Automated quality control for proton magnetic resonance spectroscopy data using convex non-negative matrix factorization

Victor Mocioiu[1, 4], Sreenath P. Kyathanahally[2], Carles Arús[1, 4], Alfredo Vellido[3, 4], Margarida Julià-Sapé[4, 1]

[1] Universitat Autònoma de Barcelona- Departament de Bioquímica i Biologia Molecular Cerdanyola del Vallès, Barcelona 08193, Spain
victor.mocioiu, carles.arus@uab.cat
[2] Depts. Radiology and Clinical Research, University of Bern, Bern, Switzerland
s.p.kyathanahally@insel.ch
[3] Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Barcelona-Tech, Campus Nord, 08034, Barcelona, Spain
avellido@cs.upc.edu
[4] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina CIBER-BBN, Cerdanyola del Vallès, Barcelona, Spain
Margarita.Julia@uab.cat

Abstract. Proton Magnetic Resonance Spectroscopy (^{1}H MRS) has proven its diagnostic potential in a variety of conditions. However, MRS is not yet widely used in clinical routine because of the lack of experts on its diagnostic interpretation. Although data-based decision support systems exist to aid diagnosis, they often take for granted that the data is of good quality, which is not always the case in a real application context. Systems based on models built with bad quality data are likely to underperform in their decision support tasks. In this study, we propose a system to filter out such bad quality data. It is based on convex Non-Negative Matrix Factorization models, used as a dimensionality reduction procedure, and on the use of several classifiers to discriminate between good and bad quality data.

Keywords: Brain tumors, magnetic resonance spectroscopy, convex non-negative matrix factorization, pattern recognition, quality control, machine learning.

1 Introduction

Proton magnetic resonance spectroscopy (^{1}H MRS, henceforth only referred to as MRS) is a magnetic resonance modality that provides metabolic information about an investigated tissue volume, thus becoming a tool for metabolomics. MRS is inherently non-invasive and can be used either on its own, or in conjunction with other MR mo-

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-31744-1_62

dalities [1] with the aim to improve diagnostic accuracy. Although MRS can be used to investigate a wide range of tissue types [2], it is mainly used for diseases of the central nervous system (CNS), and has proven a powerful tool in assessing a broad spectrum of diseases such as metabolic disorders, epilepsy, Alzheimer and Parkinson, amongst others. But, by far, its most common application is on brain tumors diagnostic assistance [2] .

MRS can be single-voxel (SV), where the signal comes from a volume of interest, or multi-voxel, using a grid/matrix of many contiguous SVs (actually, SV-like spectral vectors). MRS has several parameters of importance that should be mentioned. First of all, echo time, which can be either short (STE, lower than 40 ms), or long (LTE, higher than 40 ms), controlling what metabolites can be better seen in the spectrum. STE only allows for positive peaks in the spectrum whereas LTE can also have negative peaks. Another important parameter is field strength (measured in Tesla, T) which, in laymen terms, determines how far apart the peaks of the metabolites are, and how many metabolites can be detected. A typical field strength used in clinical routine is 1.5 T. Figure 1 shows an example of an MR spectrum.

Although MRS is a promising technique, it is not yet widely implemented in clinical routine. It can be argued that the main reason for this is the explicit need for an expert (a radiologist) to interpret the spectrum and reach a diagnostic. Previous work has been carried out to compensate for the lack of experts in MRS interpretation by using (semi-)automated classifiers or decision support systems [3-5]. The main limitation of such systems is that they assume that the spectra are of consistently good quality. Unfortunately, the definition of 'good quality' has yet to be clearly established, and by this we mean that the gold standard is human-dependent and, although some guidelines have been proposed [3, 6], it may vary from expert to expert. Furthermore, multiple types of artifacts can contaminate the signal; an extensive gallery of such artefacts is presented in [7]. Figure 2 shows an example of a bad quality spectrum (compare to Figure 1, which corresponds to a good quality spectrum).

In a previous work by van der Graaf *et al.* [8], a semi-automatic filtering procedure, based on the signal-to-noise ratio (SNR) of the spectrum and the water bandwidth (WBW - defined as the line width at half of the maximum intensity of the water peak), was proposed. However, this system relies on a board of experts for validating the final decision.

A fully automated system, trained on a subset of the eTUMOR [9, 10] and INTERPRET [11, 12] SV spectra (144 SVs, 72 acceptable and 72 unacceptable), was proposed by Wright *et al.* [13]. The system consists of a least squares support vector machine [14] with a radial basis function kernel and fastICA[15], which was used for dimensionality reduction. The test set comprised of 98 SVs (58 acceptable and 40 unacceptable) from the eTUMOR database. The results of this study were encouraging (an accuracy of 88%), but we argue that this might be due to several reasons that might not hold up in a real clinical setting. First of all, because the number of cases analyzed was relatively low, something that can be seen as a detrimental for the generalization power of the system. Another and more important aspect to consider is that the two classes are fairly balanced for training and testing – again affecting generalization but also performance metrics.

In this article, we propose a bad quality data filtering system based on convex non-negative matrix factorization (cNMF), used as a dimensionality reduction model as well as an artifact identification tool. The results from cNMF, specifically the mixing matrix (details in the Methods section), are then used as features for the following classifiers: Naïve Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), AdaBoost, and Random Forests (RF). We evaluate these classifiers based on accuracy, area under the receiver operator characteristic curve (AUC), sensitivity, specificity, F_1 score, and balanced error rate. Performance metrics are averaged over a 10-fold stratified cross validation.

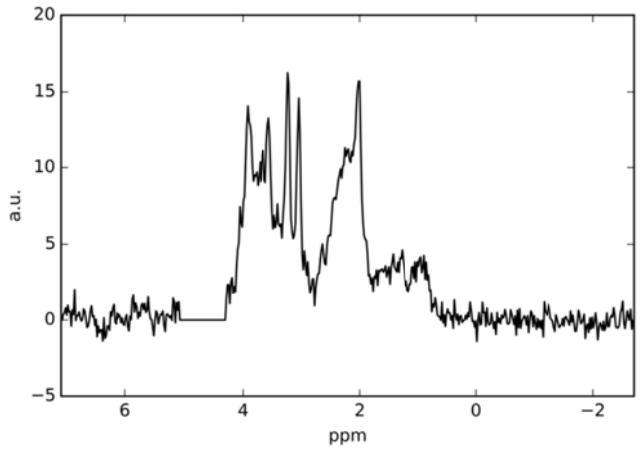

Fig. 1. Example of an MR spectrum from the eTumour database[10], considered of good quality. The acquisition parameters are described in the Materials section. Note that it has already undergone specific preprocessing procedures as described in the materials section. The x-axis of the spectrum is in parts per million (ppm) and the y-axis usually in arbitrary units. Note also that, traditionally in spectroscopy, the x-axis is reversed when compared to the Cartesian system [11].

2 Materials and Methods

2.1 Materials

The investigated dataset comes from the INTERPRET [11, 12] and eTUMOUR [10] multicentre databases and comprises of 1,196 SV STE spectra, out of which 982 were labeled as good quality and 198 as bad quality spectra acquiered at 1.5 T. The INTERPRET data came from seven clinical centers, whereas the eTUMOUR data came from eleven different centers – in addition some of the centers were working with more than one MR scanner from different manufacturers. The spectrum-labeling panel consisted of three experts and the labeling procedure was as follows: if at least two experts accepted a spectrum, then it was labeled as 'good' and if at least two re-

jected the spectrum, then it was labeled as 'bad'. Since spectra of each multi-center project came from different clinical centers, and in order to avoid expert bias, the judging system was set so that experts from one center did not judge spectra from their own center. Therefore, different combinations of experts were set to judge spectra from each clinical center [10, 12].

Prior to building our system, the data underwent a well-defined pre-processing procedure [4]. Residual water filtering using the HLSVD algorithm with 10 Lorentzians was initially applied, followed by apodization using a 1Hz bandwidth with a Lorentzian lineshape. Then the values in the [5.11, 4.31] interval were set to 0. Afterwards, baseline offset was corrected using two ranges – [-2 -1] ppm and [9 11] ppm. The spectrum was then normalized to unit length and multiplied by 100. Finally, the spectrum was aligned according to the algorithm presented in [4]. After preprocessing, each spectrum had 512 points in the 7.1 to -2.7 ppm range; this range was used for the remainder of our study. As a result, our analyzed data matrix was 1,196×512.

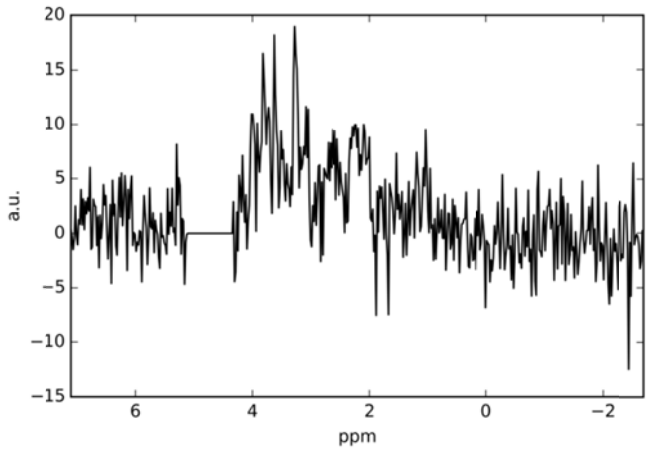

Fig. 2. An example of a bad quality spectrum due to low SNR from the eTumour database[10]. The acquisition parameters are described in the Materials section. Note that it has already undergone specific preprocessing procedures as described in the Materials section.

2.2 Methods

Convex non-negative matrix factorization

Convex non-negative matrix factorization (cNMF) [16] is a blind source separation method that, among other NMF algorithms [17], has been extensively used for MRS data analysis [17, 18]. NMF methods rely on factorizing an initial data matrix, **X** (*m* rows -data dimensionality, and *n* columns -samples/spectra) into two matrices **F** (*m* rows and *k* columns – sources/basis vectors) and **G** (*k* rows, *n* columns). Furthermore,

in cNMF, **F** is constrained to lie in the column space of the input data **X**, so that the cNMF formula can be written as:

$$\mathbf{X} \approx \mathbf{AXG} \tag{1}$$

where **A** fully determines **F**. **G** is also called the mixing matrix, as it holds the coefficients to recompose a specific data sample. It is a well-known fact that the results of the cNMF algorithm are dependent on the initialization scheme; in our implementation, we use the k-means++ algorithm [19]. It should be mentioned that there is no fully established method for choosing the optimal number of sources/basis vectors.

Implementation details

Some of the classifiers used in this study are not parameterless. LR was used in its basic variant, as well as with regularization, namely l2 – LRCV. AdaBoost was built using fifty decision trees as estimators. In the case of RF, fifty estimators were used; bootstrap samples of the training set were used to build the trees and the maximum number of features considered when looking for the best split was the square root of the total number of features. All nodes were expanded until all leaves were rendered pure. LR, LRCV and RF versions that take into account class imbalance by assigning a proportional weight to the less represented class were also built – they will be in turn be named LRA, LRCVA and RFA.

Regarding the details of our proposed system, we extracted from three up to eleven sources. A maximum of eleven sources was extracted because, as reported in [6], one should account usually for up to 9 technical requirements to make the spectrum clinically interpretable. The 2 remaining sources should account for the variability that is present in the 'good' class. We then used a 10-fold stratified cross-validation loop to split the mixing matrix in training and test subsets; trained the classifiers; and computed the accuracy, sensitivity, specificity, F1 score, and balanced error rate. Final results are reported as averaged performance metrics.

3 Results

We start by presenting the eleven extracted sources in Figure 3. Several aspects should be mentioned: first, that, even though the extracted sources are akin to the input space, they are not true spectra as such and, therefore, the *x* and *y* axes are left unlabeled (however the *x* axis would correspond to the [7.1 -2.7] ppm range and the *y* axis to the [-20,40] a.u. range). Another thing to stress is that some of the negative peaks are not completely shown in this figure, for two reasons: first, because in STE, negative peaks are not possible and are directly regarded as artefacts; second, a more practical reason was not to undermine the amplitude of the positive peaks. An important aspect to note is that we have found five sources pertaining to good spectra, and only six artefactual sources.

The classification results for the investigated classifiers are presented in Table 1. It is worth highlighting that RF shows the best performance in terms of accuracy, AUC, BER, and F1 score, whereas LR achieves the best specificity and LRA the best sensitivity.

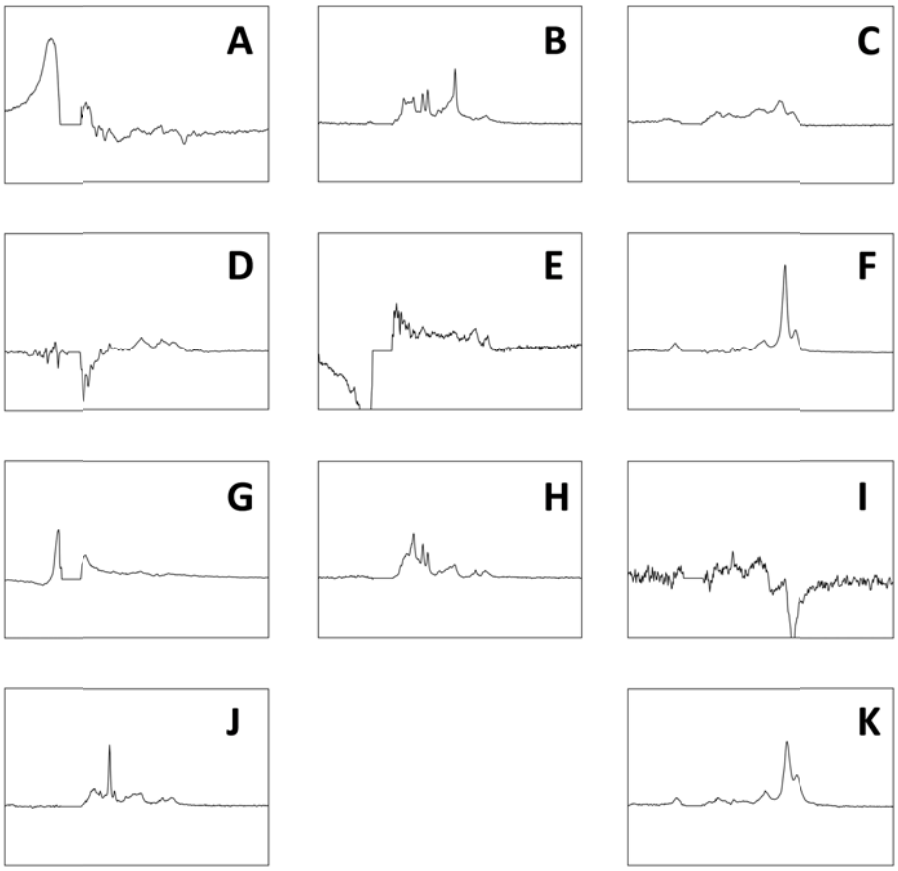

Fig. 3. Gallery of extracted sources: source A corresponds to bad water suppression (bad), B corresponds to normal brain tissue (good), C is low SNR (bad), D is badly phased (bad), E is a combination of bad water suppression and low SNR (bad), F is necrotic tissue (good), G corresponds to bad water suppression and bad homogeneity(bad), H would correspond to the spectral pattern of a low grade glial tumour (good), I is bad phasing and low SNR (bad), J represents the spectral pattern of meningioma (good), and K represents a source necrosis (good).

4 Discussion

From Figure 3, we see that we have obtained different results to those expected according to our hypothesis; 5 sources represent good spectra. They correspond to normal brain tissue, necrotic tissue, low grade glial tumor, and meningioma, and necrosis again (Figure 3: B, F, H, J, K).

These results are in line with previous work that showed that a clear separation could be drawn between these classes when dealing only with good quality spectra [11, 18]. The two sources corresponding to necrosis are in line with the results reported in [18, 20]. Regarding the artifactual sources, we observe that four of them (Figure 3: A, C, D, G) can be regarded as pure artifacts, while the rest can be regarded as a combination of two superimposed artefacts (Figure 3: E, I). Because five sources corresponding to the good class were found, we re-run the source extraction experiment for 13 sources (the five *good* sources plus the eight *bad* sources from our initial hypothesis; we subtracted one because we saw that water suppression tends to mix with other artefacts). In this iteration we

	Accuracy	*AUC*	*Sensitvity*	*Specificity*	*BER*	*F1 score*
LR	0.85	0.84	0.25	**0.97**	0.21	0.91
LRA	0.77	0.84	**0.72**	0.78	0.32	0.85
LRCV	0.85	0.85	0.27	0.96	0.22	0.91
LRCVA	0.78	0.85	0.68	0.80	0.30	0.85
RF	0.85	**0.86**	0.37	0.95	0.21	0.91
RFA	**0.86**	**0.86**	0.34	0.96	**0.20**	**0.92**
LDA	0.85	**0.86**	0.35	0.95	0.21	0.91
ADA	0.85	0.85	0.42	0.94	0.22	0.91
NB	0.83	0.82	0.45	0.91	0.27	0.90

Table 1. Performance metrics for the investigated classifiers. Best performances are shown in bold.

found the same five sources corresponding to good spectra, plus eight artifactual sources. The nature of the rest of the artifactual sources was the same as that of the ones previously extracted; they mix differently, however. Classification results improve marginally with the addition of the 2 extra sources and are not reported herein.

Regarding the classification results, it is important to take into account that the class imbalance was approximately 8 to 2 in favor of the *good* class, which implies that accuracies in Table 1 could have been achieved just by always predicting the good class. Because of this, the most important metric to take into account is the sensitivity (how good the model is at telling if a predicted bad spectrum is really a bad spectrum), where almost all classifiers perform poorly – indicating that the problem at hand cannot be well-described by our dimensionality reduction scheme. However, LRA and LRCVA exhibit good sensitivity, which can be justified by the fact that they give higher weights for the *bad* class. This would imply that our feature space is linearly separable, but does not obey a normal distribution (otherwise LDA would have also had a high sensitivity, which is not the case).

As we previously mentioned, the original labeling of the analyzed data was performed by two or three expert spectroscopists; some cases were labeled as *bad*, but still one of the experts deemed it to be of acceptable quality. As such, we removed those 'borderline' spectra from the testing phase of our system, in order to see if the performance metrics would improve. The results, presented in Table 2, show that sensitivity improved for all classifiers, while other metrics did not change significantly.

One limitation of our system is that the cNMF optimization process is known to fall into local minima and, up to date, there is no way to assess whether the algorithm converged to the optimal minimum or not. In our case, this would translate into obtaining slightly different sources depending on the run of the algorithm (and thus on initialization). This issue is meant to be addressed in future work.

5 Conclusion

We have presented a system that tries to address an important issue in MRS for brain tumour analysis as a metabolomics problem: data quality control. Our system used cNMF as a dimensionality reduction and artifact-identification scheme and then investigated a range of classifiers in the task of discriminating between good and bad quality spectra.

By using LRA, a sensitivity of 0.72 and a specificity of 0.78 was achieved, and by taking out the 'borderline' cases, sensitivity was increased to a value of 0.76. Our results indicate that proper separation between the two classes can be achieved, but further investigation is needed.

	Accuracy	*AUC*	*Sensitvity*	*Specificity*	*BER*	*F1 score*
LR	0.88	0.86	0.30	**0.97**	**0.20**	**0.93**
LRA	0.78	0.86	**0.76**	0.78	0.33	0.86
LRCV	0.88	**0.88**	0.32	0.96	0.21	**0.93**
LRCVA	0.79	0.87	0.73	0.80	0.32	0.87
RF	0.88	**0.88**	0.44	0.95	0.21	**0.93**
RFA	**0.89**	**0.88**	0.44	0.92	0.21	**0.93**
LDA	0.88	**0.88**	0.40	0.92	0.21	**0.93**
ADA	0.88	0.86	0.46	0.94	0.22	**0.93**
NB	0.86	0.84	0.50	0.91	0.27	0.91

Table 2. Classification metrics after removing the samples that were labeled good by one expert spectrosopist and bad by the other two.

Bibliography

1. Julia-Sape, M., Coronel, I., Majos, C., Candiota, A.P., Serrallonga, M., Cos, M., Aguilera, C., Acebes, J.J., Griffiths, J.R., Arus, C.: Prospective diagnostic performance evaluation of single-voxel 1H MRS for typing and grading of brain tumours. Nmr Biomed 25, 661-673 (2012)
2. Stagg, C., Rothman, D.L.: Magnetic resonance spectroscopy: tools for neuroscience research and emerging clinical applications.Ch. 3. Academic Press (2013)
3. 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 19, 22-33 (2006)
4. Pérez-Ruiz, A., Julià-Sapé, M., Mercadal, G., Olier, I., Majós, C., Arús, 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 Bioinform 11, 581 (2010)
5. Ortega-Martorell, S., Olier, I., Julia-Sape, M., Arus, C.: SpectraClassifier 1.0: a user friendly, automated MRS-based classifier-development system. BMC Bioinform 11, 106 (2010)
6. Oz, G., Alger, J.R., Barker, P.B., Bartha, R., Bizzi, A., Boesch, C., Bolan, P.J., Brindle, K.M., Cudalbu, C., Dincer, A., Dydak, U., Emir, U.E., Frahm, J., Gonzalez, R.G., Gruber, S., Gruetter, R., Gupta, R.K., Heerschap, A., Henning, A., Hetherington, H.P., Howe, F.A., Huppi, P.S., Hurd, R.E., Kantarci, K., Klomp, D.W., Kreis, R., Kruiskamp, M.J., Leach, M.O., Lin, A.P., Luijten, P.R., Marjanska, M., Maudsley, A.A., Meyerhoff, D.J., Mountford, C.E., Nelson, S.J., Pamir, M.N., Pan, J.W., Peet, A.C., Poptani, H., Posse, S., Pouwels, P.J., Ratai, E.M., Ross, B.D., Scheenen, T.W., Schuster, C., Smith, I.C., Soher, B.J., Tkac, I., Vigneron, D.B., Kauppinen, R.A., Group, M.R.S.C.: Clinical proton MR spectroscopy in central nervous system disorders. Radiology 270, 658-679 (2014)
7. Kreis, R.: Issues of spectral quality in clinical 1H-magnetic resonance spectroscopy and a gallery of artifacts. Nmr Biomed 17, 361-381 (2004)
8. van der Graaf, M., Julia-Sape, M., Howe, F.A., Ziegler, A., Majos, C., Moreno-Torres, A., Rijpkema, M., Acosta, D., Opstad, K.S., van der Meulen, Y.M., Arus, C., Heerschap, A.: MRS quality assessment in a multicentre study on MRS-based classification of brain tumours. Nmr Biomed 21, 148-158 (2008)
9. García-Gómez, J.M., Luts, J., Julià-Sapé, M., Krooshof, P., Tortajada, S., Robledo, J.V., Melssen, W., Fuster-García, E., Olier, I., Postma, G.: Multiproject–multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. Magnetic Resonance Materials in Physics, Biology and Medicine 22, 5-18 (2009)
10. Julia-Sape, M., Lurgi, M., Mier, M., Estanyol, F., Rafael, X., Candiota, A.P., Barcelo, A., Garcia, A., Martinez-Bisbal, M.C., Ferrer-Luna, R., Moreno-Torres, A., Celda, B., Arus, C.: Strategies for annotation and curation of translational databases: the eTUMOUR

project. Database : the journal of biological databases and curation 2012, bas035 (2012)
11. Tate, A.R., Underwood, J., Acosta, D.M., Julia-Sape, M., Majos, C., Moreno-Torres, A., Howe, F.A., van der Graaf, M., Lefournier, V., Murphy, M.M., Loosemore, A., Ladroue, C., Wesseling, P., Bosson, J.L., Cabanas, M.E., Simonetti, A.W., Gajewicz, W., Calvar, J., Capdevila, A., Wilkins, P.R., Bell, B.A., Remy, C., Heerschap, A., Watson, D., Griffiths, J.R., Arus, C.: Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR in biomedicine 19, 411-434 (2006)
12. Julià-Sapé, M., Griffiths, J.R., Tate, R.A., Howe, F.A., Acosta, D., Postma, G., Underwood, J., Majós, C., Arús, C.: Classification of brain tumours from MR spectra: the INTERPRET collaboration and its outcomes. NMR in biomedicine 28, 1772-1787 (2015)
13. Wright, A.J., Arus, C., Wijnen, J.P., Moreno-Torres, A., Griffiths, J.R., Celda, B., Howe, F.A.: Automated quality control protocol for MR spectra of brain tumors. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 59, 1274-1281 (2008)
14. Van Gestel, T., Suykens, J.A., Lanckriet, G., Lambrechts, A., De Moor, B., Vandewalle, J.: Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel Fisher discriminant analysis. Neural computation 14, 1115-1147 (2002)
15. Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council 10, 626-634 (1999)
16. Ding, C., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE transactions on pattern analysis and machine intelligence 32, 45-55 (2010)
17. Sauwen, N., Sima, D.M., Van Cauter, S., Veraart, J., Leemans, A., Maes, F., Himmelreich, U., Van Huffel, S.: Hierarchical non-negative matrix factorization to characterize brain tumor heterogeneity using multi-parametric MRI. Nmr Biomed 28, 1599-1624 (2015)
18. Ortega-Martorell, S., Lisboa, P.J., 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 13, 38 (2012)
19. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035. Society for Industrial and Applied Mathematics, (Year)
20. Tate, A.R., Griffiths, J.R., Martinez-Perez, I., Moreno, A., Barba, I., Cabanas, M.E., Watson, D., Alonso, J., Bartumeus, F., Isamat, F., Ferrer, I., Vila, F., Ferrer, E., Capdevila, A., Arus, C.: Towards a method for automated classification of 1H MRS spectra from brain tumours. Nmr Biomed 11, 177-191 (1998)

Acknowledgements. 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.