# Exploratory Characterization of Outliers in a Multi-centre $^1$H-MRS Brain Tumour Dataset

Alfredo Vellido[1,⋆], Margarida Julià-Sapé[2,3], Enrique Romero[1], and Carles Arús[3,2]

[1] Dept. de Llenguatges i Sistemes Informàtics - Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3. 08034, Barcelona, Spain
{avellido,eromero}@lsi.upc.edu
http://www.lsi.upc.edu/~websoco/AIDTumour
[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Spain
[3] Grup d'Aplicacions Biomèdiques de la RMN (GABRMN)
Departament de Bioquímica i Biología Molecular (BBM). Unitat de Biociències
Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
marga@carbon.uab.es,carles.arus@uab.es

**Abstract.** As part of the AIDTumour research project, the analysis of MRS data corresponding to various tumour pathologies is used to assist expert diagnosis. The high dimensionality of the MR spectra might obscure atypical aspects of the data that would jeopardize their automated classification and, as a result, the process of computer-based diagnostic assistance. In this paper, we put forward a method to overcome this potential problem that combines automatic outlier detection, visualization through dimensionality reduction, and expert opinion.

**Keywords:** Proton Magnetic Resonance Spectroscopy, Brain Tumours, Outlier Detection, Data exploration, Data Visualization, Dimensionality Reduction; Medical Decision Support Systems.

## 1 Introduction

Decision making in oncology is a sensitive matter, and even more so in the specific area of brain tumour oncologic diagnosis, for which the direct and indirect costs - both human and financial - of misdiagnosis are very high. In this area, in which most diagnostic techniques must be non-invasive, clinicians should benefit from the use of an at least partially automated computer-based medical Decision Support System (DSS).

AIDTumour (Artificial Intelligence Decision Tools for Tumour diagnosis [1]) is a research project for the design and implementation of a medical DSS to assist experts in the diagnosis of human brain tumours on the basis of data

⋆ A. Vellido is a Spanish M.E.C. Ramón y Cajal researcher. The authors acknowledge funding from M.E.C. projects TIN2006-08114 and SAF2005-03650, and, from 1$^{st}$ January 2003, Generalitat de Catalunya (grant CIRIT SGR2005-00863).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 189–196, 2008.

obtained by Magnetic Resonance Spectroscopy (MRS). This is a technique that can shed light on cases that remain ambiguous after clinical investigation. The MRS data used in AIDTumour and analyzed in this paper belong to a complex multi-centre set containing cases of several brain tumour pathologies [2]. These data have undergone a rigorous pre-processing quality control that validates them from the viewpoint of the radiologists. Nevertheless, and for their use in an automated computer-based DSS, the various origins of these spectra and the complexity of their pre-processing make further data exploration advisable.

It might be problematic to include some of the spectra in an automated DSS without further ado for three different reasons: Firstly, some may contain measurement or acquisition artifacts that, even if not completely precluding diagnosis by visual inspection, might induce errors in computer-based diagnosis: these are what we call here *artifact-related outliers*. Secondly, atypical cases that do not contain artifacts but are nevertheless unrepresentative of the main distributions of the whole dataset: herein, these will be referred to as *distinct outliers* [3]. Thirdly, some cases with a clear biopsy-based diagnosis (tumour type attribution) may yield spectra that are quantitatively similar to those of other tumour types, misleading a computer-based classification system. Even if representative of the data as a whole, they are still unrepresentative of their own tumour type: these we will call *class outliers*. Note that these three reasons are not always mutually exclusive.

In this paper, we show the effectiveness of a method to identify and characterize potentially conflicting MRS data that combines techniques of dimensionality reduction, exploratory visualization, and outlier detection, with expert knowledge. The introduction of the latter is paramount, as it will help to skim those cases truly conflictive out of those shortlisted by blind quantitative criteria. Overall, this method is conceived as a preliminary step to data classification in the DSS. Dimensionality reduction is not trivial in this setting, as the available MRS data are scarce and high dimensional. Sammon's mapping [4] is used to this end. Generative Topographic Mapping (GTM [5]), a manifold learning model, is used to quantify spectra atypicality.

## 2 MRS Data

The analysed MRS data correspond to 217 short-echo time (SET) and 195 long-echo time (LET) single voxel $^1$H MR spectra acquired in vivo from brain tumour patients. They include 58 (SET) and 55 (LET) meningiomas (*mm*), 86 (SET) and 78 (LET) glioblastomas (*gl*), 38 (SET) and 31 (LET) metastases (*me*), 22 (SET) and 20 (LET) astrocytomas grade II (*a2*), 6 (SET and LET) oligoastrocytomas grade II (*oa*), and 7 (SET) and 5 (LET) oligodendrogliomas grade II (*od*). For details on data acquisition and processing, see [2]. Class labelling was performed according to the World Health Organization (WHO) system for diagnosing brain tumours by histopathological analysis of a biopsy sample. For the reported analysis, spectra were bundled into three groups, namely: G1: *low grade gliomas* (*a2*, *oa* and *od*); G2: *high grade malignant tumours* (*me* and *gl*); and

G3: *meningiomas*. The clinically-relevant regions of the spectra were sampled to obtain 195 frequency intensity values (measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector), from 4.25 parts per million (ppm) down to 0.56 ppm, which become data attributes.

## 3 Methods

### 3.1 MRS Data Visualization through Sammon's Mapping

In order to allow the visualization of the data through dimensionality reduction, the spectra were mapped onto a 3-D space through Sammon's mapping [4]. The non-linear mapping is constructed as to minimize the inter-point distortions it introduces, quantified by Sammon's error measure:

$$\frac{1}{\sum_{i<j}\delta_{ij}}\sum_{i<j}\frac{(\delta_{ij}-\xi_{ij})^2}{\delta_{ij}}, \tag{1}$$

where $\delta_{ij}$ is the Euclidean distance between spectra $i$ and $j$ in the original data space and $\xi_{ij}$ is the Euclidean distance between the projections of these spectra in the 3-D space. In this study, the minimization of the Sammon's error was performed by the Newton method. A collection of models was obtained by varying the initial points (100 different random values) and the step size (9 different values), for a total of 900 runs. The models with lowest Sammon's error were selected for further analysis.

### 3.2 Outlier Detection Using *t*-GTM

Generative Topographic Mapping (GTM [5]) is a non-linear latent variable model defined as a mapping from a low dimensional latent space onto the multivariate data space. The mapping is carried through by a set of basis functions and is defined as a generalized linear regression model:

$$\mathbf{y} = \phi(\mathbf{u})\mathbf{W}, \tag{2}$$

where $\mathbf{W}$ is a matrix of adaptive weights that defines the mapping, and $\mathbf{u}$ is a point in latent space. $\phi$ are $M$ basis functions that, in the original formulation, were chosen to be spherically symmetric Gaussians. For this Gaussian GTM, the presence of outliers is likely to negatively bias the estimation of its adaptive parameters. In order to overcome this limitation, the GTM was recently redefined [3] as a constrained mixture of Student's $t$ distributions: the $t$-GTM. The mapping described by Equation (2) remains, with the basis functions now being Student's $t$ distributions. As a byproduct of this reformulation of GTM, and following [7], a statistic quantifying to what extent $t$-GTM considers a data case to be an outlier can be defined as $O_n = \sum_k p(\mathbf{u}_k|\mathbf{x}_n)\beta\|\mathbf{y}_k - \mathbf{x}_n\|^2$, where $\beta$ is the inverse of the noise variance. The larger the value of this statistic the more likely the case is to be an outlier. Notice that $p(\mathbf{u}_k|\mathbf{x}_n)$ is the responsibility assumed by a latent point $k : 1, \ldots, K$ for the data case $n$ and, the same as for the standard GTM, it is obtained as part of the maximum likelihood estimation of the model's parameters.

### 3.3 Shortlisting Outlier Cases of Interest

The free software package KING [6] is used to visualize in 3-D the Sammon's mapping of the spectra described in section 2, enabling a preliminary data exploration. The data projections obtained with Sammon's mapping were then modelled by $t$-GTM, obtaining a value of $O_n$ for each data case, indicating the corresponding degree of atypicality. Histograms of $O_n$ were generated to shortlist potentially conflictive cases of the three types described in the introduction. Loose thresholds of the statistic were set for the selection of the lists of outlier candidates. Using all this information, an expert in MRS then singled out those spectra she/he considered to be truly atypical in any sense and compared them to the characteristic spectra corresponding to their tumour type.

## 4 Experimental Results and Discussion

### 4.1 Short Echo Time $^1$H MRS Data

The histogram in Fig. 1 displays the distribution of the value of the statistic $O_n$, first calculated for the complete SET MRS dataset. A threshold of $O_n = 20$ was set to shortlist outlier candidate spectra. This yielded 23 potential outliers, which were inspected by an expert who decided that only 19 of them (4 *distinct outliers* and 15 *artifact-related outliers*) qualified as such, for different causes listed in Table 1 (left). Notice that there are plenty of *low grade gliomas* (37% of all outliers, while only 16% of all data). Six different types of artifacts were found in the data, namely: spectra heavily contaminated by noise; bad water signal suppression as part of the data pre-processing; incorrect spectrum alignment of the ppm reference; incorrect baseline; the existence of polispiculated artifact; and signal distorsion due to eddy currents (induced as a result of field gradient switching in signal acquisition).

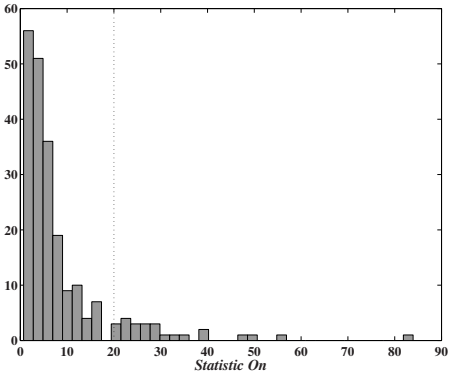


**Fig. 1.** Histogram of statistic $O_n$ for the SET dataset. The selected threshold at value 20 is represented as a vertical dotted line.

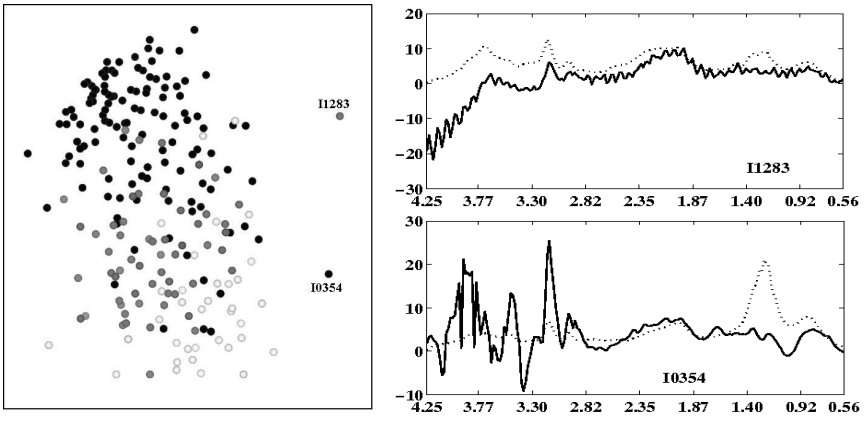


**Fig. 2.** 3-D Sammon's mapping view of two cases of interest (with groups of tumours displayed in different shades of gray), on the left column, and their corresponding individual spectra (solid lines) and mean spectra (dotted lines) of the tumour groups they belong to, on the right column. The abscissa axis displays frequency in ppm.

To illustrate the visualization of the high-dimensional spectra through Sammon's mapping, Fig. 2 displays SET cases I1283 (a meningioma, which the expert described as being contaminated by noise, and affected by bad water suppression, polispiculated effect and eddy currents), and I0354 (a glioblastoma, which the expert described as being affected by a polispiculated effect). Their atypicality is clearly captured by the visualization.

Spectra can also be atypical specifically with respect to their group of tumours. These are what we call *class outliers*. The histograms of $O_n$ for each group of tumours are omitted here for the sake of brevity. Five *low grade gliomas*, 20 *high grade malignant tumours*, and 13 *meningiomas* where shortlisted and inspected by the expert, who considered that, out of these, none of the *low grade gliomas*, only 9 *high grade malignant tumours*, and 8 *meningiomas* should be tagged as *class outliers*. Some of them also contain artifacts, given that, as mentioned in the introduction, *artefact-related outliers* and *class outliers* are not mutually exclusive characterizations. They are described in Table 2 (left). It is very interesting that, even though *low grade glioma* outliers are plentiful, as seen in Table 1 (left), there is no *class outlier* amongst them in the SET spectra, suggesting a well-defined against the rest but less-than-compact structure in this group of tumours.

### 4.2 Long Echo Time $^1$H MRS Data

The histogram in Fig. 3 displays the distribution of $O_n$ for the complete LET MRS dataset. A threshold of $O_n = 15$ was set to shortlist outlier candidate spectra. This yielded 21 potential outliers, which were again inspected by an

**Table 1.** Outlier characterization of the SET (left) and LET (right) $^1$H MRS datasets. Columnwise, *Id* is an anonymized case identifier; star superscripts indicate that there are artifacts that do not preclude the expert's correct interpretation of the case. *Tum* refers to tumour type (see labels in section 2). *Dis* refers to *Distinct outliers*. Six types of artifacts were found: *noi* stands for noise; *wat*, for bad water signal suppression; *ali*, for alignment; *lin*, linebase; *pol*, for the polispiculated effect; and *edd* for eddy currents. See main text for details.

| *Id* | *Tum* | *Dis* | *Artifact-relat. outl.* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | noi | wat | ali | bas | pol | edd |
| I0335 | G1(a2) | X | | | | | | |
| I1052* | G1(a2) | | X | | | | | |
| I1087* | G1(a2) | | | X | | | | |
| I0060 | G1(oa) | | X | | | | | |
| I0069 | G1(oa) | | | | X | | | |
| I0450 | G1(oa) | X | | | | | | |
| I0179 | G1(od) | X | | | | | | |
| I0135* | G2(gl) | | | | | X | | |
| I0172* | G2(gl) | | | X | | X | | |
| I0354* | G2(gl) | | | | | | X | |
| I0421* | G2(gl) | | | X | | | | |
| I1024* | G2(gl) | | | X | | | | |
| I0055 | G2(me) | | X | | | | | |
| I0244* | G3(mm) | X | | | | | | |
| I0375 | G3(mm) | | X | | | | | |
| I0381* | G3(mm) | | | X | | | | |
| I0390* | G3(mm) | | | | | | X | |
| I0393* | G3(mm) | | X | | | X | X | |
| I1283* | G3(mm) | | X | X | | | X | X |

| *Id* | *Tum* | *Dis* | *Artifact-relat. outl.* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | noi | wat | ali | bas | pol | edd |
| I1061 | G1(a2) | | | | X | | | |
| I0062* | G2(gl) | | X | X | | X | | |
| I0105* | G2(gl) | X | | | | | | |
| I0172 | G2(gl) | | | X | | X | | |
| I0175* | G2(gl) | | X | | | | X | |
| I0354* | G2(gl) | | | X | | | X | |
| I0428* | G2(gl) | | | X | | | X | |
| I1044* | G2(gl) | | | | | | X | |
| I1057* | G2(gl) | | X | X | | | X | |
| I1379* | G2(gl) | | X | | | | | X |
| I0027 | G2(me) | | X | | X | | | |
| I0368* | G2(me) | | | X | | | X | |
| I1070 | G2(me) | X | | | | | | |
| I0390* | G3(mm) | | | | | | X | |
| I0420 | G3(mm) | | | | | | X | |
| I1074 | G3(mm) | | X | | | | | |
| I1090 | G3(mm) | X | | | | | | |
| I1378 | G3(mm) | | X | | | | X | |

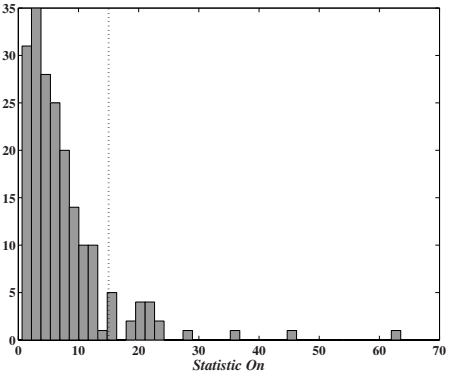


**Fig. 3.** Histogram of statistic $O_n$ for the LET dataset. The selected threshold at value 15 is represented as a vertical dotted line.

expert, who decided that only 18 of them qualified as such (3 *distinct outliers* and 15 *artifact-related outliers*). The corresponding characterization is presented in Table 1 (right). Interestingly, in this case there is almost no *low grade glioma* outlier and, instead, *high grade malignant* outliers predominate (67% of all outliers, while only 56% of all data).

Turning now our attention to *class outliers*, 9 *low grade gliomas*, 7 *high grade malignant tumours*, and 10 *meningiomas* were shortlisted and inspected by the expert, who considered that, out of these, none of the *low grade gliomas*, only 2 *high grade malignant tumours*, and 5 *meningiomas* should be tagged as *class outliers*. Some of them also contain artifacts, and they are characterised in Table 2 (right). It is worth noting that there are far less *class outliers* in the LET dataset than in the SET one, suggesting a much more compact definition of the tumour groups in the former representation. It is also interesting that, again, there is no *class outlier* amongst the *low grade gliomas*. Together with the almost complete lack of outliers in this tumour group shown in Table 1 (right), this indicates that they have a much more compact and well-defined structure in the LET representation.

**Table 2.** *Class outlier* characterization of the SET (left) and LET (right) $^1$H MRS datasets, by groups of tumours. Label description as in Table 1.

| *Id* | *Tum* | *Artifacts* | | | | | |
|---|---|---|---|---|---|---|---|
| | | noi | wat | ali | bas | pol | edd |
| ***Low grade gliomas*** (G1) | | | | | | | |
| ∅ | | | | | | | |
| ***High grade malignant*** (G2) | | | | | | | |
| I0021* | gl | | | | | | |
| I0358* | gl | | | | | X | |
| I0200* | gl | | | | | X | |
| I1390 | gl | | | X | | | |
| I0168* | gl | | | | | | |
| I1098* | gl | | | | | | |
| I1076* | me | | | | X | | |
| I0352* | me | | | | X | | |
| I1377* | me | | | | | | |
| ***Meningiomas*** (G3) | | | | | | | |
| I0160* | mm | | | | | | |
| I1090* | mm | | | | | | |
| I1073* | mm | | | | | X | |
| I0009 | mm | | | | | | |
| I0390* | mm | | | | | X | |
| I1378* | mm | | | | X | | |
| I0375 | mm | X | | | | X | |
| I1149* | mm | | | | | | |

| *Id* | *Tum* | *Artifacts* | | | | | |
|---|---|---|---|---|---|---|---|
| | | noi | wat | ali | bas | pol | edd |
| ***Low grade gliomas*** (G1) | | | | | | | |
| ∅ | | | | | | | |
| ***High grade malignant*** (G2) | | | | | | | |
| I0105* | gl | | | | | | |
| I1070 | me | | | | | | |
| ***Meningiomas*** (G3) | | | | | | | |
| I0114* | mm | | | | | | |
| I1090 | mm | | | | | | |
| I1378 | mm | X | | | | X | |
| I0002* | mm | | | | | | |
| I0009* | mm | | | | | | |

## 5 Conclusion

In this paper, we have defined a method to identify and characterize potentially conflicting MRS multi-center data corresponding to several brain tumour pathologies, which combines dimensionality reduction, outlier detection and exploratory visualization techniques with expert knowledge. This combination of data-based analysis and human expertise is one of the distinctive hallmarks of Evidence-Based Medicine (EBM) for healthcare practice [8]. This method will be embedded in a medical DSS resulting from the AIDTumour [1] research project.

Several research questions would require further research. First, the usefulness of outlier detection and characterization for the improvement of automated tumour diagnostic classification should be assessed. For instance, does the fact the LET MRS data include less *class outliers* mean that we should expect better classification of tumour groups using these and not SET data? Second, only 2 glioblastomas and 1 meningioma tagged as *artifact-related outliers*, and 3 meningiomas tagged as *class outliers* appear both for SET and LET data. How can we explain this level of mismatch? Specific policies to deal with this wide variety of situations should be carefully implemented in the projected DSS.

**Acknowledgments.** Authors gratefully acknowledge the former INTERPRET (EU-IST-1999-10310) European project partners. Data providers: Dr. C. Majós (IDI), Dr.À. Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Dr. W. Gajewicz (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier and Mr. A. Pérez (all from GABRMN-UAB). C. Arús and M. Julià-Sapé are funded by the CIBER of Bioengineering, Biomaterials and Nanomedicine, an initiative of the *Instituto de Salud Carlos* III (ISCIII) of Spain.

## References

1. Artificial Intelligence Decision Tools for Tumour diagnosis research project, `http://www.lsi.upc.edu/~websoco/AIDTumour`
2. Julià-Sapé, M., et al.: A Multi-Centre, Web-Accessible and Quality Control-Checked Database of in Vivo MR Spectra of Brain Tumour Patients. Magn. Reson. Mater. Phy. 19, 22–33 (2006)
3. Vellido, A., Lisboa, P.J.G.: Handling Outliers in Brain Tumour MRS Data Analysis through Robust Topographic Mapping. Comput. Biol. Med. 36, 1049–1063 (2006)
4. Sammon Jr., J.W.: A nonlinear mapping for data structure analysis. IEEE T. Comput. C-18, 401–409 (1969)
5. Bishop, C.M., Svensén, M., Williams, C.K.I.: The Generative Topographic Mapping. Neural Comput. 10(1), 215–234 (1998)
6. KING visualization software, `http://kinemage.biochem.duke.edu/software/king.php`
7. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Stat. Comput. 10, 339–348 (2000)
8. Dickersin, K., Straus, S.E., Bero, L.A.: Evidence Based Medicine: Increasing, not Dictating. Choice. Brit. Med. J. 334(suppl.1), s10 (2007)