

Temporal and Cross-Subject Probabilistic Models for fMRI Prediction Tasks

May 13, 2006

Abstract

We present a probabilistic model applied to the fMRI video rating prediction task of the Pittsburgh Brain Activity Interpretation Competition. Using a Dynamic Gaussian Markov Random Field, we model the relationship between the subjects' fMRI voxel measurements and the rated properties of the videos, such as presence of language or subject amusement. Also included in our model are dependencies of the ratings across time steps, and between subjects. Rather than grouping voxels into regions, we chose to use individual voxels as features. However, for some of the ratings, we found that using a prior with a bias toward similar parameters for neighboring voxels improved our predictions. Our model performed significantly better than a baseline regression model on held out training data, and displayed good performance in predicting the scored ratings across the three subjects in the training and test data sets. We chose a general model that is applicable to all the rating categories, rather than specially attacking single rating types at a time.

1 Introduction

The Pittsburgh Brain Activity Interpretation Competition presented fMRI measurements [7], with the goal of predicting the ratings made by three subjects on short videos, ranging from emotional qualities to the presence of specific actors.

Our approach utilized graphical probabilistic models, which allowed us to represent many relevant relationships, including evolution of ratings over time, the likelihood of different subjects rating experiences similarly, and the relationship between voxels and the ratings. The basic features in our model were individual voxels, as we chose not to aggregate them into regions.

This model demonstrated good performance on many of the ratings, showing that a probabilistic approach to fMRI prediction tasks holds promise. Further, we found that feature selection and regularization were effective in handling the noisy voxel measurements. We present results showing improvement over simpler models, and evaluate performance on training and test movies.

2 Method

We constructed a graphical probabilistic model of brain voxel activations and video ratings. The model described below can be learned from training data, and subsequently applied to new fMRI sessions to predict unseen ratings.

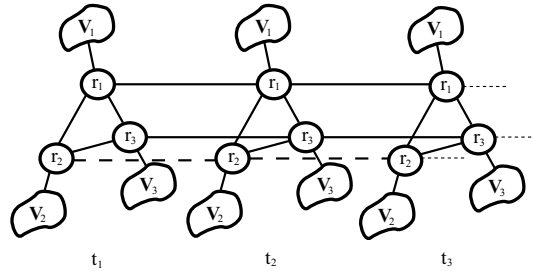


Figure 1: GMRF model for one rating r , over three subjects and three time steps.

2.1 Model

We model the interactions of voxel measurements and video ratings with Dynamic Gaussian Markov Random Field (GMRF) [3] [5]. This is an undirected graphical probabilistic model which expresses a large joint distribution using a small number of parameters. This compactness facilitates the reliable learning of models over many variables even with limited data.

Specifically, we employ a standard representation of a GMRF derived from the inverse covariance $Q = \Sigma^{-1}$ of the underlying Gaussian distribution. Note that Q only has as many non-zero entries as nodes and edges in our graph, reflecting the reduction in parameter space gained by our representation. We assume a stationary distribution, with parameters shared for components which are repeated across time steps. For instance, the parameters describing the dependence of a rating on voxels is the same at all times $t = 1, \dots, T$.

The variables in our GMRF model, Fig. 2.1, are activations V of voxels across time, and movie ratings r for all subjects across time. For each subject $s \in 1, 2, 3$ and every voxel i in that subject's brain, we introduce a set of variables representing that voxel's BOLD measurements across all time steps. For every subject s and each rating j , our model includes another set of variables representing the values the subject assigned to that rating over time.

Each rating j has edges connecting it to a set of relevant voxels. Also within a time slice rating j of subject s is connected to rating j of all other subjects. Last, across time we include edges between nodes for specific rating j and subject s .

2.2 Procedural Summary

The steps necessary to learn and use such a model are outlined as follows:

1. Use preprocessed, motion corrected functional MRI data for all subjects in Analyze format, and remove non-brain voxels, normalize remaining voxels
2. Learning 1: Feature selection and regularized regression for predicting each rating from relevant voxels
3. Learning 2: Find maximum likelihood estimate of parameters for time step edges and cross-subject edges
4. Model selection: Cross validation of feature selection and regularization
5. Prediction: Find most likely rating values jointly across time and subjects given functional data

2.3 Learning

To learn parameters for our model, we take a maximum likelihood approach - for a training data set, including the voxels *and* all the subjective ratings, and find parameters that maximize the probability of these observations. In our case, we find the non-zero entries of our precision matrix Q to maximize the likelihood of the training data.

Considering the noise in the BOLD signal and the large number of voxels, we chose to train the model conditionally [1]. Rather than maximizing $P(R, V, \theta)$, we maximize $P(R, \theta|V)$. This allows us to avoid modeling the voxel data.

We split parameter learning into two phases, an approximation that was computationally more efficient, and facilitated some estimation techniques.

First, we find parameters to predict each rating given relevant voxels from each subject as if independent. We employed Pearson correlation feature selection to identify a set of predictive voxels for each rating. Ridge regression was used to obtain regression coefficients, which are converted into precision matrix entries. For some ratings, we also learned a covariance prior over regression coefficients as a function of spatial distance between voxels [4], which can be applied as an extension to ridge regression.

Second, learning the maximum likelihood parameters for the rest of the edges (cross time and cross subject) is a convex problem for a Gaussian distribution. This sub-problem is also small enough to solve with standard optimization packages. We then combine all learned parameters from both steps, reusing them across time, to create the final joint model.

Other settings needed for our model were chosen using cross-validation. These include how many voxels to choose, regularization parameters, and whether to include the spatial prior.

2.4 Inference and Prediction

Prediction of ratings given novel functional data can be obtained from our learned model. We incorporate the observed functional data from all three subjects, and find the conditional distributions of the ratings. The means μ of this distribution are exactly the mostly likely assignments to the ratings, and have a closed form solution, which can be computed explicitly for small models. For larger models we used Gaussian Belief Propagation [6].

2.5 Other Approaches

Besides the model described in depth, we considered non-Gaussian distributions for the ratings before settling on the Gaussian model. We also explored ways to connect ratings over time and between subjects. We tried a Maximum Entropy Markov Model [2], which seemed to under-emphasize the use of voxel measurements. We also evaluated a local directed model which, rather than using global inference over the time series, added voxel data from surrounding time slices in predicting a rating. This model demonstrated improvement over single-time slice models, but was significantly worse than our global model.

3 Results and Discussion

To evaluate our model, we tested it on held out data from movies 1 and 2. We analyzed the benefit of various components of our model components. We also demonstrate that our model achieved reasonable correlations on a large number of the ratings.

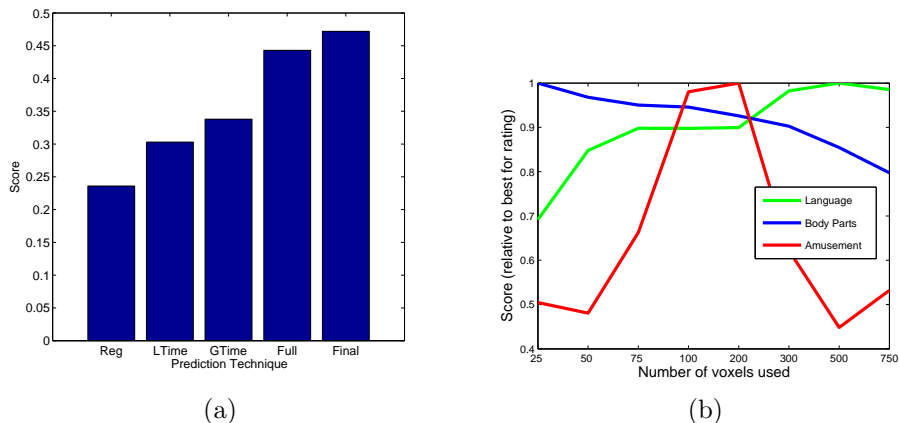


Figure 2: (a) Performance of various models on training movies (b) Performance with varying number voxels for Language, Amusement, and Body Parts

Fig. 3 shows that performance significantly improved as we moved from 1) regularized linear regression to 2) local probabilistic model incorporating only nearby time steps to 3) a global probabilistic model across time, and finally to 4) our full global model and 5) cross-validated variations. This graph shows models trained on movie1, tested on movie2, with a best score of 0.473. We observed significant variation in ratings types. For instance, for both movies 1 and 2, **Language**, **Faces**, and **Motion** scored higher than **Food** and **Amusement**. Such orderings were also largely consistent for the other methods tried.

We also analyzed our feature selection technique. We used cross-validation on training data from both movies 1 and 2 to choose the number of voxels to be used for each rating. The optimal number did vary significantly (Fig. 3) - for instance, the **Body Parts** rating does well with small voxel sets, while the **Language** rating does well with larger sets of voxels, and **Amusement** uses an intermediate number. This may reflect the actual number of voxels activated by different mental activities, but likely also reflects the voxel noise and difficulty of the predictions. Fig. 3 shows images of sample voxels chosen by feature selection. **Language** selective voxels appear in a recognizable area, but others such as **Body Parts** were hard to interpret.

We also evaluated the spatial prior used in regression. First, the prior learned from unbiased regression across the ratings tasks does indeed suggest that nearby voxels should have similar regression parameters. When we then incorporate this prior, we see some improvement, particularly in the **Actor** and **Location** ratings. The results were mild and mixed for the core ratings. The spatial prior may have worked better had we used an inflated cortex or even just gray matter.

Finally, we submitted movie3 predictions. Our numbers were slightly better than the numbers from movies 1 and 2, which may be due simply to the use of the entire training set. Our best overall score from the first two submissions was .488, and the relative performance on different ratings was also consistent with movies 1 and 2.

Overall, this work demonstrated that a wide range of predictions can be made reliably from fMRI data. Our model demonstrated the value of time-series and multi-subject data in improving fMRI prediction, especially combined with global probabilistic reasoning. Further, while single voxels are noisy, appropriate regularization and informative priors improved their predictive ability. We would like to compare this to

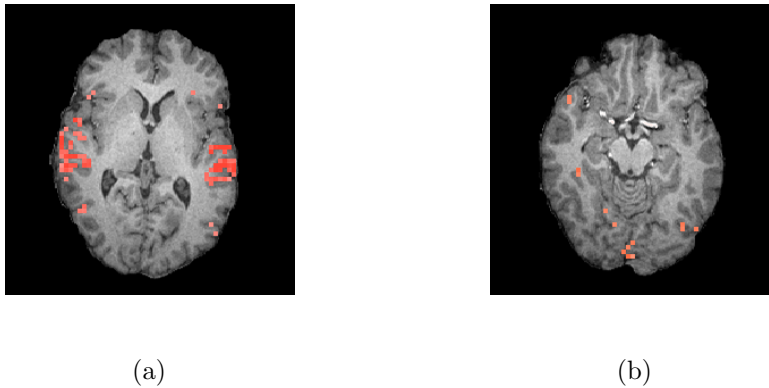


Figure 3: (a) Voxels correlated with Language, in functional slice 15, subject 1. (b) Voxels correlated with Body Parts from slice 11, subject 1

clustering, other informative priors, and RIO analysis. Other potential extensions to our work include focusing on the hardest ratings, and incorporating the relationship between rating types.

References

- [1] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.
- [2] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning*, pages 591–598. Morgan Kaufmann, San Francisco, CA, 2000.
- [3] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
- [4] R. Raina, A. Ng, and D. Koller. Transfer learning by constructing informative priors. *To appear ICML*, 2006.
- [5] M.J. Wainwright, E. B. Sudderth, and A. S. Willsky. Tree-based modeling and estimation of gaussian processes on graphs with cycles. In *Advances in Neural Information Processing Systems*, pages 661–667. MIT, 2001.
- [6] Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, 13:2173–2200, 2001.
- [7] E. Zarahn, G. K. Aguirre, and M. D’Esposito. Empirical analyses of bold fmri statistics. i. spatially unsmoothed data collected under null-hypothesis conditions. *Neuroimage*, 5(3):179–198, 1997.

4 Appendix Materials

4.1 Detailed Training Results

Below we include correlations for our best training run on movie2 for each scored rating.

Total score Subj1: 0.453, not using Actor or Location
Total score Subj2: 0.589, not using Actor or Location
Total score Subj3: 0.501, not using Actor or Location
Final score: 0.473

	Type	Used	Sub1	Sub2	Sub3
Amusement	Base	Yes	0.328	0.436	0.513
Attention	Base	Yes	0.054	0.533	0.340
Arousal	Base	Yes	0.114	0.599	0.576
BodyParts	Base	Yes	0.501	0.487	0.415
EnvSounds	Base	Yes	0.342	0.639	0.401
Faces	Base	Yes	0.665	0.690	0.687
Food	Base	Yes	0.032	0.049	-0.029
Language	Base	Yes	0.764	0.769	0.800
Laughter	Base	Yes	0.555	0.592	0.593
Motion	Base	Yes	0.561	0.511	0.600
Music	Base	Yes	0.471	0.462	0.400
Sadness	Base	Yes	0.527	0.559	0.101
Tools	Base	Yes	0.276	0.309	0.246
Kitchen	Place	Yes	0.349	0.349	0.349
LivingRoom	Place	Yes	0.356	0.356	0.356
ToolTime	Place	Yes	0.050	0.050	0.050
Tim	Actor	Yes	0.055	0.055	0.055
Jill	Actor	Yes	0.380	0.380	0.380
Al	Actor	Yes	0.253	0.253	0.253

4.2 Model Details

For completeness, we describe the formulation of a Gaussian Markov Random Field in detail. As described, the parameters are derived from the inverse of the covariance matrix, known as the *precision matrix* and written $Q = \Sigma^{-1}$. Note that in a precision matrix Q of a joint Gaussian over $X = \{x_1, \dots, x_n\}$, $Q(i, j) = 0$ exactly when variables x_i and x_j are conditionally independent given all other variables. The same independence property corresponds to the absence of an edge between x_i and x_j in our GMRF, so we can use the precision matrix to form node potentials from $Q(i, i)$ and edge potentials from $Q(i, j)$. Note that this representation assumes the mean of each variable is $\mu_i = 0$.

If E is the set of edges in our graph, the joint distribution can then be written:

$$P(X) = \frac{1}{Z} \prod_i \exp\left(-\frac{1}{2}Q(i, i)x_i^2\right) \prod_{i, j \in E} \exp(-Q(i, j)x_i x_j) \quad (1)$$