Multivariate analysis with the Brede Toolbox

Finn Årup Nielsen

The Lundbeck Foundation Center for Integrated Molecular Brain Imaging;
Department of Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark;
Neurobiology Research Unit, Rigshospitalet, Copenhagen, Denmark.

Summary

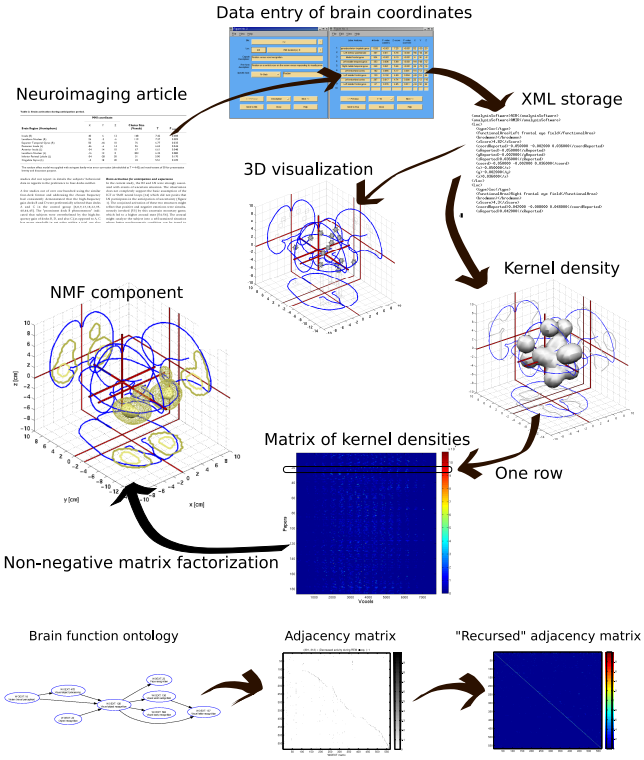
The Brede Toolbox is a Matlab toolbox that implements some of the standard multivariate analysis algorithms, such as K-means, ICA and NMF. The Matlab functions operate on annotated matrices and plotting and HTML generation functions can use this annotation when reporting analysis results, e.g., such that results of hierarchical NMF can be plotted with a single function. The Brede Toolbox has been used for text mining PubMed abstracts, analysis of Wikipedia citations and for analyzing the content of the neuroinformatics Brede Database. The latter analysis involves the formation of multiple kernel density estimates of brain coordinates and subsequent multivariate analyses of these densities with the results containing 3D visualization in brain space added to the Brede Database homepage.

Matlab structures

Most of the functions in the Brede Toolbox typically operates with structures, and, e.g., the matrices are represented in a "mat" structure where rows and columns are annotated either with simple types or with other structures. Example:

```
      type: 'mat'
    matrix: [586x7752 double]
      rows: {586x1 cell}
   columns: [3x7752 double]
description: 'Voxelized experiments'
```

Here the rows are structures of functional neuroimaging "experiments".[1]



Multivariate analysis

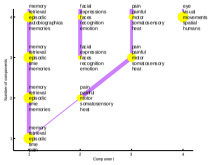
The Matlab structures take care of bookkeep the context information, so that corner cube v alization[2] of results from non-negative matrix torization[3] on a stack of kernel density estima (KDE) of brain coordinates[4] can be done sim e.g. by

```
load wobibs
E = brede_bib_bib2exp(B);
M = brede_exp_exp2mat(E);
[W,H] = brede_mat_nmf(M);
brede_mat_2mat2html(W,H);
```

Several multivariate analysis algorithms are im mented in the Brede Toolbox: NMF, ICA, S K-means clustering, canonical correlation. In tain contexts we apply "non-negative partial l squares" which is simply NMF of the inner proc of two matrices:

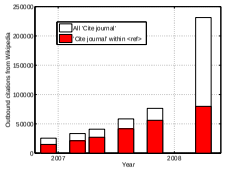
$$\mathbf{WH} = \mathrm{nmf}(\mathbf{X}^{\mathrm{T}}\mathbf{Y})$$

One matrix may be from KDE, the other a bag words matrix of article abstract words.[5] "Hie chical NMF" is simply independent varying si NMFs, which result may be visualized with a *ter bush* visualization.[6] Here with a bag-of-w matrix:



Wikipedia citation clustering

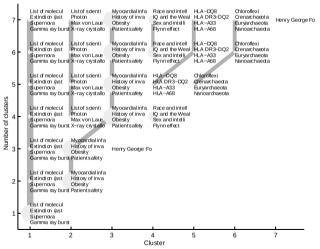
Wikipedia articles may cite scientific articles via the `cite journal` template. From XML dumps of Wikipedia it is possible to extract citation data from Wikipedia pages where the citations occur and match journal title to variation recorded in an ontology of journals.[7]



With the extracted citations a data matrix can be built

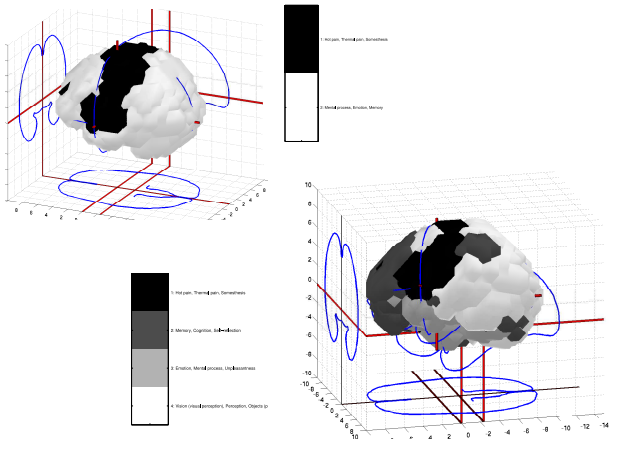
$$\mathbf{X}(\text{Wikipedia pages} \times \text{cited journals}),$$

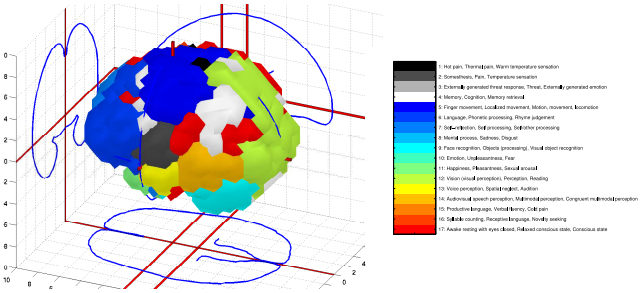
and this matrix can subsequently be subjected to hierarchical NMF and visualized with a *cluster bush* plot as implemented in the Brede Toolbox:[6,8]



These clusters may display groups of journals and groups of Wikipedia articles, e.g., in neurology.

Multivariate analysis with the Brede Database brain function ontology





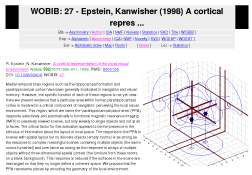
Hierarchical non-negative partial least squares with two matrices: One "recursed" adjacency matrix constructed from the brain function ontology, and another matrix constructed from kernel density estimates of the brain coordinates.

Data is available

Data from the Brede Database is available or Web:

http://neuro.imm.dtu.dk/services/brededatal

Also results from multivariate analysis is avai on that site as well as pages for the individual i in the Brede Database, e.g., here part of the for a paper by Epstein et al.:



Also the extracted data from the Wikipedia tering is available for download.

Acknowledgment

Thanks to the Lundbeck Foundation for fun and Lars Kai Hansen and Daniela Balslev.

References

[1] Fox PT, et al. BrainMap: A database of human function brain mapp Thacher RW, et al., eds., *Functional Neuroimaging: Technical Foun* Academic Press, San Diego, California, 1994; pages 95–105.

[2] Rehm K, et al. A symbolic environment for visualizing activated foci in fu neuroimaging datasets. *Medical Image Analysis*, 1998;2:215–226.

[3] Lee DD and Seung HS. Algorithms for non-negative matrix factorization. TK, Dietterich TG, and Tresp V, eds., *Advances in Neural Information ing Systems 13: Proceedings of the 2000 Conference*. MIT Press, Ca Massachusetts, 2001; pages 556–562.

[4] Nielsen FA and Hansen LK. Modeling of activation data in the Brainl database: Detection of outliers. *Human Brain Mapping*, 2002;15:146–1

[5] Nielsen FA, Hansen LK, and Balslev D. Mining for associations betwe and brain activation in a functional neuroimaging database. *Neuroinfo* 2004;2:369–380.

[6] Nielsen FÅ, Balslev D, and Hansen LK. Mining the posterior cingulate gation between memory and pain component. *NeuroImage*, 2005;27:52

[7] Nielsen FÅ. Scientific citations in *Wikipedia*. *First Monday*, 2007;12.

[8] Nielsen FÅ. Clustering of scientific citations in Wikipedia. In *Wikiman*