Mining the BrainMap™ database: Detection of outliers

Finn Årup Nielsen, Lars Kai Hansen and Ulrik Kjems
Department of Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, DENMARK
(fn,lkhansen,kjems)@imm.dtu.dk

Abstract

We describe a system for modeling BrainMap — a neuroscience database describing activation foci reported from many neuroimaging studies. We apply machine learning techniques in the form of probability density models and use them in order to spot "novelty" among activation foci, and find several outliers that are most likely entry errors. The system can be incorporated into the database entry program.

1 Introduction

Neuroimaging experiments based on positron emission tomography (PET), or functional magnetic resonance imaging (fMRI), are accumulating vast spatio-temporal databases at a rate that calls for new innovative informatics tools. Neuroscience databases are conceptually and physically linked in complex socio-scientific networks of human relations, publications, and funding programs. The neuroinformatics challenge is to organize these networks and make them transparent and useful for the neuroscience community [7]. An important part of neuroinformatics concerns the process of relating different functional neuroimaging studies to each other.

A functional neuroimaging study based on fMRI or PET examines the neural correlate of a mental process as it affects cerebral blood flow. Under the functional segregation paradigm the brain image consists of activation hot spots that each are related to a cognitive component. When discussing the results of a functional study the observed set of activation foci is compared to foci from other studies reported in the literature. Such discussions are naturally driven primarily by the informal neuroscientific insight of the researchers involved. Description of the set of foci can be given in terms of lobes or gyri (e.g. dorsolateral prefrontal cortex), in terms of an informal set of functional areas (such as visual areas V1, V2, ...), or in cytoarchitectural areas — labeled by the so-called Brodmann numbers [2]. An increasingly popular, formal and quantitative alternative is to report foci with reference to the Talairach system [9] — a standardized Euclidean system of reference.

The BrainMap™ database is an extensive collection of papers containing Talairach coordinates from human brain mapping studies maintained by the Research Imaging Center, University of Texas Health Science Center at San Antonio [4]. The

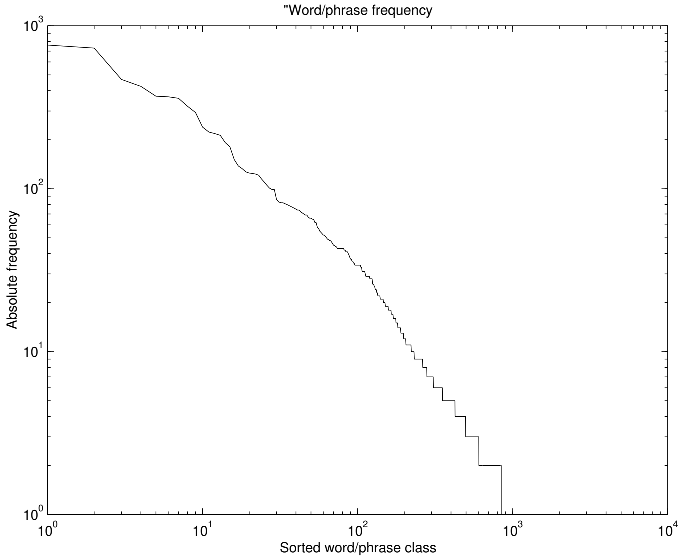


Figure 1: The 1231 classes sorted according to frequency. The most occurring words are "gyrus", "cortex" and "frontal". The most occurring phrase is "frontal gyrus" as the 12th most frequent class. There are approximately 600 classes with more than two examples.

access to the database is provided by either a graphical user interface application [6] or a web-based interface, see http://ric.uthscsa.edu. The database contains bibliographic information about the paper, formal descriptions of the study (e.g. modality, behavioral domain, response type) and 3D coordinates and associated anatomical labels.

It is important for meta-analysis that the underlying database is correct. If it contains examples that are somehow "controversial", they should be labeled so. We use simple and fully automatic machine learning algorithms to model the three-dimensional densities of Talairach coordinates and identify such outlying entries and report these as ranked according to their novelty.

2 Methods

We downloaded most of the paper, experiment and location webpages from the BrainMap. From the location webpages (each containing one Talairach coordinate) we obtained the values from two fields: "Coordinates in Talairach, 1988 space:" and "Lobar anatomy". The values (the strings) from "Lobar anatomy" were tokenized using non-letter characters as separators. All words and all phrases where recorded and given their own class, e.g., the string "midline occipital lobe" generated an event in the classes of "midline", "occipital", "lobe", "midline occipital", "occipital lobe" and "midline occipital lobe".

We were able to download 7263 location webpages. 3935 of these locations had an associated anatomical label, thus went into one or several of the word/phrase classes. There were 1231 word/phrase classes, see figure 1.

We next construct a probability density model in the three-dimensional Talairach space $\mathbf{x}$ by conditioning on the word/phrase w class: $p(\mathbf{x}|w)$.

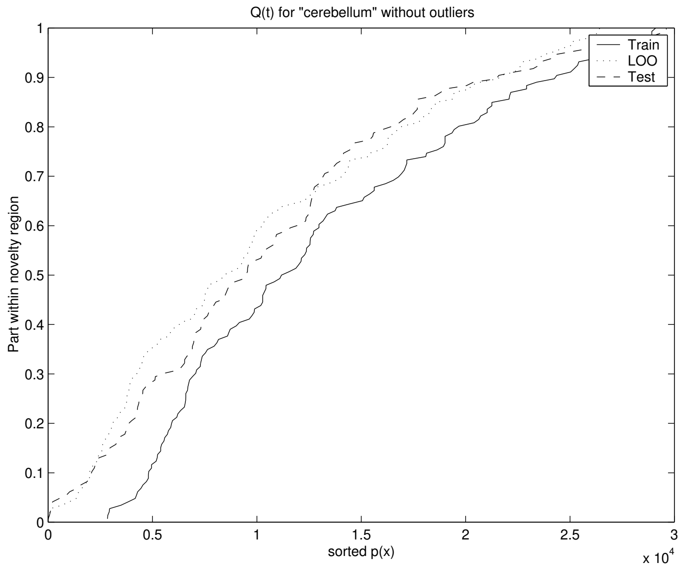


Figure 2: Accumulated histograms for the probability density for reported locations labelled as "cerebellum".

We use a relative simple estimator to model the probability density: A variation of a Specht kernel estimator [8], where the width of the Gaussian kernel (σ^2) is optimized by leave-one-out (LOO) cross-validation. Our implementation is based on a fast one-dimensional Newton optimization of the leave-one-out cost function (negative log probability),

$$E(\sigma^2, w) = -\sum_{n=1}^{N_w} \log p_{-n}(\mathbf{x}_n | \sigma^2, w), \tag{1}$$

where $\mathbf{x}_n$ is the three-dimensional Talairach coordinate of the n'th location with label w, N_w is the number of locations labeled w, and the density based on all examples except the n'th is given by

$$p_{-n}(\mathbf{x}|\sigma^2, w) = \frac{1}{N_w - 1} \sum_{n' \neq n}^{N_w} \left(2\pi\sigma^2\right)^{-3/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{x}_{n'})^2\right). \tag{2}$$

As the probability density estimation will be affected by outliers we use a two-stage strategy: In the first stage we use all coordinates with the class to obtain the probability density. In the second stage we exclude the 5% most unlikely coordinates and estimate the probability density on the remaining 95%. This will introduce a bias in the final probability density but as the main goal here is outlier detection we have not made an attempt to correct this bias for the present experiments.

Once we have a suitable probability density model we are able to evaluate new studies. Novelty detection is implemented using the estimated density value $p(\mathbf{x})$ as test statistic. We simply rank the locations according to their densities with potential outliers among the low density values.

A p-value for novelty can be assigned by converting the densities by:

$$Q(t) = \mathrm{Prob}(\mathbf{x} \in \mathcal{R}(t)), \quad \mathcal{R}(t) = \{\mathbf{x} :\ p(\mathbf{x}) < t\} \tag{3}$$

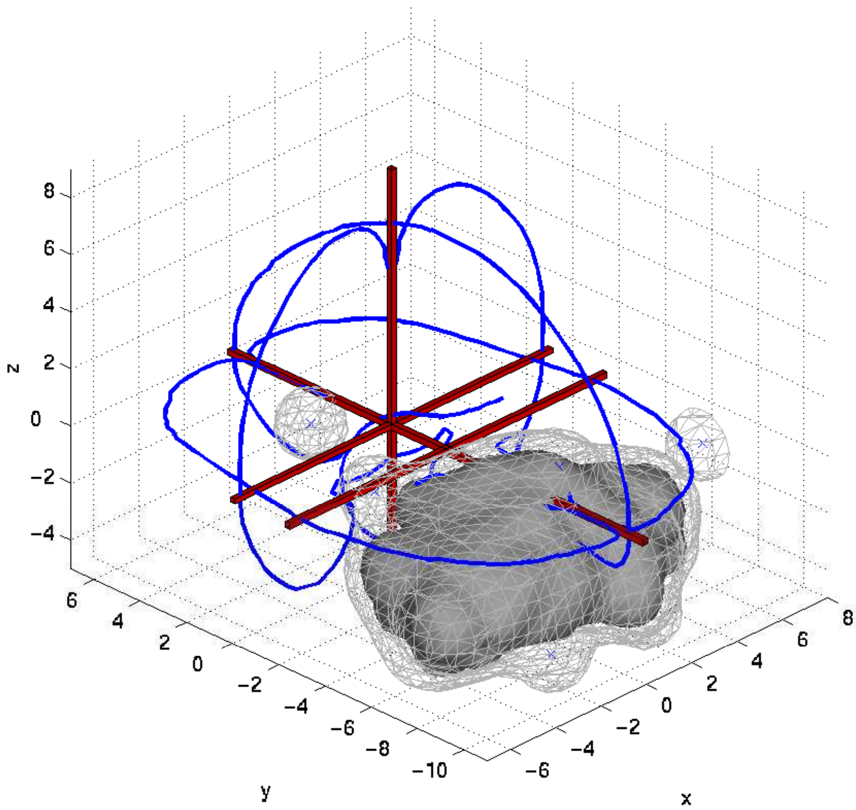
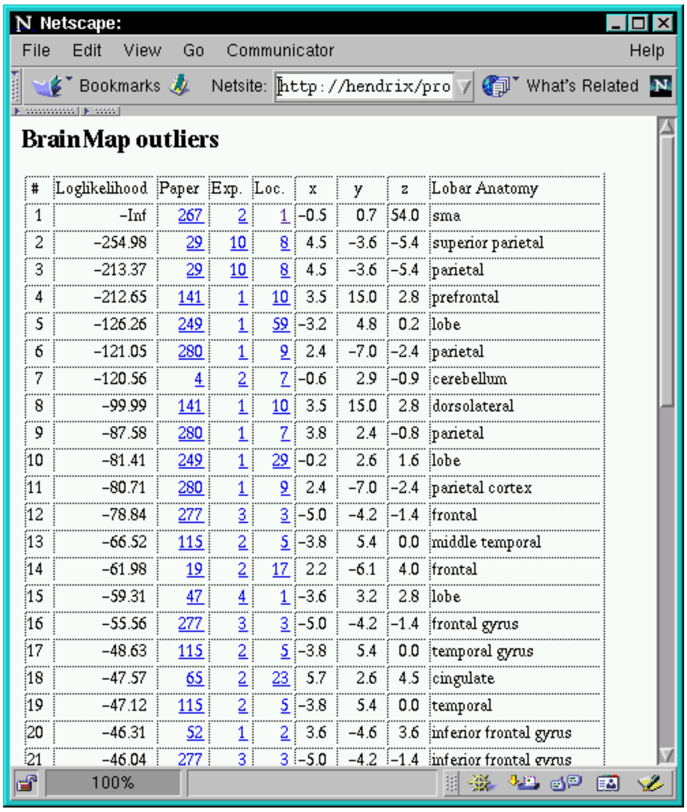


Figure 3: Probability density estimate of the "cerebellum" class in Talairach space. The wireframe-model is the first stage probability density estimation where all the locations are included and the polygon model is the second stage probability density estimate where the 5% most extreme are excluded. Note that two isolated "blobs" created by isolated, outlying, locations were eliminated going from the first to the second level density

The function $Q(t)$ is estimated on a test set of locations not used for estimating the density, see [5] and references herein. For a new set of locations the p-value is simply the relative rank of the new $Q(t)$ values with respect to the test set, e.g., if a location ranks 10'th among 1000 test foci, the p-value is 0.01. In figure 2 we show the $Q(t)$-functions for the training set, a test-set, and the leave-one-out estimate based on the training set. The two latter curves are very similar, suggesting that we in fact need not sacrifice test-data for evaluation of the $Q(t)$ function.

3 Results

In figure 3 we present the density formed by modeling the 294 entries labelled "cerebellum". The first level density is shown as wireframe, and the second level trimmed density is shown as a rendered surface. Note that two isolated "blobs" that were created by isolated outlying locations were eliminated going from the first to the second level density.



BrainMap outliers

#	Loglikelihood	Paper	Exp.	Loc.	x	y	z	Lobar Anatomy
1	-Inf	267	2	1	-0.5	0.7	54.0	sma
2	-254.98	29	10	8	4.5	-3.6	-5.4	superior parietal
3	-213.37	29	10	8	4.5	-3.6	-5.4	parietal
4	-212.65	141	1	10	3.5	15.0	2.8	prefrontal
5	-126.26	249	1	59	-3.2	4.8	0.2	lobe
6	-121.05	280	1	9	2.4	-7.0	-2.4	parietal
7	-120.56	4	2	7	-0.6	2.9	-0.9	cerebellum
8	-99.99	141	1	10	3.5	15.0	2.8	dorsolateral
9	-87.58	280	1	7	3.8	2.4	-0.8	parietal
10	-81.41	249	1	29	-0.2	2.6	1.6	lobe
11	-80.71	280	1	9	2.4	-7.0	-2.4	parietal cortex
12	-78.84	277	3	3	-5.0	-4.2	-1.4	frontal
13	-66.52	115	2	5	-3.8	5.4	0.0	middle temporal
14	-61.98	19	2	17	2.2	-6.1	4.0	frontal
15	-59.31	47	4	1	-3.6	3.2	2.8	lobe
16	-55.56	277	3	3	-5.0	-4.2	-1.4	frontal gyrus
17	-48.63	115	2	5	-3.8	5.4	0.0	temporal gyrus
18	-47.57	65	2	23	5.7	2.6	4.5	cingulate
19	-47.12	115	2	5	-3.8	5.4	0.0	temporal
20	-46.31	52	1	2	3.6	-4.6	3.6	inferior frontal gyrus
21	-46.04	277	3	3	-5.0	-4.2	-1.4	inferior frontal gyrus

Figure 4: List of those locations estimated to have the highest novelty. "Paper", "Exp." and "Loc." correspond to the identifiers used in the BrainMap database. X, y, z and "Lobar anatomy" are the associated fields in the database with the coordinates in centimeter and the "loglikehood" is our novelty measure.

Figure 4 shows the highest ranked outliers. The most extreme outlier is termed "SMA" and clearly has a z-coordinate that is wrong: $z = 54.5$ cm. This would correspond to an activation almost half a meter outside the brain. This is probably an entry error where the reported coordinates are given in millimeter and have been interpreted as centimeter during data entering. Other locations with similar errors are entries 4 and 8. These highly abnormal locations could, of course, easily be spotted simply by plotting all the data in the same three-dimensional plot.

A location that could not be spotted this way is the 7'th in figure 4. This location is inside the brain (see also figure 3) and is not an outlier with respect to the complete set of locations. However, when conditioning on the label "cerebellum" it is a strong outlier. A systematic manual screening would then require that the locations are plotted conditioned on may be 100+ classes.

An example where a phrase provides more information than a single word is the 2nd and 3rd entry in figure 4 — both referring to the same BrainMap location. The z-coordinate presumably has the wrong sign, adding "superior" in front of "parietal" makes the location yet more unlikely.

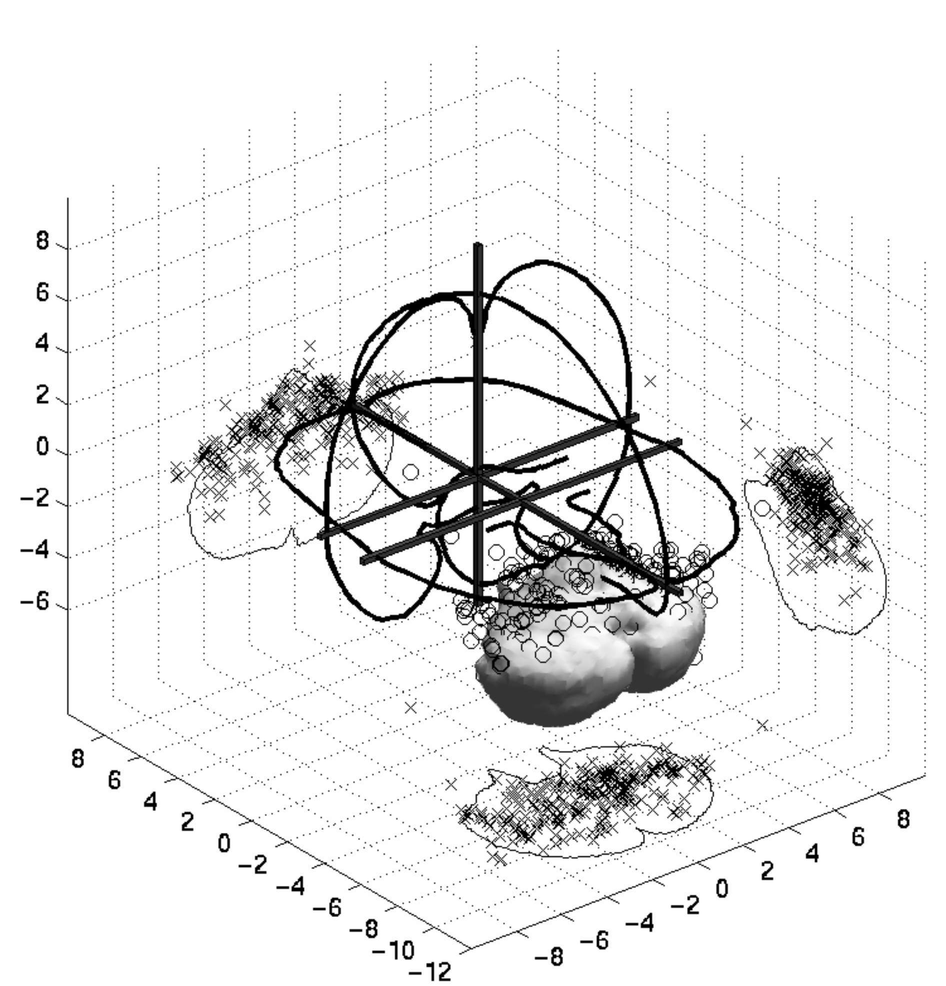


Figure 5: The ICBM cerebellum with all the cerebellum locations from BrainMap. Many of the BrainMap locations are *above* the ICBM cerebellum.

An alternative scheme for detecting outliers in the BrainMap database would be to use an anatomical atlas, i.e., labeled probability volumes. The so-called ICBM atlas [3] being a prominent example. In this atlas the cerebellum has been identified and each voxel in a large cubic box is given a probability for being "cerebellum", see figure 5.[1] Evaluating the BrainMap location in this model we find that many of the locations have zero probability of being "cerebellum": $P(w = \text{"cerebellum"}|\mathbf{x}) = 0$. This is mostly due to variation in the template coordinate system[2], the ICBM template brain simply being bigger than the standard Talairach brain. We suggest as future work to fuse the two volumes by three-dimensional warps and thus provide even better founded models of the accumulated knowledge of neuroimaging.

[1]The volume is a *probability* volume $P(w|\mathbf{x})$, rather than *density* volume $p(\mathbf{x}|w)$, and thus cannot be directly compared with the activation focus densities.

[2]The volume was aligned with the Talairach reference system using an approximation to the Meyer-Lindenberg affine correction [1].

4 Conclusion

We have described a system that can assist the neuroscientists in quality control of neuroimaging data. The system is based on a simple non-parametric kernel density estimator equipped with a scheme for ranking novelty. We found that a leave-one-out based estimate of the density statistic gave a good approximation to the unbiased statistic offered by a test set.

Using the system on the BrainMap database we found a number of entries that were erroneous.

Our approach makes it possible to screen the database for non-trivial errors, such as conditional outliers. We can automatically construct hundreds of conditional densities only limited by the amount of labelled data provided.

The approach can also be used for other meta-analytic tasks, such as evaluating the information content in a set of foci.

Acknowledgments

We thank Research Imaging Center, University of Texas Health Science Center at San Antonio for access to the BrainMap database. This paper has been supported by the Danish research councils through "THOR center for Neuroinformatics" and "Interdisciplinary Neuroresearch" and the American NIH "Human Brain Project" grant R01 DA09246 and P20 MH57180.

References

[1] Brett, M. The MNI brain and the Talairach atlas. Internet: http://www.mrc-cbu.cam.ac.uk/Imaging/mnispace.html, August 1999.

[2] Brodmann, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues.* Barth, Leipzig, 1909.

[3] Evans, A. C., Collins, D. L., and Holmes, C. J. Automatic 3D regional MRI segmentation and statistical probability anatomy maps. In Myers, R., Cunningham, V. J., Bailey, D. L., and Jones, T., editors, *Quantification of Brain Function Using PET*, chapter 25, pages 123–130. Academic Press, Inc, 1996.

[4] Fox, P. T. and Lancaster, J. L. Neuroscience on the net. *Science*, 266(5187):994–996, November 1994.

[5] Hansen, L. K., Kolenda, T., Nielsen, F. Å., Kjems, U., and Larsen, J. Modeling text with generalizable gaussian mixtures. In *Proceedings of ICASSP'2000*. Turkey, June 2000.

[6] Lancaster, J. L., Chan, E., Mikiten, S. A., Nguyen, S., and Fox, P. T. BrainMap^{TM} search and view. In *NeuroImage*, volume 5, page S634, 1997.

[7] Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., Nadkarni, P. M., and Miller, P. L. The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends in Neurosciences*, 21(11):460–468, November 1998.

[8] Specht, D. F. Probabilistic neural networks. *Neural Networks*, 3(1):109–118, 1990.

[9] Talairach, J. and Tournoux, P. *Co-planar Stereotaxic Atlas of the Human Brain.* Thieme Medical Publisher Inc, New York, 1988.