Functional topographic mapping for robust handling of outliers in brain tumour data

Alfredo Vellido[1]*, Paulo J.G. Lisboa[2]

[1]Universitat Politècnica de Catalunya (UPC). Soft Computing Group.
C. Jordi Girona, 1-3, 08034, Barcelona, Spain.
[2]Liverpool John Moores University (LJMU). Neural Computation Group.
Byrom St, L3 3AF, Liverpool, U.K.

Abstract. Magnetic Resonance spectra comprise finite frequency measurements sampled from a continuous frequency distribution and, therefore, are amenable to Functional Data Analysis (FDA) techniques. In this paper, MR spectral data are considered with the purpose of discriminating between brain tumour types. Models to fit these data can be affected by the uncertainty associated to the presence of outliers. A functional variation on a model for data clustering and visualization, the *t*-GTM, is introduced. It is defined as a mixture of Student t-distributions that is robust towards outliers. The effectiveness of this model for outlier detection and tumour type visualization is compared for raw and functional data.

1 Introduction

Magnetic Resonance Spectroscopy (MRS) is a non-invasive tool capable of providing a detailed fingerprint of the biochemistry of living tissue. Decisions made on the basis on Magnetic Resonance Imaging (MRI) information can sometimes be uncertain and somehow biased by the subjectivity of the expert. The additional information conveyed by MR spectra can help the clinical expert by disambiguating diagnostic and prognostic decisions [1].

In this paper, we deal with the decision problem of brain tumour discrimination from single-voxel MRS spectra. Spectrometric data are sampled at discrete frequencies over a finite range. Under this premise, MRS data can be analyzed using Functional Data Analysis (FDA) techniques [2]. Several neural network models, including Self-Organizing Maps (SOM:[3]), have recently been adapted to functional data [4,5,6]. FDA, in this context, becomes a tool for feature extraction and offers an alternative to variable selection for dimensionality reduction. This is a priority in MRS, where spectra are usually characterised by their high dimensionality.

A potential source of uncertainty in diagnosis and prognosis based on MRS is the existence of data outliers. The decision support neural network-based model that is subject of this study is a redefinition of the standard Generative Topographic Mapping (GTM:[7]) for multivariate data clustering and visualization. The GTM can be understood both as a probabilistic alternative to SOM and as a constrained mixture of distributions. It was originally defined as a mixture of Gaussian distributions but, especially for small sample sizes, this is prone to lack robustness in the presence of

[*] Alfredo Vellido is a Ramón y Cajal Research Fellow under a program of the Spanish Ministry of Education and Science.

outliers. Some recent studies have suggested multivariate Student *t*-distributions as a robust alternative to Gaussians for mixture models. In this paper, the GTM is redefined as a constrained mixture of Student *t*-distributions, termed *t*-GTM [8], which is capable of processing functional data. The experiments reported in this exploratory study compare the performance of the *t*-GTM, as method for outlier detection and visualization of grouped data, when using either a selection of the original variables or several functional data projections.

2 *t*-GTM as a constrained mixture of *t*-distributions

The GTM is a non-linear latent variable model that defines a mapping from a low dimensional latent space onto the multivariate data space (as opposed to the projection from the data space to the visualization space performed by the SOM). The mapping is carried through by a set of basis functions generating a (mixture) density distribution, and it is defined as a generalized linear regression model:

$$\mathbf{y} = \boldsymbol{\Phi}(\mathbf{u})\mathbf{W} \qquad (1)$$

where $\boldsymbol{\Phi}$ is a set of M basis functions $\boldsymbol{\Phi}(\mathbf{u}) = (\phi_1(\mathbf{u}),...,\phi_M(\mathbf{u}))$, $\mathbf{W}$ is a matrix of adaptive weights w_{md}, and $\mathbf{u}$ is a point in latent space. This latent visualization space can be discretized as a regular grid of K latent points $\mathbf{u}_k$, similar to that of the SOM. A probability distribution for the data can then be defined, leading to an expression for the complete log-likelihood $L_c(\mathbf{W},\beta|\mathbf{X})$, and the Expectation-Maximization (E-M) algorithm can be used to obtain the Maximum Likelihood (ML) estimates of the adaptive parameters $\mathbf{W}$ and β. Details can be found in [7].

For the Gaussian GTM, the presence of outliers is likely to negatively bias the estimation of parameters $\mathbf{W}$ and β, especially for small sample sizes. It is also likely to result in extreme estimates of the posterior probabilities of component membership [9], leading to distortions in the mapping that will affect the data clustering and its visualization. To overcome this limitation, the GTM can be redefined as a constrained mixture of Student *t*-distributions: the *t*-GTM. Assuming now that the basis functions $\boldsymbol{\Phi}$ are Student *t*-distributions, the data probability can be defined as

$$P(\mathbf{x}|\mathbf{u},\mathbf{W},\beta,\nu) = \frac{\Gamma\left(\nu/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu/2\right)(\nu\pi)^{D/2}}\left(1 + \beta/\nu\,\|\mathbf{y} - \mathbf{x}\|^2\right)^{-\frac{\nu+D}{2}}, \qquad (2)$$

where $\Gamma(\cdot)$ is the gamma function and ν can be understood as a tuner that adapts the level of robustness (divergence from normality) for the mixture. This leads to a new complete log-likelihood:

$$L_c(\mathbf{W},\beta,\nu|\mathbf{X}) = \sum_{n=1}^{N} log\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{\Gamma\left(\nu_k/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu_k/2\right)(\nu_k\pi)^{D/2}}\left(1 + \beta/\nu_k\,\|\mathbf{y}_k - \mathbf{x}_n\|^2\right)^{-\frac{\nu_k+D}{2}}\right\}.(3)$$

From this expression, ML estimates of the adaptive parameters $\mathbf{W}$ and β can be calculated, using the E-M algorithm. For details, see [10].

According to [9], a given data instance could be identified as an outlier if the value of

$$O_n^* = \sum_k \hat{z}_{kn} \beta \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2 , \tag{4}$$

which is explicitly calculated at the M-step of the E-M algorithm for the *t*-GTM, was sufficiently large.

2.1 Functional MRS data analysis using the *t*-GTM

MR spectroscopic data are sampled at discrete frequencies, usually selected as the location of spectral peaks of known metabolic indicators. For the data analyzed in this study, detailed in the next section, choice of metabolites may be important for the accurate discrimination of tumour types. As sampled functions, MRS data can be analysed using FDA techniques [2]. A SOM model, adapted to work on functional data, was recently proposed in [6]. In FDA, given an observed data point $\mathbf{x}_n \in \Re^D$, where D is the data dimensionality, we assume the existence of $\mathbf{x}_n^* \in \Re^D$ and a smooth function g, so that, $\forall n \in \{1,\ldots,N\}$, $\mathbf{x}_n = g(\mathbf{x}_n^*) + \varepsilon_n$, where ε_n is the measurement error for data point n. This function g can be represented by its projection on the space spanned by a number P of basis functions φ_i, such that

$$\sum_{d=1}^{D} \left(\mathbf{x}_n - \sum_{p=1}^{P} \alpha_i \varphi_i \left(\mathbf{x}_n^* \right) \right)^2 \tag{5}$$

is minimized. In practice, each data point $\mathbf{x}_n$ can be replaced by a vector of coefficients $\boldsymbol{\alpha}$ of lower dimensionality. The lower the chosen P, the less flexible the data representation becomes, which is equivalent to a coarse smoothing of the assumed underlying function. If all data points share the same dimensionality, this becomes a feature extraction method for dimensionality reduction. Following [6,11], functions are represented in this study on a fixed truncated basis, using B-splines. When these functional data are fed to the *t*-GTM, the model effectively performs curve clustering and visualization on a low-dimensional space.

3 MRS and brain tumour data

The data used in this study consist of 98 single voxel PROBE (PROton Brain Exam system) spectra acquired *in vivo* for five viable tumour types: Astrocytes, Glioblastomas, Metastases, Meningiomas, and Oligodendrogliomas, as well as cystic regions from tumours, which are likely to be outliers due to their differences in composition from the tumours themselves. A description of the automated protocol used for data acquisition can be found in [12]. The spectra were digitised, sampling the region known to contain clinically relevant metabolic information, into 194 frequency intensity values. Such high dimensionality makes either feature extraction or variable selection necessary. In [12], a process based on Multivariate Bayesian Variable Selection was shown to provide a good description of the data set in the form of a selection of 6 frequency intensities, corresponding to Fatty Acids, Lactate, a compound-unassigned peak, Glutamine, Choline, and Taurine-Inositol. Feature extraction is an alternative to variable selection for dimensionality reduction. An example of this is Independent Component Analysis (ICA) that was used in [12] to extract features that were independent MRS signal sources. A further alternative method is FDA, described in section 2.1, which makes use of the intrinsic properties of spectrometric data. In this framework, the raw 194 variables are projected onto a basis of B-spline functions using 6, 12, and 18 basis functions to account for increasing levels of complexity. These sets of features will be compared with the aforementioned selection of 6 variables for their accuracy in outlier identification. Their relative ability to discriminate between tumour groups will also be explored.

4 Experiments

Figure 1 displays histograms of the statistic (4) for the following data sets: a selection of 6 variables, as described in the previous section, on the top row; a description of the spectra using a 6 B-spline basis, on the bottom row, left; using a 12 B-spline basis, on the bottom row, centre; and using a 18 B-spline basis, on the bottom row, right. The results for the selection of 6 variables are extracted from [8]: the 7 data instances with largest values of (4) are cystic regions and, in more detail, 14 out the 17 cystic regions in the data set fall within the three highest decile intervals of (4). In comparison, for the 6 B-spline projection, 5 out of the 7 data instances with largest values of (4) are cystic regions, and 9 out the 17 cystic regions are included in the three highest deciles. For the 12 B-spline projection, 4 out of the 7 data instances with largest values of (4) are cystic regions, and 13 out the 17 cystic regions are included in the three highest deciles. Finally, for the 18 B-spline projection, 6 out of the 7 data instances with largest values of (4) are cystic regions, and 13 out the 17 cystic regions are included in the three highest deciles. In summary, functional approximations of the data, obtained by projection onto a low-dimensional subspace of basis functions, preserve sufficient information for the *t*-GTM to identify outliers with a reasonable accuracy. This accuracy, though, is lower than the one provided by the variable selection, which is also easier to interpret.

Some preliminary results, concerning the ability of the *t*-GTM to discriminate between different tumour types using the different data descriptions, can be found in Figure 2, where Glioblastomas are visualized against cystic regions (more thorough comparisons are omitted, for the sake of brevity). The selection of 6 variables yields a clear discrimination between spectra from the two tissue types. Most strikingly, the discrimination for the 6 B-spline projection is almost as good, whereas those for the 12 and 18 B-spline projections are rather poor. From this point of view, the 6 B-spline projection is filtering the irrelevant information in the full data set more efficiently than the higher dimensional projections.

5 Conclusion

The presence of outliers is a potential source of uncertainty in diagnostic and prognostic decision making based on MR spectroscopic information. MR spectra can be considered discrete instances of generalized functional forms and, as such, can be analyzed using FDA techniques. Functional forms of Artificial Neural Networks have

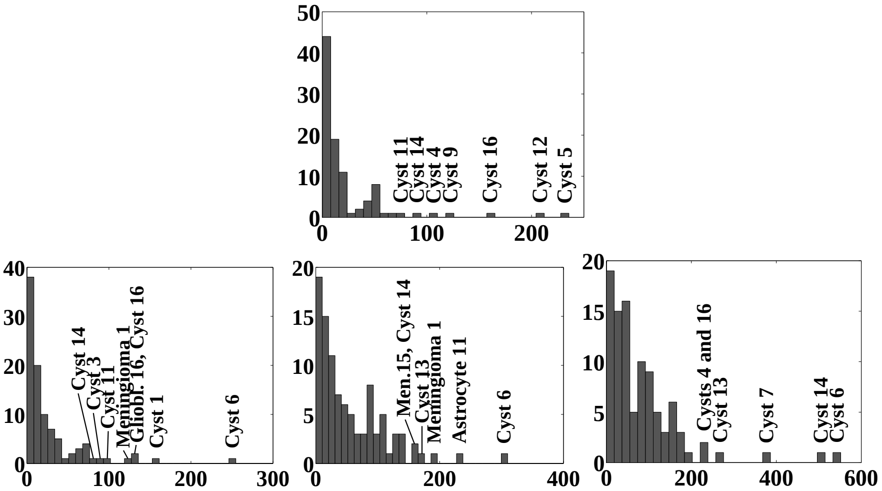

Figure 1: Histograms of the statistic (4) for several data sets resulting from different dimensionality reduction strategies. On the top row, results for a set of six variables selected in [12]. On the bottom row results for several FDA solutions: From left to right, 6 B-splines, 12 B-splines and 18 B-splines. In order to allow comparisons, the most extreme outliers have been labelled.

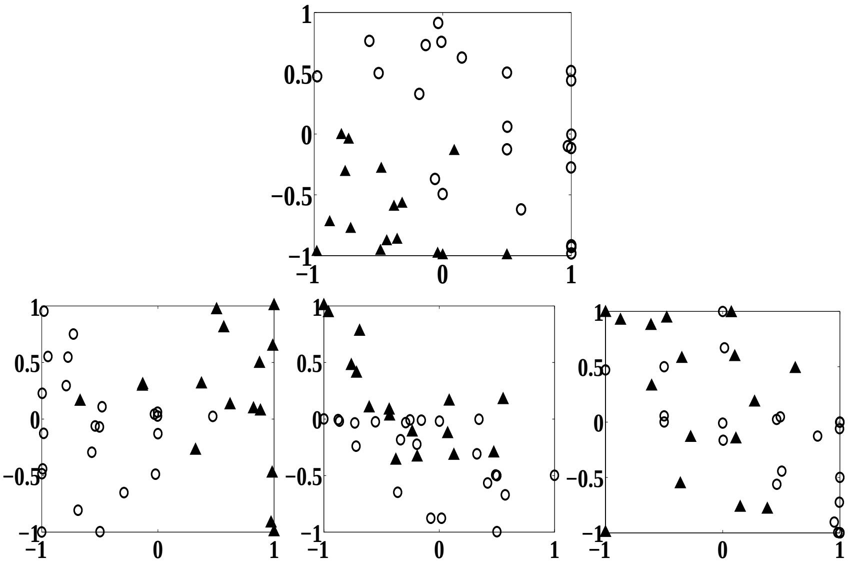

Figure 2: The GTM allows visualizing multivariate data in a low-dimensional latent space. This is a visualization, on the *t*-GTM 2-D space, of Glioblastomas (circles) and cystic regions (triangles). On the top row: visualization of the results for a selection of 6 relevant variables [12]. On the bottom row: from left to right, visualization of the results for FDA solutions with 6, 12, and 18 B-splines.

recently been proposed, and they are likely to bear considerable potential as tools for automated decision support in medical applications. FDA, as applied to the MRS data used in this study can be regarded as a feature extraction process.

A functional variant of the GTM defined as a constrained mixture of *t*-distributions has been defined. This was shown to behave robustly in the presence of functional outliers for a parsimonious description of the data set that reduces its dimensionality by more than 90%. However, the functional descriptions of the data do not compare positively with the very restrictive variable selection from [12]. This result seems to suggest that a dimensionality reduction procedure based on a well-reasoned variable selection strategy preserves relevant data information better than a FDA-based strategy.

References

[1] P.J.G. Lisboa, W. El-Deredy, Y.Y.B. Lee, Y. Huang, A.R. Corona Hernandez and P. Harris, Characterisation of Brain Tissue from MR Spectra for Tumour Discrimination. In in H. Yan, editor, *Signal Processing for Magnetic Resonance Imaging and Spectroscopy*, pages 569-588, Marcel Dekker, New York, 2002

[2] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, New York, Springer-Verlag, 1997.

[3] T. Kohonen. *Self-organizing Maps (3rd ed.)*, Springer-Verlag, Berlin, 2000.

[4] F. Rossi and B. Conan-Guez, Functional multi-layer perceptron: A non-linear tool for functional data analysis, *Neural Networks*, in press, 2005.

[5] N. Delannay, F. Rossi, B. Conan-Guez and M. Verleysen, Functional radial basis function networks (FRBFN), In M. Verleysen, editor, *proceedings of the 12th European Symposium on Artificial Neural Networks* (ESANN 2004), D-Side Pub., pages 313-318, Bruges (Belgium), 2004.

[6] F. Rossi, B. Conan-Guez and A. El Golli, Clustering Functional Data with the SOM algorithm, In M. Verleysen, editor, *proceedings of the 12th European Symposium on Artificial Neural Networks* (ESANN 2004), D-Side Pub., pages 305-312, Bruges (Belgium), 2004.

[7] C.M. Bishop, M. Svensén and C.K.I. Williams, GTM: The Generative Topographic Mapping, *Neural Computation*, 10:215-234, 1998.

[8] A. Vellido, P.J.G. Lisboa and D. Vicente, Handling outliers and missing data in brain tumor clinical assessment usign *t*-GTM, submitted for publication, ESANN 2005.

[9] D. Peel and G.J. McLachlan, Robust mixture modelling using the t distribution, *Statistics and Computing*, 10:339–348, 2000.

[10] A. Vellido. Generative Topographic Mapping as a constrained mixture of Student *t*-distributions: Theoretical developments, Technical Report LSI-44-7-R, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, 2004.

[11] C. Abraham, P.-A. Cornillon, E. Matzner-Lober and N. Molinari, Unsupervised curve clustering using b-splines, *Scandinavian Journal of Statistics*, 30:581-595, 2003.

[12] Y. Huang, P.J.G. Lisboa and W. El-Deredy, Tumour grading from Magnetic Resonance Spectroscopy: A comparison of feature extraction with variable selection, *Statistics in Medicine,* 22:147-164, 2003.