Title

Extraction of artefactual MRS patterns from a large database using non-negative matrix factorization

Running head (70 characters)

Unsupervised artefact detection in in vivo brain tumor MRS data

Authors and institutions

Yanisleydis Hernández-Villegas[1,2,3], Sandra Ortega-Martorell[5,2], Carles Arús[1,2,3], Alfredo Vellido[4,2*], Margarida Julià-Sapé[2,1,3]*.

[1]Departamento de Bioquímica y Biología Molecular, Universidad Autónoma de Barcelona (UAB); [2] Centro de Investigación Biomédica en Red (CIBER); [3]Instituto de Biotecnología y de Biomedicina (IBB), Universidad Autónoma de Barcelona (UAB); [4]SOCO research group at Intelligent Data Science and Artificial Intelligence Research Center (IDEAI-UPC), Universitat Politècnica de Catalunya-BarcelonaTech; [5]Department of Applied Mathematics, Liverpool John Moores University.

Abstract (300 w)

Despite the success of automated pattern recognition methods in problems of human brain tumor diagnostic classification, limited attention has been paid to the issue of automated data quality assessment in the field of MRS for neuro-oncology. Beyond some early attempts to address this issue, the current standard in practice is MRS quality control through human (expert-based) assessment. One aspect of automatic quality control is the problem of detecting artefacts in MRS data. Artefacts, whose variety has already been reviewed in some detail and some of which may even escape human quality control, have a negative influence in pattern recognition methods attempting to assist tumor characterization. The automatic detection of MRS artefacts should be beneficial for radiology as it guarantees more reliable tumor characterizations, as well as the development of more robust pattern recognition-based tumor classifiers and more trustable MRS data processing and analysis pipelines. Feature extraction methods have previously been used to help distinguishing between good and bad quality spectra to apply subsequent supervised pattern recognition techniques. In this study, we apply feature extraction differently and use a variant of a method for blind source separation, namely Convex Non-Negative Matrix Factorization, to unveil MRS signal sources in a completely unsupervised way. We hypothesize that, while most sources will correspond to the different tumor patterns, some of them will reflect signal artefacts. The experimental work reported in this paper, analyzing a combined short and long echo time ^{1}H-MRS database of more than 2000 spectra acquired at 1.5T and corresponding to different tumor types and other anomalous masses, provides a first proof of concept that points to the possible validity of this approach.

Keywords (3-6, from menu provided)

Post-acquisition Processing < Methods and Engineering

MR Spectroscopy (MRS) and Spectroscopic Imaging (MRSI) Methods < Methods and Engineering

Spectroscopic quantitation < MR Spectroscopy (MRS) and Spectroscopic Imaging (MRSI) Methods < Methods and Engineering

Artifacts and corrections < Acquisition Methods < Methods and Engineering

Acronyms:

a.u: Arbitrary units
AQC: Automated quality control
BSS: Blind Source Separation
CC: Coding coefficients.
CNMF: Convex Non-Negative Matrix Factorization
CNN: Convolutional Neural Network
GQ: Good quality
ICA: Independent Component Analysis
INTERPRET: International network for Pattern Recognition of Tumors Using Magnetic Resonance.
MRS: Magnetic Resonance Spectroscopy
MRSI: MRS Imaging
LTE: Long echo time
NMF: Non-negative Matrix Factorization
ppm: Parts per million
PR: Pattern Recognition
PRESS: Point- resolved spectroscopy sequence
RF: Random Forest
SNR: Signal-to-noise ratio
STEAM: Stimulated echo acquisition mode sequence
STD (+/-): Standard deviation (plus/minus)
STE: Short echo time
SV: Single voxel

Word count: 5653

Introduction

Scant attention has been paid to the issue of automated data quality assessment in the field of MRS for neuro-oncology (1) and, although recent studies have started addressing this issue, often using supervised pattern recognition (PR) approaches, the current standard in practice is quality control through human assessment (2). One reason for this may be the lack of the type of biocuration standards that begin to be common in other life sciences fields such as genomics and, to a lesser extent, proteomics (3). Further reasons include the fact that MRS data in this area are scarce and fragmented. Fragmentation is both geographical and institutional, as the effort of gathering multi-center and international data is hindered by different barriers. The clinical centers who are ultimately responsible for data acquisition have few obvious incentives to even partially transfer the control of their data to third parties, and such parties, who should be responsible for managing multi-center data, either do not exist or lack the ability to sustain such role in a long-term basis. Furthermore, efforts to gather and manage international databases often collide with local legal limitations for the transfer and sharing of this type of personal medical information.

Having said this, it is also true that some research efforts have been made in order to address the problem of MRS automated quality control (AQC) and that this problem has been approached from different perspectives. Early concerns about issues of spectral quality in clinical MRS and the lack of standards for the definition of what makes a spectrum acceptable or not were, for instance, raised in (4). In this review, a list of possible artefacts, many of them difficult to detect even by expert visual inspection, was compiled; several quality assessment quantitative measures were put forward and a number of criteria for spectra rejection were formulated. The need for the definition of quality requirements and goals for ^{1}H-MRS data, as well as for the implementation of measures to guarantee quality standards and the sustained management of data quality have recently been stressed in (2).

Part of the spectra in the current paper were analyzed at a first level in (1), where the quality assessment concerned the immediate step after data acquisition by automatic determination of the signal-to-noise ratio (SNR) in a water-suppressed spectrum and of the line width of the water resonance (water band width, WBW) in the corresponding non-suppressed spectrum. Threshold criteria for the selection of spectra were then empirically determined and additional artefact detection was carried out by human visual inspection.

In recent research (5), AQC was taken to a second level that uses previously validated databases (6-8) as a starting point. In that study, a range of different PR classifiers were trained to mimic human decision making about the quality of spectra from data transformed according to different feature extraction methods. To learn this task, the classifiers used original human quality ratings from both multi-center and local experts as training labels. Classifier performance was subsequently compared with variance in human judgment. This work was in turn inspired by a previous smaller-scale study (9) in which a least squares support vector machine was trained from features extracted by independent component analysis (ICA) to learn to distinguish acceptable from unacceptable spectra. This AQC approach has been recently extended to clinical ^{1}H-MRSI information in (10), where a random forest (RF) classifier was trained on MRSI grids previously labeled as acceptable or non-acceptable by two expert spectroscopists and where, in order to account for potential intra-expert reliability effects, each of the spectra was labeled three times by each expert. A similar approach, also using RF as the classifier of choice, was earlier presented in (11). Note that all these approaches aim to replicate human decision in a data-based automated form, but do not attempt to assess quality dispensing with human prior assessment.

An alternative approach to AQC attempted to distinguish potentially problematic spectra using an outlier analysis (12). A fully unsupervised manifold learning technique was used to model the data distribution and a shortlist of spectra that did not conform to it was obtained. This shortlist of quantitatively atypical cases was inspected by experts to distinguish between naturally atypical spectra and spectra with artefact related anomalies. The categorization of the artefacts in those singled-out cases was subsequently carried out individually and in detail by human experts. The purpose of our present study was to apply a totally unsupervised PR approach on the largest multicenter collection of single voxel (SV) spectra of brain tumors available to date, to identify artefactual MRS patterns in a way which is expert-interpretable.

In this study, we use feature extraction in a different manner for the purpose of MRS AQC. The proposed approach is based on a method of the blind source separation family (to which ICA also belongs), namely Non-negative Matrix Factorization: NMF (13), and, more specifically, one of its variants known as Convex NMF: CNMF (14). NMF was originally developed (13) as a method for the estimation of the latent (unobservable) sources of image, but it can be used with any kind of signal assumed to consist on a combination of such sources. If applied to an MR spectrum, the goal is discovering the hidden signal sources whose weighted combination constitute it., be it tissue types or artefactual patterns.

The rest of the paper is structured as follows: we first describe the dataset used in the experiments, which is the largest multicenter collection to date of SV brain tumor spectra at short and at long TE, obtained at 1.5T. Next, we report the experimental design, with a brief description of how the CNMF algorithm works, and how we designed the descriptive study and evaluated it. Then results for short time of echo (STE) and long time of echo (LTE) are shown separately and discussed. Finally, some conclusions are drawn, and possible future lines of research are outlined.

Materials and methods

Data acquisition and processing

The data analyzed in this study are the same that were reported in detail in (5). In brief, these are SV spectra from human brain tumors, acquired in 1.5T scanners from three different manufacturers (GE, Siemens and Philips) and different scanner models during the period 1994-2009. They were downloaded from the multi-center INTERPRET (6,8,15) and eTUMOUR (7) databases and processed with the INTERPRET data manipulation software (8,16) and parameters, with a further realignment correction as reported in (5). Note that this processing included setting the region between [4.2, 5.1] ppm to zero values, and the final processed spectrum consisted of 512 frequency points. The total number of STE (20-32ms) spectra acquired with PRESS or STEAM, processed and available for further analysis was 1,180. The corresponding total number of LTE (135-144 ms) spectra acquired with PRESS was 977. For this study, the original quality ratings by expert spectroscopists were not used, although they were available with the data matrices from (5). Regarding the quality as assessed by the expert spectroscopists' panels for STE, 982 spectra were deemed to be good and 198 bad quality spectra, whereas for LTE, 828 were deemed to be good and 149 bad (5) - see Table 1 for details.

The available spectra correspond to the variety of pathologies gathered in the databases. The distribution of spectra by tumor type and echo time is shown on Table 1. Some of the artefacts known to be present in the spectra include (although are not limited to) low SNR and/or bad water suppression (5). For evaluation (see section further on), seven classes or superclasses (brain tumor groupings) were considered: low grade gliomas (including astrocytoma, oligodendroglioma and oligoastrocytoma of WHO grade II), aggressive tumors (which included glioblastoma and metastasis), meningioma, lymphoma, primitive neuroectodermal tumors (PNET), astrocytoma WHO grade III, abscess as well as normal brain, as in (8,16).

Experimental design

Sources or archetypical spectral patterns were extracted using CNMF (14). This method generalizes NMF by admitting negative values in the observed data. Note that some of the spectra in the database include inverted peaks with such negative values. The optimal number of sources to be extracted is not known a priori (17). Although this would be a relevant problem in a more general experimental setting, it is not a relevant one in this study, as we are interested in the exploration of the existence of signal artefacts across a wide range of source number values. For this reason, a descriptive study extracting from 4 to 20 sources per TE was set up. Extractions start at 4 sources as the minimum necessary to maintain a correspondence between the sources (or groups of sources) and the main types of tissue, according to (17).

CNMF works by factorizing the observed data matrix X (of dimensions $D \times N$, where D is the dimension of the data -512 points or spectral frequencies in our case- and N is the number of samples: 1,180 spectra at STE plus 977 at LTE) into two matrices: F (the matrix of extracted sources, of $D \times K$ dimensions, where K is the number of sources -from 4 to 20 in the reported experiments-) and G (the mixture or coding matrix, of dimensions $N \times K$, where the values in a column are the weights associated with a source or base vector for each spectrum). The product of these two matrices provides a good approximation to the original data matrix. It is important to note that the values in G are all non-negative and, therefore, each spectrum can be seen as a weighted combination of sources acting as data centroids. Therefore, we are making the important assumption that an MR spectrum is the measurable manifestation of the weighted combination of non-directly measurable (hidden or latent) signal sources. Furthermore, F is constrained to lie in the column space of the input data X, so that the CNMF formula can be written as in Eq. 1:

$$X_{\pm} \approx F G_{+}^{T} , \qquad \textbf{Equation 1}$$

where $F = X_{\pm} W_{+}$. This leads to $= G(G^T G)^{-1}$; the ± subscript represents a mixed-sign data matrix and the + subscript indicates that the matrix is non-negative. W (of dimensions $N \times K$) is an auxiliary adaptative weight matrix that fully determines G.

Matrix G is also called the *mixing matrix*, as it holds the coefficients (or coding coefficients, CC) to recompose a specific data sample. The CC value of each column in the mixing matrix therefore provides us with an estimation of the degree of contribution of each of the sources to each reconstructed spectrum. Each spectrum i (of N) is represented as the linear combination of the k^{th} source (out of K) and the CC G_i, as described by Eq. 2:

$$X_i = F_1 G_{i1} + \cdots + F_k G_{ik} + \cdots + F_K G_{iK} \qquad \textbf{Equation 2}$$

NMF methods unavoidably converge to local minima. As a result, the NMF bases will be different for different initializations. In this study, we use the *k*-means++ algorithm (18) for initialization. CNMF is based on iterative update algorithms, just like the original NMF, in which the factors are updated alternately until convergence (19). The algorithm works as follows:

Step 1: Initialize G and W. This is achieved here with the *k*-means++ algorithm, as in (18), aiming to ensure that the algorithm starts from values close to the actual data centroids.

Step 2: Update G, leaving W fixed, using the rule in Eq. 3:

$$G_{ik} \leftarrow G_{ik} \sqrt{\frac{[(X^T X)^+ W]_{ik} + [G W^T (X^T X)^- W]_{ik}}{[(X^T X)^- W]_{ik} + [G W^T (X^T X)^+ W]_{ik}}} \qquad \textbf{Equation 3}$$

Where $(\cdot)^+$ is the positive part of the matrix, where all negative values become zeros; and $(\cdot)^-$ is the negative part of the matrix, where all positive values become zeros.

Step 3: W is updated, leaving G fixed using the rule in Eq. 4:

$$W_{ik} \leftarrow W_{ik} \sqrt{\frac{[(X^T X)^+ G]_{ik} + [(X^T X)^- W G^T G]_{ik}}{[(X^T X)^- G]_{ik} + [(X^T X)^+ W G^T G]_{ik}}} \qquad \textbf{Equation 4}$$

Ten repetitions were carried out for each of the 17 source extractions (from 4 sources to 20) at both TEs, since the extracted sources may vary because of the *k*-means++ initialization. This number of repetitions was considered to be enough to calculate the mean and standard deviation (STD) of the sources extracted.

In order to calculate the mean and STD of the sources, we first grouped them by similarity. For this, the Pearson correlation coefficients between each source and all the sources at each repetition were calculated, and those with the highest coefficient values at each repetition were grouped together. The first extraction was chosen as starting point. The obtained sources were graphically represented to allow a first intuitive visual verification of their characteristics. As mentioned in the introduction, we hypothesize that some of the sources would be identified as artefacts, while others will describe prototypical tumor patterns or normal tissue, as the databases from which the spectra are obtained comprise spectra of both good and poor quality.

CNMF was implemented in Python language (20) and run either via Google Cloud Platform, or at the computer cluster at the *Institut de Biotecnologia i Biomedicina* (IBB) in Barcelona, Spain.

Evaluation

The obtained sources were first qualitatively explored by two members of the team who are expert spectroscopists (CA and MJS) and then quantitatively assessed according to different calculated measures with the purpose of finding an automated way to distinguish artefact sources. The quantitative measures include:

- Pearson product-moment correlation coefficients (matrix R in Eq. 5) between the means of each of the matrices created with the sources obtained over 10 repetitions (matrix Y) and the means of the different tumor classes, abscesses and normal tissue from the INTERPRET validated database (matrix Z) (6).

  $$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}}, \qquad \textbf{Equation 5}$$

  where c_{ii}, c_{jj}, c_{ij} are elements of the covariance matrix C of (Y, Z). The values of R belong to the closed interval $[-1,1]$. This measure evaluates whether the extracted sources correlate with known prototypical spectra of different pathologies, or with healthy tissue.

- Euclidean distances between the means of each of the matrices created with the sources obtained over 10 repetitions (Y) and the mean spectra of different tumor classes in the INTERPRET validated database (Z), calculated as $\|Y - Z\|^2$, evaluate the similarity between the extracted sources and the different prototypical spectra of different pathologies, or healthy tissue.

- The CC of the mixing matrix (G) of the means of each of the matrices created with the sources obtained over 10 repetitions (Y). These can be understood as estimates of the concentration/abundance of the constituent signals or sources in the conformation of each spectrum. These will help us to determine how well the sources obtained through convex NMF represent the artefacts.

Results

Here, we report some of the experts' interpretations of the extracted sources. For the sake of brevity, only part of the complete set of results is reported, with some detailed results moved to the supplementary materials.

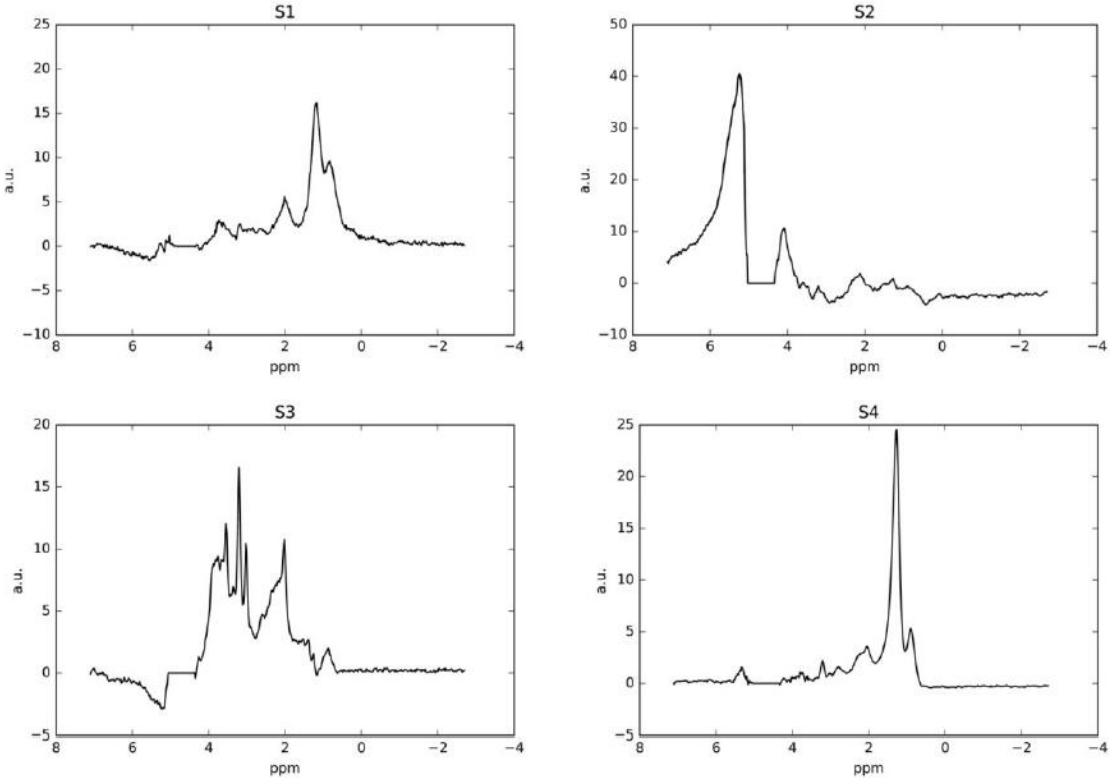


Figure 1 shows the mean and standard deviation (STD) of sources extracted for $K = 4$ (minimum number of sources) at STE. Sources *S1* and *S4* show patterns that resemble those of high-grade glial tumors, characterized by the predominance of mobile lipids (0.9, 1.3 ppm). Source *S3* is similar to low grade glial tumor spectra, in which there is an increase in the Choline peak, a decrease in Creatine and N-acetyl aspartate, and an increase in the Myo-inositol/Glycine peak, with respect to normal brain parenchyma pattern. Source *S2*, instead, can be considered as an artefact due to poor water suppression, which can be observed in the residual water signal around the offset area (4.2-5.1ppm).

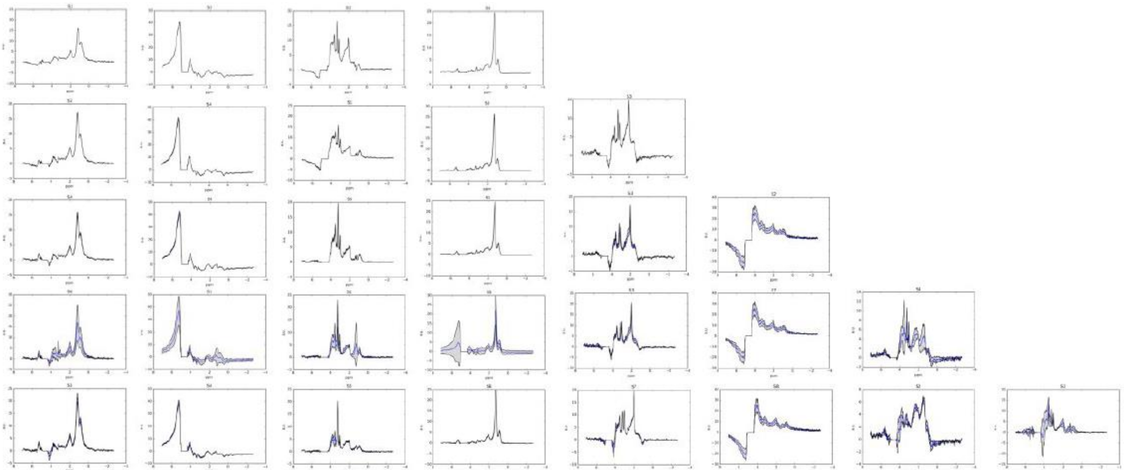


Figure 2 widens the scope and shows the extractions from $K = 4$ to $K = 8$ (by rows) at STE. Sources in column 2 show the poor water suppression artefact, whereas in column 6 poor water suppression and negative intensities/bad water phasing can be observed. This should be considered as an artefactual source, given that spectra at STE are not supposed to have negative values. Column 8 shows a source that is compatible with a combination of artefacts: poor water suppression and spurious echoes (4).

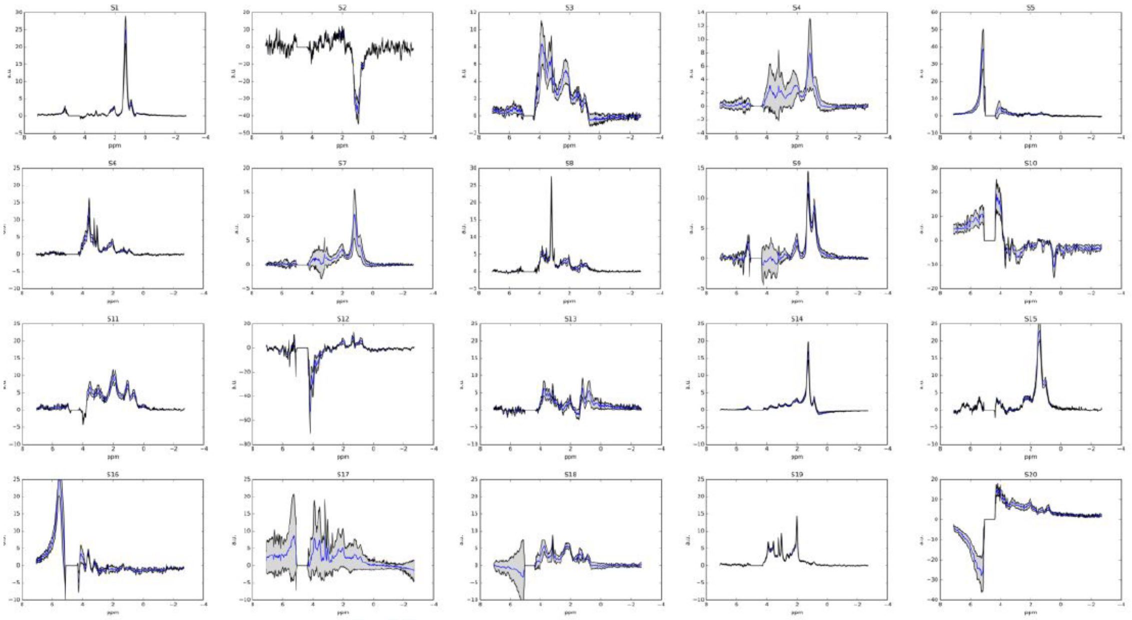


Figure 3 shows the extraction for the maximum of twenty sources. Table 2 displays the consensus expert spectroscopists' evaluation. It can be observed that sources *S1, S4, S7, S9, S14* and *S15* are compatible with high-grade tumors, which is related to the presence of mobile lipid peaks at 0.9 and 1.28 ppm. Amongst these, *S9* shows an uncommon high methyl resonance at ca. 0.9 ppm, compatible with the spectral pattern of some oligodendrogliomas (21,22). *S11*, *S17* and *S18*, even if still interpretable, contain artefactual patterns mainly due to insufficient water suppression- in particular for *S2, S5, S10, S12, S13, S16* and *S20* show clear artefactual patterns, and *S18* is borderline regarding this aspect. It appears that the problem in most them is bad water suppression (*S2, S5, S10, S12, S16, S20*), sometimes only in the downfield side of the suppressed water signal, rarely used for classifier development. It can also be seen that more than one artefact coexists in some instances, for example low SNR (*S2*, *S13*) and spurious echoes (*S2, S13*). The remaining sources have characteristics that match the type of patterns of known tumors, as in *S3* or *S18*, which are compatible with meningioma; *S6*, with low grade glioma; S8, with PNET or astrocytoma grade III, and *S19*, with normal brain. Importantly, all these sources consistently appear and also show little variability throughout all extractions ($K = 4, \ldots, 20$).

With

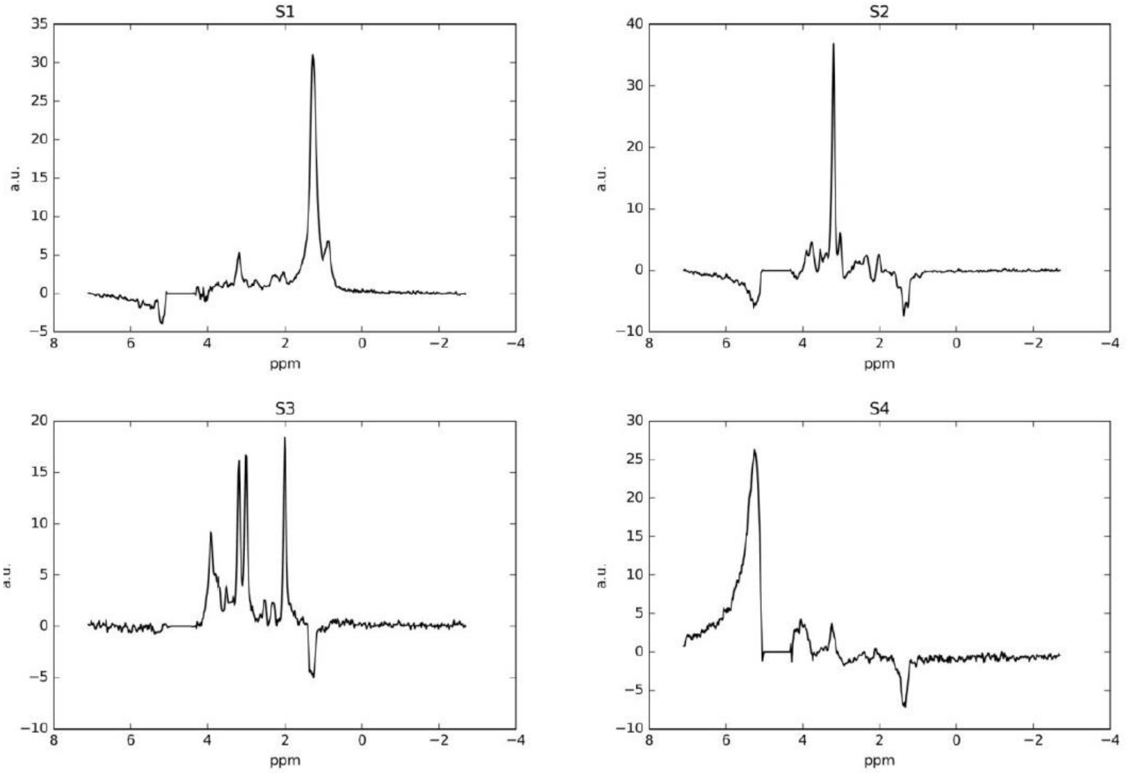


Figure 4, we now move to similar experiments for LTE data sources. It includes the results for the $K = 4$ extraction, where *S1, S2* and *S3* display good quality patterns, while *S4* clearly corresponds to a bad water suppression artefact. *S1* and *S2*, though, also show a small contribution from incomplete water suppression.

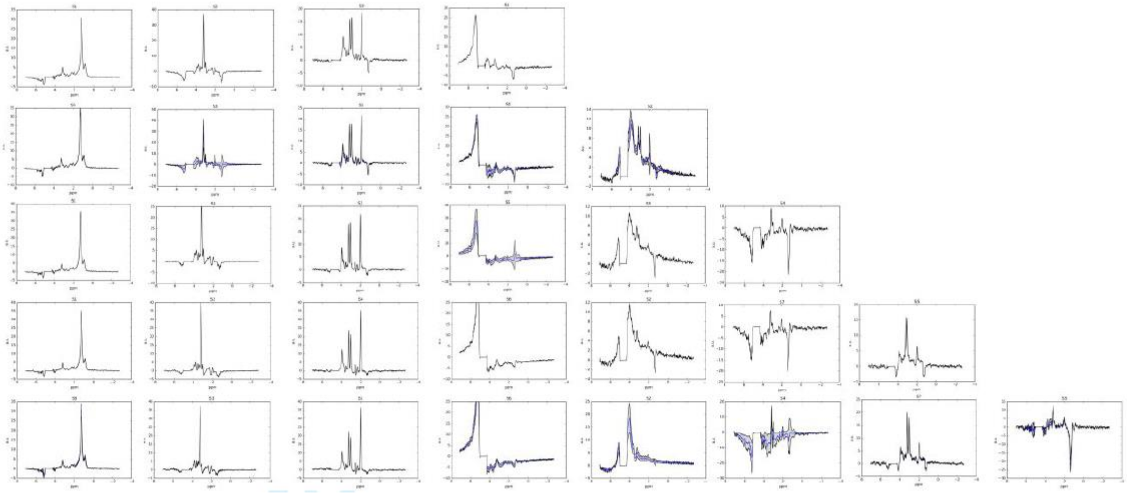


Figure 5 displays extractions from four to eight sources. It can be observed again that some of the sources appear consistently in the different extractions and are the less variable, and that the variability in the solutions increases

when more sources are extracted.

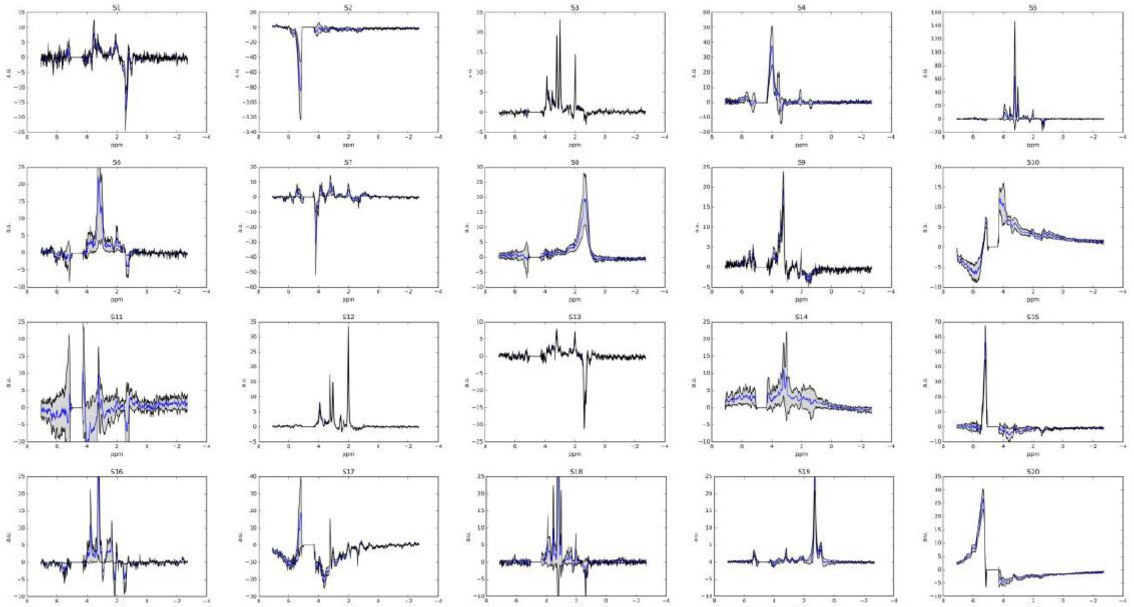


Figure 6 shows the extraction for $K = 20$ at LTE, where it can also be appreciated that only *S3*, *S12* and *S13* show low variation, while the rest of sources show different degrees of variability. Such variability can be assessed in detail from

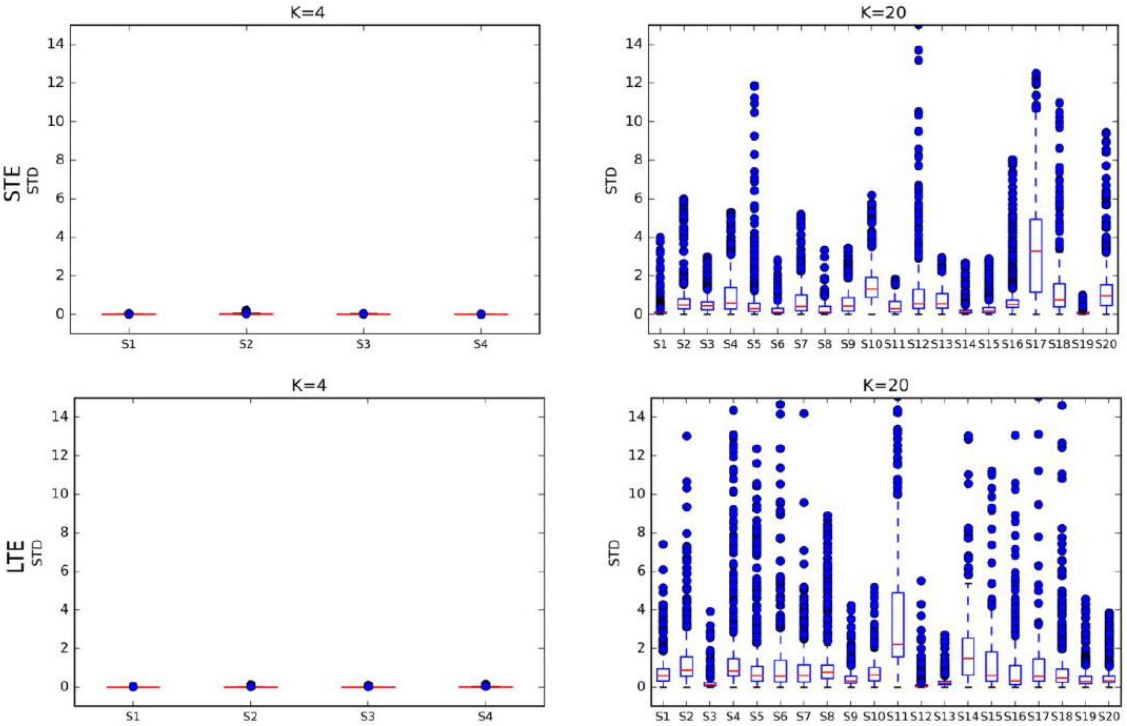
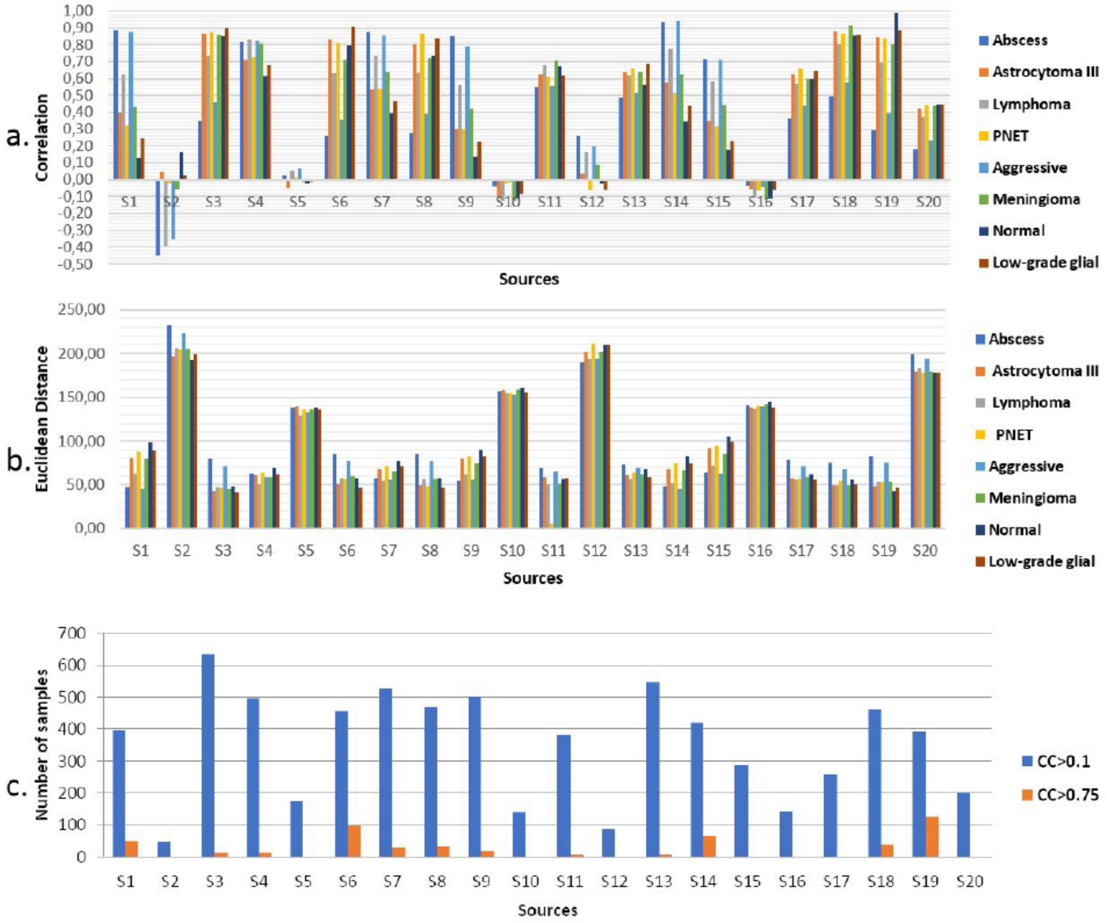


Figure 7, which shows, for STE and LTE, the standard deviation of the different sources in the form of box-plots. These plots provide evidence that the 4-source extraction is the less variable whereas the solutions obtained with the 20-source extraction are rather unstable, although there is a gradient, best seen for the STE sources, between low variability (S19, S11, S13) to large variability (S12, S17, S5, S20). Additionally, the standard deviation of the 20-source extraction solutions at STE is clearly lower than at LTE.

Supplementary Figures 1 to 22 provide the details of the standard deviation for all the extractions at the different TEs, where it can be noted that either 4 or 5 sources at STE and 4 at LTE are optimal in terms of source stability. In general, extractions at STE are more stable than at LTE.



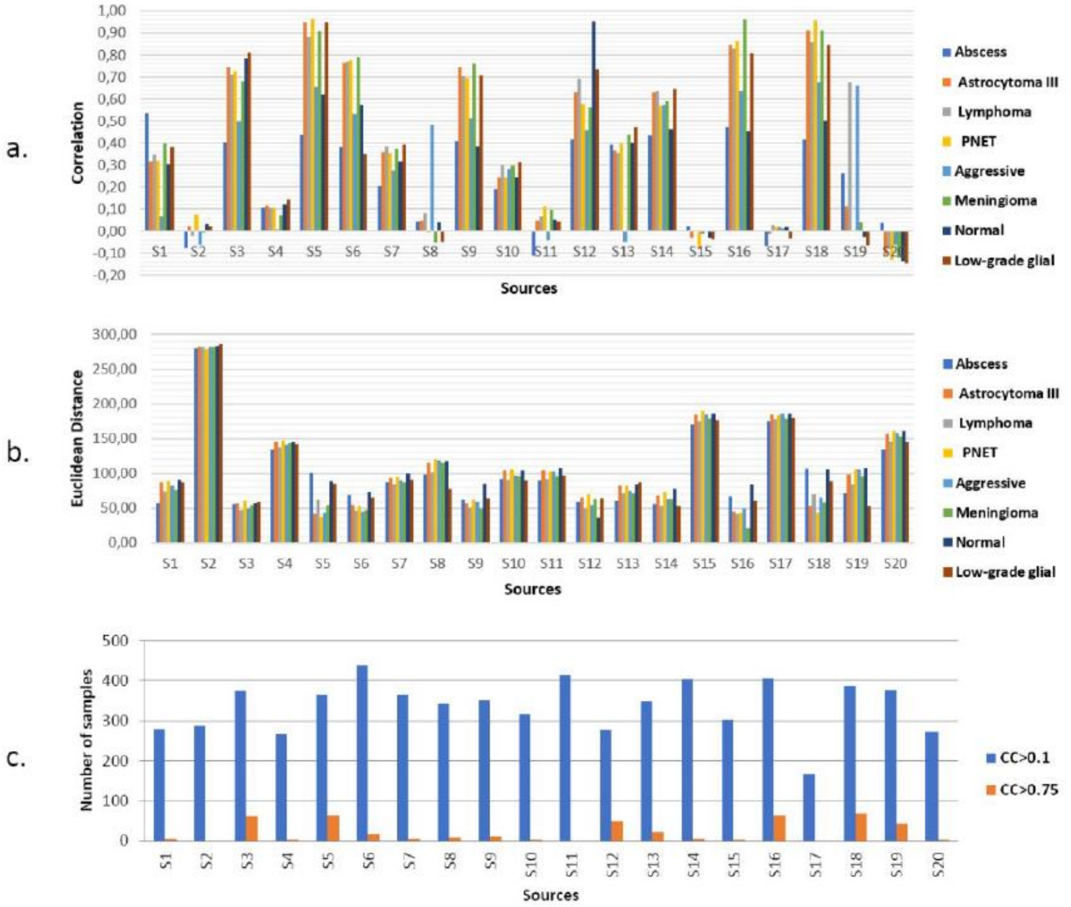


Figure 8 and Figure 9, for STE and LTE respectively, show the correlations and Euclidean distances between the sources obtained at $K = 20$ and the different mean spectra from the INTERPRET database, as well as the CCs. Tables 2 and 3 summarize the results of the different criteria for $K = 20$ from, in turn, data acquired at STE and LTE. As it can be observed, most artefactual sources do not correlate (Pearson < 0.50) with at least one of the compared types; there is a high Euclidean distance between the sources and the compared types and there are no samples with CCs higher than 0.75. The experts also considered that the above-mentioned patterns were artefactual or contained artefacts, in particular for STE.

Figures in the Supplementary materials show the equivalent results for $K = 9, \ldots, 19$, at STE and LTE.

Discussion

In this study, we extracted characteristic spectral patterns in a wholly unsupervised way, i.e. disregarding instrumental quality or tumor type labels. The mathematical approach chosen was CNMF, on the assumption that the observed spectra are the result of a combination of unobserved signal sources.

An alternative approach could have been to apply a technique such as ICA. ICA restricts the sources to be statistically independent from each other (i.e. the occurrence of one does not affect the probability of occurrence of the other), leading to MRS sources that poorly resemble the tissue types involved (23). For this reason, even when ICA has been extensively used to remove artefacts from electroencephalographic recordings (24), we did not consider it our first choice for extracting the kind of artefacts that can be found in MRS data. The non-negativity constraints of NMF, instead, lead to a parts-based representation because they allow only additive, not subtractive, combinations. This parts-based representation is key to explain the success of this BSS method in MRS data. ICA learns holistic (i.e. the whole rather than the sum of its parts) instead of parts-based representations. Amongst NMF variants, we chose to use CNMF as 1) it applies to both nonnegative and mixed-sign data matrices (key for long time of echo –LTE- MRS data),), 2) it has proven to represent better the underlying signals in the data (25,26) as the sources must lay in the convex hull of the data, and 3) CNMF is bound to generate sparse mixing matrices (with many elements taking values close to zero), which is a very useful property that can be exploited in future work in the artefact removal process. The use of NMF and CNMF for the analysis of MRS has already been reported in the field of neuro-oncology (25-28). These methods have mostly been used to detect sources that might be related to specific tissue types in and around the tumor, accounting for the spatial co-existence of tissue types.

Here, the use of CNMF had quite different goals. We hypothesized that, should some of the analyzed MRS data be contaminated by errors in the form of artefacts of different type, some of the sources extracted by CNMF should mostly reflect such artefacts, while the rest of sources would mainly reflect true tissue information. If this hypothesis holds, it follows that the MRS data could be adequately reconstructed from only those sources containing true signal, by removing the artefactual sources from the reconstruction.

As the number of underlying sources in the dataset is not known *a priori*, we performed a descriptive study extracting from four to twenty different sources from the available spectra. Note that the criteria to choose the most appropriate number of sources may be based on strictly quantitative measures, on the radiological interpretability of the extracted sources, or on a trade-off between both approaches. This was not the objective of the current study and, therefore, such number remains to be determined. To address this problem, for example, Laruelo (29) used vertex component analysis (30), Vilamala *et al*. used a Bayesian NMF variant (31), and, in (32), the authors proposed an approach to automatically discard irrelevant sources during the iterative process of matrices decomposition. However, in terms of source extraction stability and according to the reported results, choices of $K = 4 - 5$ for STE and $K = 4$ for LTE seem optimal to represent major tissue and artefact classes.

The experiments were carried out on the largest multicenter SV MRS brain tumor patient database available to date. The results reported in the previous section clearly indicate that some of the sources appear consistently across extractions, no matter the number of sources extracted, and that they correspond to well-defined sources (in the sense that they clearly correspond to either tumor types or to artefacts). The artefactual patterns are mostly different shapes of bad water suppression, as well as low SNR. The bad water suppression artefact is the most conspicuous and appears even in the extraction of only four sources. A recent work by Kyathanahally et al. (33) used a convolutional neural network (CNN, a variant of deep learning model) to detect the ghosting artefact (4), which is very difficult to classify with conventional methods. It is difficult to ascertain whether CNMF is as good as deep learning in detecting this kind of artefact. The spectra we used in this work were already defined on the frequency domain, so a detailed analysis of the cause of each artefact was out of the scope of our study. Also, the dataset we used contains a wide variety of artefacts, sometimes more than one in each spectrum (e.g. bad water suppression and ghosting artefact), in contrast to (33), where the authors used simulated and *in vivo* volunteers' spectra in which, purposely, the only artefact was the ghosting one. It remains to be tested whether a deep learning approach would also be as good as CNMF to chase other kinds of artefacts, but at any rate these two approaches seem to be complementary. Recent work by Gurbani *et al*. (34), using CNN, seems to suggest so, as their algorithm was able to pick artefactual patterns of different origins with remarkable efficiency (AUC of 0.95 in the test set). Their dataset was composed of 8,894 spectra from only nine patients.

One of the hypotheses in our study was that some of the sources extracted by CNMF should mostly reflect known artefacts, while the rest of sources would mainly reflect true tissue information. The results reported in figures 1 to 6 support this hypothesis to a large extent, as artefactual sources were easily identified and characterized by spectroscopy experts. Furthermore, these sources repeatedly and consistently appeared with small variants in every extraction from 4 to 20 sources. Most importantly, the quantitative measures support the experts' proposals. The results for data acquired at STE reported in Figure 8 provide us with a detailed picture. Out of the 20 extracted sources, S2, 5, 10, 12 and 16, identified as artefactual, have very low correlations and corresponding high Euclidean distances with all types included in the databases (tumors, abscesses and normal tissue). They also show low CC values, which is consistent with the fact that they only weight strongly on a limited number of spectra. On the other hand, non-artefactual sources show overall high correlations and low Euclidean distances. Moreover, some sources correlate highly with specific profiles. For instance, S1, 7, 9, 14 and 15 highly correlate with both abscesses and aggressive tumors, while S19 correlates highly with normal tissue. Note that the CC values offer some further interesting insight: those sources with the highest number of values over the 0.75 threshold are precisely the less variable and best-defined ones, corresponding quite neatly to database types. A similar analysis could be presented for the data acquired at LTE, but we omit it here for the sake of brevity.

When only a few sources are extracted, they are more likely to be combinations of more basic sources and these combinations tend to break into more basic components as the number of sources increase. Related to this, we found that the instability of the sources globally increased as the number of extracted sources increased. This is no surprise, as the uncertainty of the results is bound to increase for more sources when the number of spectra remains the same. Note though that this variability is by no means homogeneous over the extracted sources, with some of them showing very low variability. What is more, some sources show high variability in some frequency ranges and low variability in others. This is visually clear from figures 1 to 6, but also quantitatively from the boxplots of Figure 7. Artefactual sources have, in general, more variability. The likely reason for that is that these sources are present in a limited number of spectra and have limited leverage on the rest. A few of the non-artefactual sources also show high variability, which might be a sign of their low impact in the overall signal.

In the past, most efforts towards quality control of MRS data have been based on supervised approaches that are known to have some limitations. Each spectrum had always been treated as either being of good quality or bad quality. Then a bad quality spectrum would be so, irrespective of the cause (the artefact) and the magnitude of the problem: as an extreme example, a slightly badly phased spectrum could end up in the same category as an extremely noisy spectrum, or one with bad water suppression and a very important problem with the phasing as well as with small peaks in the frequency region of interest, all artefacts at the same time. Therefore, one limitation to this approach is the evident fact that labelling depends on experts, and different experts may have different thresholds for accepting a spectrum based on its quality. This was extensively recorded in the same source database where the current dataset has been taken from (6-8), but never systematically studied. Nevertheless, the fact has always been duly acknowledged in all previous studies (for example in (5), to cite just one recent study).

Another related limitation to supervised approaches is the mere existence of a diversity of artefacts, ranging from low SNR to bad water suppression, ghosting, bad or imperfect phasing. Kyathanahally *et al.* demonstrate this fact graphically in Figure 1 of their publication (5), where it can be seen that the means and standard deviations of good quality spectra and bad quality spectra clearly overlap, leaving approaches such as those based on linear discriminant analysis unsuitable for the task, a fact known since early work (1), where a quadratic discriminant classifier was employed instead.

Supervised approaches, in the end, require a simplified labeling setting to which an unsupervised approach such as CNMF is not restricted to. For this reason, sophisticated classifiers such as those from the deep learning family (34) are only suitable for such simplified setting, in which they can achieve very competitive results. A word of caution must be given though, as deep learning methods are only meant to provide a neat advantage in data rich settings, which are uncommon in the MRS(I) domain. An example of that are the excellent results recently obtained by alternative classifiers in a similar setting (5) without resorting to deep model architectures, but to a boosting and data sampling method (RUSBoost (35)) specifically suited to class-imbalanced data sets.

An unexpected finding of our study has been that, when there is a sufficiently high number of sources, we begin to observe patterns that are partly usable and partly unusable (for example see Figure 3, STE, source 18, region downfield from water). In fact, for 20 sources extracted at STE, there appears to be a total coincidence when the experts consider a source as artefactual and, 1) its Pearson's correlation with at least one of the compared classes is higher than 0.50, 2) the Euclidean distance between this source and all the means of the different classes is lower than 100 and, 3) none of the spectra in the database has a CC higher than 0.75. However, results for LTE are not as clear-cut, mainly because there are some examples of these "partially artefactual sources".

Altogether, evaluating the sources with three different quantitative measures appears to be a valuable approach, as in clear-cut artefacts all measures would agree, while in partially valid spectra there might be disagreement between these measures, should a threshold for decision be established. Gurbani *et al*. (34) used an approach named GRAD-cam (36), and they were able to identify that the most artefactual regions (approximately [0, 1.6] and [3.7, 4.5] ppm) were those out of the main interesting metabolite regions. Despite their spectra having a narrower spectral range than ours ([0, 4.5] ppm vs [-2.7, 7.1]), their results point to their CNN being able to at least recognize bad water suppression and bad homogeneity, although exclusion of spectra with a metabolite linewidth greater than 18 Hz had been performed before the experiment.

The fact that NMF methods "pick" artefacts, as well as metabolically-interesting patterns, has been known since the first application of this technique to MRS data of humans (figure 9 in (37)), and has recently been corroborated (figure 7.7 in (29)). However, this fact is usually overlooked, other than for the need of getting rid of the artefacts. One simple strategy used by Sajda *et al*. (37) was to remove artefactual sources (recognized by the expert spectroscopists) from subsequent analyses by a masking procedure. Another useful approach when artefact detection is not the objective is to discard bad quality spectra before performing further data analyses, for instance using well-established threshold criteria as in (17,19,27,38,39), and/or by using integrated peak areas of selected metabolite intensities (40,41).

As for our results, artefacts are conspicuous, indefectibly appearing when asking even for the lowest number of sources ($K = 4$). In this sense, unsupervised CNMF is shown to be a powerful tool for this kind of imbalanced datasets (a high number of good quality spectra and a low number of bad quality spectra), for which the adoption of an oversampling schema for the bad quality spectra class (5,34) is advisable for supervised approaches to perform optimally.

Another question that can be raised in view of the results presented in this study and others addressing similar issues is: are some PR approaches best suited to detect one particular type of artefact than others? This question merits further in-depth research.

Authors' contributions

AV and MJS conceived the study, wrote the drafts and supervised YHV's work. YHV performed the experiments, prepared the figures and helped to draft the manuscript. SOM assisted YHV with code and use of CNMF. MJS and CA performed the expert spectroscopists' role. All authors read and approved the final version of the submitted draft.

Acknowledgements

We thank the eTUMOUR and INTERPRET consortia for providing data, and also Sreenath P. Kyathanahally for the processed spectral matrices and Victor Mocioiu for the Python code. This work was funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n◦ ITN-GA-2012-316679 – TRANSACT. This work was also partially funded by CIBER-BBN, which is an initiative of the VI National R&D&I Plan *2008-2011, CIBER Actions* and financed by the *Instituto de Salud Carlos III* with assistance from the European

Regional Development Fund. CA and MJ were also funded by MOLIMAGLIO (SAF2014-52332-R) from MINECO, Spain. AV was funded by Spanish MINECO TIN2016-79576-R research project. YHV was funded by 2017FI_B_00758.

References

1. van der Graaf M, Julia-Sape M, Howe FA, Ziegler A, Majos C, Moreno-Torres A, Rijpkema M, Acosta D, Opstad KS, van der Meulen YM, Arus C, Heerschap A. MRS quality assessment in a multicentre study on MRS-based classification of brain tumours. NMR Biomed 2008;21(2):148-158.
2. Pedrosa de Barros N, Slotboom J. Quality management in in vivo proton MRS. Anal Biochem 2017;529:98-116.
3. König C, Shaim I, Vellido A, Romero E, Alquézar R, Giraldo J. Using machine learning tools for protein database biocuration assistance. Scientific Reports 2018;8(1):10148.
4. Kreis R. Issues of spectral quality in clinical 1H-magnetic resonance spectroscopy and a gallery of artifacts. NMR Biomed 2004;17(6):361-381.
5. Kyathanahally SP, Mocioiu V, Pedrosa de Barros N, Slotboom J, Wright AJ, Julia-Sape M, Arus C, Kreis R. Quality of clinical brain tumor MR spectra judged by humans and machine learning tools. Magn Reson Med 2018;79(5):2500-2510.
6. Julia-Sape M, Acosta D, Mier M, Arus C, Watson D, consortium I. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. MAGMA 2006;19(1):22-33.
7. Julia-Sape M, Lurgi M, Mier M, Estanyol F, Rafael X, Candiota AP, Barcelo A, Garcia A, Martinez-Bisbal MC, Ferrer-Luna R, Moreno-Torres A, Celda B, Arus C. Strategies for annotation and curation of translational databases: the eTUMOUR project. Database (Oxford) 2012;2012:bas035.
8. Tate AR, Underwood J, Acosta DM, Julia-Sape M, Majos C, Moreno-Torres A, Howe FA, van der Graaf M, Lefournier V, Murphy MM, Loosemore A, Ladroue C, Wesseling P, Luc Bosson J, Cabanas ME, Simonetti AW, Gajewicz W, Calvar J, Capdevila A, Wilkins PR, Bell BA, Remy C, Heerschap A, Watson D, Griffiths JR, Arus C. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed 2006;19(4):411-434.
9. Wright AJ, Arus C, Wijnen JP, Moreno-Torres A, Griffiths JR, Celda B, Howe FA. Automated quality control protocol for MR spectra of brain tumors. Magn Reson Med 2008;59(6):1274-1281.
10. Pedrosa de Barros N, McKinley R, Knecht U, Wiest R, Slotboom J. Automatic quality control in clinical (1)H MRSI of brain cancer. NMR Biomed 2016;29(5):563-575.
11. Menze BH, Kelm BM, Weber M-A, Bachert P, Hamprecht FA. Mimicking the human expert: Pattern recognition for an automated assessment of data quality in MR spectroscopic images. Magnetic Resonance in Medicine 2008;59(6):1457-1466.
12. Vellido A, Romero E, González-Navarro FF, Belanche-Muñoz LA, Julià-Sapé M, Arús C. Outlier exploration and diagnostic classification of a multi-centre 1H-MRS brain tumour database. Neurocomputing 2009;72(13-15):3085-3097.
13. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401(6755):788-791.
14. Ding C, Li T, Jordan MI. Convex and semi-nonnegative matrix factorizations. IEEE transactions on pattern analysis and machine intelligence 2010;32(1):45-55.

15. Julia-Sape M, Griffiths JR, Tate AR, Howe FA, Acosta D, Postma G, Underwood J, Majos C, Arus C. Classification of brain tumours from MR spectra: the INTERPRET collaboration and its outcomes. NMR Biomed 2015;28(12):1772-1787.
16. Perez-Ruiz A, Julia-Sape M, Mercadal G, Olier I, Majos C, Arus C. The INTERPRET Decision-Support System version 3.0 for evaluation of Magnetic Resonance Spectroscopy data from human brain tumours and other abnormal brain masses. BMC Bioinformatics 2010;11:581.
17. Ortega-Martorell S, Lisboa PJ, Vellido A, Julia-Sape M, Arus C. Non-negative matrix factorisation methods for the spectral decomposition of MRS data from human brain tumours. BMC Bioinformatics 2012;13:38.
18. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. New Orleans, Louisiana: Society for Industrial and Applied Mathematics; 2007. p 1027-1035.
19. Ortega-Martorell S, Lisboa PJ, Vellido A, Simoes RV, Pumarola M, Julia-Sape M, Arus C. Convex non-negative matrix factorization for brain tumor delimitation from MRSI data. PLoS One 2012;7(10):e47824.
20. Mocioiu V, Kyathanahally SP, Arús C, Vellido A, Julià-Sapé M. Automated Quality Control for Proton Magnetic Resonance Spectroscopy Data Using Convex Non-negative Matrix Factorization. Bioinformatics and Biomedical Engineering; 2016; Cham. Springer International Publishing. p 719-727. (Bioinformatics and Biomedical Engineering).
21. Garcia-Gomez JM, Tortajada S, Vidal C, Julia-Sape M, Luts J, Moreno-Torres A, Van Huffel S, Arus C, Robles M. The effect of combining two echo times in automatic brain tumor classification by MRS. NMR Biomed 2008;21(10):1112-1125.
22. Garcia-Gomez JM, Luts J, Julia-Sape M, Krooshof P, Tortajada S, Robledo JV, Melssen W, Fuster-Garcia E, Olier I, Postma G, Monleon D, Moreno-Torres A, Pujol J, Candiota AP, Martinez-Bisbal MC, Suykens J, Buydens L, Celda B, Van Huffel S, Arus C, Robles M. Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. Magma Magn Reson Mater Phys Biol Med 2009;22(1):5-18.
23. Ortega-Martorell S, Vellido A, Lisboa PJG, Julia-Sape M, Arus C, Ieee. Spectral decomposition methods for the analysis of MRS information from human brain tumors. New York: Ieee; 2011. 3279-3284 p.
24. Tzyy-Ping J, Colin H, Te-Won L, Scott M, Martin JM, Vicente I, Sejnowski TJ. Extended ICA Removes Artifacts from Electroencephalographic Recordings. 1998:894--900.
25. Vilamala A, Lisboa PJG, Ortega-Martorell S, Vellido A. Discriminant Convex Non-negative Matrix Factorization for the classification of human brain tumours. Pattern Recognition Letters 2013;34(14):1734-1747.
26. Ortega-Martorell S, Ruiz H, Vellido A, Olier I, Romero E, Julia-Sape M, Martin JD, Jarman IH, Arus C, Lisboa PJ. A novel semi-supervised methodology for extracting tumor type-specific MRS sources in human brain data. PLoS One 2013;8(12):e83773.
27. Yuqian L, M. SD, Van CS, R. CSA, Uwe H, Yiming P, Sabine VH. Hierarchical non-negative matrix factorization (hNMF): a tissue pattern differentiation method for glioblastoma multiforme diagnosis using MRSI. NMR in Biomedicine 2013;26(3):307-319.
28. Ortega-Martorell S, Ruiz H, Vellido A, Olier I, Romero E, Julià-Sapé M, Martín JD, Jarman IH, Arús C, Lisboa PJG. A Novel Semi-Supervised Methodology for Extracting Tumor Type-Specific MRS Sources in Human Brain Data. PLOS ONE 2013;8(12):e83773.
29. Laruelo Fernandez A. Integration of magnetic resonance spectroscopic imaging into the radiotherapy treatment planning: Université Paul Sabatier - Toulouse III; 2016.
30. Nascimento JMP, Dias JMB. Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 2005;43(4):898-910.

31. Vilamala A, Vellido A, Belanche LA. Bayesian semi non-negative matrix factorisation. 2016; Bruges, Belgium. i6doc.com publication.

32. Ortega-Martorell S, Olier I, Julià-Sapé M, Arús C, Lisboa P. Automatic relevance source determination in human brain tumors using Bayesian NMF. 2014 9-12 Dec. 2014. p 99-104.

33. Kyathanahally SP, Doring A, Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. Magn Reson Med 2018;80(3):851-863.

34. Gurbani SS, Schreibmann E, Maudsley AA, Cordova JS, Soher BJ, Poptani H, Verma G, Barker PB, Shim H, Cooper LAD. A convolutional neural network to filter artifacts in spectroscopic MRI. Magnetic Resonance in Medicine 2018;0(0).

35. Seiffert C, Khoshgoftaar TM, Hulse JV, Napolitano A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 2010;40(1):185-197.

36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 22-29 Oct. 2017. p 618-626.

37. Sajda P, Du S, Brown TR, Stoyanova R, Shungu DC, Mao X, Parra LC. Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. IEEE Trans Med Imaging 2004;23(12):1453-1465.

38. Li Y, Pi Y, Liu X, Liu Y, Van Cauter S. Data Analysis and Tissue Type Assignment for Glioblastoma Multiforme. BioMed research international 2014;2014:10.

39. Ghasemi K, Khanmohammadi M, Saligheh Rad H. Accurate grading of brain gliomas by soft independent modeling of class analogy based on non-negative matrix factorization of proton magnetic resonance spectra. Magnetic Resonance in Chemistry 2016;54(2):119-125.

40. Sauwen N, Acou M, Van Cauter S, Sima DM, Veraart J, Maes F, Himmelreich U, Achten E, Van Huffel S. Comparison of unsupervised classification methods for brain tumor segmentation using multi-parametric MRI. NeuroImage : Clinical 2016;12:753-764.

41. Li Y, Liu X, Wei F, Sima DM, Van Cauter S, Himmelreich U, Pi Y, Hu G, Yao Y, Van Huffel S. An advanced MRI and MRSI data fusion scheme for enhancing unsupervised brain tumor differentiation. Comput Biol Med 2017;81:121-129.

Figures

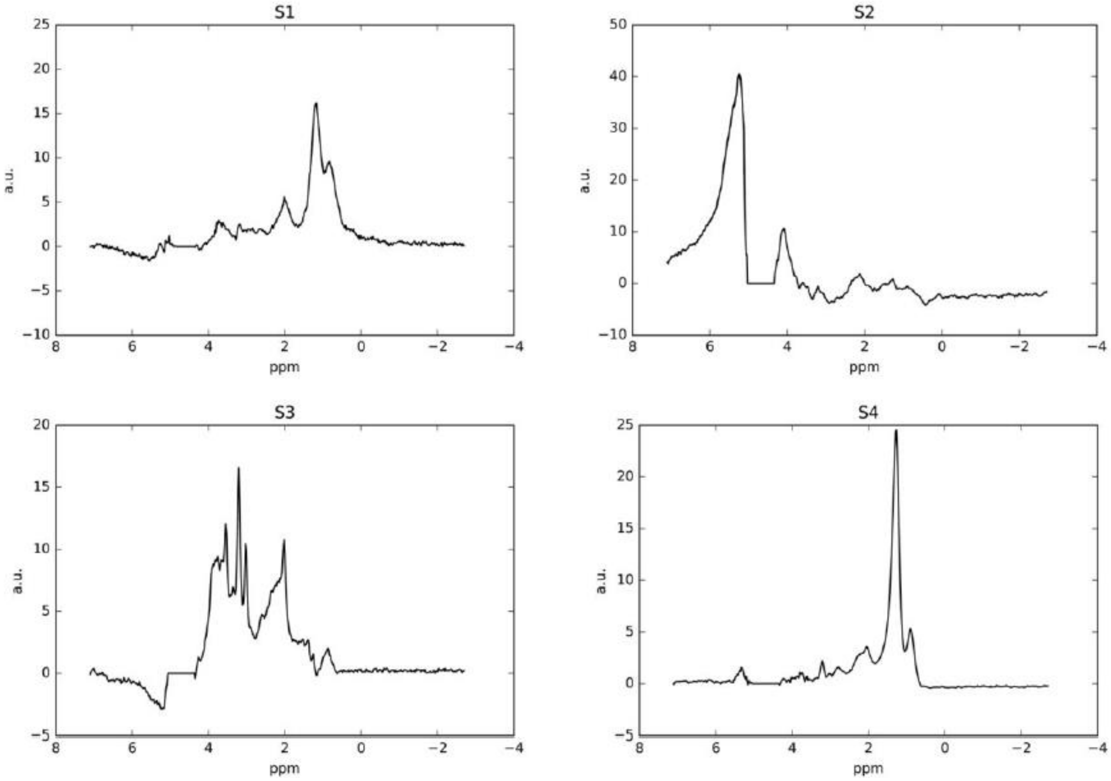


Figure 1. Mean and STD (+/-) of sources (S) extracted for *K=4* (*K* being number of sources) from spectra acquired at STE (*N*=1,180). S1, S2, S3 and S4 stand for source number 1, 2, 3 and 4, respectively. The *x*-axis of the graph is represented in parts per million (ppm), while the *y*-axis represents the intensities in arbitrary units (a.u.). The mean is represented by a blue line and the variability described as STD (+/-) is displayed in gray shade, enclosed by a black line. In this source extraction, variability is extremely low, which explains why only a single black line seems to represent the source. The sources closely resemble characteristic spectra of different types, and could even be taken by a mean spectrum if no more information was given. As the original spectra had been processed with the INTERPRET pipeline {Tate, 2006 #15942}, which includes a residual water suppression from points between 4.2 to 5.1 ppm set to zero prior to unit length normalisation, sources also display this characteristic of the processing pipeline. The zeroing of the 4.2-5.1 ppm interval was incorporated into the INTERPRET pipeline because if there were any remnants of water signal, the intensity of the rest of the spectrum would be affected when performing the unit length normalization.

The first and third sources (S1 and S3) have a typical pattern of necrosis with high lipids at 0.9, 1.28 and 2 ppm, with S4 additionally showing choline-containing compounds at 3.21 ppm and lipids at 5.3 ppm, and a different methyl/methylene (0.9ppm/1.28ppm) ratio than for S1. S2 shows a typical pattern of bad water suppression, that the zeroing between 4.2 and 5.1 could only partially eliminate, therefore the appearance of these two “tails”, from the incompletely suppressed water signal, appearing between 3.9 and 4.2 approximately and between 5.1 and 7.1 ppm. No other metabolite signals can be identified in this S2. The third source (S3), shows the typical pattern for an infiltrative, low-grade glial tumour, in particular the high choline-containing compounds / creatine ratio (3.21ppm/3.03ppm) is indicative of high proliferation, whereas the decrease in the intensity of the N-acetyl-containing compounds at 2.01 ppm (it should be about twice the height of the creatine peak in a normal brain) is indicative of a decreased amount/functionality of neurones.

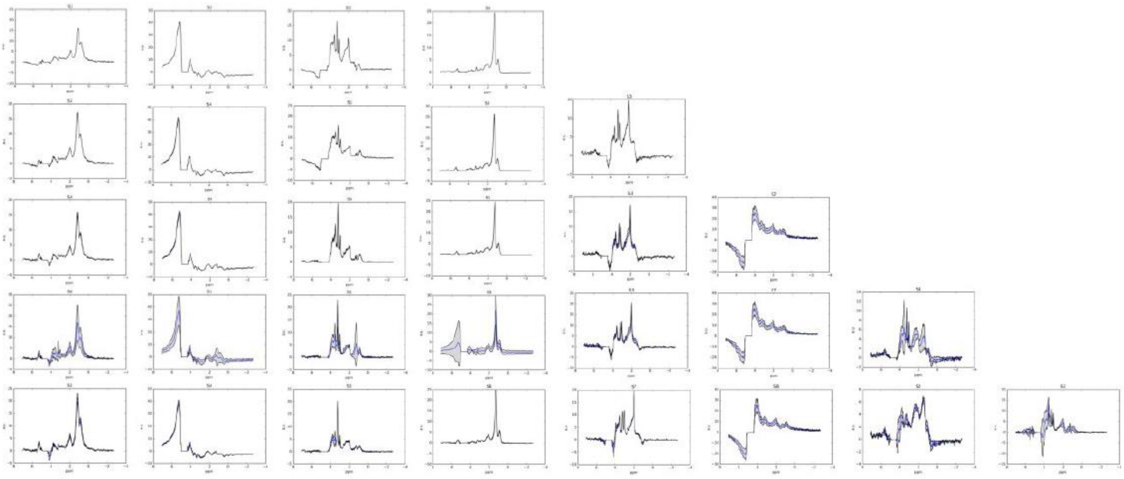


Figure 2. Mean and STD (+/-) of sources extracted at STE, for *K=4* to *K=8*, from spectra acquired at STE (*N*=1,180). Each row corresponds to a different source extraction, starting with *K=4* at the top. Columns were organized according to the similarity of the sources. Columns 1, 2, 3 and 4 correspond to sources that have similar characteristics to the ones for *K=4*. Other features as in figure 1 legend.

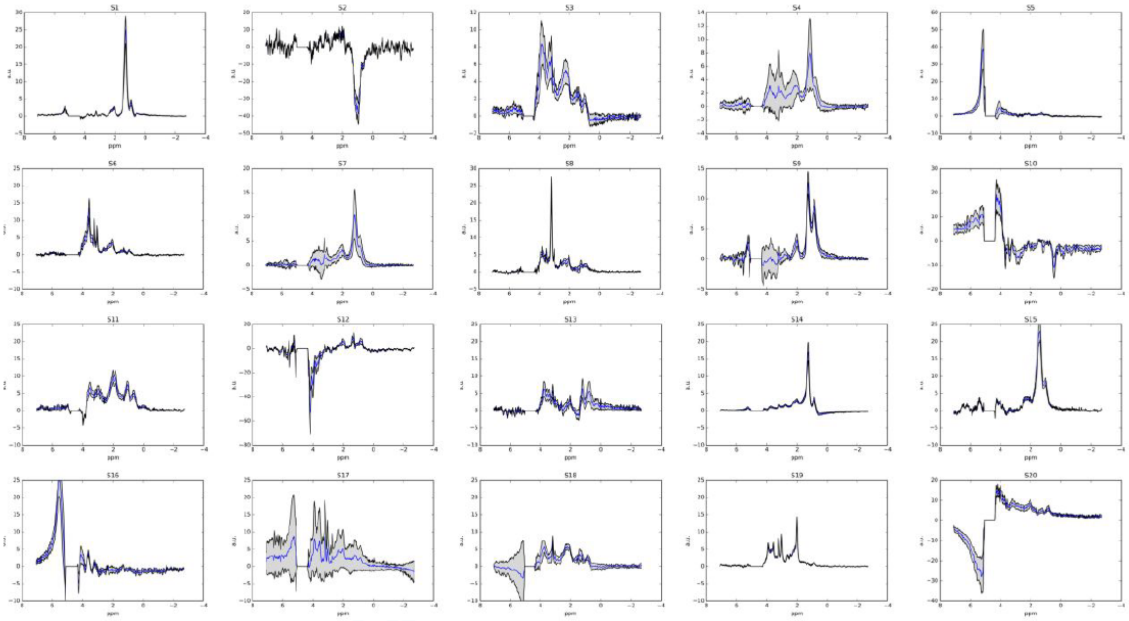


Figure 3. Mean and STD (+/-) of sources extracted for *K=20* from data acquired at STE (*N*=1,180). Again, the mean is represented by a blue line, while variability described as STD (+/-) is shaded in gray, bounded by black lines.

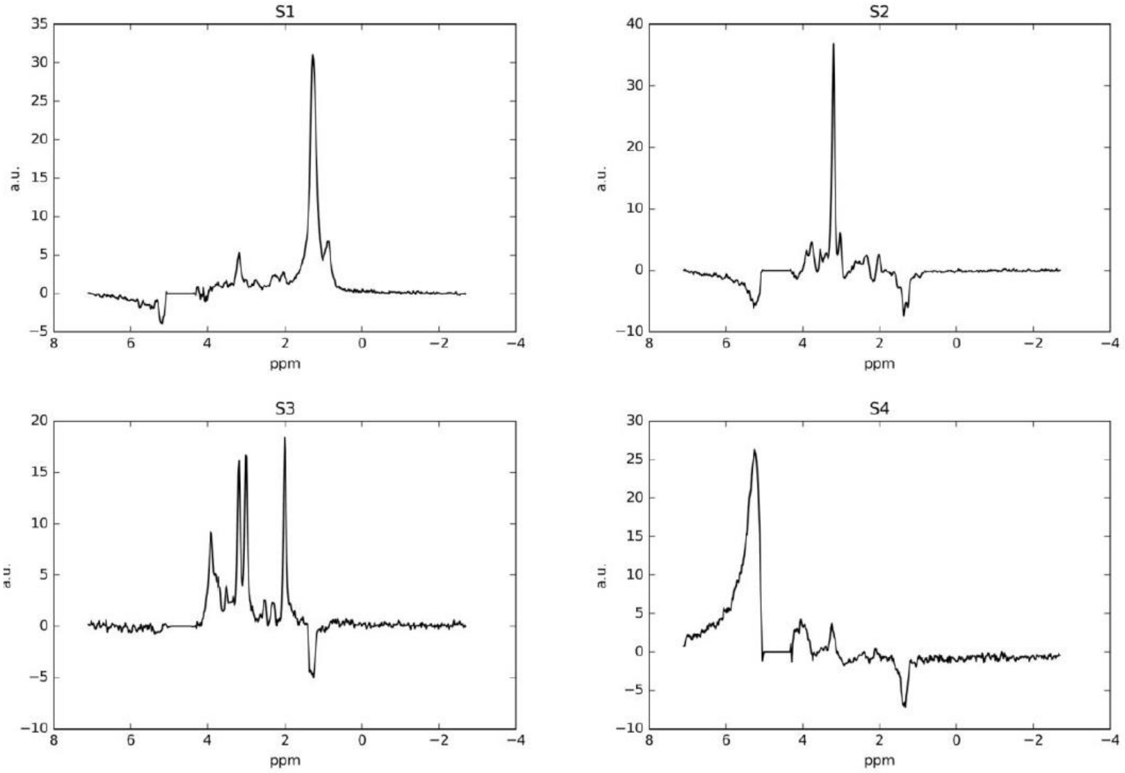


Figure 4. Mean and STD (+/-) of sources extracted for *K=4* from data acquired at LTE (*N*=977). Representation as in previous figures.

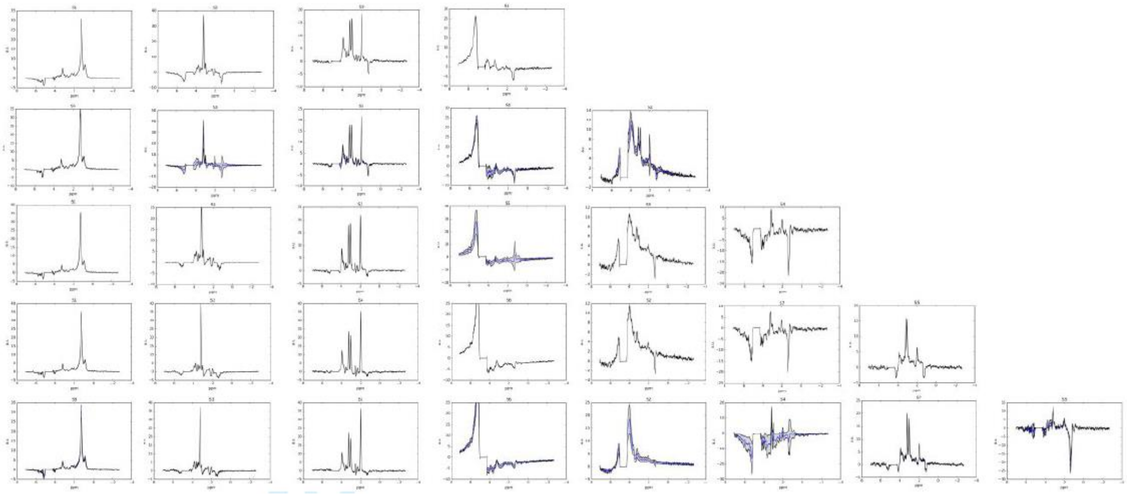


Figure 5. Mean and STD (+/-) of sources extracted from data acquired at LTE (*N*=977). Each row corresponds to a different source extraction from *K=4* to *K=8*. Columns were again organized depending on the similarity of the sources. Other features as in figure 1 legend.

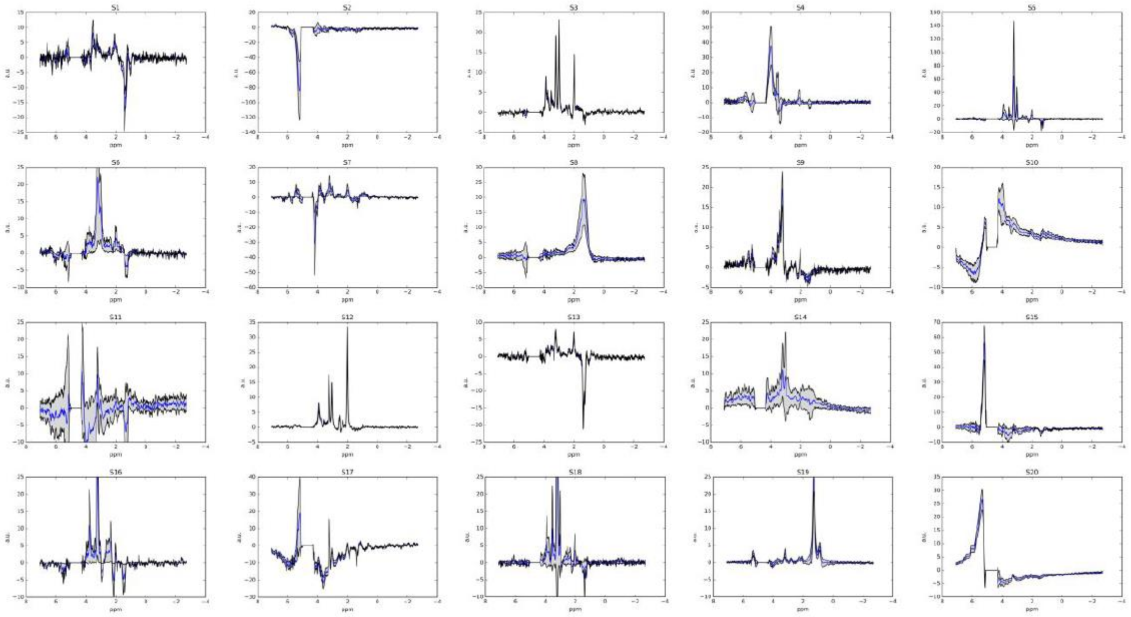


Figure 6. Mean and STD (+/-) of sources extracted for *K=20* from data acquired at LTE (*N*=977). Representation as in previous figures.

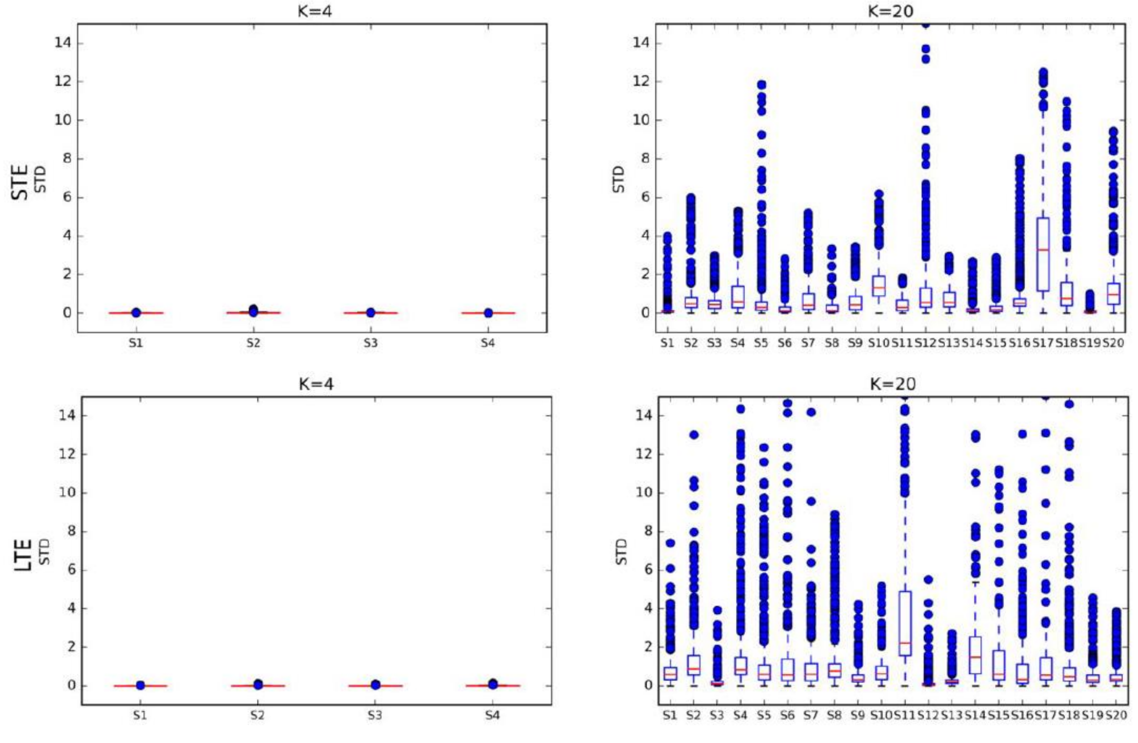


Figure 7. Boxplots of the STD values for the ten algorithm run repetitions and for each of the sources in two different extractions (*K = 4* and *K = 20*) from data acquired at STE and LTE; (The box extends from the lower to upper quartile values of the STD, with a line at the median. The whiskers extend from the box to show the range of the data. (Outlier points are those past the end of the whiskers). STD was calculated from the matrix in which there are ten rows (corresponding to the ten extractions) and 512 points (corresponding to the number of points of each source).

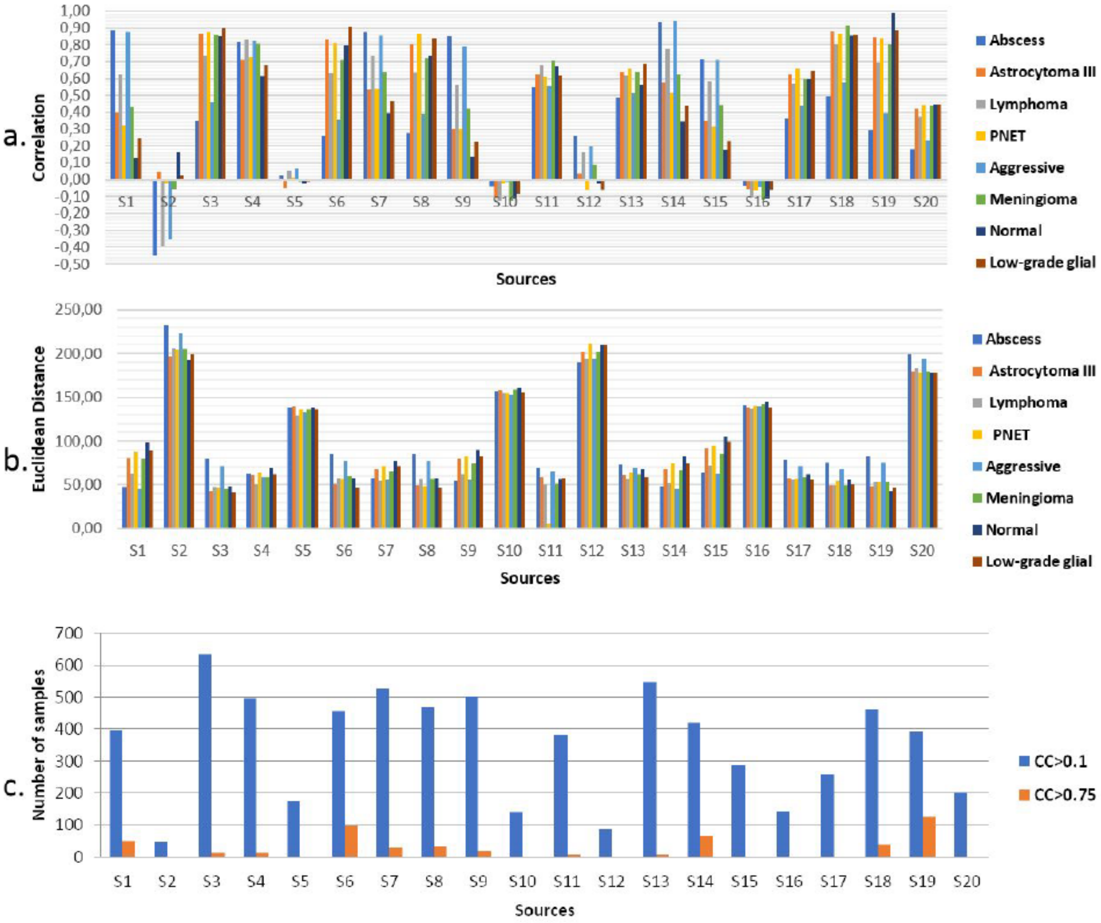


Figure 8. a) Correlation between sources extracted for *K=20* (from data acquired at STE) and mean spectra from the types included in the INTERPRET validated database (6), where the *x*-axis corresponds to the source number and the *y*-axis to the values of the correlations. b) Euclidean distance between each source for the *K=20* extraction (from data acquired at STE) and mean spectra from the types included in the INTERPRET validated database (6), where the *x*-axis again corresponds to the source number, while the *y*-axis corresponds to Euclidean distances. c) CC of the mixing matrix for *K = 20*, where the *x*-axis corresponds to the source number and the *y*-axis corresponds to the number of samples.

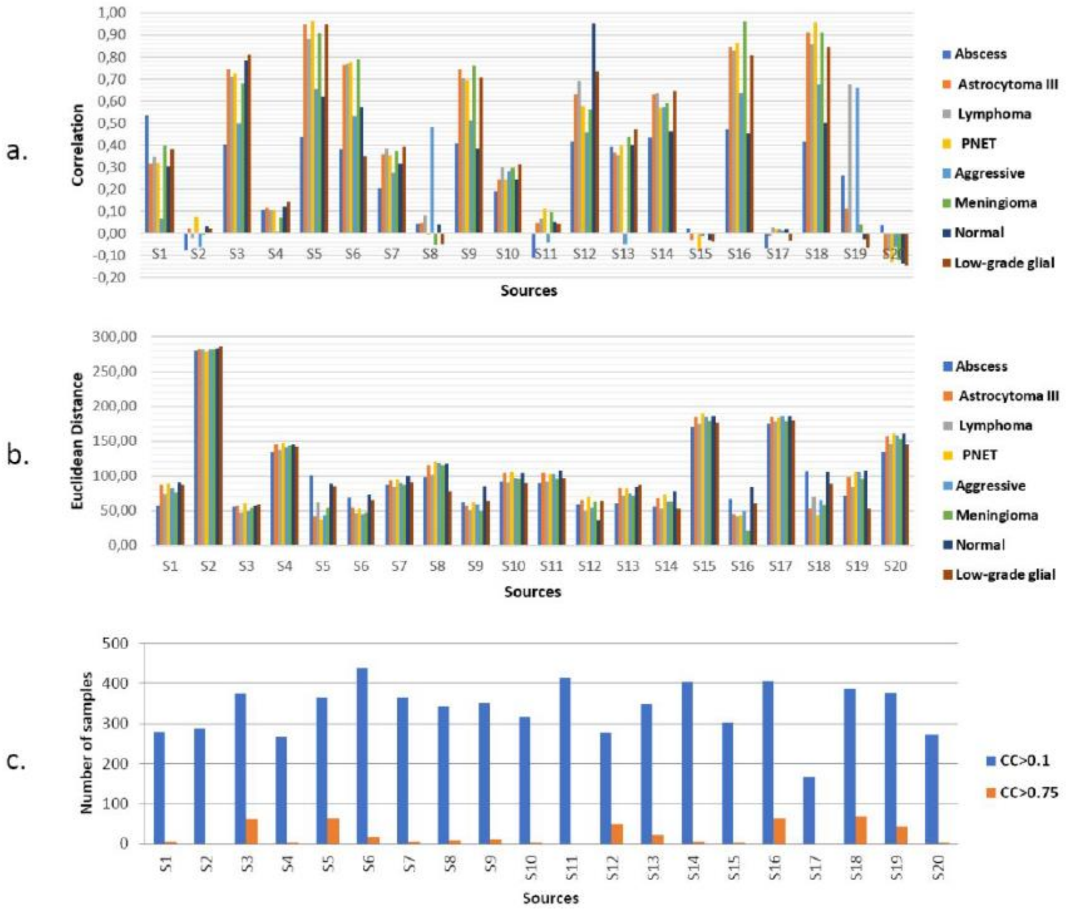


Figure 9. Correlations, Euclidean distances and CC for data acquired at LTE, represented as in Figure 8.

		STE				LTE			
Type		**GOOD**	**POOR**	**BAD**	**Total**	**GOOD**	**POOR**	**BAD**	**Total**
Abscess		9	0	2	11	8	1	2	11
Astrocytoma WHO grade III		7	0	0	7	7	0	0	7
Lymphoma		16	2	1	19	15	0	3	18
PNET		11	0	0	11	8	0	0	8
Glioblastoma	Aggressive	189	5	18	212	215	9	34	258
Metastasis		87	1	7	95	78	4	10	92
Meningioma		100	4	23	127	87	5	13	105
Astrocytoma	Low grade glial (WHO grade II)	68	2	7	77	60	7	4	73
Oligodendroglioma		27	0	2	29	39	2	2	43
Oligoastrocytoma		12	0	3	15	22	1	1	24
Pilocytic astrocytoma		27	1	9	37	37	1	9	47
Other Pathologies		100	19	10	129	156	25	3	184
Not available		304	23	55	382	77	1	12	88
Total		982	49	149	1180	828	38	111	977

Table 1. Number of spectra, acquired at STE and LTE, available per tumor type and quality label. The GOOD, POOR and BAD labels are taken from the data matrix from study (5), in which the the intermediate label of "poor quality" was assigned to the rejected spectra that had been seen by three experts and had been accepted by one of them. *Not available* corresponds to cases lacking definitive/consensus diagnosis in the database.

Source number	Consensus expert spectroscopists' evaluation	Pearson correlation > 0.50 at least with one of the compared classes	Euclidean distance with all the compared classes, at least > 100	Number of samples with CC > 0.75
1	Good quality	Yes	No	Several
2	Artefactual pattern	No	Yes	None
3	Good quality	Yes	No	Several
4	Good quality	Yes	No	Several
5	Artefactual pattern	No	Yes	None
6	Good quality	Yes	No	Several
7	Good quality	Yes	No	Several
8	Good quality	Yes	No	Several
9	Good quality	Yes	No	Several
10	Artefactual pattern	No	Yes	None
11	Good quality	Yes	No	Several
12	Artefactual pattern	No	Yes	None
13	Artefactual pattern	Yes	No	Several
14	Good quality	Yes	No	Several
15	Good quality	Yes	No	None
16	Artefactual pattern	No	Yes	None
17	Partly artefactual pattern	Yes	No	None
18	Partly artefactual pattern	Yes	No	Several
19	Good quality	Yes	No	Several
20	Artefactual pattern	No	Yes	None

Table 2. Summary of the evaluations for the 20-source extraction, at STE.

Source number	**Consensus expert spectroscopists' evaluation**	**Pearson correlation > 0.40 at least with one of the compared classes**	**Euclidean distance with all the compared classes, at least > 100**	**Number of samples with CC > 0.75**
1	Good quality	Yes	No	None
2	Artefactual pattern	No	Yes	None
3	Good quality	Yes	No	Several
4	Artefactual pattern	No	Yes	None
5	Artefactual pattern but source too variable to be sure	Yes	No	Several
6	Good quality but source too variable to be sure	Yes	No	Several
7	Partly artefactual pattern	No (close for low grade glial)	No	Several
8	Artefactual pattern	No	Yes	Several
9	Artefactual pattern	Yes	No	Several
10	Artefactual pattern	No	No	None
11	Artefactual pattern	No	No	None
12	Good quality	Yes	No	Several
13	Good quality	Yes	No	Several
14	Artefactual pattern, but source too variable to be sure	Yes	No	Several
15	Artefactual pattern	No	Yes	None
16	Partly artefactual pattern but source too variable to be sure	Yes	No	Several
17	Partly artefactual pattern	No	Yes	None
18	Partly artefactual pattern but source too variable to be sure	Yes	No	Several
19	Good quality	Yes	No	Several
20	Artefactual pattern	No	Yes	None

Table 3. Summary of the evaluations for the 20-source extraction, at LTE. For some sources, the evaluation is uncertain (source too variable), because there is so much variability that one of the 10 solutions may be the actual reverse of the evaluation.