Text Analysis, Generalizable Gaussian Mixture and Neuroinformatics

Finn Årup Nielsen

Section for Digital Signal Processing
Department of Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Email: fn@imm.dtu.dk
WWW: http://www.imm.dtu.dk/~fn

TEXT ANALYSIS APPLICATION

Machine text analysis

- Document level
  - Document retrieval (usually called "information retrieval")
    - From keywords or from latent meaning $p(\mathcal{X}|\mathcal{C})$
    - "Find similar documents" $p(\mathcal{X}|\mathbf{x})$
  - Organize documents by subject (automatic librarian).
    - Supervised (document classification): $p(\mathcal{C}|\mathcal{X})$
    - Unsupervised (document clustering): $p(\mathcal{X})$,
  - Filtering (electronic mail from customers, spam filtering, survailence, fraud detection). $p(\mathcal{C}|\mathcal{X})$
  - Determine authorship. $p(\mathcal{C}|\mathcal{X})$
- With-in text level
  - Automatic generation of index
  - Finding inconsistencies between documents
  - "Extract key information from text"
- Hybrid
  - Automatic hyperlinking, cross-referencing [Salton and Buckley, 1989].

WHAT WE WOULD LIKE TO DO IN CONNECTION WITH NEUROINFORMATICS

- Application of machine learning techniques on texts to do something useful
  - Independent Component Analysis (ICA) [Kolenda and Hansen, 1999]
  - Generalizable Gaussian Mixture (model) (GGM) [Hansen et al., 1999]
  - ...
- Web-site for functional neuroimaging, e.g., Web-search facility for webpages, personal homepage, journal article.
- Make machines understand functional neuroimaging studies.
  - Data —> Information —> Knowledge
  - Quantitative and automatic meta-analyses
  - Quantitative and automatic priors for the analysis of functional neuroimaging studies.

REPRESENTATION

- Language modelling or "Syntactic".
- Vector space, or "Statistic" (Gerard Salton, Cornell, [Salton et al., 1975]) Discrete/continues, Hierachical?

Vector space

- Bag-of-words. [Sahami, 1999]
  Count the frequency of words, — disregarding the order and language structure.
  - Eliminate most frequent words — the so-called "stop-words": the, of, ... Approximately 500 words in the stop word list
  - Eliminate least frequent words, keeping only co-occuring words.
  - Word stemming, eliminate word suffix (computer, computers, computing —> compute)
- N-grams

N-GRAMS

- Example (bigram, 2-gram): aa: 1st variable; ab: 2nd variable; ...
- Used to determined authorship and literary style [Kjell and Frieder, 1992, Soboroff et al., 1997]
- Probably usable to determine the quality of a web-page.

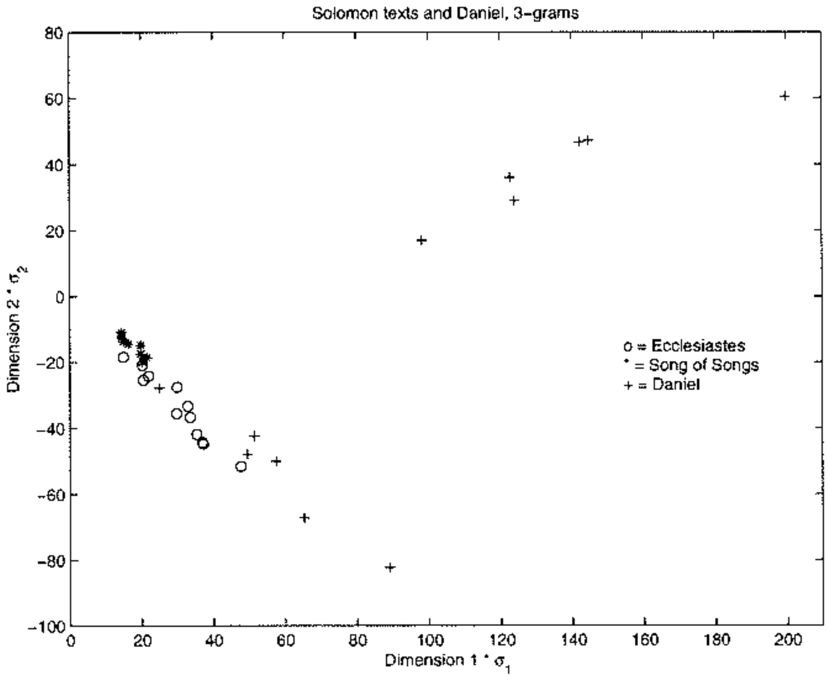
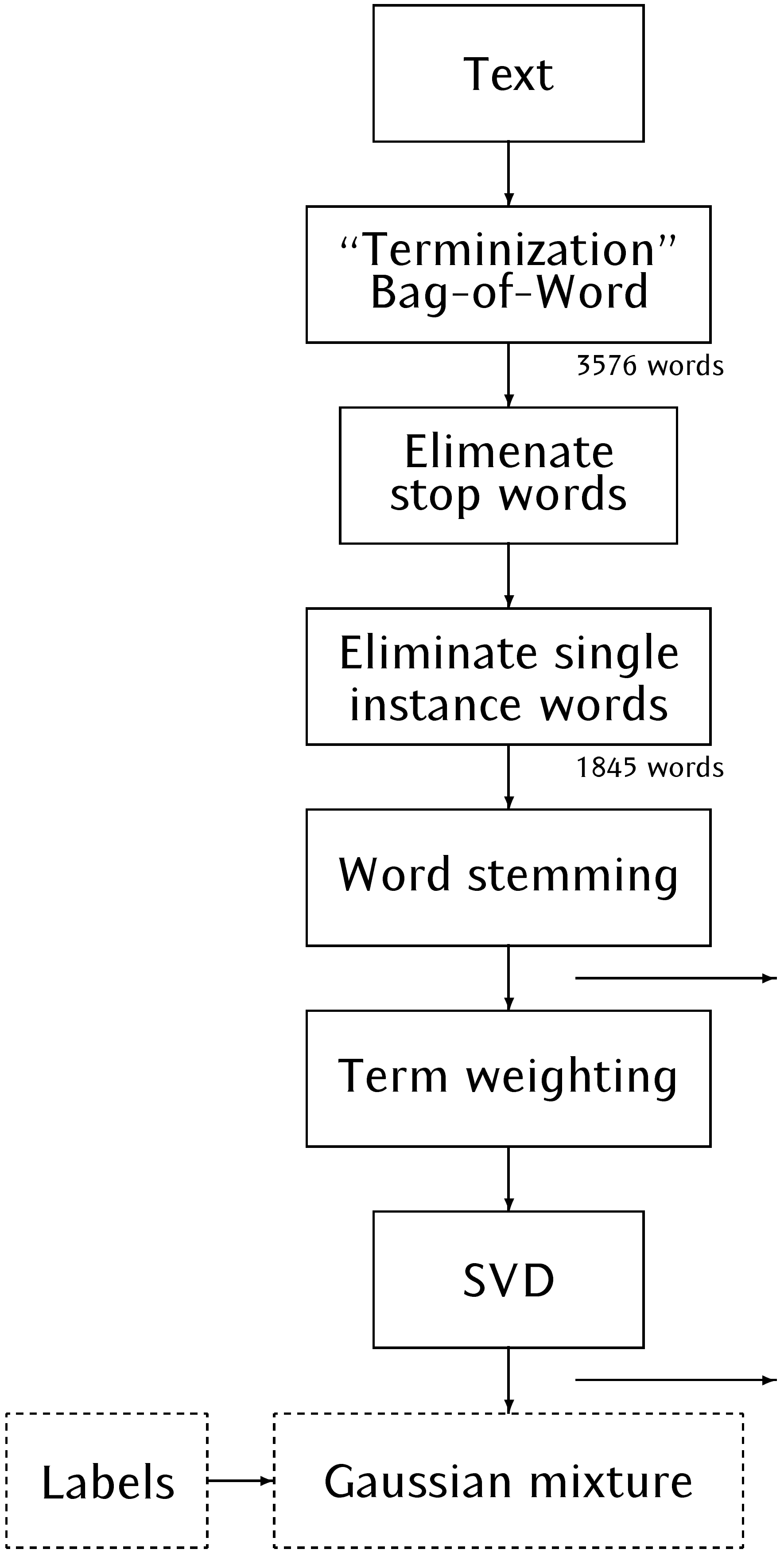
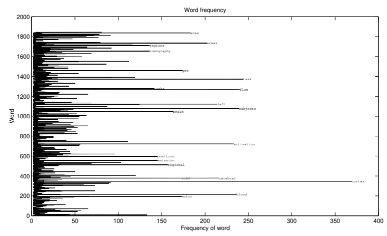


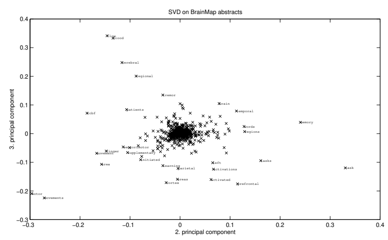
Figure 1: The first two scaled dimensions of the SVD.

Figure 1: Figure from [Soboroff et al., 1997]

BAG-OF WORDS PROCESSING



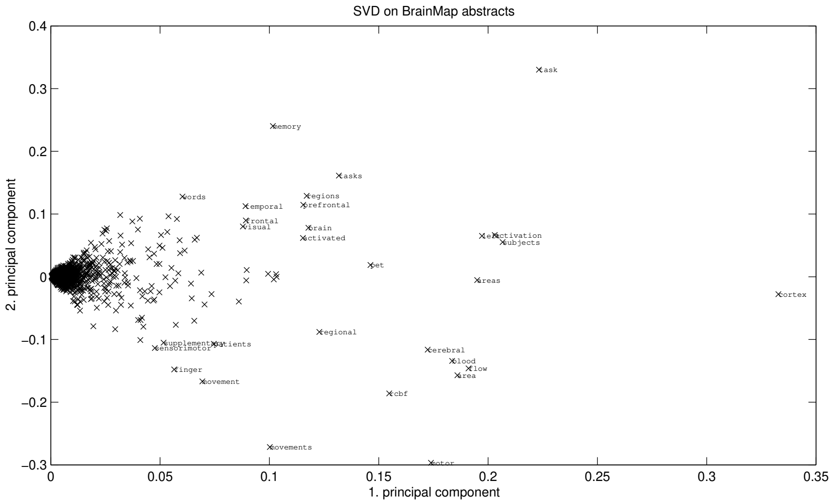


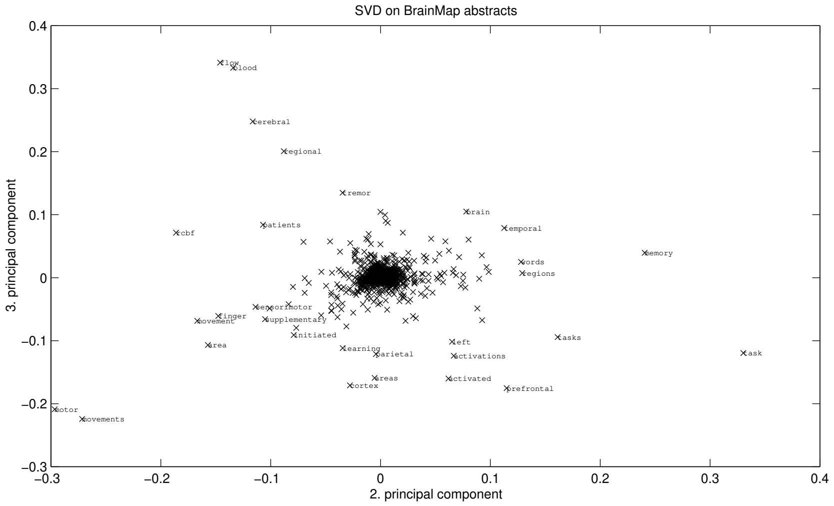


TERM WEIGHTING

- Bag-of-word data matrix is sparse.
- [Salton and Buckley, 1988]: "These [text retrieval] results depend crucially on the choice of effective term weighting systems."
- Term weight [Singhal et al., 1995]
  - Term frequency (*tf*)
  - Inverse document frequency (*idf*)
  - Document length normalization
- Many (ad hoc) schemes
  - Binary, Frequency
  - Smart: $0.5 + 0.5\ tf/\max(tf)$
- No (unifying) probabilistic formulation of the problem

SINGULAR VALUE DECOMPOSITION ON BAG-OF-WORDS DATA MATRICES





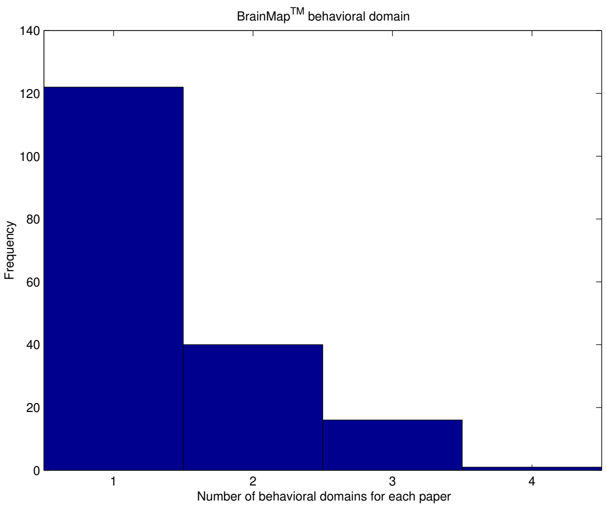
- PCA/SVD on word matrix: "Latent semantic analysis" [Deerwester et al., 1990].
- Groups not separated in the principal components.
- Independent Component Analysis seems to be better to separate the groups [Kolenda and Hansen, 1999].

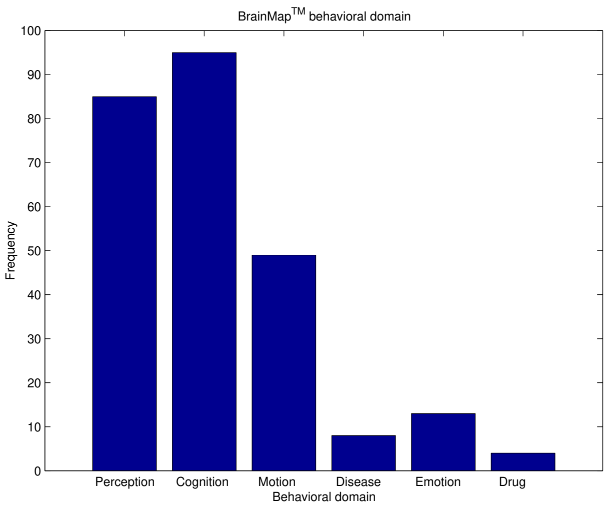
BRAINMAP

- BrainMap database `http://ric.uthscsa.edu`
- Contains data from functional neuroimaging studies with Talairach coordinates.
  - Original peer-reviewed papers
  - Meta-analyses
- Paper (title, author, abstract), structured description of the study (behavioral domain, modality, tracer) and Talairach coordinates

Behavioral domain

- Hierarchical structure
- On the highest level: 6 classes that are not exclusive



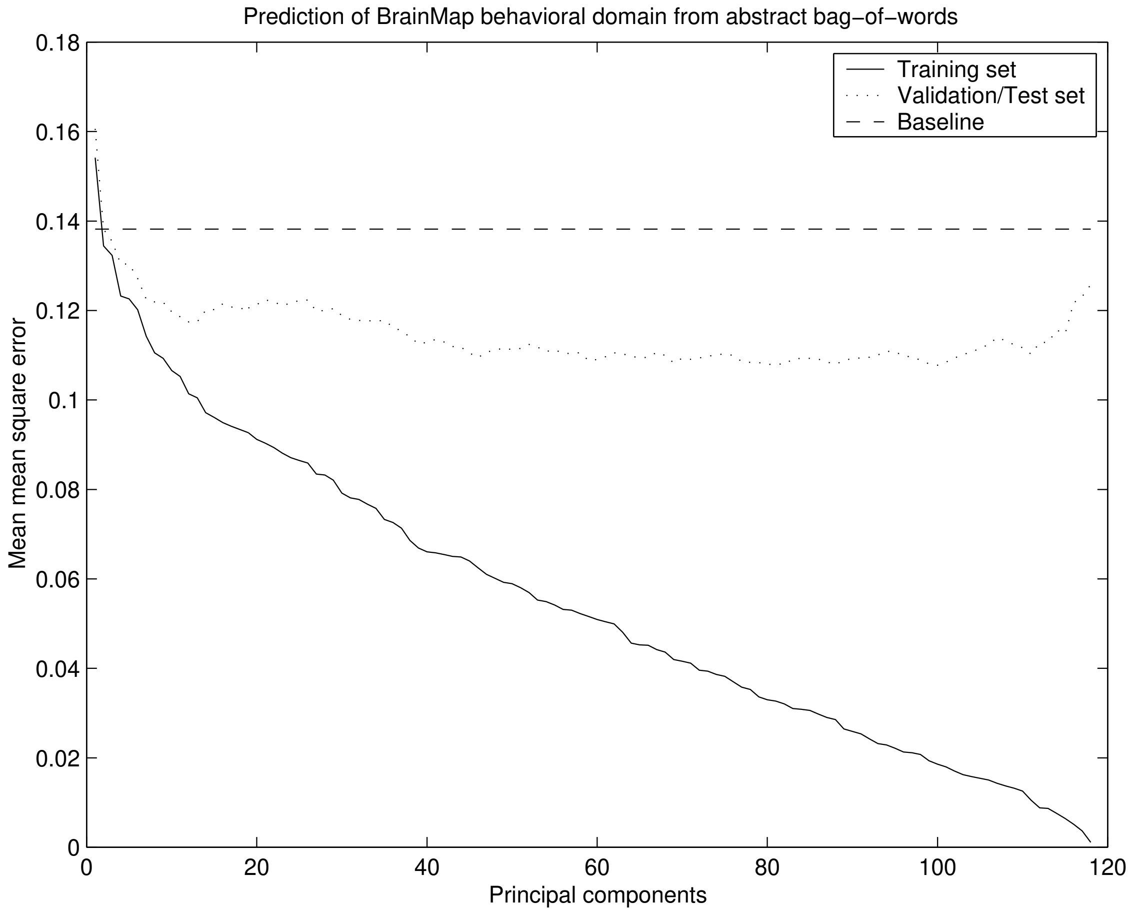


MULTIVARIATE REGRESSION ANALYSIS ON BRAINMAP

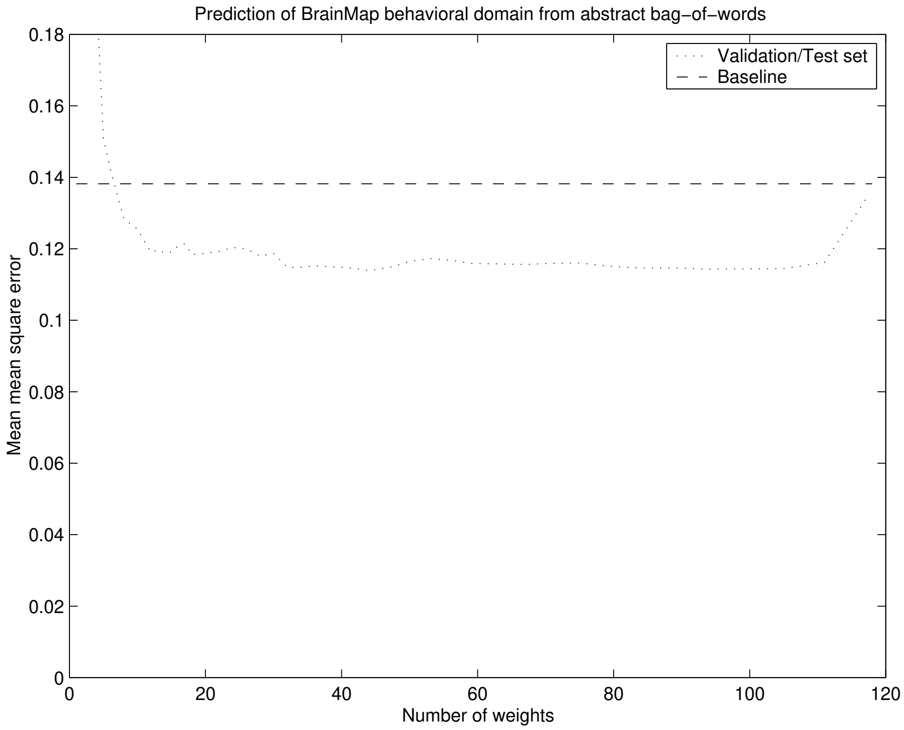
$$\mathbf{X}: \quad \text{Bag-of-words from abstract of BrainMap} \tag{1}$$
$$\mathbf{Y}: \quad \text{“Behavioral Domain” entry from BrainMap} \tag{2}$$

Multivariate regression analysis with a matrix of degenerate rank:

$$\mathbf{Y} = \mathbf{X_1 B_1^*} + \mathbf{U} \tag{3}$$



NEURAL NETWORK ANALYSIS



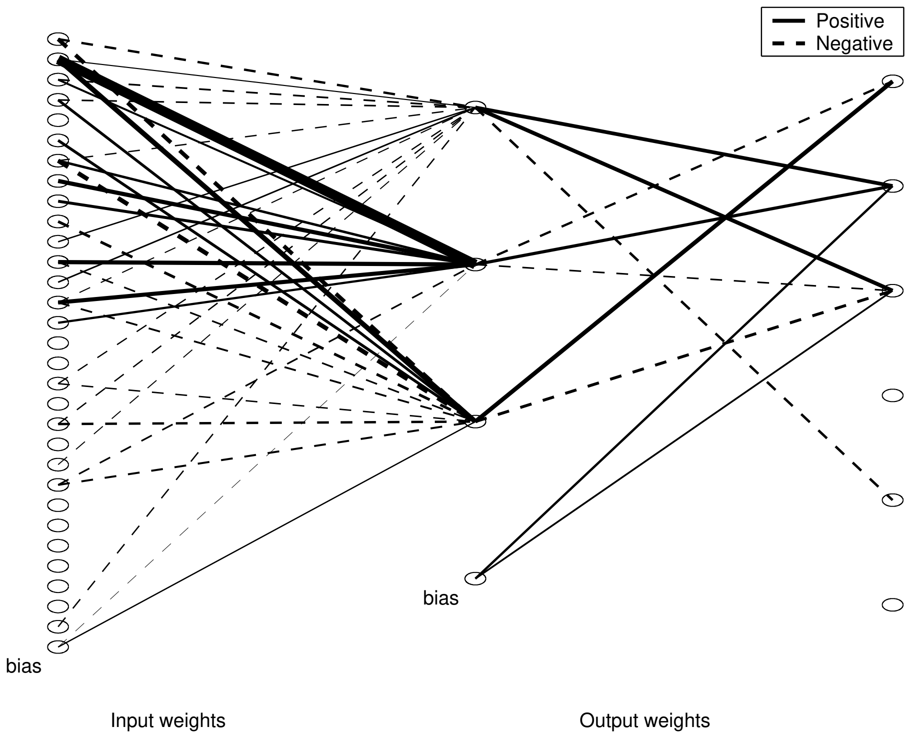


Figure 2: Neural network with 44 weights.

GENERALIZABLE GAUSSIAN MIXTURE

Algorithm: Generalizable Gaussian Mixture

1. Initialization
   (a) Compute the mean vector $\boldsymbol{\mu}_0 = N^{-1} \sum_n \boldsymbol{x}_n$.
   (b) Compute the covariance matrix of the data set: $\boldsymbol{\Sigma}_0 = N^{-1} \sum_n (\boldsymbol{x}_n - \boldsymbol{\mu}_0)(\boldsymbol{x}_n - \boldsymbol{\mu}_0)^\mathsf{T}$.
   (c) Initialize $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.
   (d) Initialize $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0$.
   (e) Initialize $P(k) = 1/K$.
2. Repeat until convergence
   (a) Compute $p(k|\boldsymbol{x}_n)$ and assign $\boldsymbol{x}_n$ to the most likely component.
   (b) Split the data set in two equal parts $D_{\boldsymbol{\mu}}$, $D_{\boldsymbol{\Sigma}}$.
   (c) For each k estimate $\boldsymbol{\mu}_k$ on the points in $D_{\boldsymbol{\mu}}$ assigned to component k.
   (d) For each k estimate $\boldsymbol{\Sigma}_k$ on the points in $D_{\boldsymbol{\Sigma}}$ assigned to component k. If the number of datapoints assigned to the k'th component is less than the dimension, let $\boldsymbol{\Sigma}_k = c \cdot \boldsymbol{\Sigma}_0$, where c is determined so that the total variance of the component corresponds to the variance of the datapoints associated with it.
   (e) Estimate $P(k)$ as the frequency of assigments to component k.

GGM ON BRAINMAP DATA

- Generalizable Gaussian Mixture
- Inputs ($\mathbf{x}$): Talairach coordinates.
- Labels (c): Three "behavioral domain" (only the list taken from the list)

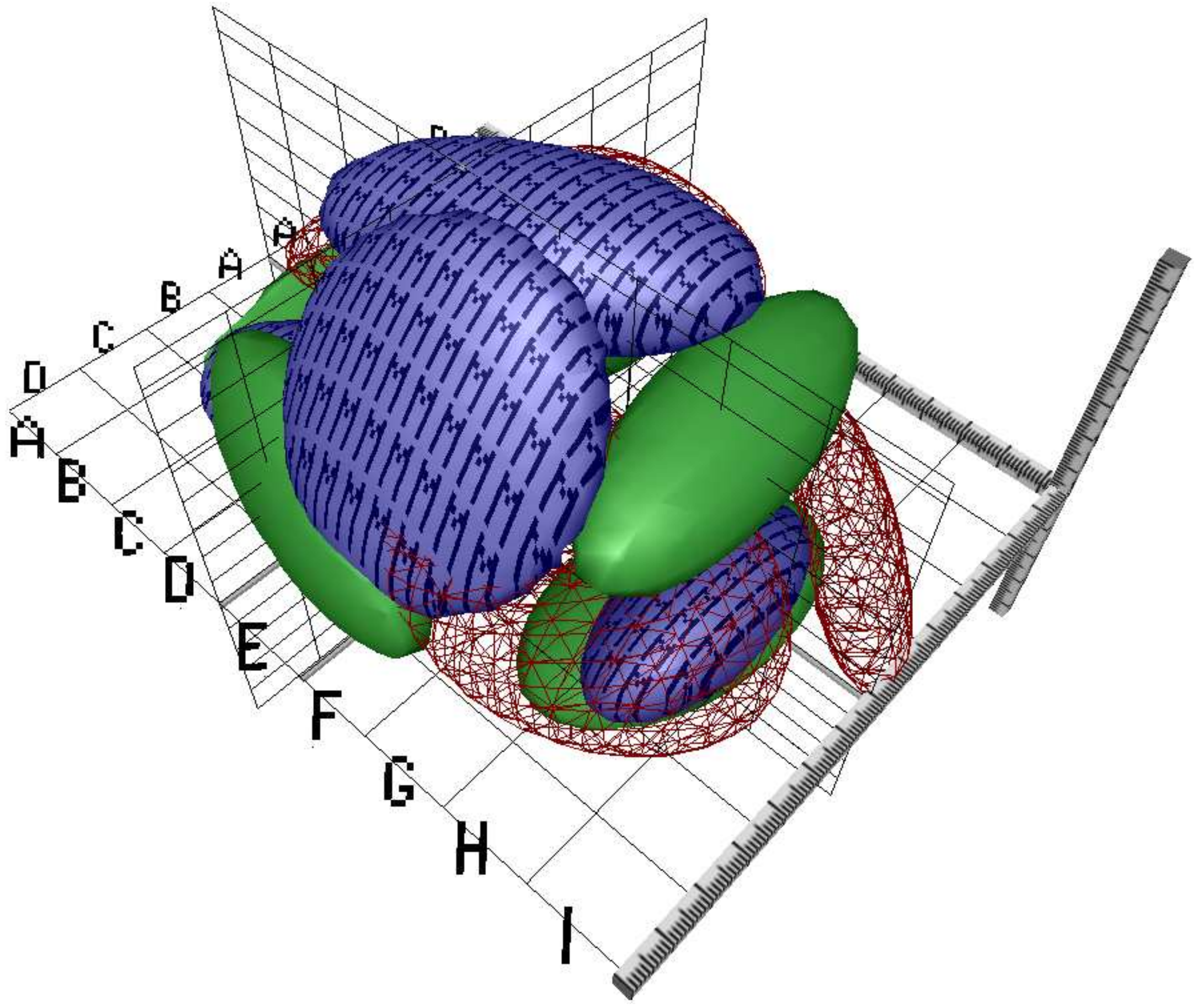


Figure 3: GGM on Talairach coordinates and "behavioral domain" from the BrainMap database. Wireframe is "perception", non-textured is "cognition" and M-textured is "motion".

NOVELTY DETECTION

- Novelty detection: Take a new study and what the probability for obtaining such a study.
- Cannot see the difference between a bad and a good study
- Easy with the modeling of $p(\mathbf{x}|c)$

Automatic labeling

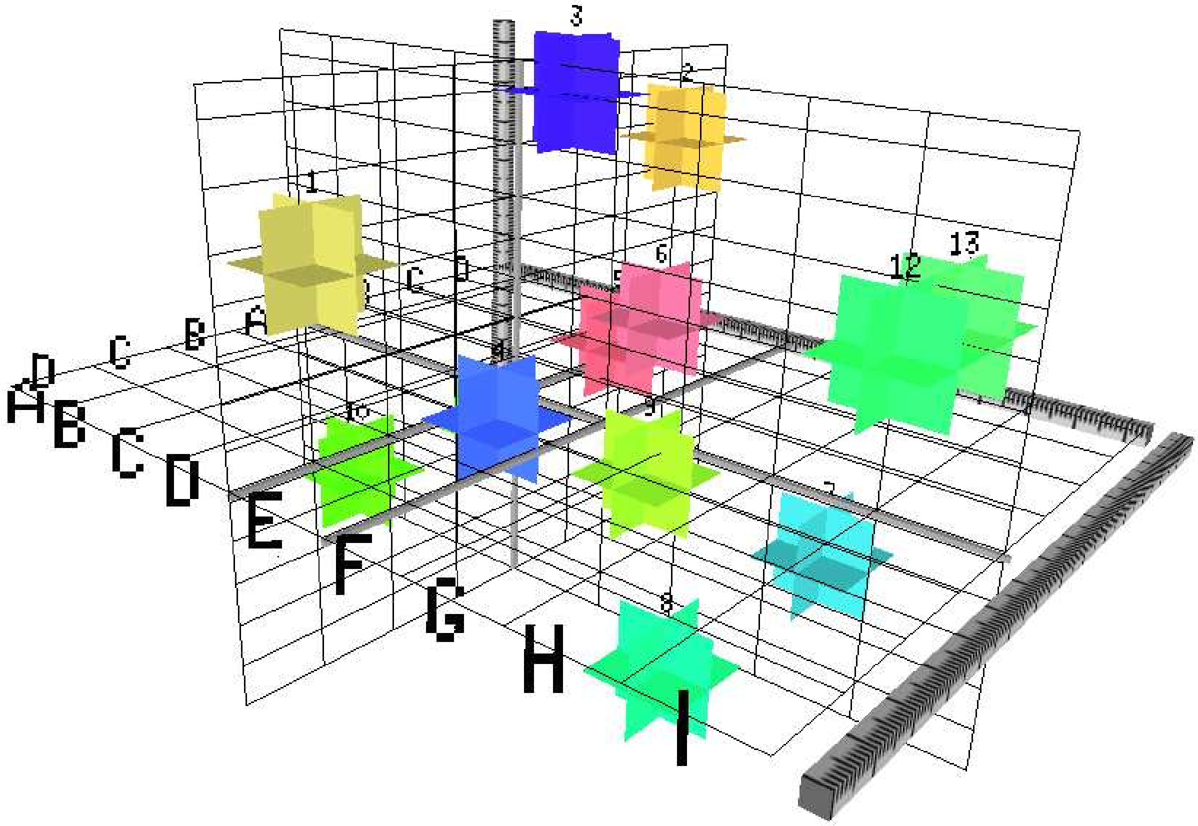


Figure 4: Automatic labeling detection on each individual Talairach coordinate from the paper [Law et al., 1998].

REFERENCES

References

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[Hansen et al., 1999] Hansen, L. K., Sigurdsson, S., Kolenda, T., Nielsen, F. Å., Kjems, U., and Larsen, J. (1999). Modeling text with generalizable gaussian mixtures. Submitted to NIPS'99.

[Kjell and Frieder, 1992] Kjell, B. and Frieder, O. (1992). Visualization of literary style. In *IEEE International Conference on Systems, Man and Cybernetics*.

[Kolenda and Hansen, 1999] Kolenda, T. and Hansen, L. K. (1999). Independent components in text. Submitted to NIPS'99.

[Law et al., 1998] Law, I., Svarer, C., Rostrup, E., and Paulson, O. B. (1998). Parieto-occipital cortex activation during self-generated eye movements in the dark. *Brain*, 121(11):2189–2200.

[Sahami, 1999] Sahami, M. (1999). *Using Machine Learning to Improve Information Access*. Ph.d. thesis, Computer Science Department, Stanford University.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

[Salton and Buckley, 1989] Salton, G. and Buckley, C. (1989). On the automatic generation of content links in hypertext. Technical Report TR89–993, Department of Computer Science, Cornell University.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM*, 18:613–620.

[Singhal et al., 1995] Singhal, A., Buckley, C., Mitra, M., and Salton, G. (1995). Pivoted document length normalization. Technical Report TR95–1560, Department of Computer Science, Cornell University.

[Soboroff et al., 1997] Soboroff, I. M., Nicholas, C., and Kukla, J. M. (1997). Visualizing document authorship using N-grams and latent semantic indexing. In Ebert, D. S. and Nicholas, C. K., editors, *Workshop on New Paradigms in Information Visualization and Manipulation*. ACM Press.