samples that are analyzed in metabolomics experiments. Also, a novel OSD approach using principal component analysis as an alternative to the traditional least squares approach is introduced, allowing the extraction of refined spectra when compounds elute under the influence of biological matrices, compound co-elution or other types of noise.

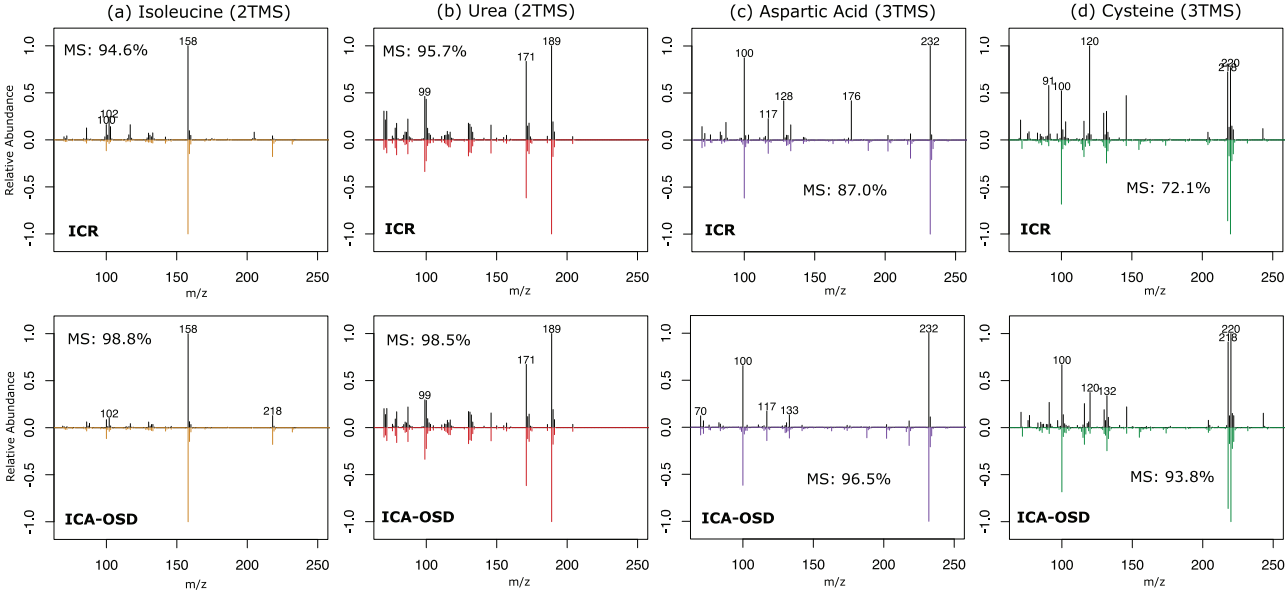

Fig. 6. Comparison of the extracted spectra (black) and the reference GMD spectra (color) in the biological samples. Significant qualitative and quantitative differences can be appreciated between least squares (ICR) and OSD (ICA–OSD) approaches. The extracted spectra by ICR (top row) and ICA–OSD (bottom row) are shown in black for (a) isoleucine, (b) urea, (c) aspartic acid and (d) cysteine. The reference spectra are shown in the same axis for a better visual appreciation. The match score (MS) is noted in each plot.

ACKNOWLEDGMENTS. The authors acknowledge the Centre for Omic Sciences (COS) for providing the standards and the samples, and especially to Dr. R. Ras and Mrs. S. Marine, from COS, for the invaluable scientific counselling and lab assistance. The authors also acknowledge Dr. M. Vinaixa, Dr. O. Yanes and Dr. S. Samino from YanesLab (URV), Dr. F. Fernandez-Albert from Polytechnical University of Catalonia (UPC), Mr. M. Woldegebriel and Mr. M. Lopatka from Universiteit van Amsterdam for their support and for providing constructive advice and comments. This research was partially funded by MINECO grant TEC2012–31074, TEC2013–44666–R and TEC2014-60337-R (to AP). CIBER–BBN is an initiative of the Spanish ISCIII. X.D. also acknowledges the University Rovira and Virgili Martí and Franquès grants for the financial support.

Availability: The *osd* R package is freely available on the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=osd.

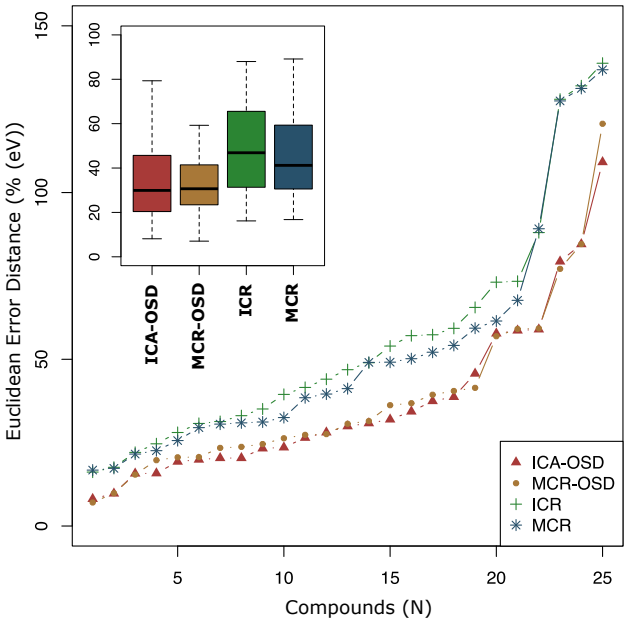

Fig. 4. Euclidean error distance curves. This shows how close each compound is to the original spectrum in terms of relative error. This graphic assists the evaluation of the deconvolution capability between the methods compared. Outliers in the boxplot are not shown. The ρ-values for the euclidean error distances between LS and OSD approaches show that those differences are statistically significative (ρ-value $<$ 0.0005).

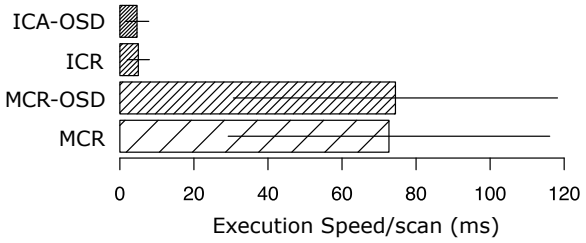

Fig. 5. Time comparison between methods. The barplot shows the mean and standard deviation speed of execution, in miliseconds, necessary to proces one scan of data by each method.

1. G.J. Patti, O. Yanes, G. Siuzdak Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews Molecular Cell Biology, 13 (2012) 63–269.
2. Aihua Zhang, Hui Sun, and Xijun Wang. Serum metabolomics as a novel diagnostic approach for disease: a systematic review. Analytical and Bioanalytical Chemistry, 404 (2012) 1239–1245.
3. C.Ruckebusch, L.Blanchet. Multivariate curve resolution: a review of advanced and tailored applications and challenges. Analytical Chimica Acta, 765 (2013) 28–36.
4. A. de Juan, J. Jaumot, R. Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Anal. Methods, 6 (2014) 4964.
5. P. Gemperline Practical Guide to Chemometrics, second ed., CRC Press, 2012.
6. J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. Radar and Signal Processing, IEE Proceedings F, 140 (1993) 362–370.
7. G. Wang, Q. Ding, Y. Sun, L. He, X. Sun. Estimation of source infrared spectra profiles of acetylspiramycin active components from troches using kernel independent component analysis, Spectrochim. Acta A: Mol. Biomol. Spectrosc, 70 (2008) 571–576
8. M. Toiviainen, F. Corona, J. Paaso, P. Teppola. Blind source separation in diffuse reflectance NIR spectroscopy using independent component analysis, Journal of Chemometrics 24 (2010) 514–522.
9. I. Schelkanova, V. Toronov. Independent component analysis of broad- band near-infrared spectroscopy data acquired on adult human head, Biomed. Opt. Express 3 (2012) 64–74
10. Y.B. Monakhova, S.S. Kolesnikova, S.P. Mushtakova. Independent component analysis algorithms for spectral decomposition in UV/VIS analysis of metalcontaining mixtures including multimineral food supplements and platinum concentrates, Anal. Methods 5 (2013) 2761–2772
11. I. Toumi, S. Caldarelli, B. Torrsani. A review of blind source separation in NMR spectroscopy. Prog. Nucl. Magn. Reson. Spectrosc. 81 (2014) 37–64
12. G.Wang, Q.Ding, Z.Hou Independent component analysis and its applications in signal processing for analytical chemistry. TrAC – Trends Anal. Chem 27 (2008) 368–376.
13. Guoqing Wang, Wensheng Cai, and Xueguang Shao. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. Chemometrics and Intelligent Laboratory Systems, 82 (2006) 137–144.
14. Guoqing Wang, Wensheng Cai, and Xueguang Shao. A post-modification approach to independent component analysis for resolution of overlapping GC/MS signals: from independent components to chemical components. Sci. China Ser. B: Chem. 50 (2007) 530–537.
15. Zhichao Liu, Wensheng Cai, and Xueguang Shao. Sequential extraction of mass spectra and chromatographic profiles from overlapping gas chromatography-mass spectroscopy signals. Journal of Chromatography A 1190 (2008) 358–364.
16. X. Shao, Z. Liu, W. Cai. Extraction of chemical information from complex analytical signals by a non-negative independent component analysis. Analyst 134 (2009) 2095–2099.
17. Xueguang Shao, Zhichao Liu, and Wensheng Cai. Resolving multi-component overlapping GC-MS signals by immune algorithms. TrAC Trends in Analytical Chemistry, 28 (2009) 1312–1321.
18. J.V. Stone. Independent Component Analysis: A Tutorial Introduction, A Bradford Book, Cambridge, MA, 2004
19. H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry? Trends Anal. Chem. 31 (2012) 134–143.

20. X. Shao, W. Wang, Z. Hou, W. Cai A new regression method based on independent component analysis. Talanta 69 (3) (2006) 676–680
21. Jan Hummel, Nadine Strehmel, Joachim Selbig, Dirk Walther, and Joachim Kopka. Decision tree supported substructure prediction of metabolites from GC-MS profiles. Metabolomics, 6 (2010) 322–333.
22. N.Psychogios,D.D.Hau,J.Peng,A.C.Guo,R.Mandal,S.Bouatra,I.Sinelnikov,R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T.L. Pedersen, S.R. Smith, F. Bamforth, R. Greiner, B. McManus, J.W. Newman, T. Goodfriend, D.S. Wishart, The human serum metabolome, PLoS ONE 6 (2011) e16957
23. W.B.Dunn,D.Broadhurst,P.Begley,E.Zelena,S.Francis-McIntyre,N.Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Human Serum Metabolome (HUSERMET) Consortium, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat. Protoc. 6 (2011) 1060–1083.
24. Y. Ni, Y. Qiu, W. Jiang, K. Suttlemyre, M. Su, W. Zhang, W. Jia, X. Du, ADAP- GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies, Anal. Chem, 84 (2012) 6619–6629.
25. A de Juan, Y Vander Heyden, R Tauler, and D. L Massart. Assessment of new constraints applied to the alternating least squares method. Analytica Chimica Acta, 346 (1997) 307–318.
26. Katty X. Wan, Ilan Vidavsky, and Michael L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. Journal of the American Society for Mass Spectrometry, 13 (2002) 85–88.
27. A. Savitzky, M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures Anal. Chem, 36 (1964) 1627–1639.
28. D.N.Rutledge, D. Jouan-Rimbaud Bouveresse. Independent components analysis with the JADE algorithm, TrAC – Trends Anal. Chem. 50 (2013) 22–32.
29. Ivo H. M. van Stokkum, Katharine M. Mullen, and Velitchka V. Mihaleva. Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS. Chemometrics and Intelligent Laboratory Systems, 95 (2009) 150–163.
30. Y.B. Monakhova, A.M. Tsikin, T. Kuballa, D.W. Lachenmeier, S.P. Mushtakova. Independent component analysis (ICA) algorithms for improved spectral deconvolution of overlapped signals in 1H NMR analysis: application to foods and related products. Magn. Reson. Chem. 52 (2014) 231–240.
31. S. Peters, H.-G. Janssen, G. Vivo-Truyols. A new method for the automated selection of the number of components for deconvolving overlapping chromatographic peaks. Anal. Chim. Acta 799 (2013) 29–35.