A Bayesian model of memory in a multi-context environment

Dave F. Kleinschmidt (dave.kleinschmidt@rutgers.edu)
Pernille Hemmer (pernille.hemmer@rutgers.edu)
Department of Psychology, Rutgers University, New Brunswick
152 Frelinghuysen Road, Piscataway, NJ 08854

Abstract

In a noisy but structured world, memory can be improved by enhancing limited stimulus-specific memory with statistical information about the context. To do this, people have to learn the statistical structure of their current environment. We present a Sequential Monte Carlo (particle filter) model of how people track the statistical properties of the environment across multiple contexts. This model approximates non-parametric Bayesian clustering of percepts over time, capturing how people impute structure in their perceptual experience in order to more efficiently encode that experience in memory. Each trial is treated as a draw from a context-specific distribution, where the number of contexts is unknown (and potentially infinite). The model maintains a finite set of hypotheses about how the percepts encountered thus far are assigned to contexts, updating these in parallel as each new percept comes in. We apply this model to a recall task where subjects had to recall the position of dots (Robbins, Hemmer, & Tang, 2014). Unbeknownst to subjects, each dot appeared in one of a few pre-defined regions on the screen. Our model captures subjects' ability to learn the inventory of contexts, the statistics of dot positions within each context, and the statistics of transitions between contexts—as reflected in both recall and prediction.

Keywords: Bayesian modeling; memory; learning; belief updating

Introduction

Every cognitive function—perceptual inference, learning, memory, decision making, etc.—takes place in *context*, and understanding these cognitive functions requires understanding the role that the context plays. When cognitive functions are considered in isolation, context can appear to be a source of errors, distraction, or added uncertainty. For example, Roediger and McDermott (1995) induced "false recall" by having subjects study lists of near associates of a word but not the critical word itself. However, when considered ecologically, larger-scale regularities in the environment mean that context can function as a source of additional *information*, reducing the amount of information that must be stored about particular instances. Evidence abounds that people draw on the *context* an item occurred in as an additional source of information (e.g., DuBrow, Rouhani, Niv, & Norman, 2017; Huttenlocher, Hedges, & Duncan, 1991; Orhan & Jacobs, 2013; Schulz, Franklin, & Gershman, 2018; Qian & Aslin, 2014). In this view, so-called "false recall" is really a reflection of the mis-match between the *experimenter's* defined context and the *subject's* inferred context.

However, this raises the question of what *is* a context, and how do people know? For instance, Huttenlocher et al. (1991) found that immediate spatial recall of a location in a circular area is biased towards the average radius of all locations in the experiment. They proposed that memory for an individual item's location is encoded at two levels: the item itself, and the *category* it was assigned to. However, their proposed model does not address what constitutes a category or how subjects decide, and instead simply defines the category based on the long-run statistics of locations encountered in their experiment. However, Robbins et al. (2014) discovered that in a similar task with multiple (implicit) contexts, subjects recall draws on *context*-level statistics, rather than the long-run (experiment-level) statistics.

Here, we propose a Bayesian model of learning and memory in multi-context environments, and apply this model to the data from Robbins et al. (2014) human spatial memory experiment. The model treats the problem of identifying latent contexts as a sequential non-parametric clustering problem, where agents must update their beliefs about which context they are in and the properties of that context *online*, with one data point at a time. This model thus captures psychological constraints on the discovery of latent contexts which is not captured by previous Bayesian models.

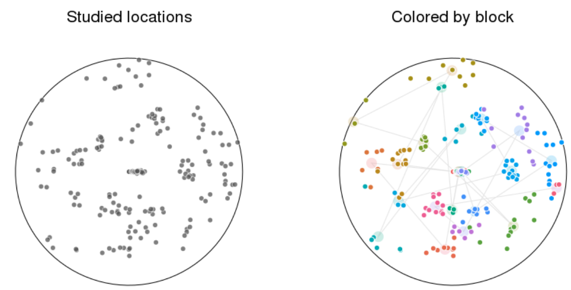

Figure 1: All locations that subject 4 studied (left), color-coded by their block (right), large dots show the average location for each block, and the gray lines show the sequence of blocks

Data

The data we model is described in detail in Robbins et al. (2014), but we provide a brief summary of the procedure here. In this experiment, 8 participants were asked to record the location of a dot presented in a circle (see Figure 1) and reconstruct that location from memory. Participants were given a cover story in order to keep the task engaging; they were told that the circle was a garden and the dots were moles. In order to save their garden, they had to "catch" the moles by clicking on the locations where they saw them.

After an initial presentation of 20 dots at the center of the circle, dots were presented in blocks (3, 6, 9, or 12 presentations in a cluster), sampled from a multinomial normal distribution with a mean of a given radius and one of three variances (0.01, 0.04, and 0.06 in a unit circle). There was no explicit signal to the subject when one block ended and the next began. The mean angles and radii were informed by Huttenlocher et al. (1991). There were 24 angle measures including the axes, and the measures consisted of the same relative angles in each quadrant. Four different distances measuring out from the center of the circle to the circumference were chosen

Each dot was viewed for one second followed by a combined visual mask and distractor task designed to remove the dot from participants' visual field and introduce uncertainty in the memory process. This mask consisted of a grid of black and white squares; after this mask was removed, an "X" appeared on the screen and participants were asked to report the color of the square (black or white)previously in that location. Data from the distractor task was recorded but not analyzed. After the completion of the distractor task, participants were asked to **recall** the location of the dot from memory by clicking a spot in the circle. After every three trials, participants were asked to make a **prediction** about a future dot location. Prediction trials alternated between prediction for the next trial and prediction for five trials from now. Each block (defined as a cluster of trials at one mean) was followed by a prediction for the expected dot location 10 trials from the current trial. This resulted in a total of 280 trials: 80 prediction trials and 200 recall trials.

Modeling

Our model has three components. First, we model how people infer the assignment of stimuli to contexts as nonparametric Bayesian clustering, approximated sequentially with a particle filter. Second, we model encoding and recall of locations as Bayesian cue combination with a prior from the context (much like Huttenlocher et al., 1991). Third, we model subjects' predictions about future locations via the posterior predictive distribution of the context model.

Context model

We modeled learners inferences about the underlying context on each trial as a sequential Bayesian non-parametric clustering problem. The goal of the learner in this model is to infer the cluster assignment z_i of observation x_i, given the previous observations $x_{1:i-1}$ and their labels $z_{1:i-1}$:

$$p(z_i = j|x_{1:i}, z_{1:i-1}) \propto p(x_i|z_i = j, z_{1:i-1}, x_{1:i-1})p(z_i = j|z_{1:i-1})$$

The sequential prior $p(z_i = j|z_{1:i-1})$ is a "Hibachi Grill Process" (Fox, Sudderth, Jordan, & Willsky, 2011, 2A; Qian & Aslin, 2014), which is like the standard Chinese Restaurant Process (CRP) with an added (constant) probability assigned to the previous state. This corresponds to the following generative model: with probability $0 < \rho < 1$ the previous state is picked, $j = z_{i-1}$, and with probability $1-\rho$ a component is chosen from a Chinese Restaurant Process with concentration α, which assigns probability to each state proportional to the number of observations assigned to it already,[1] and creates a new state with probability proportional to $\alpha > 0$. We refer to the ρ parameter as the "stickiness" because it controls how likely, a priori, the model is to stick to the same state.

The likelihood $p(x_i|z_i = j, z_{1:i-1}, x_{1:i-1}) = p(x_i|x_{\{k;z_k=j\}})$ is computed by marginalizing over the mean and covariance of a multivariate normal distribution given the data points previously assigned to that cluster and a conjugate Normal-Inverse Wishart prior (Gelman, Carlin, Stern, & Rubin, 2003). This has the advantage that it only requires tracking the sufficient statistics of the previous observations from the cluster (sample mean and covariance), and not the individual observations.

Inference: Sequential Monte Carlo

Instead of a standard batch inference technique, we use an online, Sequential Monte Carlo/particle filter technique. This method approximates the posterior beliefs after $i-1$ observations $p(z_{1:i-1}|x_{1:i-1})$ as a weighted population of K particles, each of which is one possible value of the $i-1$ labels, denoted $z^{(k)}_{1:i-1}$. This population of particles represents an *importance sample* from the posterior. When a new observation x_i comes in, the population moves to target the updated posterior $p(z_{1:i}|x_{1:i})$. There are many algorithms to do this, and the effectiveness of a particular algorithm will depend on the problem. We use the algorithm of Chen and Liu (2000), as described in, Fearnhead (2004): for each particle k, a state assignment is sampled for x_i according to $p(z_i|x_{1:i}, z^{(k)}_{1:i-1})$, and the weight $w^{(k)}_i$ is updated by the ratio of

$$\frac{\sum_j p((z^{(k)}_{1:i-1}, j)|x_{1:i})}{p(z^{(k)}_{1:i-1}|x_{1:i-1})}$$

to ensure that each particle's weight reflects its ability to *predict* the point x_i, rather than just *explain* it. When too much of the total weight for the population (constrained to sum to 1) is captured by a small number of particles (measured by the ratio of the variance of the weights to their mean being greater than 0.5), a new population is resampled (with replacement) and the weights are set to be uniform.

This is for two reasons. First, because we wish to query the model's beliefs about the current context at every point throughout the experiment, an online approximation is much more computationally efficient. A batch algorithm like Gibbs sampling or Hamiltonian Monte Carlo requires one full sweep through the data for each sample, which must be done independently for each data point, so drawing K samples for each of N data points is $O(KN^2)$. A particle filter propagates uncertainty with a fixed population of K particles, updating each particle in parallel as each data point comes in, meaning the complexity is only $O(KN)$. This means it is possible to effectively model longer experiments.

[1] One important difference from a standard CRP is that only non-sticky transitions count for the purposes of sampling new states from the CRP.

Second, an online learning algorithm better approximates *psychological* constraints on learning, and in particular unlike batch MCMC algorithms does not assume that learners can go back and revisit each observation and their decisions about it.[2] This class of models thus provides a possible bridge between computational and algorithmic level approaches to modeling learning and memory (Kleinschmidt, 2018; Sanborn, Griffiths, & Navarro, 2010).

Encoding and recall

The noisy memory trace is modeled as a normal distribution centered at the studied location x with an isometric covariance matrix Σ_x, whose diagonal elements are all equal to σ_x^2, which is a free parameter of the model. This noisy memory trace is combined with a *context prior*, which is approximated by the population of particles. Specifically, each particle k represents one possible assignment of the observations $x_{1:i}$ to clusters $z_{1:i}^{(k)}$. We can thus model each particle's context as the expected mean and covariance matrix for all the points that particle k has assigned to the same cluster as the studied point $z_i^{(k)}$:

$$\mu_c^{(k)}, \Sigma_c^{(k)} = E(\mu, \Sigma)_{p(\mu,\Sigma|x_{1:i}, z_{1:i}^{(k)})}$$

Then the best guess of the studied location under particle k's model of the context is the combination of a normal likelihood (from the noisy trace of the studied item) and a normal prior (from the context), which works out to be the inverse variance-weighted average of the two means:

$$\hat{x}^{(k)} = (\Sigma_c^{(k)^{-1}} + \Sigma_x^{-1})^{-1} (\Sigma_c^{(k)^{-1}} \mu_c^{(k)} + \Sigma_x^{-1} x)$$

Prediction

To model subjects predictions about future locations, we sample 100 locations from the posterior predictive distribution of the population of particles. To sample one predicted location at a n trials in the future, we sample a particle from the population according to their weights, draw a sample of n future states from that particle's Hibachi Grill Process, and then sample one point from the posterior predictive distribution of the resulting cluster. In the case that the predicted cluster is a new cluster, we sample from the prior predictive.

Procedure

To evaluate this model, we simulated the data from Robbins et al. (2014) with a range of parameter values. The concentration parameter α was set to $0.01, 0.1, 1$, or 10, and the stickiness parameter ρ was set to $0.1, 0.5$, or 0.9. The memory noise standard deviation parameter σ_x varied along $0.01, 0.1, 1$, (for a circle with a radius of 1), although only results from $\sigma_x = 0.1$ are presented here. The prior for the cluster parameters was based on the distribution of true block means/covariances. In principle, this could be inferred as well but we leave that enhancement for future work. We ran 10 repetitions with each of the 36 combinations of parameters, all of which used 100 particles for each subject's data.

[2]These approaches also do not *preclude* revising previous decisions, they just do not *require* it.

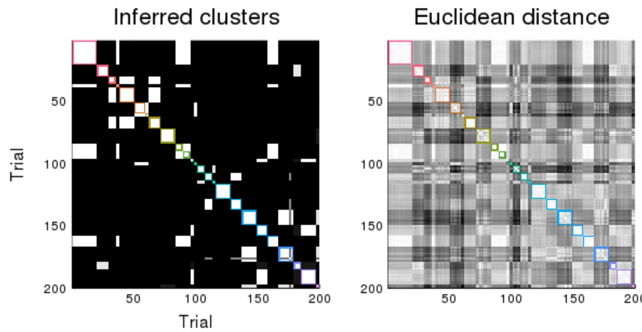

Figure 2: Cluster assignment similarity matrix for clusters inferred by one population of particles from subject 4's studied locations (left), with the true (experimenter-defined) blocks outlined in colors (see Figure 1). The similarity matrix based on the Euclidean distance between each location is shown for comparison (right) and to show that the model groups some similar locations into the same cluster even though they are from different blocks.

The particle filter algorithm was implemented in Julia 1.1 (Bezanson, Edelman, Karpinski, & Shah, 2017). The code, simulation results, and Weave.jl (Pastell, 2017) source for this paper is available from osf.io/dqz73/

Results

Clustering

First, how well does this algorithm do at recovering the underlying cluster structure? This is not a straightforward question to answer: each particle in the population represents a potentially different assignment of observations to clusters, and the cluster indices used in one particle might not align with those in another particle. To get around this we look at the assignment similarity matrix, which is an $N \times N$ matrix, where element (i, j) is the probability that trials i and j are assigned to the same cluster. This probability is calculated by averaging across all particles in the population according to their weight.

Figure 2 shows the assignment similarity matrix for one subject, based on a 100-particle filter with $\alpha = 0.01, \rho = 0.9$ (left) with the true, experimenter-defined block structure is outlined in the colors from Figure 1, and the pairwise Euclidean distance between the locations for comparison (right). This example shows a number of important features of the model's inferences about the underlying changes in context. First, relative to the experimenter-defined blocks, the model occasionally undersegments, grouping adjacent blocks together into a single context. Second, the model also sometimes infers that it has *returned* to a previous context, instead of creating a new context when it infers that the block has

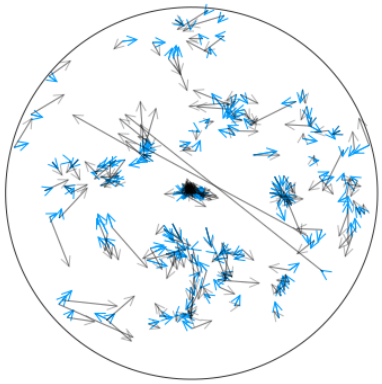

Figure 3: Subject 4's recalled locations (gray arrows, pointing from studied to recalled location) compared with model simulation (blue arrows; $\alpha = 0.01, \rho = 0.9, \sigma_x = 0.1$)

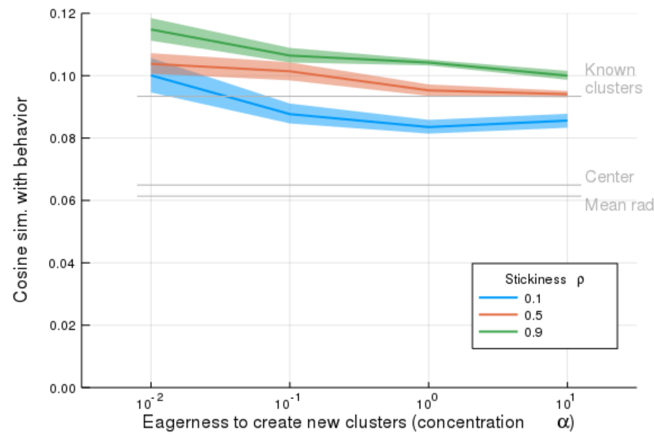

Figure 4: Mean cosine-similarity of model predicted and actual recall deviations across parameter values (ribbons show 95% bootstrapped CIs over model runs). Gray lines show baselines: always deviate toward center, average radius, and center of true clusters

changed. This can be seen from the off-(block)-diagonal entries in the assignment similarity matrix (Figure 2, left). As the Euclidean similarity matrix (Figure 2, right) shows, this tends to happen when the points in two blocks are close together. Third, because of the online nature of the model, it maintains relatively less uncertainty about the clustering of early trials. Note though that Figure 2 shows the beliefs of the model at the *end* of the experiment, which reflect the totality of the locations it has encountered.

Recall

Next, we assess how well the inferred contexts can predict recall. Figure 3 shows one subject's actual deviations from studied to recalled locations (gray arrows) versus the model's predicted deviations (blue arrows). To quantify goodness of fit, we use the cosine similarity of the model's and subject's recall deviation (i.e., blue and black arrows in Figure 3), which ranges from 1 (deviations perfectly aligned) to -1 (deviations in opposite directions), with 0 corresponding to orthogonal deviations. We chose this metric because it is less sensitive to large outlier responses than mean-squared error, and because approximations of the likelihood of a subject's response given the model is highly sensitive to free parameters and difficult to reliably estimate. Moreover, the baseline models we compare against also do not have straightforward likelihood models, but they *do* make straightforward predictions about the directions of recall deviations.

Figure 4 shows the cosine similarity with of all subjects' responses with the multi-context Bayesian model. The ribbons show the 95% bootstrapped confidence intervals over model runs, which indicate that the approximate inference strategy leads to reasonably consistent inferences for a given set of parameters. At all parameter settings, the model performs better than chance, predicting subjects' recall deviation directions at a cosine similarity of around 0.1 (relative to a chance level of 0). The model performs best for high ρ stickiness and low α concentration.

We also compare the model's performance against three baselines. First, we compare it against a "known clusters" model, which uses the true (experimenter defined) clusters with the same Bayesian cue combination model of encoding and recall. Second, we compare it to two baselines based on previous literature on similar memory tasks (Huttenlocher et al., 1991): one that always biases recall towards the center (the average location of all trials), and one that biases recall towards the mean radius.

First, at the whole range of parameters explored, the multi-context model performs better than the center- or mean-radius-biased baselines. Second, except for low stickiness $\rho = 0.1$, our model provides a better fit to human behavior than the "known clusters" baseline, which differs from our model only in that the true cluster labels are provided for each data point, rather than being inferred. This suggests that, at least according to the cosine similarity metric, our context-inference model better captures how people combine information about the current context during recall than the "ground truth" clusters.

However, an important caveat is that there is substantial variability across *subjects*. The cosine similarity for $\alpha = 0.01, \rho = 0.9$ has a 95% boostrapped CI across subjects of $[0.05, 0.17]$, which while significantly better than chance is not significantly better than the baseline models, even when taking into account the substantial variability in the cosine similarity for the baseline models themselves. With only 8 subjects in this dataset it is unclear how well the model's performance will generalize to other datasets, and future work with better-powered designs is required.

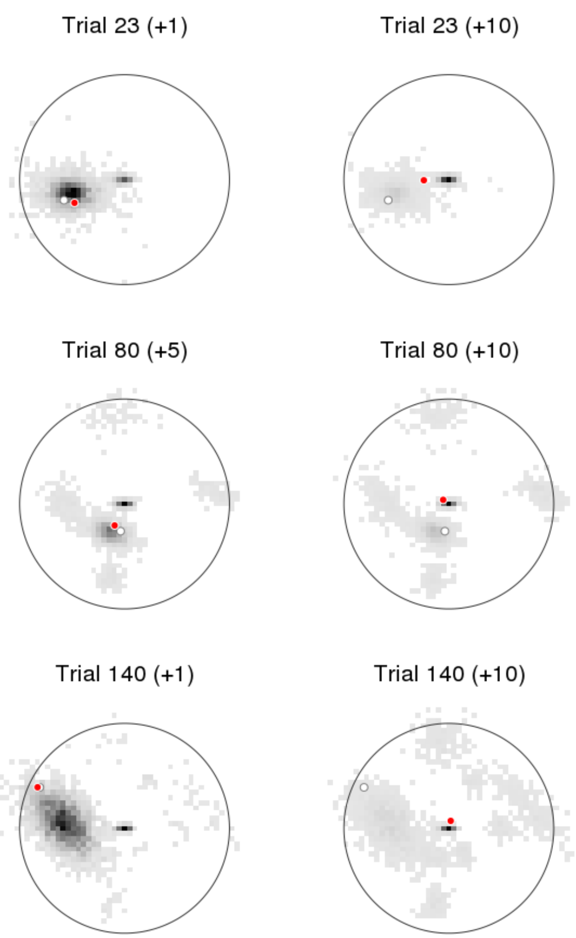

Figure 5: Subject 7's (red points) and model's (gray regions) predictions about upcoming locations at various points throughout the experiment and various prediction horizons. The white points show the last recalled location.

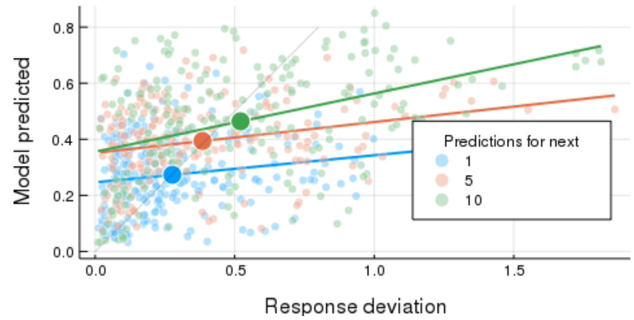

Figure 6: Model predicted ($\alpha = 0.01, \rho = 0.9$) and actual deviations from last studied point for prediction task. Small points show deviations of predictions for each trial, and large points show average deviations for each lag (1, 5, or 10 trials).

Prediction

Subjects also, every three recall trials, predicted the location where points would appear in 1, 5, or 10 trials in the future. This is a more explicit probe of what subjects know about the cluster structure than the recall task. Figure 5 shows six examples of how the model's prediction about upcoming locations capture subjects' behavior. For +1 trial predictions, the model's distribution of predicted locations primarily reflects its beliefs about the *current* cluster (as reflected by the higher density of predictions near the white studied point), because of the "sticky" Hibachi Grill Process prior on states. At +10 trials, the model is much more likely to predict the center cluster, which recurs frequently throughout the experiment (see also Figure 2). Likewise, subjects also have picked up on this pattern and are more likely to predict locations close to the center on +10 prediction trials.

Our model also captures how the average distance from the last studied point increases as subjects are asked to predict the location of points +1, +5, and +10 trials into the future (Figure 6, large points). Moreover, the model also captures variation *within* these delay levels: after removing the effect of delay level by centering, the model's and subjects' prediction deviations are correlated at $\rho = 0.31$ (95% bootstrapped CI: $[0.25, 0.38]$, and significant at $p = 0.014$ in a mixed model with random intercepts and slopes by subject).

Discussion

We have demonstrated that human recall and prediction in a multi-context spatial memory task can be modeled by a Bayesian model that infers the latent contexts via nonparametric clustering. This model updates its beliefs *online*, one observation at a time, with Sequential Monte Carlo. Exploring a range of parameters for the state transition prior, we found that subjects recall behavior is best captured with high "stickiness" (prior probability of remaining in the same cluster) and low concentration (prior probability of creating a new cluster). Together, this suggests that people expect—

until they receive evidence to the contrary—that contexts will continue for a number of trials, and that old contexts will return in the future.

While we treated these parameters as free when fitting our model, this was merely a simplifying assumption that we made to make the model easier to implement. It is possible—and conceptually fairly straightforward in a Bayesian model like this—that they could be *inferred* from the same data that the model uses to infer the contexts themselves. It is thus possible that our interpretation of what these parameter values mean for people's expectations about the latent cluster structure actually reflect what people have *learned* from their experience in this particular experiment, where contexts *do* tend to go on for a number of trials and recur multiple times (at least for the central cluster). Future work is required to tease these possibilities apart.

The possibility that people might be inferring the hyperparameters that govern how contexts change raises the question of what kind of changes people expect in the structure of contexts across environments. That is, are people's models of contexts nested hierarchically, in a way that allows for variation not only in the specific features of each context (e.g., the location of dots in space) but also the properties of how contexts *change* within a larger context/environment (e.g., the stickiness of contexts)? This calls for future experiments that manipulate the generative model for the contexts themselves, within subjects and over time.

More work is also needed to assess whether people actually are remembering and revisiting old contexts, as our model assumes. It is possible that people are really just detecting *changes* in context, and creating a fresh representation of a context every time they detect such a change. One way to address this is by simulating such a change-point model, which is the limiting case of our model when the concentration parameter α goes to infinity. Another way is to collect more empirical data with changes in context explicitly designed to elicit anticipation for returning to old contexts.

Finally, the strategy of our model—inferring discrete changes in context and remembering contexts—presupposes a particular underlying structure for how contexts actually tend to change in the world. A number of different strategies could be optimal, given different environments, and it is an ecological question as to which strategies are likely to be useful in the kinds of environments people tend to find themselves in. For instance, environments where latent variables don't change suddenly but rather drift slowly and continuously call for a very different family of strategies. So while our model describes behavior well in *this* particular experimental environment, that does not necessarily mean that it would also describe behavior well in an environment that does not follow the structural assumptions that the model makes.

Conclusion

In a structured world, local context—either simultaneous or temporally extended—can provide a great deal of information about how to interpret or remember stimuli. We have proposed a Bayesian model that infers latent context variables from unlabeled data, and uses that context to encode and retrieve information from memory. This model processes data *online*, one observation at a time, and captures people's behavior in a multi-context spatial memory task.

Acknowledgments

This work was supported by a National Science Foundation Grant 1453276 awarded to Pernille Hemmer.

References

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. doi:10.1137/141000671

Chen, R., & Liu, J. S. (2000). Mixture Kalman Filters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(3), 493–508. JSTOR: 2680693

DuBrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017). Does mental context drift or shift? *Current opinion in behavioral sciences*, *17*, 141–146. doi:10.1016/j.cobeha.2017.08.003. pmid: 29335678

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, *14*(1), 11–21. doi:10.1023/B:STCO.0000009418.04621.cd

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, *5*, 1020–1056. doi:10.1214/10-AOAS395

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second). Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*(3), 352–376. doi:10.1037/0033-295X.98.3.352

Kleinschmidt, D. F. (2018). Learning distributions as they come: Particle filter models for online distributional learning of phonetic categories. In T. T. Rogers, X. Rau, X. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1933–1938). doi:10.31234/osf.io/dymc8

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*(2), 297–328. doi:10.1037/a0031541

Pastell, M. (2017). Weave.jl: Scientific Reports Using Julia. *The Journal of Open Source Software*, *2*(11), 204. doi:10.21105/joss.00204

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences*, *111*(40), 14400–14405. doi:10.1073/pnas.1416109111

Robbins, T., Hemmer, P., & Tang, Y. (2014). Bayesian Updating: A Framework for Understanding Medical Decision Making. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (p. 6). Quebec City: Cognitive Science Society.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. doi:10.1037/0278-7393.21.4.803

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–67. doi:10.1037/a0020511. PMID: 21038975

Schulz, E., Franklin, N. T., & Gershman, S. J. (2018). Finding structure in multi-armed bandits. *bioRxiv*, 432534. doi:10.1101/432534