WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia IJCNN

Towards interpretable classifiers with blind signal separation

Héctor Ruiz
Department of Mathematics and Statistics
Liverpool John Moores University
Liverpool, United Kingdom
H.Ruiz@2010.ljmu.ac.uk

Sandra Ortega-Martorell
Departament de Bioquímica i Biología Molecular
Universitat Autònoma de Barcelona
Cerdanyola del Vallès, Spain
Sandra.Ortega@uab.cat

Ian H. Jarman
Department of Mathematics and Statistics
Liverpool John Moores University
Liverpool, United Kingdom
I.H.Jarman@ljmu.ac.uk

Alfredo Vellido
Department of Computer Languages and Systems
Universitat Politècnica de Catalunya
Barcelona, Spain
avellido@lsi.upc.edu

José D. Martín
Departamento de Ingeniería Electrónica
Universidad de Valencia
Burjassot, Spain
jose.d.martin@uv.es

Enrique Romero
Department of Computer Languages and Systems
Universitat Politècnica de Catalunya
Barcelona, Spain
eromero@lsi.upc.edu

Paulo J.G. Lisboa
Department of Mathematics and Statistics
Liverpool John Moores University
Liverpool, United Kingdom
P.J.Lisboa@ljmu.ac.uk

Abstract—**Blind signal separation (BSS) is a powerful tool to open-up complex signals into component sources that are often interpretable. However, BSS methods are generally unsupervised, therefore the assignment of class membership from the elements of the mixing matrix may be sub-optimal. This paper proposes a three-stage approach using Fisher information metric to define a natural metric for the data, from which a Euclidean approximation can then be used to drive BSS. Results with synthetic data models of real-world high-dimensional data show that the classification accuracy of the method is good for challenging problems, while retaining interpretability.**

Blind signal separation; non-negative matrix factorisation; Fisher information; Riemannian metric; data mapping; magnetic resonance spectroscopy; brain tumour

I. Introduction

Blind signal separation (BSS) is a well-known family of tools to separate complex signals into linear combinations of sources whose joint distribution is close to factorised into a product of independent univariate density functions for the individual sources. This approach is rendered even more interpretable when it is applied in the convex space of positive semi-definite mixing and unmixing matrices [1]. Both the sources themselves and the partial membership of each source class can then be evaluated against prior knowledge.

In our example, synthetic data models are built from single voxel magnetic resonance spectroscopy (MRS) signal corresponding to a neuro-oncology problem. The sources will ideally approximate prototypes for each brain tissue class and the maximal values in each row of the mixing matrix will correspond to the correct binary classification of that observation. In this data set the correct prototype is taken to be the mean of the generating distribution.

In a previous work [2], the authors investigated the application of non-negative matrix factorisation (NMF) methods [3,4] for the extraction of tissue type-specific MRS signal sources in a fully unsupervised mode, to the analysis of an international, multi-centre database that incorporates MRS data corresponding to several types of human brain tumours. The accuracies of the labels inferred for each patient case where comparable to traditional supervised classifiers.

However, there is some instability in the classification arising from mixing in data space, especially for challenging differential assignments such as the discrimination of low astrocytic tumours from high grade and from metastatic growths. This limitation arises because the method is fully unsupervised. This is reflected in the generating modes by wide standard deviations for each class in addition to high dimensionality of the data.

Recently, the Fisher information matrix has been proposed as an effective way to build a meaningful, well determined metric in primary data space with which to disaggregate class-labelled data [5]. This metric defines the natural geometry of data space taking into account both the position and class labelling of each data point. However, the Fisher information metric is non-Euclidean, i.e. it does not respect the triangle inequality, with the consequence that projective methods such as BSS cannot be applied directly in this space.

The natural way to bridge the gap between the non-Euclidean metric and projective spaces is to map the data onto an approximate, rigorous Euclidean metric space. This may be done using data projection methods such as the Sammon mapping, multidimensional scaling (MDS) or the iterative majorisation algorithm (IMA).

In this paper we use synthetic data to study the hypothesis that the three-stage approach consisting on first defining a Fisher Information metric, then approximating the empirical data distribution with a Euclidean projective space onto which, subsequently, Convex-NMF [1] can be applied, results in a natural decomposition of the data with sources that are closer to the true class prototypes and in higher classification accuracies than those obtained using a purely unsupervised implementation of NMF.

II. METHODOLOGY

A. Convex non-negative matrix factorisation

In this study, we use a variant of the non-negative matrix factorisation (NMF) family [3,4], namely Convex-NMF [1]. In conventional NMF methods a non-negative data matrix V (of dimensions d-by-n, where d is the data dimensionality and n is the number of observations) is approximately factorised into two non-negative matrices: the matrix of sources or data basis W (of dimensions d-by-k, where k is the number of sources, and $k < d$) and the mixing matrix H (of dimensions k-by-n, each of whose columns provides the encoding of a data point: the spectrum of an observation in this study). The product of these two matrices provides a good approximation to the original data matrix, in the form $V \approx WH$.

To achieve interpretability, Convex-NMF imposes a constraint that the vectors (columns) defining W must lie within the column space of V, i.e. $W = VA$ (where A is an auxiliary adaptative weight matrix that fully determines W), so that $V \approx VAH$. By restricting W to convex combinations of the columns of V we can, in fact, understand each of the basis or sources as weighted sums of data points. This NMF variant applies to both non-negative and mixed-sign data matrices, and only H and A are constrained to be non-negative. These factors are updated as follows:

$$\begin{aligned} H^T &\leftarrow H^T \sqrt{\frac{(V^TV)^+A \;+\; H^TA^T(V^TV)^-A}{(V^TV)^-A \;+\; H^TA^T(V^TV)^+A}} \\ A &\leftarrow A \sqrt{\frac{(V^TV)^+H^T \;+\; (V^TV)^-AHH^T}{(V^TV)^-H^T \;+\; (V^TV)^+AHH^T}} \end{aligned} \quad , \qquad (1)$$

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. H and A are initialised using K-means clustering, as proposed in [1]. This multiplicative algorithm minimises the reconstruction error given by $\|V\text{-}VAH\|^2$, while ensuring that the elements of the matrices H and A remain nonnegative.

B. Interpretation of Convex-NMF in the context of MRS data

Given that the observed MRS data are of mixed sign, their sources should also be of mixed sign. Thus, understanding W as the source spectra matrix, the sources will be intuitively interpretable and no pre-processing of the spectra will be required in order to make them non-negative, thus preventing any unnecessary loss of information (in the case of our data, losing the information in the negative peaks of the LTE MRS spectra). Due to the fact that it is non-negative by definition, the mixing matrix H can be understood as estimates of the concentration/abundance of the constituent signals.

C. Fisher information metric

The Fisher information (FI) is a measure of the amount of information that a variable x carries about a magnitude θ upon which its probability depends [6]. This is obtained by deriving the logarithm of the conditional probability $p(x|\theta)$ with respect to θ and then calculating the conditional expectation over x with respect to the said probability, which results in the measure being independent on x.

In this work, an alternative definition is used where the roles of x and θ are swapped, and therefore the derivative is of $p(\theta|x)$ with respect to x [7]. This FI is now a function of x and takes the form of a square matrix of the same dimensionality as x, that is, the dimensionality of the data space:

$$\begin{aligned} FI(x) &= E_{p(\theta|x)}\{(\nabla_x \log p(\theta|x))(\nabla_x \log p(\theta|x))^T\} \\ &= -E_{p(\theta|x)}\{\nabla_x^2 \log p(\theta|x)\} \end{aligned} \quad , \qquad (2)$$

where $E_{p(\theta|x)}$ denotes the expectation over the values of θ with respect to $p(\theta|x)$ and ∇_x is the gradient with respect to x.

The motivation behind this modification of the original FI is to use it in data mining applications, where x would be a point in the space of the data and θ would be an auxiliary class variable c. In this scenario, it is easy to define a differential metric using the FI matrix:

$$d(x, x + \Delta x)^2 = \Delta x^T FI(x) \Delta x \; . \qquad (3)$$

This gives the distance between two neighbouring points x and $x+\Delta x$ under the metric defined by the FI matrix.

The interesting property of this metric is that it automatically scales each dimension of the data space according to its degree of relevancy with respect to class membership, expanding directions along which $p(c|x)$ changes rapidly and compressing those where the variation is little. The result is a Riemannian space where the posterior class membership probability changes evenly in all directions.

D. Multi-layer perceptron

A crucial stage in the development of the FI metric is the estimation of $p(c|x)$. The ability of the metric to precisely reflect similarity between data points into distances is conditioned by how accurate the probability function on which the FI is based is.

Our choice for an estimator is a multi-layer perceptron (MLP), a feedforward artificial neural network whose versatility makes it ideal for highly non-linear data distributions. The MLP is present in the initial learning step of the process, where its internal weights are trained with the labelled dataset. The perceptron can then estimate probabilities for unlabelled instances in this semi-supervised manner.

E. Dataset projection

After estimating the class membership probability it is possible to compute distances between any two points x_A and x_B in the data space by solving the following path integral along the geodesic path:

$$d(x_A, x_B) = \left| \int_{x_A}^{x_B} \sqrt{\dot{x}(t)^T FI\big(x(t)\big)\dot{x}(t)}\, dt \right|, \qquad (4)$$

where $x(t)$ is the shortest path that goes from x_A to x_B in the space defined by the Fisher metric.

This integral is not directly solvable for the non-linear case; one way around this is to approximate it by dividing the geodesic path into a number of segments whose distance can be obtained using the linear solution to the integral [5].

At this point, a pairwise distance matrix is produced that contains the Fisher distances between every pair of points in the dataset. This will be used to map the dataset from the original primary data space into a Euclidean feature space of the desired dimensionality. We compare several projection methods, namely the Sammon mapping, MDS and IMA.

1) Sammon mapping: This algorithm is used to analyse multivariate data by mapping the data points from an original high dimensional space to a space of lower dimensionality [8]. This non-linear mapping is based on the preservation of the original pairwise distances between data points when moving to the new data space.

The algorithm considers the situation of having N points in a space of dimensionality L that we want to map to another space of dimensionality D, and defines another set of N vectors in this D-space. The distance between points x_i and x_j in the original space is given by d_{ij}*, and the distance between their corresponding maps in the D-space is denoted by d_{ij}. An initial set of mapped points is generated, which result in a value for the following error function, also known as Sammon's stress:

$$E = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j}^{N} \frac{\left(d_{ij}{}^{*} - d_{ij}\right)^2}{d_{ij}{}^{*}} \,. \qquad (5)$$

The position of the points in the D-space is then iteratively adjusted to reduce the error. In most cases, the method used to minimize this error function is gradient descent.

The distance measure for d_{ij}* is usually Euclidean. However, since we want to take into account the prior information that we have about the data in the form of class labels, we use the Fisher metric to compute these distances.

2) Metric multidimensional scaling: The idea on which metric MDS is based, as in Sammon mapping, is the preservation of the original pairwise distances in the projected space. The error function usually has a similar shape as (5) [9]. In this work, we use

$$E = \frac{1}{N^2} \sum_{i<j}^{N} \left(d_{ij}{}^{*} - d_{ij}\right)^2 . \qquad (6)$$

The main difference with (5) lies in the absence of normalisation of the squared differences of the distances. Similar to Sammon mapping, the position of the set of mapped points is iteratively adjusted by gradient descent so as to minimise the error function. Again, the distances d_{ij}* are computed with the Fisher metric.

3) Iterative majorisation algorithm: The last algorithm in this section expresses the mapping from an original L-space to a D-space as a function $f(x;W)=W^T\cdot\Phi(x)$, where W is a P-*by*-D matrix containing the free parameters and $\Phi(x)=(\Phi_1(x), \ldots , \Phi_P(x))^T$ contains the values of the P basis functions $\Phi_i(x)$. The mapping $f(x;W)$ is a linear combination of these basis functions, which can be linear or non-linear. In this work, we have used $P=N$ with $\Phi(x_i)=(d_{i1}{}^{*}, d_{i2}{}^{*}, \ldots , d_{iN}{}^{*})^T$, where $d_{ij}{}^{*}$ is the Fisher distance between points x_i and x_j. The method tries to minimise the error function

$$E = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(d_{ij}{}^{*} - q_{ij}(W)\right)^2 , \qquad (7)$$

where $q_{ij}(W)=\|W^T(\Phi(x_i)-\Phi(x_j))\|$. This is minimised with respect to the weights W using the iterative majorisation algorithm. More detail on this can be found in [10].

III. Experimental Results

A. Description of the data

The data analysed in this study are modelled from samples extracted from a database used in a previous publication [2]. Class (tumour type) labelling was used to generate posterior distributions of the data density, i.e. p(data|class) using single multivariate normal models fitted to the mean and variance/covariance matrices of class specific cohorts of single-voxel proton MRS (SV-^{1}H-MRS) acquired at two different echo times (short, 20-32 ms (STE) and long, 135-144 ms (LTE)) from brain tumour patients.

The analysed data set included, at LTE, samples of the generated data for 20 astrocytomas grade II (A2), 78 glioblastomas (GL), and 31 metastases (ME); at STE, it included 22 A2, 86 GL, and 38 ME. The data dimensionality is 195 reflecting the clinically-relevant frequency intensity values measured in parts per million (ppm) that are typically sampled from each spectrum in the [4.24,0.50] ppm interval.

A second dataset was generated for the validation of the methods. In this dataset, each class has 50 samples generated using the same means and covariance matrices used for the training set. Note that, during the experiments, the metric is derived using the training dataset labels only, and then it is applied to calculate pairwise distance matrices of the training and validation datasets, so no usage of the validation class labels is made, as is expected from a semi-supervised learning method.

All parameters mirror the actual data as closely as possible. The aim of using generated data is to be able to test the proposed methodology against known ground truth.

B. *Empirical results*

Pairwise classification between types of brain tissue were performed, paying attention to the accuracy of the results and the quality of the sources obtained.

Accuracy is measured as the ratio of correctly classified cases out of the total number of instances, and the quality of the sources is determined in terms of how similar they are compared to the mean spectrum of the corresponding class. Similarity is assessed using the correlation between the resulting sources and mean spectra. Each classification problem is run 20 times to average the effect of the variation that the initialisation of Convex-NMF causes on the results.

The classification results on training data (Table I) show a general improvement on the performance of the original approach when we use the Fisher metric pre-processing before applying Convex-NMF, whether it is using the Sammon mapping, MDS or IMA. The increase of the accuracy on the validation dataset (Table III) is smaller, as one would expect, but still significant.

There are some exceptions where Convex-NMF performs better than the alternatives, as in GL vs. ME at STE, when using MDS with training data; in A2 vs. ME at STE, when using Sammon with validation data; and in A2 vs. GL at STE, where Convex-NMF outperforms all three mappings with validation data. Despite that, there is an overall tendency of accuracy increase.

Regarding the quality of the sources (Tables II and IV), all four approaches yield very good sources in general. The most interesting case is GL against ME both at STE and LTE, where plain Convex-NMF does not perform as well as in the other classifications. This is because GL and ME types have a very similar spectral pattern (their mean spectra have correlations of 0.9891 at STE and 0.9211 at LTE between each other), which also explains why the classification accuracies are so low for those two cases. The Fisher metric alternatives manage to obtain very high source correlations and accuracies even for the GL vs. ME problem, due to the additional information that they bring into Convex-NMF from the auxiliary data labels through the Fisher metric.

In the accuracy tables I and III, each cell contains the amount of well classified samples in percentage and also in fraction form in brackets. The column *Original* refers to the results obtained applying plain Convex-NMF, and the other subheaders identify the projection method used to map the data. In the source correlation tables, the two scores within each cell correspond to the correlation between the source of the corresponding tissue type and the true mean spectrum of that class.

TABLE I. CLASSIFICATION ACCURACIES FOR THE TRAINING DATASET

		STE				**LTE**			
		Original	***Sammon***	***MDS***	***IMA***	***Original***	***Sammon***	***MDS***	***IMA***
A2 vs. GL	total	83.3% (90/108)	96.8% (104.6/108)	96.3% (104/108)	99.1% (107/108)	60.2% (59/98)	99% (97/98)	99% (97/98)	99% (97/98)
	A2	100% (22/22)	90.5% (19.9/22)	100% (22/22)	100% (22/22)	100% (20/20)	100% (20/20)	100% (20/20)	100% (20/20)
	GL	79.1% (68/86)	98.4% (84.7/86)	95.3% (82/86)	98.8% (85/86)	50% (39/78)	98.7% (77/78)	98.7% (77/78)	98.7% (77/78)
A2 vs. ME	total	98.3% (59/60)	99.8% (59.9/60)	100% (60/60)	100% (60/60)	86.3% (44/51)	100% (51/51)	99.9% (50.9/51)	100% (51/51)
	A2	100% (22/22)	99.5% (21.9/22)	100% (22/22)	100% (22/22)	100% (20/20)	100% (20/20)	100% (20/20)	100% (20/20)
	ME	97.4% (37/38)	100% (38/38)	100% (38/38)	100% (38/38)	77.4% (24/31)	100% (31/31)	99.8% (30.9/31)	100% (31/31)
GL vs. ME	total	69.8% (86.6/124)	82.2% (102/124)	69.6% (86.4/124)	88.7% (110/124)	60.6% (66/109)	95.6% (104.2/109)	94.5% (103/109)	97.2% (106/109)
	GL	61.1% (52.6/86)	77.3% (66.5/86)	57.4% (49.4/86)	86% (74/86)	55.1% (43/78)	95.6% (74.6/78)	92.3% (72/78)	96.2% (75/78)
	ME	89.5% (34/38)	93.4% (35.5/38)	97.4% (37/38)	94.7% (36/38)	74.2% (23/31)	95.5% (29.6/31)	100% (31/31)	100% (31/31)

TABLE II. SOURCE CORRELATIONS FOR THE TRAINING DATASET

		STE				LTE			
		Original	***Sammon***	***MDS***	***IMA***	***Original***	***Sammon***	***MDS***	***IMA***
A2 vs. GL	A2 GL	0.98 0.96	0.89 1	0.99 1	0.94 1	0.98 0.70	0.98 1	0.97 1	0.98 1
A2 vs. ME	A2 ME	0.98 0.99	0.93 1	0.94 1	0.99 1	0.99 0.88	0.99 1	0.98 0.99	0.99 1
GL vs. ME	GL ME	0.95 0.98	0.99 1	0.97 1	0.97 1	0.73 0.86	0.99 0.99	0.99 1	0.99 1

TABLE III. CLASSIFICATION ACCURACIES FOR THE VALIDATION DATASET

		STE				LTE			
		Original	***Sammon***	***MDS***	***IMA***	***Original***	***Sammon***	***MDS***	***IMA***
A2 vs. GL	total	94% (94/100)	85.5% (85.5/100)	90% (90/100)	90% (90/100)	80% (80/100)	98% (98/100)	95% (95/100)	97% (97/100)
	A2	98% (49/50)	82.8% (41.4/50)	88% (44/50)	84% (42/50)	100% (50/50)	100% (50/50)	92% (46/50)	100% (50/50)
	GL	90% (45/50)	88.1% (44.1/50)	92% (46/50)	96% (48/50)	60% (30/50)	96% (48/50)	98% (49/50)	94% (47/50)
A2 vs. ME	total	97% (97/100)	93.8% (93.8/100)	99% (99/100)	99% (99/100)	88% (88/100)	99.9% (99.9/100)	99.9% (99.9/100)	100% (100/100)
	A2	98% (49/50)	92.9% (46.5/50)	98% (49/50)	98% (49/50)	100% (50/50)	100% (50/50)	100% (50/50)	100% (50/50)
	ME	96% (48/50)	94.6% (47.3/50)	100% (50/50)	100% (50/50)	76% (38/50)	99.9% (49.9/50)	99.9% (49.9/50)	100% (50/50)
GL vs. ME	total	67.7% (67.7/100)	72.8% (72.8/100)	75% (75/100)	74% (74/100)	62% (62/100)	81.8% (81.8/100)	82% (82/100)	83% (83/100)
	GL	64% (32/50)	71.3% (35.7/50)	58.1% (29.1/50)	64% (32/50)	52% (26/50)	84.8% (42.4/50)	92% (46/50)	86% (43/50)
	ME	71.4% (35.7/50)	74.2% (37.1/50)	91.9% (45.9/50)	84% (42/50)	72% (36/50)	78.7% (39.4/50)	72% (36/50)	80% (40/50)

TABLE IV. SOURCE CORRELATIONS FOR THE VALIDATION DATASET

		STE				LTE			
		Original	***Sammon***	***MDS***	***IMA***	***Original***	***Sammon***	***MDS***	***IMA***
A2 vs. GL	A2 GL	0.96 0.96	0.98 1	0.99 1	0.99 1	0.99 0.80	1 0.99	1 0.99	1 0.98
A2 vs. ME	A2 ME	0.96 0.99	0.93 0.94	0.98 0.99	0.99 1	1 0.94	0.99 0.96	0.99 0.97	1 1
GL vs. ME	GL ME	0.94 0.98	1 1	0.98 1	0.98 1	0.68 0.89	0.99 0.99	0.99 0.99	0.99 0.99

Figures 1 to 5 represent the sources involved in a particular classification problem, GL against ME at LTE for the validation dataset. We have chosen to illustrate this case because it is where Convex-NMF struggles most to achieve good classification rates and sources due to the prototypical spectra of the two tissue types being very similar, as can be seen in figure 1. It is therefore for a problem like this that the Fisher metric approaches will show a more significant performance improvement.

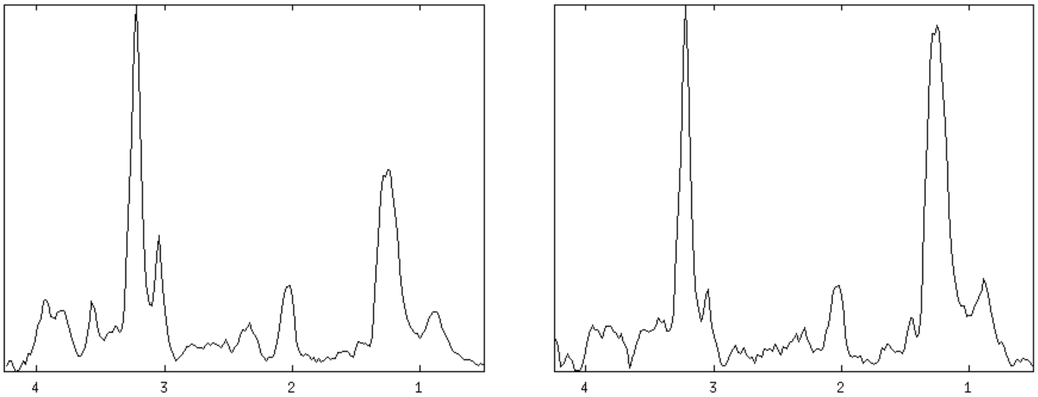

Figure 1. True GL (left) and ME (right) prototype spectra. X-axis: frequencies in ppm scale. Y-axis: Intensities normalised to unit length (UL2)

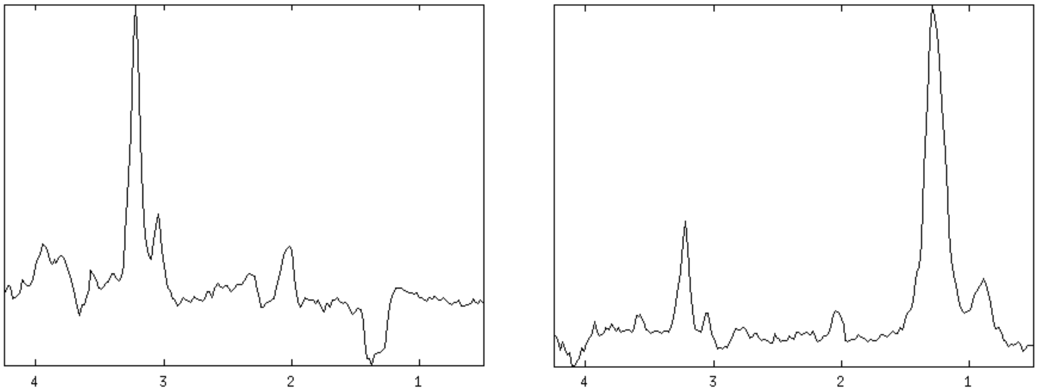

Figure 2. GL (left) and ME (right) sources retrieved using Convex-NMF. X-axis: frequencies in ppm scale. Y-axis: Intensities normalised to unit length (UL2)

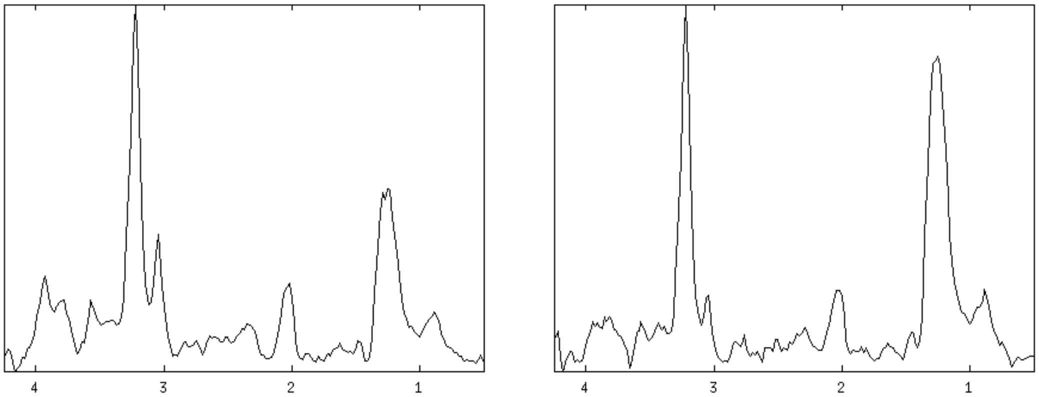

Figure 3. GL (left) and ME (right) sources retrieved using Sammon map. X-axis: frequencies in ppm scale. Y-axis: Intensities normalised to unit length (UL2)

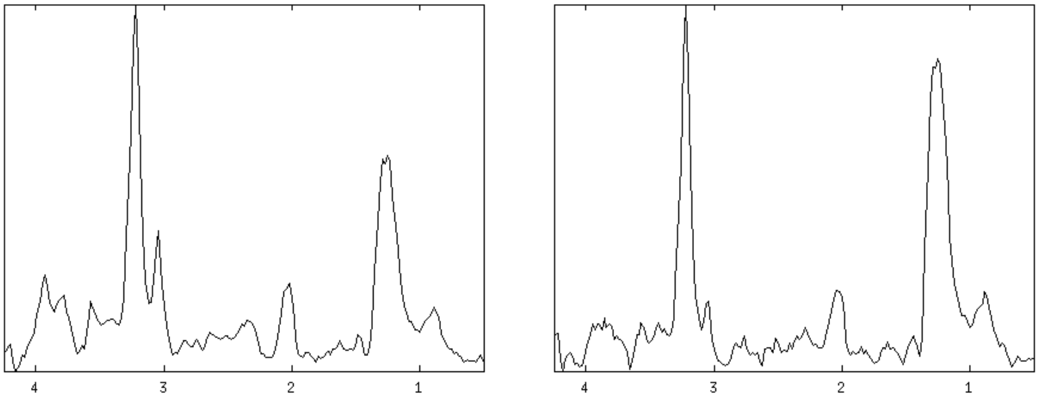

Figure 4. GL (left) and ME (right) sources retrieved using MDS. X-axis: frequencies in ppm scale. Y-axis: Intensities normalised to unit length (UL2)

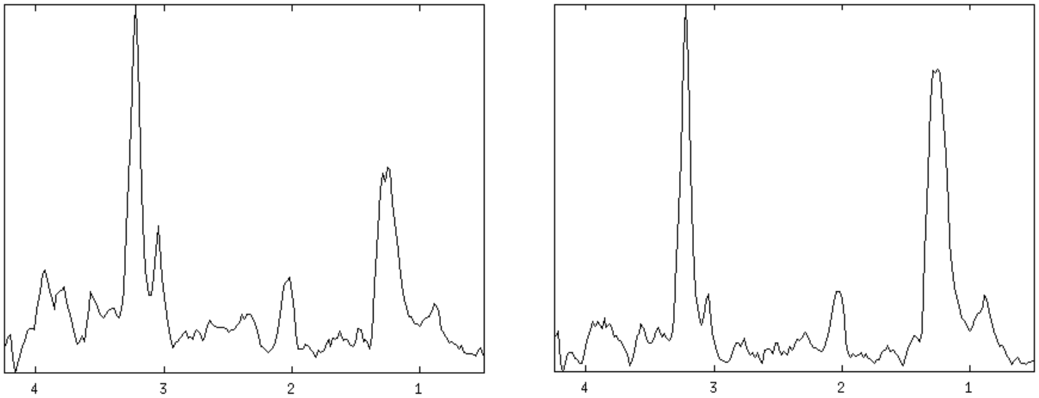

Figure 5. GL (left) and ME (right) sources retrieved using IMA. X-axis: frequencies in ppm scale. Y-axis: Intensities normalised to unit length (UL2)

The resemblance between the true spectra of the two classes is very high in terms of the position and height of the peaks. The only clear differences between them are the height ratio of the two main peaks and the height of the small one immediately to the right of the left main peak. We can see in figure 2 how the basic Convex-NMF does not manage to get both peaks right at the same time. Instead, it identifies each class with one of the two main peaks. The rest of the methods, however, pick up correctly the proportion of the height of the peaks for both classes and produce sources very similar to the originals.

IV. Discussion

The results of the experiments performed confirm the hypothesis that an unsupervised interpretable method for blind classification/signal separation, namely as Convex-NMF, can benefit from the use of known data labels and result in a more accurate classifier without any loss in the interpretability of the results.

Moreover, a mechanism is provided that achieves blind signal separation with a semi-supervised approach, by first finding a natural metric to describe the class assignments, followed by a mapping of the data into an approximate distribution in a Euclidean space where the blind signal separation can be applied with standard projective methods.

Furthermore, for the data analysed in this work, not only was the accuracy of the classification of the samples generally better, but the sources extracted were also of higher quality than those obtained using the original unsupervised method, both in the training and validation stages. In our opinion, the improvement on these two aspects, especially in complex classification problems, justifies the additional pre-processing steps that precede the original approach.

Future work is to replicate this methodology on the original medical data.

References

[1] C. Ding, T. Li, and M.I. Jordan, “Convex and semi-nonnegative matrix factorizations,” IEEE Transactions on pattern analysis and machine intelligence, vol. 32, no. 1, pp. 45-55, 2010.

[2] S. Ortega-Martorell, P.J.G. Lisboa, A. Vellido, M. Julià-Sapé, and C. Arús, “Non-negative Matrix Factorisation methods for the spectral decomposition of MRS data from human brain tumours,” BMC Bioinformatics, vol. 13, no. 38, 2012

[3] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” Environmetrics, vol. 5, no. 2, pp. 111-126, 1994.

[4] D. Lee and H. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788-791, 1999.

[5] H. Ruiz, I.H. Jarman, J.D. Martín, and P.J.G. Lisboa, “The role of Fisher information in primary data space for neighbourhood mapping,” European Symposium on Artificial Neural Networks (ESANN) 2011 proceedings.

[6] S. Amari, “Information geometry on hierarchy of probability distributions,” IEEE Information theory, vol. 47, no. 5, pp. 1701-1711, 2001.

[7] S. Kaski and J. Sinkkonen, “Metrics that learn relevance,” International Joint Conference on Neural Networks (IJCNN) 2000 proceedings, vol. 5, pp. 547-552, 2000.

[8] J.W. Sammon, “A nonlinear mapping for data structure analysis,” IEEE Transactions on computers, vol. C-18, no. 5, pp. 401-409, 1969.

[9] J.A. Lee and M. Verleysen, “Nonlinear Dimensionality Reduction”, Springer-Verlag, NY, 2007.

[10] Z. Zhang, “Learning metrics via discriminant kernels and multidimensional scaling: Toward expected Euclidean representation”, International Conference on Machine Learning (ICML) 2003 proceedings, pp. 872-879, 2003.