THE COEVOLUTION OF DATA AND HYPOTHESES IN BAYESIAN CULTURAL EVOLUTION

VANESSA FERDINAND

Santa Fe Institute, Santa Fe, USA
vanessa@santafe.edu

Culture is a special evolutionary system that is composed of two types of replicators: public structures in the world, such as artifacts and behaviors, and private structures in the mind, such as brain states or grammars. In this paper, I utilize a mathematical equivalence between replicator dynamics and Bayesian inference to specify a model where cultural artifacts and learners' hypotheses co-evolve.

1. A Replicator Dynamics Model of Bayesian Cultural Evolution

1.1. *Model A: Artifact Evolution*

The replicator dynamics equation (Hofbauer & Sigmund, 1998) is a general model of natural selection that can be used to describe the evolution of a population of cultural artifacts or behaviors over time, where $p(d_i)$ is the proportion of the ith artifact in the population of artifacts and $p(d_i)'$ is its proportion in the next generation. The relationship between $p(d_i)$ and $p(d_i)'$ is

$$p(d_i)' = \frac{p(d_i)f(d_i)}{\sum_{k=1}^{n} p(d_k)f(d_k)} \tag{1}$$

where $f(d_i)$ is the fitness of the ith artifact and $\sum_{k=1}^{n} p(d_k)f(d_k)$ is the mean fitness of all the artifacts in the population. $f(d_i)$ can be unpacked into the expression $f(d_i|e_j)$, which specifies that the fitness of d is dependent upon the environment, e, at any given time. The population of artifacts can adapt to a fixed environment, where $f(d_i)$ is the same each generation, or a changing environment, where $f(d_i)$ is different each generation.

DOI:10.12775/3991-1.028

1.2. *Model B: Bayesian Hypotheses Evolution*

Bayesian inference can be used to describe learning, where an agent entertains a set of hypotheses about the state of the world and then updates the probability of these hypotheses after observing some data from the world. The agent assigns each hypothesis a certain probability of being correct before learning, $p(h_j)$, and after learning, $p(h_j)'$. The relationship between $p(h_j)$ and $p(h_j)'$ is

$$p(h_j)' = \frac{p(h_j)p(d_i|h_j)}{\sum_{k=1}^{n} p(h_k)p(d_i|h_k)} \tag{2}$$

where $p(d_i|h_j)$ is the likelihood of the observed data under the jth hypothesis and $\sum_{k=1}^{n} p(h_k)p(d_i|h_j)$ is the mean probability of the data under all of the agent's hypotheses. Bayesian updating can be iterated over a sequence of observations. These observations can be the same each generation, where the hypotheses adapt to a fixed environment, or they can can differ each generation, where the hypotheses adapt to a moving target.

1.3. *The Fitness Landscape of Learning*

Even though Bayesian updating is a model of learning, we can talk about the evolution of hypotheses because the replicator dynamics equation and Bayesian inference are formally equivalent, as noted by (Shalizi et al., 2009; Harper, 2009).[1] The beauty of this equivalence lies in the interpretation of fitness in learning: the fitness of hypotheses (i.e. what causes them to gain differential support in a learner's mind) is simply the likelihood of the data under each hypothesis. Likewise, data are differentially reproduced on the basis of the support they have under each hypothesis: data that make more sense under a given hypothesis are more likely to survive than data points that make less sense under it. Therefore, the fitness of the hypotheses and the fitness of the data are both determined by the likelihood of the data given the hypotheses, such that:

$$f(d_i) = p(d_i|h^*) \tag{3}$$

$$f(h_j) = p(d^*|h_j) \tag{4}$$

Where the $*$ indicates one particular hypothesis or data type. This has the effect of yoking the fitness values of the data and hypotheses to the same fitness landscape, defined by the likelihood matrix W_{ij}. Equation 5 gives an example matrix for a population of three data points and three hypotheses:

[1] Due to this equivalence, we can conceptualize the probability distribution over hypotheses as a population of hypotheses. Therefore, this paper will refer to $p(h)$ in two ways: as the probability of a hypothesis and as the proportion of a hypothesis in a population of hypotheses.

$$W_{ij} = p(d_i|h_j) = \begin{array}{cc} \begin{matrix} h_1 & h_2 & h_3 \end{matrix} & \\ \begin{bmatrix} .8 & .2 & .3 \\ .1 & .6 & .3 \\ .1 & .2 & .4 \end{bmatrix} & \begin{matrix} d_1 \\ d_2 \\ d_3 \end{matrix} \end{array} \tag{5}$$

For example, if the population of hypotheses consists only of h_1, the relative fitness of $\{d_1, d_2, d_3\} = \{.8, .1, .1\}$ and if the population of data consists only of d_1, the relative fitness of $\{h_1, h_2, h_3\} = \{.62, .15, .23\}$. For any population of hypotheses that contain a mixture between hypotheses types, I define the fitness of each data type to be the weighted average of its fitness under each hypothesis, where each weight equals the hypothesis' proportion in the population. Likewise, the relative fitness of each hypothesis is the weighted average of its fitness under each data type, where weights equal each data type's proportion in the population.

1.4. *Model C: Coevolution of Artifacts and Hypotheses*

Model A and B can be combined to model the coevolution of data and hypotheses, where a population of artifacts adapts to a distribution over hypotheses, and vice a versa:

$$p(d_i)' = \sum_j \left[\left(\frac{p(d_i)p(d_i|h_j)}{\sum_{k=1}^{n} p(d_k)p(d_k|h_j)} \right) p(h_j) \right] \tag{6}$$

$$p(h_j)' = \sum_i \left[\left(\frac{p(h_j)p(d_i|h_j)}{\sum_{k=1}^{n} p(h_k)p(d_i|h_k)} \right) p(d_i) \right] \tag{7}$$

Equation 6 is equivalent to Equation 1, where $f(d_i)$ is specified as $p(d_i|h_j)$ and where the updated proportion $p(d_i)'$ is a sum of its updated proportion under each hypothesis, weighted by the proportion of the hypotheses at the current time step. Likewise, Equation 7 is equivalent to Equation 2, where the updated proportion $p(h_j)'$ is a sum of its updated proportion under each data type, weighted by the proportion of each data type in the current population.

2. Example Dynamics of the Model

Figure 1 provides a visualization of the model's behavior for an example set of parameters: an initial distribution over data $d_{init} = p(d_1, d_2, d_3) = \{.2, .2, .6\}$ and hypotheses $h_{init} = p(h_1, h_2, h_3) = \{.3, .4, .3\}$. The fitness values used are those in Equation 5. The top row (*Model A*) shows a population of data adapting to h_{init}, which is held fixed. The data converges to a population distribution of $\{0.56, 0.44, 0\}$, where d_1 constitutes 56% of the population, d_2 constitutes 44% of the population, and d_3 is extinct. The middle row (*Model B*) shows a population of hypotheses adapting to d_{init}, which is held fixed. The hypotheses converge

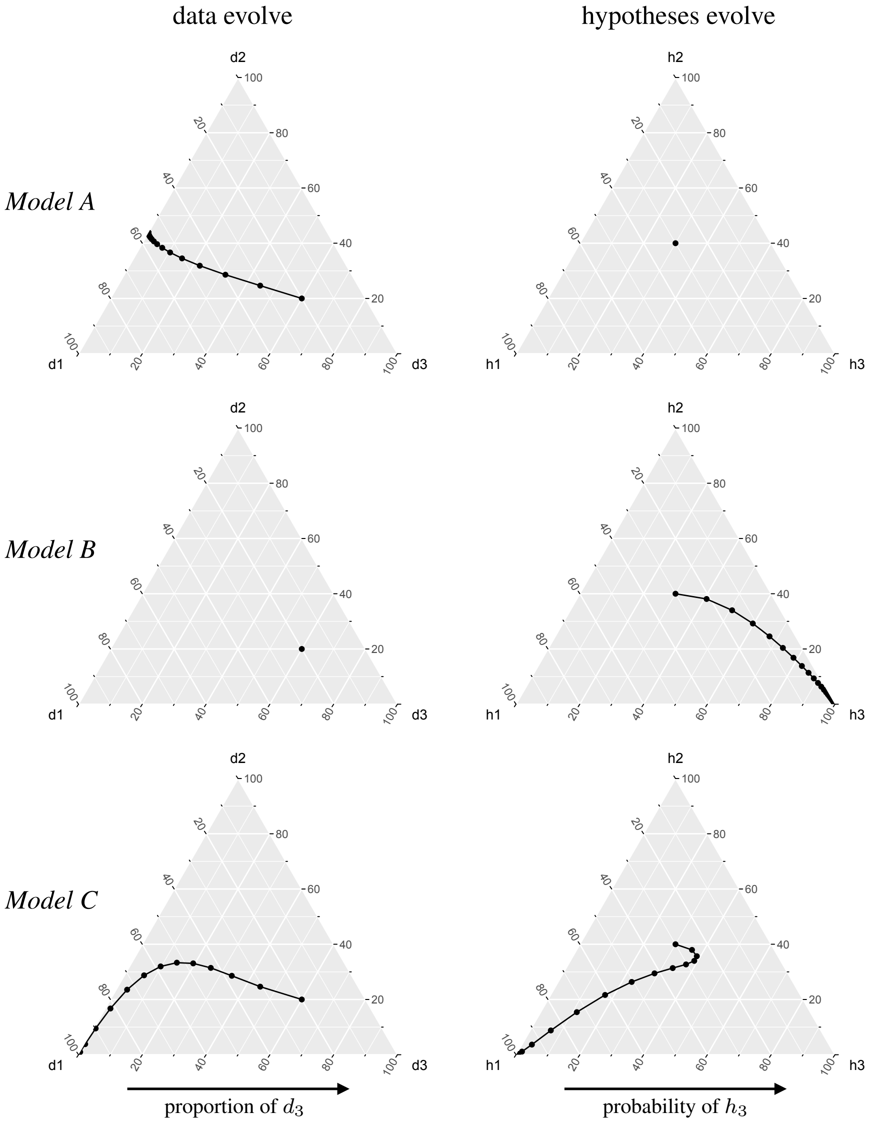

Figure 1. Example behavior of each model on a simplex plot. Axes show the proportion of each variant in the population. Corners indicate the fixation of one variant and are labeled by that variant (ex: a point in the “d1” corner means the d_1 variant comprises 100% of the population). *Model A*: Data replicate while the distribution over hypotheses remains fixed. *Model B*: Hypotheses replicate while the distribution over data remains fixed. *Model C*: Data and hypotheses coevolve, fixating in a different stable state than either variant would have on its own.

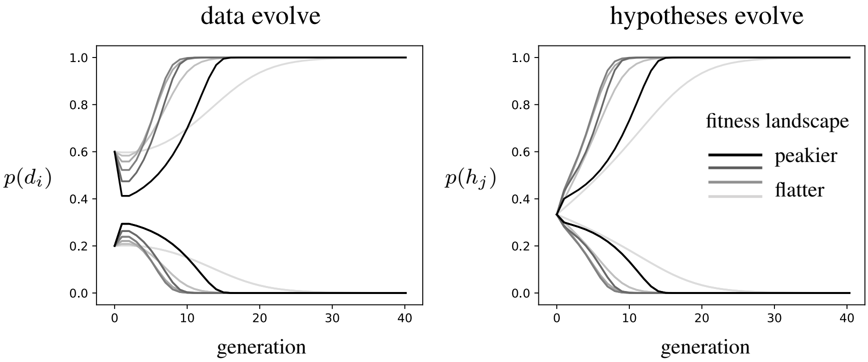

Figure 2. Example evolutionary trajectories from *Model C*, demonstrating how the length of time it takes a population to reach its stable state varies in relation to landscape typography. The y axis shows the proportion of each replicator type, which was initialized at $p(d_1, d_2, d_3) = \{.6, .2, .2\}$ and $p(h_1, h_2, h_3) = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$. Darker colors denote peakier landscapes.

to a population composed only of h_3. The bottom row shows the behavior of Model C, where the population of data (left) and the population of hypotheses (right) coevolve with one another. The data were updated first and therefore, the first step in the trajectory is identical to their first step in *Model A*. However, the first step the hypotheses take is not identical to their first step in *Model B* because the updated distribution over data caused the relative fitness among hypotheses to change, sending them in a slightly different direction. Although their trajectory is now altered, the hypotheses still continue toward h_3 until the population of data becomes sufficiently different to have changed the boundaries of the basin of attraction for h_1. At this point, the hypotheses find themselves in a new basin which converges to a different attractor: $p(h_1, h_2, h_3) = \{1, 0, 0\}$.

In the replicator dynamics equation, convergence rates are an increasing function of fitness strength: as the differential growth rate among replicators becomes larger, the population reaches fixation faster. This applies to the first two models described in this paper, but not to *Model C*. Figure 2 shows six evolutionary trajectories from *Model C* which vary only in the peakiness of their fitness landscapes (where peaky entails strong differential fitness). Flatter fitness landscapes are shown in lighter grey and peakier fitness landscapes are shown in darker grey. On the left, an initial data distribution of $p(d_i) = \{.6, .2, .2\}$ converges to $p(d_i) = \{1, 0, 0\}$. On the right, an initial uniform distribution over hyptheses $p(h_j) = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ converges to $p(h_j) = \{1, 0, 0\}$. Trajectories on the flattest landscape have the longest convergence time, reaching fixation slower than the trajectories on slightly peakier landscapes. However, as the peakiness continues to increase, we see a reversal, with the peakiest landscapes leading back to longer and longer convergence times.

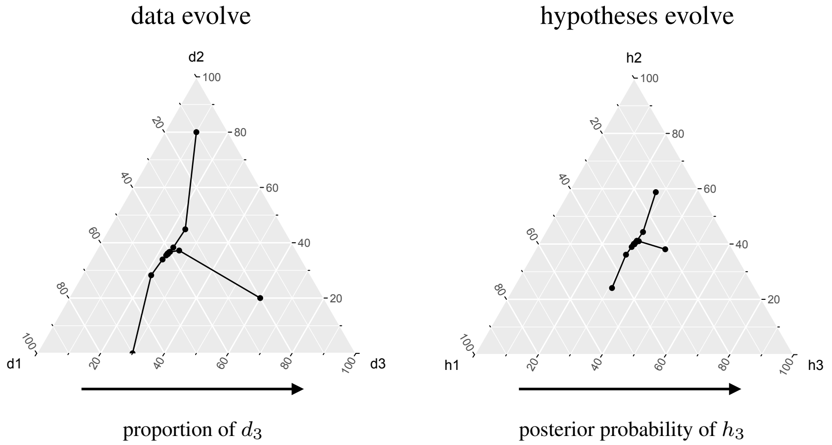

Figure 3. Example behavior of the Griffiths and Kalish (2007) model, for the likelihoods in Equation 5, a prior distribution over hypotheses $p(h_1, h_2, h_3) = \{.3, .4, .3\}$, and three different initial distributions over data. *Right*: In all cases, the posterior distribution over hypotheses converges to the prior. *Left*: In all cases, the distribution over data converges to $p(d_1, d_2, d_3) = \{.41, .36, .23\}$, which is the distribution over data that is most likely under this particular prior.

3. Comparison to Existing Models of Bayesian Cultural Evolution

In the first model of Bayesian cultural evolution, Griffiths and Kalish (2007) explored a population of Bayesian agents that shared the same, fixed prior over hypotheses. These agents receive data from one another, compute a posterior distribution, sample a hypothesis from the posterior, and then sample data from that hypothesis. Over time, the posterior distribution that agents compute converges to the prior distribution and the data converges to the integrated probability of the data under the prior distribution of hypotheses (which is the distribution over data that uniquely leads to the computed posterior being equal to the prior). Figure 3 shows an implementation of the Griffiths and Kalish model for the likelihoods in Equation 5. Three different initial distributions over data are shown. Each initial condition results in a trajectory that converges to the same point. In this model, agents produce data *de novo* each generation: in no sense are they *copying* data from the previous generation. This dynamic gradually erodes the population's information about its start state, freeing the system to converge to a unique stable state: the prior distribution over hypotheses and its corresponding data distribution. (The same behavior holds for their MAP model, with the exception that the hypotheses converge to a point near the prior.)

In another model of Bayesian cultural evolution, Beppu and Griffiths (2009) implemented a population of agents that receive data not from one another, but from a fixed source in the world. They compute a posterior distribution and then communicate complete information about the posterior to the next generation of agents, who adopt this posterior distribution as their own prior distribution. This model is formally equivalent to a Bayesian updating model of learning in a single individual and thus, is equivalent to *Model B* (see Figure 1 for example dynamics of the Beppu and Griffiths model, where hypotheses update in response to a fixed source of data).

The main difference between these two existing models and the coevolutionary model lies in their number of unique stable states and the effect of initial conditions. The models described above each converge to one stable state regardless of their initial conditions. In the coevolutionary model, there are a minimum of n stable states representing the fixation of each variant x in the set $\{x_1, x_2, ...x_n\}$. This holds on any given fitness landscape. For a certain class of fitness landscapes, where the fitness matrix is symmetrical about its diagonal, there is an additional stable state in the center of the simplex when both $d_i = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ and $h_j = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$.

In the Griffiths and Kalish model, the evolving population gradually obtains information about the prior distribution and the speed of the information gain is determined by the likelihood structure: flatter likelihoods yield faster convergence rates (and uniform likelihoods yield immediate convergence to the prior). In the Beppu and Griffiths model, the evolving population gradually obtains information about the world (via the data received) and the speed of information gain is also determined by the likelihood structure. However, the direction of the relationship here is reversed: flatter likelihoods yield slower convergence rates (and uniform likelihoods result in no change at all). In the model where data and hypotheses co-evolve, it is less clear what the populations are gaining information about. Populations appear to be gaining information about the absorbing state of the basin of attraction they are currently in. However, as the basins of attraction change, populations switch to gain information about new absorbing states, overwriting what they previously "learned". In this model, the effects of fitness typography pattern with the Beppu and Griffiths model: flatter likelihoods yield slower convergence rates and uniform likelihoods result in no change at all.

4. Discussion and Extensions

This paper introduced a new model for exploring the coevolution of data and hypotheses in cultural evolution. I believe that Bayesian models of cultural evolution provide one possible way of formalizing Sperber (1996)'s concepts of public representations (as data) and private representations (as hypotheses). Analysis of these models could also provide quantitative insight into some of the interesting dynamics described in Cultural Attraction Theory (Sperber, 1985, 1996). Early computational models of language evolution contain many examples of private replicators (as grammars in the agents' minds) and public replicators (as utterances or strings produced from the grammar) changing in consort with one another and converging to a rich diversity of semi-stable states (e.g. Kirby, 2001; Brighton, 2002). The current model was developed in order to broaden the range of evolutionary dynamics that Bayesian models of cultural evolution can capture. The next steps include fitting this model to experimental and simulation data from dynamic-rich examples of cultural evolution such as those found in social learning theory, cultural attraction theory, and the evolution of language. Various extensions to this model can be made along these lines. For example, part of the data distribution can be anchored in an unchanging world, or part of the hypotheses distribution can be anchored in an unchanging inductive bias. Furthermore, questions about causal primacy in cultural evolution could be addressed using this model. Do private replicators tend to drive the evolution of public replicators, or vice a versa? It is quite possible that certain classes of asymmetrical fitness landscapes may create dynamics where the hypotheses steer the evolution of the data and other classes of asymmetry may cause the data to steer the evolution of the hypotheses. Knowledge of this relationship could prove useful to researchers in cultural evolution who are working in domains where the likelihood relationships between public and private representations are known.

Acknowledgements

This research was supported by the Omidyar Fellowship and the Santa Fe Institute. Special thanks to Nihat Ay, Michael Lachmann, Thomas Morgan, and the postdocs at SFI for their feedback on earlier drafts of this project.

References

Beppu, A., & Griffiths, T. (2009). Iterated learning and the cultural ratchet. In *Proceedings of the cognitive science society* (Vol. 31).

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441–480.

Harper, M. (2009). The replicator equation as an inference dynamic. *arXiv preprint arXiv:0911.1763*.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *Evolutionary Computation, IEEE Transactions on*, *5*(2), 102–110.

Shalizi, C. R., et al.. (2009). Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, *3*, 1039–1074.

Sperber, D. (1985). *On anthropological knowledge.* Cambridge University Press.

Sperber, D. (1996). *Explaining culture: A naturalistic approach* (Vol. 323). Blackwell Oxford.