Conjugate Implementation of the Jern & Kemp (2013) Model with Representativeness

The representativeness model is very similar to the Jern & Kemp (2013) hierarchical Bayesian model. The essential difference here is the ultimate formulation of response probabilities. Where in the original hierarchical Bayesian model the exemplars are drawn from Gaussian distributions, in the representativeness model each exemplar is drawn in proportion to the exemplar's representativeness. The representativeness of an exemplar is defined by Tenenbaum and Griffiths (2012) as the relative evidence that is provided by the exemplar x for a given hypothesis h compared to all other hypotheses h':

$$R(x, h) = \log \frac{p(x|h)}{\sum_{h' \neq h} p(x|h')p(h')} \tag{1}$$

For simplicity, and consistency with the original hierarchical Bayesian model, we define h_C as a multivariate normal distribution for a particular category C parameterised by μ_C and Σ_C. For completeness, this document will repeat some information from the hierarchical Bayesian model document and describee how we compute these variables.

Computing μ_C

Assuming (μ_C, Σ_C) are Normal-Inverse-Wishart distributed (unknown mean, unknown variance):

$$\mu_C = \frac{\kappa\mu_0 + n_C\bar{x_C}}{\kappa + n_C} \tag{2}$$

where:

- μ_0 is the prior mean along p dimensions. Here we set it to the middle of the space.
- κ is a scalar hyper-parameter, roughly weighting the importance of μ_0. κ must be greater than zero.
- n_C is the number of observations in x_C
- $\bar{x_C}$ is the sample mean along p dimensions

In the case of a populated class, μ_C ends up lying somewhere between μ_0 and $\bar{x_C}$, depending on κ_0 and n_C. In the case of an empty class, $n_C = 0$, Equation 2 reduces to $\mu_C = \mu_0$. Because we set μ_0 to the center of the space, this outcome is the same as if we had integrated over all possible μ_C.

In practice, if $n_C = 0$, the model picks a stimulus at random from all candidates (uniform probabilities).

Computing Σ_D

Unlike μ_C, Σ_C cannot be computed considering only the members of category y. Instead, Σ_C is influenced both by the distribution of x_C and by members of other categories through Σ_D.

Σ_D is inferred based on the observed (empirical) category covariances C_y. We assume these covariances to be Wishart-distributed, and so Σ_D can be computed as:

$$\Sigma_D = \Sigma_0 + \sum_C C_C \tag{3}$$

Σ_0 is a d-by-d prior covariance matrix. We use a d-dimensional identity matrix I_d multiplied element-wise against a free parameter, λ, controlling the amount of variance assumed by the prior:

$$\Sigma_0 = \lambda I_d \tag{4}$$

Thus, categories are assumed to have some degree of variance along each feature (specified by λ), but not are assumed to possess feature-feature correlations. Differences in the assumed variance among the features, similarly to weighting in an exemplar model, can be implemented through a small change to the equation. Specifically, the variance assumed of dimension k is given by:

$$\Sigma_{0k} = \lambda w_k d \tag{5}$$

where d is the number of dimensions, and w_k ($0 \leq w_k \geq 1$, $\sum_k w_k = 1$) indicates the dimension's relative share of the total assumed variability. Under this system, evenly distributed weights result in uniformly assumed variances, equal to λ.

Computing Σ_C

Assuming (μ_C, Σ_C) are Normal-Inverse-Wishart distributed, Σ_C can be computed as:

$$\Sigma_C = [\Sigma_D \nu + C_C + \frac{\kappa n_C}{\kappa + n_C}(\bar{x_C} - \mu_C)(\bar{x_C} - \mu_C)^T](\nu + n_C)^{-1} \tag{6}$$

κ, $\bar{x_C}$, C_C, n_C, μ_0, are the same values as described above. ν is an additional free parameter, weighting the importance of Σ_D. ν must be greater than $d-1$. When x_b is empty, Equation 6 reduces to $\Sigma_C = \Sigma_D$.

Computing response probabilities $p(y|x_C)$

As mentioned earlier, this is the point where the representativeness model diverges from the hierarchical Bayesian model. Specifically, the probability of generating exemplar x is proportional to its representativeness:

$$p(x) \propto R(x, h), \tag{7}$$

where h is a multivariate normal distribution parameterised by μ_C and Σ_C.

In practice, $p(x)$ is computed by first obtaining the representativeness of every possible generation candidate x_i. The end probability is a normalization of these values:

$$p(x) = \frac{\exp(\theta \cdot R(x, h))}{\sum_i \exp(\theta \cdot R(x_i, h_i))} \tag{8}$$

where θ is a response determinism parameter.

Description of free parameters

- κ. Scalar, $\kappa > 0$. Weights the importance of μ_0 in inferring category μ_C.
- λ. Scalar, $\lambda > 0$. Sets the assumed variance in the domain prior, Σ_0.
- ν. Scalar, $\nu > d - 1$. Weights the importance of Σ_D in inferring the domain Σ_C.
- θ. Scalar, $\theta > 0$. Response determinism parameter.