A generative theory of similarity: supplementary material

Charles Kemp, Aaron Bernstein & Joshua B. Tenenbaum
{ckemp, aaronber, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Kemp et al. [1] argue that two objects are similar to the extent that they appear to have been produced by the same generative process. Here we show the stimuli used in our experiment (Figures 1 and 2) and derive the results mentioned in our paper.

Featural Model

Suppose that s^1 and s^2 are two objects represented as binary feature vectors. Let n be the number of features possessed by one or both of the objects, and let the domain D be the set of all n-place binary vectors. A generative process over D is specified by a n-place vector θ, where θ_i is the probability that an object has value 1 on feature i. We place independent beta priors on each θ_i:

$$\begin{aligned}\theta_i &\sim \text{Beta}(\alpha, \beta)\\ s_i &\sim \text{Binomial}(\theta_i),\end{aligned}$$

where s_i is the ith feature value for object s, α and β are hyperparameters and $\text{Beta}(\cdot,\cdot)$ is the beta function. Our generative theory states that

$$\text{sim}(s^1, s^2) = \frac{\int P(s^1, s^2|\theta)p(\theta)d\theta}{\int P(s^1|\theta)p(\theta)d\theta \int P(s^2|\theta)p(\theta)d\theta}. \quad (1)$$

Consider first the numerator.

$$\begin{aligned}\int P(s^1, s^2|\theta)p(\theta)d\theta &= \int P(s^1_1, s^2_1|\theta_1)\ldots P(s^1_n, s^2_n|\theta_n)d\theta_1\ldots d\theta_n\\ &= \prod_i \int P(s^1_i, s^2_i|\theta_i)d\theta_i\end{aligned}$$

Now

$$\int P(s^1_i, s^2_i|\theta_i)d\theta_i = \begin{cases}\frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} & \text{if } s^1_i = s^2_i = 1,\\ \frac{B(\alpha+1,\beta+1)}{B(\alpha,\beta)} = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} & \text{otherwise}\end{cases}$$

where $B(\cdot,\cdot)$ is the Beta function. Thus

$$\int P(s^1, s^2|\theta)p(\theta)d\theta = \left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{y^{12}} \left(\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{n-y^{12}}$$

where y^{12} is the number of features shared by both objects.

The terms in the denominator of Equation 1 can be computed similarly:

$$\int P(s^j|\theta)p(\theta)d\theta = \left(\frac{\alpha}{\alpha+\beta}\right)^{y^j} \left(\frac{\beta}{\alpha+\beta}\right)^{n-y^j}$$

where y^j is the number of features possessed by s^j.

Then

$$\text{sim}(s^1, s^2) = \left(\frac{\alpha+\beta}{\alpha+\beta+1}\right)^n \left(\frac{\alpha+1}{\alpha}\right)^{y^{12}}$$

where we use the fact that $y^1 + y^2 - y^{12} = n$. Since we are interested only in the rank order of the pairwise similarities, we can take logarithms of both sides:

$$\begin{aligned}\log(\text{sim}(s^1, s^2)) &= n \log\left(\frac{\alpha+\beta}{\alpha+\beta+1}\right) + y^{12} \log\left(\frac{\alpha+1}{\alpha}\right) \\ &= k_1 y^{12} - k_2(n - y^{12}) \\ &= k_1 |s^1 \cup s^2| - k_2(|s^1 - s^2| - |s^2 - s^1|)\end{aligned}$$

where $k_1 = \log\left(\frac{\alpha+1}{\alpha}\right) - \log\left(\frac{\alpha+\beta+1}{\alpha+\beta}\right)$, $k_2 = \log\left(\frac{\alpha+\beta+1}{\alpha+\beta}\right)$, $|s^1 \cup s^2|$ is the number of features shared by both objects, and $|s^i - s^j|$ is the number of features possessed by s^i but not s^j.

Spatial Model

Suppose that the domain D is a multidimensional space with dimension n. Consider a Gaussian generative process determined by a a mean μ and covariance matrix Σ. For simplicity, we place a uniform (hence improper) prior over μ:

$$\begin{aligned}\mu &\sim \text{Uniform}(\mathcal{R}^n) \\ s &\sim \text{Normal}(\mu, \Sigma),\end{aligned}$$

where μ and s are random variables with n dimensions, and Σ is a constant n by n matrix.

The numerator of Equation 1 becomes

$$\int P(s^1, s^2|\theta)p(\theta)d\theta \propto \int \exp((s^1 - \mu)^T \Sigma^{-1}(s^1 - \mu) - (s^2 - \mu)^T \Sigma^{-1}(s^2 - \mu))d\mu$$

Completing the square we see that

$$\begin{aligned}\int P(s^1, s^2|\theta)p(\theta)d\theta &\propto \exp\left(-\frac{1}{2}(s^1 - s^2)^T \Sigma^{-1}(s^1 - s^2)\right) \int \exp\left(-2\left(\frac{s^1 + s^2}{2} - \mu\right)^T \Sigma^{-1}\left(\frac{s^1 + s^2}{2} - \mu\right)\right) d\mu \\ &\propto \exp\left(-\frac{1}{2}(s^1 - s^2)^T \Sigma^{-1}(s^1 - s^2)\right)\end{aligned}$$

Each term in the denominator of Equation 1 takes the form

$$\int P(s^j|\theta)p(\theta)d\theta \propto \int \exp\left(-\frac{1}{2}(s^j - \mu)^T \Sigma^{-1}(s^j - \mu)\right) d\mu \propto 1$$

Thus

$$\begin{aligned}\text{sim}(s^1, s^2) &\propto \exp\left(-\frac{1}{2}(s^1 - s^2)^T \Sigma^{-1}(s^1 - s^2)\right) \\ \therefore \log(\text{sim}(s^1, s^2)) &= -\frac{1}{2}(s^1 - s^2)^T \Sigma^{-1}(s^1 - s^2)\end{aligned}$$

where the equality holds up to addition of a constant. We therefore see that similarity is inversely related to the Mahalanobis distance between s^1 and s^2.

Transformational Model

Suppose we are given a set of objects D and a set of transformations T. We assume that every transformation is reversible — if there is a transformation mapping s^1 into s^2, there must also be a transformation mapping s^2 into s^1. We use a generative process over D specified by a prototype $\theta \in D$ chosen from a uniform (and possibly improper) distribution over D. To generate an object s from the process, we sample a transformation count k from an exponential distribution, choose k transformations at random from T, then apply them to the prototype:

$$\begin{aligned}\theta &\sim \text{Uniform(D)}\\ k &\sim \text{Exponential}(\lambda)\\ t_i &\sim \text{Uniform}(T)\\ s &= t_k \cdot t_{k-1} \ldots \cdot t_1(\theta)\end{aligned}$$

where λ is a constant, and t_i is the ith transformation chosen.

We use an approximation to Equation 1:

$$\text{sim}(s^1, s^2) = \frac{P(s^1|\theta^{12})P(s^2|\theta^{12})p(\theta^{12})}{P(s^1|\theta^1)p(\theta^1)P(s^2|\theta^2)p(\theta^2)}, \tag{2}$$

where $\theta^{12} = \text{argmax}_\theta P(s^1, s^2|\theta)$, $\theta^1 = \text{argmax}_\theta P(s^1|\theta)$, and $\theta^2 = \text{argmax}_\theta P(s^2|\theta)$. We further approximate each term in Equation 2 using MAP settings of k and t:

$$P(s^j|\theta^j) = \int P(s^j|\theta^j, k, t)P(k, t)dkdt \approx P(s^j|\theta^j, \hat{k}, \hat{t})P(\hat{k}, \hat{t})$$

where θ^j, $\hat{k}$ and $\hat{t}$ are set to values that maximize $P(\theta, k, t|s^j)$. Since $\theta^j = s^j$ and $\hat{k} = 0$, $P(s^j|\theta^j, \hat{k}, \hat{t})P(\hat{k}, \hat{t}) = P(k = 0) = \lambda \propto 1$. Similarly, we use

$$P(s^1|\theta^{12})P(s^2|\theta^{12}) \approx P(s^1|\theta^{12}, \hat{k^1}, \hat{t^1})P(s^2|\theta^{12}, \hat{k^2}, \hat{t^2})P(\hat{k^1}, \hat{k^2}, \hat{t^1}, \hat{t^2}).$$

$\hat{k^1} + \hat{k^2}$ is the length of the shortest path joining s^1 and s^2 where each step along the path is a transformation from T. Since the transformations are reversible, $P(\theta|s^1, s^2)$ is the same for any θ along this path, and we can set θ^{12} to any of these values. We now see that

$$\begin{aligned}\text{sim}(s^1, s^2) \approx P(\hat{k^1}, \hat{k^2}, \hat{t^1}, \hat{t^2}) &\propto \frac{1}{|T|^{\hat{k^1}+\hat{k^2}}} \exp(-\lambda(\hat{k^1} + \hat{k^2}))\\ \therefore \log(\text{sim}(s^1, s^2)) &= -(\hat{k^1} + \hat{k^2})(\log(|T|) + \lambda) \propto -(\hat{k^1} + \hat{k^2})\end{aligned}$$

Thus the similarity of s^1 and s^2 is inversely related to the transformation distance between these objects.

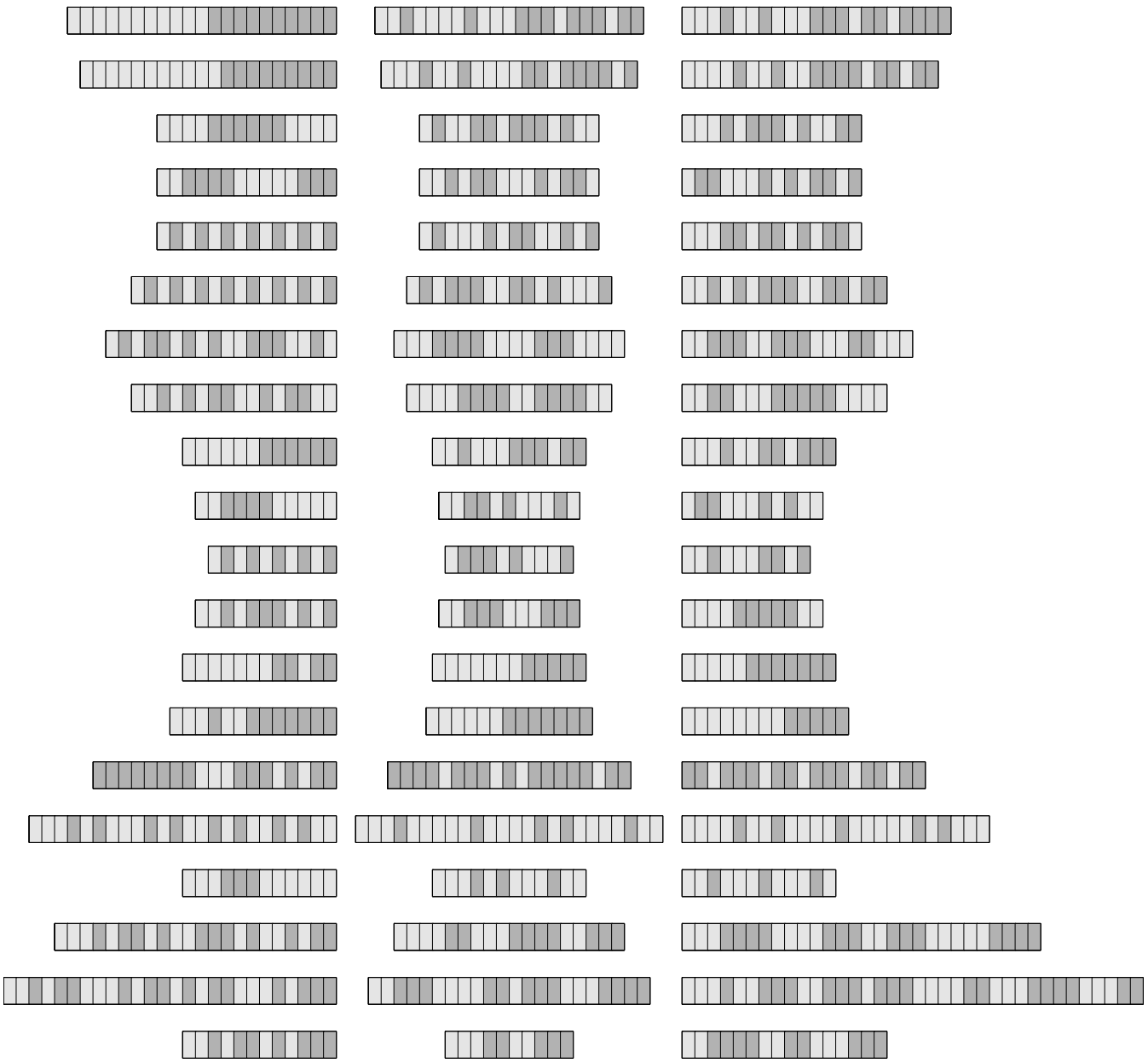

Figure 1: The 20 binary triads. Prototype strings are in the center, transformation strings are on the left, and HMM strings are on the right.

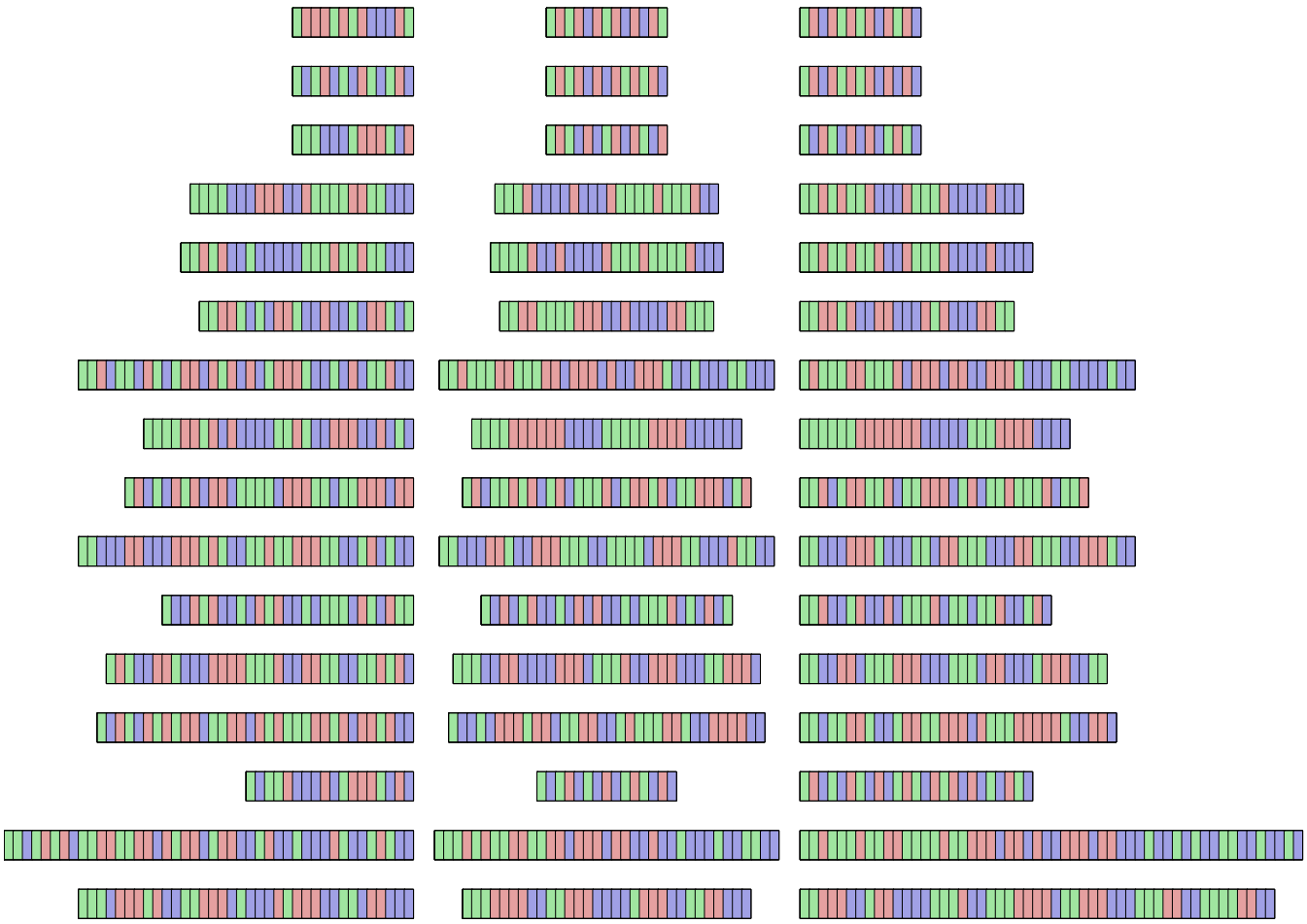

Figure 2: The 16 ternary triads. Prototype strings are in the center, transformation strings are on the left, and HMM strings are on the right.

References

[1] Kemp, C., Bernstein, A., and Tenenbaum, J. B. (Submitted). A generative theory of similarity.