Supplemental Information for

Individual Differences in Learning Predict the Return of Fear

Samuel J. Gershman and Catherine A. Hartley

Contents

1 Overview

The supplemental text is divided into 3 parts. In Section 2, we provide a technical description of the *generative process* underlying our computational model, which is a variation of the model presented in [8]. The generative process is a specification of the learner's internal model of the environment; it is generative in the sense that it describes a recipe for generating stimulus configurations from latent causes (or "states" as we will call them here for brevity).

Learning is the process of assigning observed stimulus configurations to the states that generated them. Since many possible states could have generated any given set of stimuli, this problem is rationally answered by representing and updating a distribution over states. Bayes' rule prescribes how this updating should proceed [7]. However, in our case Bayes' rule is not computationally tractable, so we must resort to approximation methods; in Section 3, we describe two such approximations, both of which are used for model-fitting. Finally, in Section 4, we provide details of how we fit the computational model to skin conductance response (SCR) data.

2 The generative process

On trial t, a learner observes a stimulus configuration, represented by a D-dimensional binary vector, $\mathbf{f}_t \in \{0,1\}^D$. Each dimension of the vector corresponds to a stimulus; for ease of exposition, we will only consider here two stimuli, the CS and US. The occurrence of stimulus d on trial t is denoted by $f_{td} = 1$. We will use $\mathbf{F}_{1:t}$ to denote the history of stimulus configurations from trial 1 to t. Likewise, we will use $\mathbf{c}_{1:t}$ to denote the state sequence from trial 1 to t.

We impute to the learner an internal model of the environment. This internal model specifies a probabilistic "recipe" for generating stimulus configurations:

$$\mathbf{c} \sim \text{CRP}(\alpha) \tag{1}$$

$$\phi_{kd} \sim \text{Beta}(a, b) \tag{2}$$

$$f_{td} \sim \text{Bernoulli}(\phi_{c_t d}) \tag{3}$$

Here $\mathbf{c}$ is a vector specifying the state that generated each trial's stimulus configuration. The distribution over states, denoted $\text{CRP}(\alpha)$, is the *Chinese restaurant process* [1, 11] with concentration parameter α. Its name comes from the following metaphor: Imagine a Chinese restaurant with an unbounded number of tables (states). The first customer (trial) enters and sits at the first table. Subsequent customers sit at an occupied table with a probability proportional to how many people are already sitting there, and at a new table with probability proportional to α. Once all the customers are seated, one has a clustering of trials into states. Mathematically, this distribution is defined by:

$$P(c_t = k|\mathbf{c}_{1:t-1}) = \begin{cases} \frac{N_k}{t-1+\alpha} & \text{if } k \leq K \\ \frac{\alpha}{t-1+\alpha} & \text{if } k = K+1, \end{cases} \tag{4}$$

where K is the total number of states generated up to trial t, and N_k is the number of trials already generated by state k (by default it is assumed that $c_1 = 1$). Intuitively, larger values of α lead to more states. In the limit $\alpha \to \infty$, every trial is generated by a unique state. In the limit $\alpha \to 0$, all trials are generated by the same state. If we were to sample T trials from this distribution, we would obtain on average $\alpha \log T$ states. Note, however, that the posterior over states (see next section) will not generally obey this law.

Each state k is associated with a "prototype" $\boldsymbol{\phi}_k$, which determines the distribution over stimulus configurations when state k is active. The prototype describes the central tendency of this distribution, namely $\mathbb{E}[f_{td}|c_t = k] = \phi_{kd}$. Each state's prototype is drawn from a Beta(a,b) distribution. When $a = b = 1$, this is a uniform distribution over the $[0, 1]$ interval, and thus all prototypes are equally likely *a priori*.

3 Approximating Bayesian inference

The computational problem facing a learner is to infer the sequence of states that gave rise to the observed stimulus configurations. On trial t, the posterior probability that state k generated stimulus configuration $\mathbf{f}_t$ is given by Bayes' rule:

$$\begin{aligned} P(c_t = k|\mathbf{F}_{1:t}) &\propto P(\mathbf{f}_t|c_t = k, \mathbf{F}_{1:t-1})P(c_t = k|\mathbf{F}_{1:t-1}) \\ &= \sum_{\mathbf{c}_{1:t-1}} P(\mathbf{f}_t|c_t = k, \mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1})P(c_t = k|\mathbf{c}_{1:t-1})P(\mathbf{c}_{1:t-1}|\mathbf{F}_{1:t-1}). \end{aligned} \tag{5}$$

The first term in Eq. 5 is the *likelihood*, expressing the match between $\mathbf{f}_t$ and the prototype of state k. More precisely, because the learner has uncertainty about the prototype, the likelihood is actually an *average* over all possible prototypes, weighted by their probability:

$$\begin{aligned} P(\mathbf{f}_t|c_t = k, \mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}) &= \int_{\phi_k} P(f_{td} = j|c_t = k, \mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}, \phi_k)P(\phi_k|\mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1})d\phi_k \\ &= \frac{M_{kd} + a}{N_k + a + b}, \end{aligned} \tag{6}$$

where $M_{kd} = \sum_{\tau=1}^{t-1} f_{\tau d}$ is the number of times stimulus d co-occurred with state k prior to trial t. The second term in Eq. 5 is the CRP prior (Eq. 4). The third term is the posterior over the state sequences from trial 1 to $t - 1$.

Because the number of unique state sequences grows exponentially with t, the sum in Eq. 5 quickly becomes computationally intractable. Below we consider two expedient approximations.

3.1 The local MAP approximation

The simplest and least expensive approximation is to choose a single high probability state sequence $\hat{\mathbf{c}}_{1:t-1}$, rather than summing over all possible state sequences. While finding the highest probability (*maximum a posteriori*, or MAP) state sequence would still require searching over all state sequences, we can iteratively construct a "local" MAP approximation:

$$\hat{c}_t = \operatorname*{argmax}_k P(\mathbf{f}_t|c_t = k, \mathbf{F}_{1:t-1}, \hat{\mathbf{c}}_{1:t-1})P(c_t = k|\hat{\mathbf{c}}_{1:t-1}). \tag{7}$$

This approximation will be reasonably accurate when there is not much ambiguity in the state sequence—i.e., the stimulus configurations fall reliably into distinct clusters. This approximation has been used in both the psychological literature [2, 4] and the statistical literature [5, 12].

3.2 Particle filtering

A more accurate approximation can be obtained by summing over a set of L "particles" (hypothetical state sequences) drawn from $P(\mathbf{c}_{1:t-1}|\mathbf{F}_{1:t-1})$ a technique known as *particle filtering* [6]. By the Law of Large Numbers, this approximation will become increasingly accurate with larger L.

Letting $\mathbf{c}_{1:t}^{(1:L)}$ denote the set of particles, the posterior over state sequences is approximated by:

$$P(\mathbf{c}_{1:t} = \mathbf{c}|\mathbf{F}_{1:t}) \approx \frac{1}{L}\sum_{l=1}^{L} \delta\left[\mathbf{c}_{1:t}^{(l)}, \mathbf{c}\right], \tag{8}$$

where $\delta[\cdot,\cdot]$ is 1 when its arguments are equal and 0 otherwise. On each trial t, we stochastically draw a set of state assignments:

$$P(c_t^{(l)} = k|\mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(1:L)}) \propto P(\mathbf{f}_t|c_t^{(l)} = k, \mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1})P(c_t^{(l)} = k|\mathbf{c}_{1:t-1}). \tag{9}$$

To normalize this distribution, we need only sum over the L particles, rather than all possible state sequences.

4 Model-fitting

In order to map the model output to SCR, we assume that the SCR on trial t, denoted by y_t, is a linear function of the model's predicted probability of the US, v_t:

$$y_t = \beta v_t + \epsilon, \tag{10}$$

where β is a scaling parameter and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a random noise term. We fix $\sigma^2 = 1$ for our analyses. Assuming that the first stimulus dimension corresponds to the US and the second dimension corresponds to the CS, the probability of a US given a CS is calculated according to:

$$\begin{aligned} v_t &= P(f_{t1} = 1|f_{t2}, \mathbf{F}_{1:t-1}) \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\sum_{k} P(f_{t1} = 1|c_t^{(l)} = k, \mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(l)})P(c_t^{(l)} = k|f_{t2}, \mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(l)}). \end{aligned} \tag{11}$$

The expression using the local MAP approximation is analogous, but with only a single particle ($L = 1$).

Using this model of the SCR, the log-likelihood of the data given parameters (α, β) is given by:

$$\mathcal{L}(\alpha, \beta) = -\frac{T}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \beta v_t)^2, \tag{12}$$

where T is the total number of trials and v_t is implicitly a function of α. This equation is derived from the logarithm of the Gaussian distribution. To compare with a model in which α is fixed to 0, we approximated the log Bayes factor [7] using:

$$\log \mathrm{BF} \approx \log \int_{\alpha} \exp\{\mathcal{L}(\alpha, \hat{\beta})\} d\alpha - \mathcal{L}(0, \hat{\beta}), \tag{13}$$

where $\hat{\beta}$ is the maximum likelihood estimate of β conditional on α.

For reasons of computational efficiency, we first found the maximum-likelihood parameter estimates using the local MAP approximation. To obtain predictions about the inferred states, we then ran the full particle filtering model with $L = 1000$ and α set to the posterior mean (for each participant separately). The model was fit to data from the acquisition phase and the first 4 blocks of extinction.

To cluster participants into two groups based on their model fits, we chose a threshold on the log BF that minimized the difference between the average differential SCR for the two groups on the second-to-last block of extinction. This ensured that the two groups extinguished to roughly the same level. The threshold minimizing this difference was 0.088, although the results were not sensitive to this precise value. Participants falling below the threshold tended to have a single state (the one-state group), while participants falling above the threshold tended to have two states (the two-state group).

5 Description of individual differences measures

The State-Trait Anxiety Inventory (STAI) is a psychological inventory based on a 4-point Likert scale and consists of 40 questions on a self-report basis. The STAI measures two types of anxiety - state anxiety, or anxiety about an event, and trait anxiety, or anxiety level as a personal characteristic. Higher scores reflect higher levels of anxiety. Candidate gene studies have linked polymorphic variation within the serotonin transporter gene (SLC6A4) to differences in fear and anxiety phenotypes. The 5-HTTLPR (serotonin- transporter-linked promoter region) is a variable repeat length polymorphism within the promoter region of SLC6A4. The 5-HTTLP has a short and a long allelic variant. The short allele has been proposed to impair emotion regulation processes, giving rise to an anxious phenotype [3, 9].

The serotonin transporter polyadenylation polymorphism (rs3813034/STPP) is a common T/G single nucleotide polymorphism that alters the use of the polyadenylation signal in which it occurs. The G allele of the STPP has been associated with decreased fear extinction retention as well as increased trait anxiety [10].

References

[1] D. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer, Berlin, 1985.

[2] J.R. Anderson. *The Adaptive Character of Thought*. Lawrence Erlbaum, 1990.

[3] Turhan Canli and Klaus-Peter Lesch. Long story short: the serotonin transporter in emotion regulation and social cognition. *Nature Neuroscience*, 10:1103–1109, 2007.

[4] Anne GE Collins and Michael J Frank. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120:190–229, 2013.

[5] Hal Daumé III. Fast search for Dirichlet process mixture models. *International Conference on Artificial Intelligence and Statistics*, 2007.

[6] A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

[7] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2004.

[8] Samuel J Gershman and Yael Niv. Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, 40:255–268, 2012.

[9] Ahmad R Hariri and Andrew Holmes. Genetics of emotional regulation: the role of the serotonin transporter in neural function. *Trends in Cognitive Sciences*, 10:182–191, 2006.

[10] Catherine A Hartley, Morgan C McKenna, Rabia Salman, Andrew Holmes, BJ Casey, Elizabeth A Phelps, and Charles E Glatt. Serotonin transporter polyadenylation polymorphism modulates the retention of fear extinction memory. *Proceedings of the National Academy of Sciences*, 109:5493–5498, 2012.

[11] J. Pitman. *Combinatorial Stochastic Processes*. Notes for Saint Flour Summer School. Techincal Report 621, Dept. Statistics, UC Berkeley., 2002.

[12] Lianming Wang and David B Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20, 2011.

6 Supplementary figures

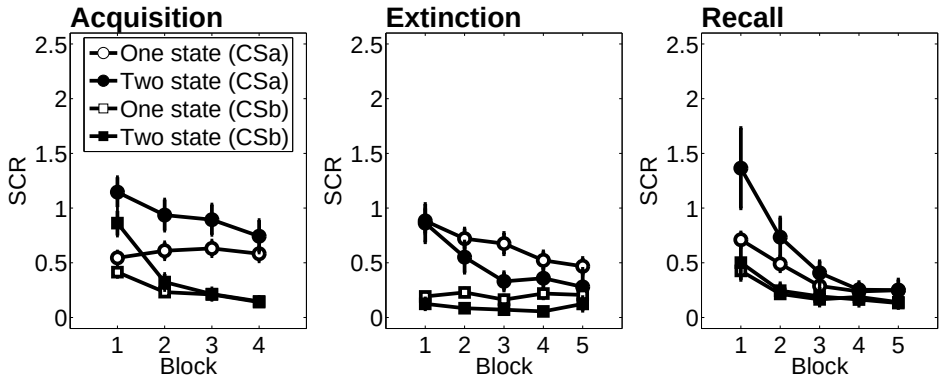

Figure 1: **Raw SCR to CSa and CSb**. Error bars show standard error of the mean.

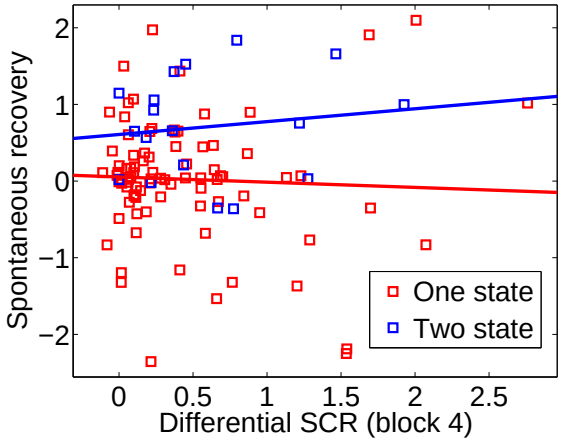

Figure 2: **Spontaneous recovery plotted against asymptotic level of responding during the acquisition phase.**