Predicting memory performance using a joint model of brain and behavior

David Halpern (david.halpern@nyu.edu), **Shannon M Tubridy (shannon.tubridy@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

Understanding the links between brain and behavior is a central goal of computational cognitive neuroscience. We present a framework for simultaneous modeling of behavioral and neuroimaging data in the context of human memory acquisition and forgetting. Using a Hidden Markov Model of memory that can account for both behavioral and functional magnetic resonance imaging (fMRI) observations, we show that we can predict memory performance in held-out data at a level well-above chance and that we can surpass the predictions made by fMRI data alone as well as those made by variants of established behavioral models. This work highlights a path for better understanding the relationship between neural data and latent cognitive processes and advances a model of memory whose predictive ability could enable model-augmented learning environments.

Keywords: memory; fmri; joint modeling; hierarchical Bayes

Introduction

The central goal of computational cognitive neuroscience is to understand the link between human behavior and the brain. In recent decades, cognitive scientists have developed and refined detailed models of complex mental processes and behaviors. Simultaneously, cognitive neuroscientists have made strides in understanding the ways in which brain patterns are related to various cognitive states. However, integration of these levels of analysis remains lacking for many higher level cognitive activities.

In the work presented here we lay out a framework for combining behavioral and physiological or neuroimaging data in a so-called "joint model" – a single generative cognitive model that is fit to both data sources. This approach holds promise as a way to more fully integrate cognitive science and cognitive neuroscience because inferred parameters must simultaneously account for the behavior and patterns of brain activity. In addition, by leveraging more types of data, these models have the potential to make better out of sample predictions about behavior than a cognitive model fit to behavior alone.

One particularly promising focus for application of joint modeling is understanding human long-term memory acquisition and forgetting. Much is known about the effects of various learning environments on the retention of information across days and weeks (Kahana, 2012).

This work has led to the development of behavioral models that are able to track the state of an individual's knowledge for a piece of information or newly learned skill and make predictions about future performance (Atkinson, 1972; Corbett & Anderson, 1995). Paralleling these efforts, neuroimaging research has identified a number of fMRI signals measurable at the time of a learning episode that are related to the future memory status of the to-be-learned material (Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980; Davachi, Mitchell, & Wagner, 2003).

Combining these insights we show that fMRI signals combined with a cognitive model can predict memory performance in held-out data with a precision that surpasses models using behavior alone.

Memory task

The behavior we seek to predict is cued-recall performance on a set of Lithuanian-English word pairs (Grimaldi, Pyc, & Rawson, 2010). Participants studied a set of 45 word pairs, one at at time, five times each. At the end of study each pair was presented again and participants give a Judgment of Learning (JOL), indicating on a 0-100 scale whether they think they will remember the association between two words. Data from two sets of participants were acquired. A large set of behavioral participants performed the study and JOL tasks as described. Then, after a delay, participants returned for a cued recall test during which they saw the Lithuanian words one at a time and had to type the English associate. Participants returned for the recall test approximately 0 hours (N=20), 24 hours (N=49), 72 hours (N=60), or 168 hours (N=49) after the study session.

A separate set of 21 participants performed the study task while undergoing fMRI scanning. All fMRI participants did the recall test at a 72 hour delay. Functional scans covering the brain were acquired at a spatial resolution of 2.5 mm^3 with a 1 second repetition time (TR) and anatomical scans were collected at a spatial resolution of .75 mm^3.

A neurally informed Hidden Markov Model of memory

In this section we lay out the structure of our model, the kinds of data used in estimating the model, and our approach to evaluating model performance.

Markov model of memory

As a starting point, we have adapted a three-state Markov model of memory originally put forward by Atkinson (1972). This work casts memory as a Markov process in which the mnemonic status of any memory is a latent state, with transitions between states dependent only on the previous state and whether a study trial is currently occurring. Each memory (e.g., for the association between two words) can be in one of three latent states: unknown (U), known with the possibility

of forgetting (K), or permanently known (P) (Figure 1A). The model contains two sets of transition parameters, *Study Transitions* reflecting the dynamics of memory acquisition (studying leads to the possibility of learning, i.e. transitioning to a stronger memory state) and *Decay Transitions* accounting for the possibility of decay or forgetting between study events (i.e. transitioning to U) (Figure 1A).

Observable emissions

While we cannot directly observe the state of a particular memory, a key feature of the models such as Atkinson (1972) was the use of a Hidden Markov statistical model to relate observable data to the latent states and transition dynamics and it is this feature that facilitates the inclusion of neuroimaging data in our work. In this work, we use three types of observations: behavioral tests of memory (*recall*), judgments of learning (*JOL*), and fMRI activation on study trials (*MRI*). The mapping between the model states and observable data is made through a set of observation distributions defining the probabilities that an observable signal takes on a value given a latent memory state or transition (Figure 1B).

Observable behavioral data There were two forms of behavioral data included in our models: Judgments of Learning (JOLs) and binary recall performance. Both kinds of behavioral data were modeled as arising from the latent model states (Figure 1B, left panel). The probability of correct recall conditioned on state was set to fixed values of $[.01, .9, .9]$ for states U, K, and P, respectively. These values were chosen to be consistent with the notions that recall of an "unknown" memory is very unlikely although there is some possibility of guessing, and the primary (behavioral) difference between the K and P states is the susceptibility to forgetting rather than the likelihood of recall.

JOLs, which were continuous ratings from a bounded scale, were modeled as truncated Gaussians with mean and variance to be estimated from the data. As with the recall observations, our models were constructed such that JOLs arose from the model states (rather than the transitions, see below).

Observable fMRI signals For a subset of participants we recorded fMRI data during the study trials. To identify candidate features for inclusion as observations we performed group Independent Component Analysis (ICA) using the ICASSO algorithm as implemented in the GIFT ICA toolbox (http://mialab.mrn.org/software/gift/) (Calhoun, Adali, Pearlson, & Pekar, 2001; Van Maanen et al., 2011). This procedure is blind to trial information and memory outcome and resulted in a set of 60 independent components that are characterized by a particular temporal (the timecourse of activation) and spatial (the loading of each component on fMRI voxels) profile for each participant. Components that were unstable across estimations (ICASSO) and components associated with signal from ventricles or motion were discarded leaving 43 independent components for inclusion as model features. Individual trial activations were calculated as the mean of timepoints 4-6 seconds post-stimulus onset, resulting in one activation value for each trial in each component for each MRI participant.

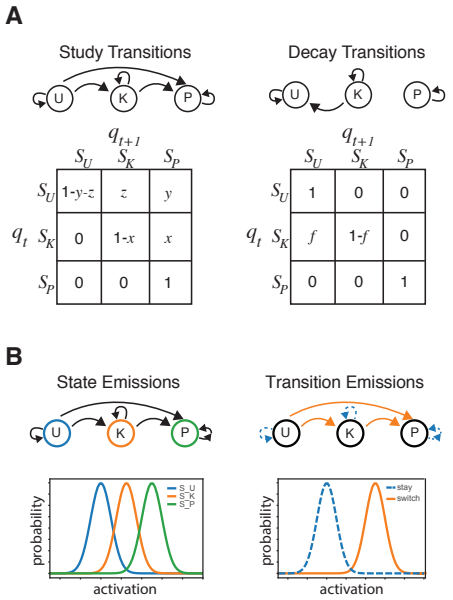

Figure 1: **A**. The matrix of transition probabilities for either study or decay events in the three state memory model. The letters within each matrix reflect the transition parameters which are estimated to data. The state labels U are "unknown", K are "known" (with possible forgetting), and P are "permanently known." **B**. Schematic of two variants for modeling observable data measured during study events: *State Emissions* arise from the latent memory states; *Transition Emissions* are reflect whether a memory stayed in same the state ("stay") or moved to a new, stronger memory state ("switch") on a study event.

For MRI observations we considered two variants of our model structure. In the ***State*** model, shown on the left of Figure 1B, the MRI signals were set up to arise from the latent states themselves. For the ***Transition*** model, we hypothesized that fMRI observations might reflect transitions between states on study trials rather than the states themselves. Given the structure of our three-state model, this means that there are observations associated with moving "up" in, or *switching*, memory state (the allowable between-state study transitions are $U \rightarrow K$ and $K \rightarrow P$) and observations associated with *staying* in the same state.

In both models the fMRI observations were modeled as Guassians with mean and variance parameters to be estimated for each component and state or transition distribution.

Model estimation

In the present work an individual's memory for each word pair was instantiated in a separate HMM. However, to get better estimates of the parameters we used a hierarchical Bayesian model that used group-level priors over the parameters to regularize the estimates. We used MCMC sampling via the NUTS algorithm as implemented in Stan (Stan Development Team, 2017a) to estimate the posterior over the parameters (4 chains of 200 iterations; 100 per chain discarded as burnin; 400 total samples per parameter). To ensure convergence, we checked

that estimates of the probability of recall had low $\hat{R}$ values (a measure of whether the sampling chains are converging to similar estimates) (Stan Development Team, 2017b; Gelman & Rubin, 1992). We estimated parameters for several different models: a *Recall* model fit to trial timing and recall performance (the binary recall success scores for each word); a model fit to trial timing, recall performance, and JOL observations (*Recall+JOL*); and two models fit to trial timing, recall performance, and fMRI observations. One of the fMRI models included fMRI observations as arising from the states (*Recall+MRI state*) and the other as reflecting transitions (*Recall+MRI transition*) (Figure 1B).

Model evaluation

We use K-fold cross validation to evaluate how well our models will predict new, unseen data, setting K to 10. Because our goal is to assess the utility of incorporating MRI signals into a memory model, the held-out data only included data from the 21 fMRI subjects. We divided up the data from these subjects into ten equally sized folds, stratified across subjects and correct trials. We then trained ten versions of each model where the training set consisted of all of the data from behavior-only subjects and nine of the ten folds of the fMRI subjects. On the held-out test set, we used the identity of the words and the trial timings (and JOL or fMRI observations, where appropriate) to generate the posterior probability of recall for each held out word at the time of test.

As we are primarily interested in our ability to classify a new piece of data as successfully recalled or not we used an area under the ROC curve metric (ROC-AUC). The model ROCs were defined by calculating, in each cross validation fold, the proportion of predicted as remembered trials that were recalled correctly (*Hits*) and the proportion of predicted as remembered trials that were not (*False Alarms*) at each level of posterior recall probability given by the model.

We also evaluated a "baseline" fMRI model that used fMRI activations without a cognitive model to predict recall. This *fMRI-baseline* model was an L2 regularized logistic regression model evaluated using the same cross validation regime as described above. The regressors in this model were the fMRI response from each study trial in each of the independent components and the to-be-predicted outcome was the probability of recall in the held-out data.

Results

Model predictions

We estimated the transition and observation parameters for several model variants and then assessed their predictions of recall performance in held out data. Comparing the ROC-AUCs across models we observed several interesting effects (Figure 2). First we note the above-chance performance of the *Recall* model (ROC-AUC = $.64\ (\pm.02)$) as a baseline against which to compare models that include other forms of observable data. The addition of JOLs as observations in the *Recall+JOL* model raised the held out ROC-AUC to $.73\ (\pm.01)$, demonstrating the utility of individual learners' metacognitive judgments in refining the model predictions.

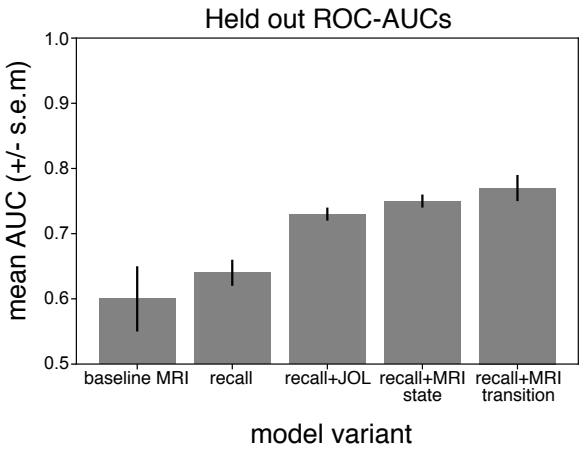

Figure 2: Mean ROC-AUC ($\pm$ s.e.m.) for memory predictions across held-out folds.

Evaluation of the MRI-based models yielded additional improvements to the ROC-AUC scores. The *Recall+MRI state* model, which included study-trial fMRI activations from a number of independent components and modeled their response as arising from the three latent memory states, achieved an AUC of $.75\ (\pm.01)$. This result, which was an improvement over the models using behavioral data alone, shows that fMRI data recorded during the course of learning can be successfully fused with a cognitive model to facilitate predictions of memory recall for individual pairs of words. This result was also a substantial improvement over the *fMRI-baseline* logistic regression model (ROC-AUC = $.60\ (\pm.05)$).

The *Recall+MRI transition* model, with fMRI activations coming from switching or staying in the same state on a particular study trial, showed the highest held-out performance, yielding a mean ROC-AUC of $.77\ (\pm.02)$.

Posterior predictive fMRI distributions

In addition to the benefits to prediction provided by joint modeling of brain and behavior, our approach enables us to inspect the model and gain insights into the ways in which different brain regions contribute to cognitive dynamics that are captured by the model. After estimation, we can examine the posterior predictive distributions for each fMRI component's activation conditioned on latent state (*MRI-state* model) or switch versus stay transition (*MRI-transition* model).

Figure 3 shows an example of posterior predictive distributions for activation in two components. The upper row shows a component associated primarily with bilateral lateral occipital and fusiform gyrus voxels along with the posterior predictive distributions from the *MRI-state* and *MRI-transition* models. This component showed activation patterns that were associated with a mean shift in magnitude for P and K states compared to U states (*MRI-state* model) as well as increased response during study trials associated with switching states relative to staying in the same state.

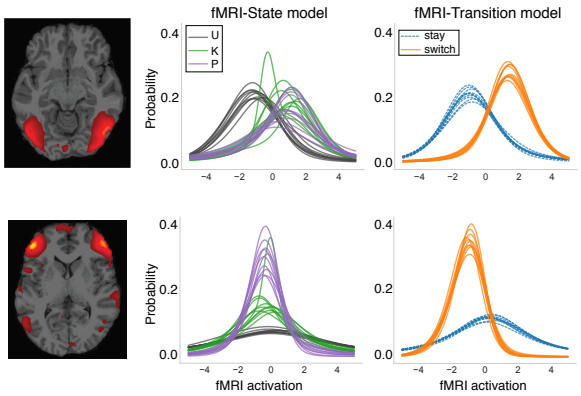

Figure 3: Posterior predictive distributions for activation in two components conditioned state or transition for the *MRI-state* and *MRI-transition* models respectively

In contrast, the component in the bottom row, primarily associated with ventro-lateral prefrontal cortical voxels, showed a different pattern. In the state model this component had similar levels of activation for each latent state (although different degrees of signal variance) whereas the transition model identified this component as having higher levels of activation and variance for *stay* transitions compared to *switch transitions*.

Conclusion

We have described a hidden Markov model of memory that can jointly model information about trial timing, behavioral observations, and fMRI measurements to predict recall performance in held-out data. The results presented here shows that when combined with a cognitive model, fMRI signals measured during the course of learning foreign language vocabulary can be leveraged to make predictions at a level surpassing that achieved by fMRI data alone or from a cognitive model using only behavioral data.

We also showed how the parameters of a generative model that captures cognitive dynamics as well as neuroimaging data can be interpreted to understand the ways in which the brain gives rise to complex cognitive behaviors unfolding over time. Although our model at this point is simple, summarizing memory in three discrete states, these analyses demonstrate the ways in which joint modeling can facilitate understanding of neural contributions to cognition beyond what is possible from analyses that are limited to considering task design or behavior without reference to the generative process.

The work described here lays the groundwork for a neurally informed model of human memory and contributes to an emerging effort to jointly model brain and behavior.

References

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

Calhoun, V., Adali, T., Pearlson, G., & Pekar, J. (2001). A method for making group inferences from functional mri data using independent component analysis. *Human Brain Mapping*, *14*(3), 140-151.

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

Davachi, L., Mitchell, J., & Wagner, A. (2003). Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proc Natl Acad Sci U S A*, *100*(4), 2157–2162.

Gelman, A., & Rubin, D. B. (1992, 11). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, *7*(4), 457–472. Retrieved from https://doi.org/10.1214/ss/1177011136 doi: 10.1214/ss/1177011136

Grimaldi, P., Pyc, M., & Rawson, K. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for lithuanian-english paired associates. *Behavior Research Methods*, *42*, 634-642.

Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press.

Paller, K., Kutas, M., & Mayes, A. (1987). Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography and clinical neurophysiology*, *67*(4), 360–71.

Ranganath, C., Johnson, M., & D'Esposito, M. (2003). Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia*, *41*(3), 378–389.

Sanquist, T., Rohrbaugh, J., Syndulko, K., & Lindsley, D. (1980). Electrocortical Signs of Levels of Processing: Perceptual Analysis and Recognition Memory. *Psychophysiology*, *17*(6), 568–576.

Stan Development Team. (2017a). *PyStan: the python interface to Stan*. Retrieved from http://mc-stan.org/ (Version 2.17.0.0)

Stan Development Team. (2017b). *Stan modeling language users guide and reference manual.* Retrieved from http://mc-stan.org/ (Version 2.17.0.0)

Van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *Journal of Neuroscience*, *31*(48), 17488-17495.