A hierarchical Bayesian approach to inferring mnemonic status from the brain

Shannon M Tubridy (shannon.tubridy@nyu.edu), **David Halpern (david.halpern@nyu.edu)**
Lila Davachi (lila.davachi@nyu.edu), **Todd M Gureckis (todd.gureckis@nyu.edu)**
New York University Department of Psychology, 6 Washington Place
New York, NY 10003

Abstract

One goal of cognitive science is to build theories of mental function that predict individual behavior. In this project we focus on predicting which word pairs in a list will be remembered at some point in the future. Contemporary approaches to this problem primarily utilize behavioral measures such as performance on quiz questions or judgements of learning. Our central hypothesis is that better prediction will come by jointly modeling both neural and behavioral data mediated by a computational cognitive model which captures the dynamics of memory retrieval over time. We lay out a framework theory for combining neural and behavioral data and present some preliminary data and simulations supportive of our approach.

Keywords: memory; fmri; cognitive model; joint modeling; hierarchical Bayesian modeling

Introduction

We develop a computational model which can predict which memories will be remembered on a later test based on observations of both behavior and neural signals. Cognitive neuroscience has identified a number of neural correlates of successful memory formation (Davachi, 2006). Complementing this work are cognitive models of memory that simulate the dynamics of learning and forgetting over time to predict whether, given a particular study history, a person is likely to have learned some piece of information (Atkinson, 1972; Corbett & Anderson, 1995). Combining such models with memory-related neural signals in a statistically optimal fashion could provide more powerful tools for understanding the links between the brain and behavior at the level of individuals (Turner et al., 2013).

To this end, we develop a neurally informed hierarchical Bayesian model of memory acquisition and decay. Our generative model adopts the three-state Markov model of mnemonic status first introduced by Atkinson (1972). Each memory could be unknown (U), in a transition state (T), or permanently learned (P) (see Figure 1). Depending on the state, we further assume an observable "emission" which reflects the pattern of brain activity associated with each latent state (similar to a hidden markov model, Rabiner, 1989). These observable emissions could accommodate many types of neural data, but here we focus on fMRI BOLD signal. Given a a prior, $\Pi_{t=0}$ over these latent memory states $S = \{s_U, s_T, s_P\}$ we can simulate a forward model of how the memory evolves. Transition probabilities, T, determine the probability of a memory moving between the different states at each point in time. Observable emissions associated with each state, E, determine the probability of observing data (behavioral or physiological) given a latent mental state. Given these parameters we can compute a posterior probability over states at any past, present, or future time and predict memory test performance. Our goal is to optimize this model to predict, at an individual level, memory performance at test for a held-out sample of participants.

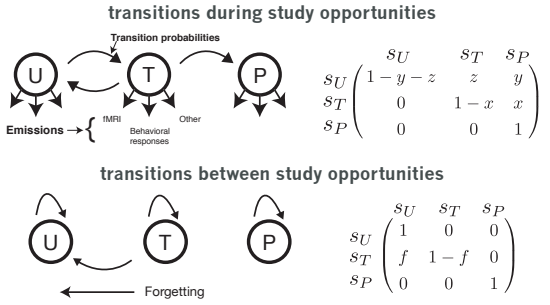

Figure 1: Structure of three state Markov model showing latent memory states, the allowable transitions between them during learning and forgetting, and the parameters governing those transitions.

Preliminary methods and results

In the follow sections we briefly report the data we have used to inform our initial modeling and some preliminary results.

Behavioral experiment

Participants (n=150) studied a set of 45 Lithuanian-English word pairs presented five times each. Memory was tested in a cued-recall test in which a Lithuanian word was presented and participants attempted to recall the associated English word. Participants were tested either 24h, 72h, or 168h after the end of the study session and there were fifty participants in each delay group. The purpose of this behavioral experiment was to help estimate some of the parameters of the hierarchical Bayesian model (e.g., the transition probabilities for individual items over time).

Preliminary MRI analyses

In addition, we collected data from nine participants who underwent fMRI scanning during study and were tested at a 72h delay. The MRI data were preprocessed in standard ways (e.g., slice time correction, motion correction, etc.). Estimates of single trial activity were pulled from anatomical maps covering 38 distinct cortical regions of interest (ROI). The average activity across voxels for each trial from each of these 38 regions were used in later analyses.

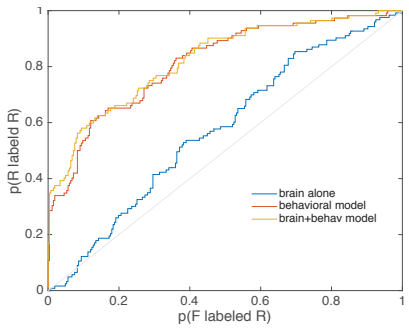

Figure 2: ROC curves for labeling memory performance (Remembered or Forgotten) using brain data alone, a model estimated using only behavioral data, and a model estimated using behavioral and fMRI data.

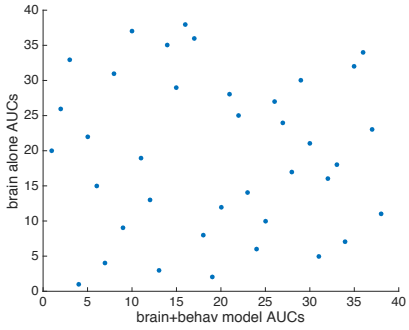

Figure 3: Scatter plot of AUCs for each brain ROI using raw data against AUCs from the models using the same ROIs.

Results

In our preliminary assessment of our approach we first considered how well activation within each ROI predicted single trial subsequent memory performance. A Receiver Operator Characteristic was created from each region which assesses how the probability of hits (correctly predicting successful memory) and false alarms (incorrectly predicting that a forgotten item will be remembered) change with different thresholds of the neural signal. Across regions we found that the most sensitive region was in left inferior frontal gyrus (IFG) with a total area under the curve (AUC) of .57 (see Figure 2, blue curve).

Next we fitted our hierarchical Bayesian model to the data from both the MRI and behavior alone participants. The model was estimated separately for each of the 38 ROIs using variational inference methods provided by STAN (Stan Development Team, 2016) and the best performing model was considered for further analysis. To compare between the fitted models (i.e., ROIs) we computed, separately for each participant, the predicted probability that a particular memory would be remembered based on the posterior mean of the model. An ROC analysis applied to these predicted probabilities from the model (for the nine fMRI subjects) is also shown in Figure 2. The AUC score of the combined model using the left IFG ROI was .82 showing the value of the combined approach. The combined model also outperformed (although marginally) a model estimated only from the behavioral data from all participants (AUC = .81).

Interestingly, the two different ways of assessing the relevance of a brain region for predicting behavior (AUC of the brain signal alone versus the posterior predictions of the model using those brain signals) disagreed about which brain regions contained the clearest signal. Figure 3 shows how the rank ordered prediction of brain regions using the two methods are uncorrelated across the two AUC measures (p=.9). This preliminary evidence indicates that the latent structure of the model is able to leverage neural information relevant for predicting memory in a different manner than would standard univariate methods.

Conclusion

In this project we have laid out the structure of a neurally informed hidden markov model of memory. Preliminary analyses have shown that combining a cognitive model with neural data marginally improves our ability to discriminate later remembered from forgotten trials and that the model was able to leverage different aspects of the ROI signal relative to a brain alone activation threshold approach.

Although our performance gains for the brain+behavior model compared to the behavioral model are modest, we find these preliminary results encouraging given limitations of this work (i.e., small data set to date). We have adopted, for now, a coarse anatomically-driven approach to identifying ROIs and summarized their activity in a simple univariate fashion. Future work will aim to identify more robust neural signatures of memory in our data set, considering the relationship between multiple ROIs and the change in signal across multiple study repetitions. In addition, this ongoing work currently contains data from small number of fMRI participants, but as we gather additional data our potential for success should improve.

Acknowledgments

Support provided by NSF grant DRL-1631436 to TMG and LD.

References

Atkinson, R. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, *96*, 124-129.

Corbett, A., & Anderson, J. (1995). Knowledge tracking: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253-278.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current opinion in Neurobiology*, *16*, 693-700.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications to speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Stan Development Team. (2016). *PyStan: the python interface to Stan.* (Version 2.14.0.0)

Turner, B., Forstmann, B., Wagenmakers, E., Brown, S., Sederbefg, P., & Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage*, *72*, 193-206.