Learning Predictive Cognitive Structure from fMRI using Supervised Topic Models

Oluwasanmi Koyejo, Priyank Patel, Joydeep Ghosh
Dept. of Electrical and Computer Engineering,
University of Texas at Austin
{*sanmi.k@, pmpatel@, ghosh@ece*}*.utexas.edu*

Russell A. Poldrack
Depts. of Psychology and Neurobiology,
University of Texas at Austin
poldrack@utexas.edu

***Abstract*—We present an experimental study of topic models applied to the analysis of functional magnetic resonance images. This study is motivated by the hypothesis that experimental task contrast images share a common set of mental concepts. We represent the images as documents and the mental concepts as topics, and evaluate the effectiveness of unsupervised topic models for the recovery of the task to mental concept mapping, We also evaluate supervised topic models that explicitly incorporate the experimental task labels. Comparing the quality of the recovered topic assignments to known mental concepts, we find that the supervised models are more effective than unsupervised approaches. The quantitative performance results are supported by a visualization of the recovered topic assignment probabilities. Our results motivate the use of supervised topic models for analyzing cognitive function with fMRI.**

***Keywords*-fMRI; mental concepts; mixed membership; topic model**

I. Introduction

Studies in the neuroimaging literature have shown that experimental tasks can be decomposed into a set of cognitive or mental concepts[1] [1], [2]. For instance, Poldrack et.al., [1] showed that images representing experimental tasks such as gambling decisions, reading aloud and risk taking were associated with mental concepts such as response inhibition, working memory, phonology and spatial processing. Based on such evidence, we investigate the hypothesis that each the image is a probabilistic superposition of latent mental concepts. The goal of this paper is to evaluate the use of unsupervised and supervised mixed membership topic models for extracting the mental concept structure in relation to the experimental tasks.

Standard clustering models assign each observation to a single cluster, which is represented by a cluster *center*. When applied to text data, the cluster representatives are also known as *topics* that describe the document corpus [3]. Recently, mixed membership topic models have been proposed for text clustering [4]. This is a more flexible approach where each word can be generated from a different topic. As a result, each document can be generated from a combination of topics. Mixed membership latent variable models have been applied to image segmentation and classification [5], among other applications.

[1]e.g. the Cognitive Atlas (http://www.cognitiveatlas.org/) is a database of associations between experimental tasks and mental concepts.

Supervised topic models are an extension of topic modeling to incorporate class labels for each document [6], [7]. The topic assignment probabilities are identified as extracted features for each document, so the overall model corresponds to joint supervised feature extraction and predictive modeling. Thus, the supervised training of topic models imposes extra constraints on the recovered topics and document assignment probabilities, as these must both describe the clustering structure of the data and predict the document labels. Researchers have observed that supervised topic models often extract topics of superior quality when compared to their unsupervised counterparts [6].

To evaluate the use of topic models, each brain image is identified as a *document*, and the mental concepts are identified as the *topics*. To capture the heterogeneity in fMRI, we incorporate mixed membership by allowing each voxel activation in a brain image to be generated from a separate topic. The proposed topic modeling approach may be contrasted with the classifier based reverse inference or "brain reading" approach for analyzing fMRI data [1], [8]–[10]. Like the classifier based methods, the proposed approach is used to analyze fMRI by interpreting the fit model parameters. However, in addition to the classifier weights, the proposed approach captures additional structure corresponding to interpretable mental concepts.

II. Data Description and Preprocessing

We analyzed two fMRI datasets, which overlapped with those presented in [1]. The *small data* consists of 130 images collected from 130 people completing one of 8 tasks, and the *large data* consists of 289 images collected from 157 people completing one of 17 tasks. The scanned brain images were preprocessed to remove motion artifacts, and a first-level statistical model was used to compute a z-score for each voxel for a specific contrast. Both datasets were collected as $(2\,mm)^3$ voxels with an image dimension of $(91, 109, 91)$ over the whole brain. See [1] for details of the fMRI collection and preprocessing methodology.

We applied dimensionality reduction by resampling the images using the `flirt applyXfm` tool [11]. The size of the resampled image was determined by the scaling $s \times (91, 109, 91)$, where $0 < s \leq 1$ is a scaling parameter.

This resulted in an increase in the volume of the corresponding image voxels from $(2\,mm)^3$ to $(2/s\,mm)^3$. We experimented with 10 linearly spaced values of s between 0.1 and 0.9. A support vector machine (SVM) [12] trained using the one-vs-all approach was used to evaluate the effect of voxel size on 5-fold stratified crossvalidation classification performance. On both datasets, we found that the performance did not improve for image sizes greater than $(17, 10, 17)$. At this size, we observed an accuracy of 68.9% on *large data* and 82.3% on *small data*. For comparison, recall that the expected performance of a random classifier is 12.5% for 8 classes and 5.8% for 17 classes. This suggests that the voxels contain useful information for discriminating between classes. The reduced dimensionality data were used for the remainder of the experiments.

We experimented with both discrete valued (*bag of words*) and continuous valued mixed membership topic models. Discrete topic models required further discretization of the image data. **Discretization:** a continuous signal can be discretized using an analog to digital converter (ADC) with predefined discrete levels. Given a continuous value, the corresponding discrete value may be defined by rounding down to the closest discrete level. We applied this approach to the z-score images, so each *word* represents a fixed unit magnitude, and the number of *words* correspond to the resulting digital level. We used a different *word* for each spatial location. We also separated the positive and negative signals. We used 20 bins each for the *positive* and *negative* signals. Taking the positive values in all the images as X, the discrete levels were defined using a uniform partition of $[0, \max(X)]$. An identical approach was used to define the negative word levels. Thus, a *word* was defined as the a triple of spatial location, positive or negative sign and a predefined unit magnitude. Our approach ensured that the words maintained the spatial information, and avoided mixing the positive and negative z-score values. We note that this approach is different from the discrete cluster codebook approach popular in the application of topic models to image data [5] where the spatial information is lost. The continuous topic models require a continuous valued *word*. Here, we used the voxel spatial location, but the quantization was not applied.

III. Methods

We evaluated a variety of mixed membership topic models for discrete or continuous data to survey the effect of different modeling assumptions on the mental concepts recovered. The models also differ in whether or not they use the target class information (supervised vs. unsupervised). We present experimental results evaluating the following models:

1) **Latent Dirichlet allocation (LDA)** [4]: is a popular model for topic modeling. LDA generates words via the mixed membership approach, but does not incorporate supervised information.
2) **Discriminative LDA (D-LDA)** [7]: augments LDA with a multinomial logistic regression model for classification.
3) **Mixed membership naïve Bayes (MMNB)** [7]: models real valued data using a Gaussian mixed membership naive Bayes model.
4) **Discriminative Mixed membership naïve Bayes (D-MMNB)** [7]: augments the MMNB model with a multinomial logistic regression for classification.
5) **Maximum margin supervised topic models for classication (MedLDA)** [6]: is a semi-generative model that enforces margin constraints on the posterior topic distributions of LDA. The result is a combination of SVM and LDA that can be trained jointly. The supervised portion of the model also inherits the sparsity and robustness properties of the SVM.

Unless otherwise noted, the outlined models were implemented using the variational inference and parameter estimation approach outlined in the respective papers. We implemented the fast inference approach for MMNB and D-MMNB [7]. We used the implementation of MedLDA provided by [6]. Of the three supervised topic models trained, only MedLDA regularizes the classifier weights. The unsupervised models do not predict class labels, so for classification results, we applied a SVM classifier to the recovered topic assignment probabilities (see [6] for a similar approach).

IV. Experiments

We compared the classification and topic recovery performance of the proposed models on *large data*. The results were computed using 5-fold stratified cross validation. The regularization hyperparameters were selected from the grid of values $C \in 10^{\{-3,-2,\ldots,3\}}$. The number of topics was selected from the set $K \in \{20, 40, 60, 80, 100\}$. The training of clustering models is prone to local minima, thus for each training set each model was randomly initialized 5 times. The result with the best fit as measured by the (lower bound) of the training log likelihood is presented.

Classification: Although the classification performance of the models is not of primary concern, the classification results can be used to evaluate the quality of the learned model. Classification results are presented in Table I. We found that the supervised models significantly outperformed their unsupervised counterparts. MedLDA achieved the best classification performance of 66.44% accuracy. The most important characteristic that seemed to determine the supervised model performance was the use of regularization for the classifier weights. The unregularized discriminative models, D-LDA (13.14%) and D-MMNB (34.59%) performed much worse that MedLDA, and also seems to overfit quickly as we used more topics. The continuous topic models, MMNB and D-MMNB failed and returned degenerate results when we used more than 40 topics.

Table I
CLASSIFICATION ACCURACY ON THE *large data* (%-ACCURACY)

MODEL	LDA	D-LDA	MMNB	D-MMNB	MEDLDA
20	13.1398(2.5678)	13.1458(2.2913)	34.5917(6.4980)	38.7598(5.5744)	59.1199(0.4076)
40	8.9958(0.7570)	8.9957(0.7569)	25.9347(4.1138)	26.6425(3.9547)	65.0400(0.0275)
60	8.9958(0.7570)	8.9957(0.7569)	–	–	65.7600(0.1758)
80	8.9958(0.7570)	8.9957(0.7569)	–	–	65.7400(0.0461)
100	8.9958(0.7570)	8.9957(0.7569)	–	–	66.4399(0.0507)

Visualization Using the best hyperparameters as determined by the cross validation performance, we re-fit the models on all of *large data*. We then embedded the posterior topic distributions extracted by each model into two dimensions using tSNE [13]. Each color/shape combination represents one of 17 classes. We expect that in an effective model, the embedded features will form tight class dependent clusters that are well separated from the clusters of other classes. The embedded features of LDA and D-LDA (Fig. 1a) did not seem to form clusters. The embeddings found for the MMNB and D-MMNB (Fig. 1b) showed cluster structure, but we found that the clustering was not class dependent. The MedLDA features recovered the most informative class dependent clustering (Fig. 1c).

Ontology Recovery: An ontology mapping each class to the set of mental concepts was extracted by Poldrack et.al., [1] for *small data*, where it was used to analyze classifier parameters. There are a total of 56 mental concepts associated with *small data*. The ontology provides an indicator vector identifying the classes associated with each mental concept. We trained all the topic models with 56 topics on *small data* with the goal of recovering the mental concepts as topic assignments. As we do not know the true matching, we computed a normalized score between the ground truth and the recovered topic assignments and applied the Hungarian matching algorithm [14] to recover a unique assignment. To compute the score, we averaged the topic assignment probabilities for each class, and normalized the scores to sum to one. We computed the score as the normalized cross-correlation between the ground truth indicator vector associated with each mental concept and the recovered topic vector. We used the $\{-1, +1\}$ representation of the indicators to compute the score. Thus, in addition to capturing the overlap, the score penalized mass assigned to the incorrect topics. Finally, we computed the normalized the score by adding 1.0 and dividing by 2.0 so the normalized score takes a value between 0 and 1. The scores are presented in Table II. MedLDA achieved the best score of 0.4863.

Table II
ONTOLOGY TOPIC MATCHING SCORE ON *small data*. SCORE LIES BETWEEN 0 AND 1. HIGHER SCORES INDICATE BETTER MATCHING.

LDA	D-LDA	MMNB	D-MMNB	MEDLDA
0.1671	0.1601	0.1821	0.1770	0.4863

V. CONCLUSION AND FUTURE WORK

In this paper, we presented an experimental evaluation of various topic models for the analysis of fMRI and compared their performance in terms of classification performance and structure recovery. We compared an array of supervised and unsupervised models and found that the MedLDA model performed the best for all our criteria. The results presented in this paper motivate the use of supervised topic models as a promising approach for inferring cognitive function from fMRI.

Our results also suggest several opportunities for improving the tested models. For instance, further regularization of the topics may improve classification performance and structure recovery. Various hypotheses related to spatial structure can be implemented as topic regularizers. Structured sparsity regularizers have been proposed for reverse inference of fMRI data [15], [16]. Other structure that may be of interest include smoothness of the image over space, or symmetry between the left and right hemispheres, The application of such constraints to regularize the topics is left as the the subject of future work.

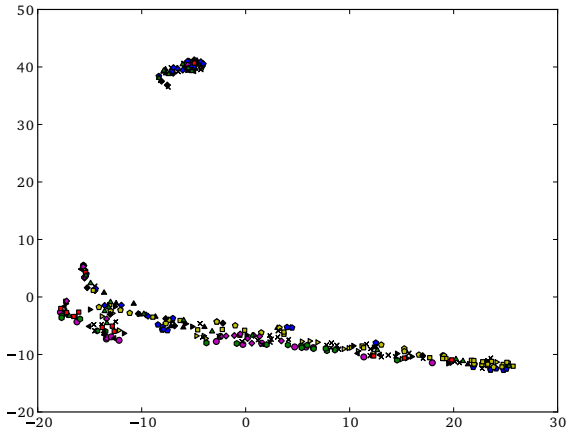
The mental concept ontology is a useful source of domain expert information for fMRI data. It is not yet clear how best to use this information. We proposed its use for independent verification of the recovered topics. Alternatively, the topic ontology can be incorporated into the training data as a supervised classifier target, or the ontology can be incorporated as known portions of the topic assignment, fixed while the unknown portions are trained. Further investigation of various options for their use is left as the subject of future work.

ACKNOWLEDGMENT

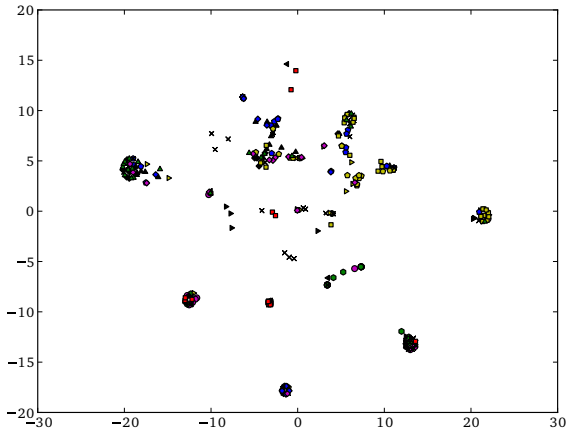
O. Koyejo acknowledges support from NSF grant IIS 1016614.

REFERENCES

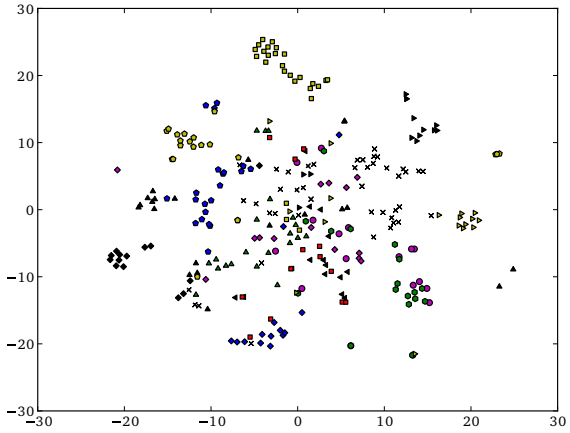
[1] R. A. Poldrack, Y. O. Halchenko, and S. J. Hanson, "Decoding the large-scale structure of brain function by classify-



(a) D-LDA $K = 20$



(b) D-MMNB $K = 20$



(c) MedLDA $K = 100$

Figure 1. Visualization of topic features for *small data*. Each color / shape combination represents one of 17 experimental task contrasts (classes). MedLDA exhibits clustering structure that is correlated with the classes. MMNB and D-MMNB exhibit clustering structure, but the clusters are not correlated with the classes. LDA and MMNB omitted due to limited space.

ing mental states across individuals," *Psychological Science*, vol. 20, pp. 1364–1372, 2009.

[2] R. A. Poldrack, A. Kittur, D. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, and R. M. Bilder, "The cognitive atlas: Towards a knowledge foundation for cognitive neuroscience," *Frontiers in Neuroinformatics*, vol. 5, 2011.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR*, 2006.

[6] J. Zhu, A. Ahmed, and E. P. Xing, "Medlda: maximum margin supervised topic models for regression and classification," in *ICML*, 2009.

[7] H. Shan and A. Banerjee, "Mixed-membership naive bayes models," *Data Mining and Knowledge Discovery*, vol. 23, pp. 1–62, 2011.

[8] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, vol. 45, pp. S199–S209, 2009.

[9] S. Ghebreab, P. W. Adriaans, and A. W. M. Smeulders, "Predictive modeling of fmri brain states using functional canonical correlation analysis," in *Artificial Intelligence in Medicine*, 2007, pp. 393–397.

[10] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns." *NeuroImage*, vol. 43, pp. 44–58, 2008.

[11] M. Jenkinson, P. Bannister, M. Brady, S. Smith *et al.*, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.

[12] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[13] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 2579–2605, 2008.

[14] H. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, pp. 83–97, 1955.

[15] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion, "Total variation regularization for fmri-based prediction of behavior," *Medical Imaging, IEEE Transactions on*, vol. 30, pp. 1328–1340, 2011.

[16] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion, "Multiscale mining of fmri data with hierarchical structured sparsity," *SIAM J. Imaging Sciences*, vol. 5, no. 3, pp. 835–856, 2012.