Statistical Inference in a Bayesian Model for Category Learning

Daniel J. Navarro
School of Psychology, University of Adelaide
daniel.navarro@adelaide.edu.au

February 4, 2006

NOTE: This is a draft, made available to allow reviewers for the corresponding *Cognitive Science* conference paper to have access to the inference scheme for purposes of the review. While the algorithm is accurately described, the note is being rewritten to improve clarity. A final version will be made public shortly.

Introduction

I describe a method for calculating the predictions made by a generative Bayesian model that describes human categories in terms of an infinite number of kinds (Navarro, submitted). For ease of exposition, the development will be limited to stimuli that can be represented in a separable metric space, but the analysis can be extended to more general representational structures in a straightforward fashion. Consequential regions are assumed to be hyper-rectangular in shape. Since space is separable, the extension of the region along different dimensions is chosen independently, in order to produce a city block distance metric. A category is assumed to entail a countably infinite number of consequences, only some finite number of which will be relevant to the learning task at hand. The model can be written:

$$
\begin{array}{rcll}
x_{ij} & | & m_{kj}, s_{kj}, r_i = k & \sim \text{Uniform}(m_{kj}, s_{kj}) \\
r_i & | & w_c, z_i = c & \sim \text{Discrete}(w_c) \\
s_{kj} & | & \lambda_j, \xi & \sim \text{Gamma}(\xi, \lambda_j) \\
m_{kj} & | & \mu_j, \tau_j & \sim \text{Normal}(\mu, 1/\tau) \\
w_c & | & \alpha & \sim \text{Stick}(\alpha) \\
\lambda_j & | & \beta_1, \beta_2 & \sim \text{Gamma}(\beta_1, \beta_2) \\
\mu_j & | & \mu_0, \tau_0, \tau_j & \sim \text{Normal}(\mu_0, 1/\tau_j\tau_0) \\
\tau_j & | & \phi_1, \phi_2 & \sim \text{Gamma}(\phi_1, \phi_2) \\
z_i & | & \zeta & \sim \text{Discrete}(\zeta) \\
\zeta & | & \eta & \sim \text{Dirichlet}(\eta)
\end{array}
$$

In these expressions, i indexes a stimulus, j indexes a dimension, k indexes a region, and c indexes a category. This model is described in detail by Navarro (submitted),

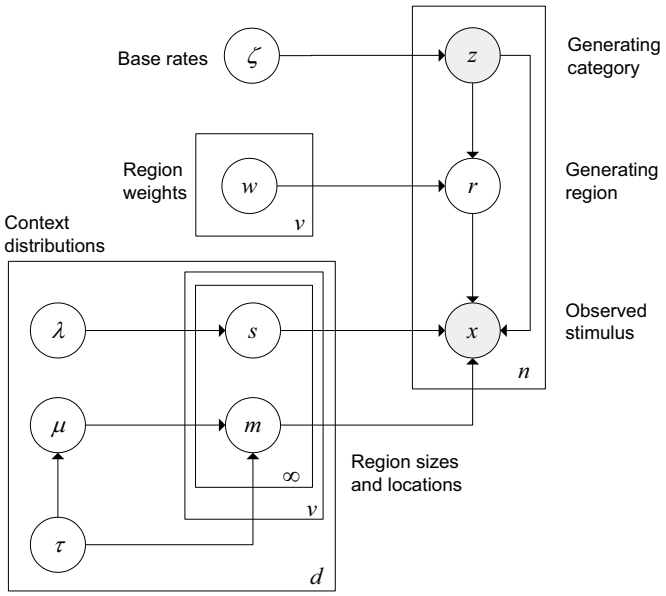

Figure 1: Graphical representation of the Bayesian categorization model. Shaded circles denote observed variables, white rounded circles refer to fixed values, and white circles indicate latent, unobserved variables. Arrows indicate dependencies between variables, while plates enclose a set of independent replications. In this figure, n is the number of stimuli, v is the number of categories, and d is the number of dimensions. Since this paper is concerned with supervised learning, the category variables z are shaded. Naturally, it is only the previous category labels that are observed, not a novel one.

and illustrated graphically in Figure 1. In general, most of the "parameter" values are fixed *a priori*. I treat the Gamma hyper-priors as standard exponentials, so β_1, β_2, ϕ_1 and ϕ_2 are all fixed at 1. I set a diffuse prior over the location parameters, with $\mu_0 = 0$ and $\tau_0 = .001$. Unless there is a reason for people to have an *a priori* bias for one category over another, the prior over ζ is uniform, so I fix $\eta = 1$. The only psychologically important values are ξ and α, which should be treated as genuine parameters.

Structure of the Inference Algorithm

On trial t in a supervised category learning experiment, people are shown a v-dimensional stimulus $x_t = (x_{t1}, \ldots, x_{tv})$, and asked to predict the corresponding category label z_t. Feedback is provided, so the category labels for earlier trials $\boldsymbol{z}_{-t} = (z_1, \ldots z_{t-1})$ are available to the learner (hence the shading in the graphical model). Inference takes place by finding $p(z_t|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta})$, the conditional distribution for the category label, given the preceding experimental history. The parameter vector in this model consists of the variables describing the prior $\boldsymbol{\theta} = (\xi, \alpha, \beta_1, \beta_2, \phi_1, \phi_2, \mu_0, \tau_0, \eta)$. The model has a number of latent, hidden variables

$\boldsymbol{h} = (\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{m}, \zeta, w, \tau, \mu, \lambda)$ that need to be integrated out in order to calculate $p(z_t|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta})$. However, this integral is intractable so we turn to approximate inference methods. Our approach is firstly to rewrite $p(z_t|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta})$ as a marginalization over the joint posterior distribution over the collection of variables $z_t, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{m}$:

$$p(z_t|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta}) = \int p(z_t, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{m}|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta}) \; d(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{m})$$

Written in this form, it will be straightforward to develop numerical methods for approximating the category membership probability. Given a large enough set of $\mathcal{L}$ samples $z_t^\ell, \boldsymbol{r}^\ell, \boldsymbol{m}^\ell, \boldsymbol{s}^\ell$ such that,

$$z_t^\ell, \boldsymbol{r}^\ell, \boldsymbol{m}^\ell, \boldsymbol{s}^\ell \sim p(z_t, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{m}|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta})$$

we may adopt the approximation:

$$p(z_t|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta}) \approx \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \delta(z_t - z_t^\ell).$$

where $\delta(\cdot)$ is an indicator function that takes on value 1 when the argument is 0, and is 0 otherwise. Therefore, the category assignment probabilities may be found by drawing samples from the joint posterior $p(z_t, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{m}|\boldsymbol{z}_{-t}, \boldsymbol{x}, \boldsymbol{\theta})$ integrating out the values of ζ, w, τ, μ and λ in the process. I now discuss a Markov chain Monte Carlo method (e.g., Gilks, Richardson & Spiegelhalter, 1995) for sampling z_t^ℓ, $\boldsymbol{r}^\ell$, $\boldsymbol{m}^\ell$, and $\boldsymbol{s}^\ell$ from the joint posterior distribution.

The basic approach is a combination of Gibbs and Metropolis-Hastings sampling, in which we repeatedly cycle through each of the variables, resampling them from their conditional posterior distribution given the sampled values for all the other variables (i.e., Gibbs sampling). However, when the conditional posterior is intractable, we substitute a Metropolis-Hastings scheme that leaves the stationary distribution invariant. The relevant conditional posteriors are

$$\begin{array}{rcl} m_{kj} & | & \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}_{-kj}, \boldsymbol{r}, \boldsymbol{\theta} \\ s_{kj} & | & \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}_{-kj}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{\theta} \\ r_i & | & \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{r}_{-i}, \boldsymbol{\theta}, i \neq t \\ r_t, z_t & | & \boldsymbol{x}, \boldsymbol{z}_{-t}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{r}_{-t}, \boldsymbol{\theta}, i = t \end{array}$$

As before, i indexes a stimulus, j indexes a dimension and k indexes a region. When $t = 1$, all variables are initialized randomly, and then the sampler is run until the the values of z_t, $\boldsymbol{r}$, $\boldsymbol{s}$ and $\boldsymbol{m}$ converge to samples from the joint posterior. After the sampler converges, the values of these variables are recorded to produce the draws for z_t^ℓ, $\boldsymbol{r}^\ell$, $\boldsymbol{m}^\ell$, and $\boldsymbol{s}^\ell$, with a sufficient lag between successive draws to minimize the correlation between these draws. For subsequent trials (i.e., $t > 1$) the same procedure is adopted, but the initial values are set by copying the state of the sampler from the previous trial, assigning values to new variables randomly. This allows faster convergence for subsequent trials. In what follows, I describe procedures for drawing from the conditional distributions.

Resampling Means

In the first part of the sampler, we sample the mean m_{kj} along dimension j for region k from its conditional posterior $p(m_{kj}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}_{-kj}, \boldsymbol{r}, \boldsymbol{\theta})$. If the region memberships $\boldsymbol{r}$ and sizes $\boldsymbol{s}$ are all known along with all the category memberships $\boldsymbol{z}$, and m_{kj} is the only unknown location variable, then by applying Bayes' theorem and dropping variables that do not affect the relevant probabilities, we obtain:

$$\begin{aligned} & p(m_{kj} \,|\, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}_{-kj}, \boldsymbol{r}, \boldsymbol{\theta}) \\ & \quad \propto \; p(\boldsymbol{x}_k \,|\, \boldsymbol{x}_{-k}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{\theta}) \; p(m_{kj} \,|\, \boldsymbol{x}_{-k}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}_{-kj}, \boldsymbol{r}, \boldsymbol{\theta}) \\ & \quad = \; p(\boldsymbol{x}_k \,|\, m_k, s_k) \; p(m_{kj} \,|\, \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2), \end{aligned}$$

where $\boldsymbol{x}_k$ denotes the set of stimuli that are currently assigned to region k, and $\boldsymbol{x}_{-k}$ refers to the remaining stimuli. This expression applies equally to all regions, irrespective of the category of which they form a part. Since every point inside region is equally likely, the likelihood term is given by

$$p(\boldsymbol{x}_k \,|\, m_k, s_k) = \prod_{i \in k} p(x_i \,|\, m_k, s_k),$$

where the sum is taken over all the stimuli that are assigned to region k, and

$$p(x_i \,|\, m_k, s_k) = \begin{cases} \prod_j {s_{kj}}^{-1} & \text{if } x_i \in r_k \\ 0 & \text{otherwise} \end{cases} .$$

In this expression, $x_i \in r_k$ is true if and only if x_i falls inside the part of the space corresponding to region r_k. The second term in the expression for the conditional posterior is a little more complex, since

$$\begin{aligned} & p(m_{kj} \,|\, \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2) \\ & \quad = \int p(m_{kj} \,|\, \mu_j, \tau_j) \; p(\mu_j, \tau_j \,|\, \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2) \; d(\mu_j, \tau_j). \end{aligned}$$

Since the normal-gamma prior over μ_j and τ_j is conjugate to the normal distribution for locations that they describe, the posterior distribution $p(\mu_j, \tau_j \,|\, \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2)$ is also normal-gamma, with updated parameters.

$$\begin{array}{rcll} \mu_j & | & \mu_0, \tau_0, \tau_j, \boldsymbol{m}_{-kj} & \sim \; \text{Normal}(\mu_n, 1/\tau_j\tau_n) \\ \tau_j & | & \phi_1, \phi_2, \boldsymbol{m}_{-kj} & \sim \; \text{Gamma}(\phi_{1n}, \phi_{2n}) \end{array}$$

The updated parameters are given as follows. Firstly, we define

$$\bar{m} = {n_m}^{-1} \sum_{q \neq k} m_{jq}$$

to be the average value of all previously observed location parameters on dimension j, and note that since the parameters μ_j and τ_j are shared by all categories, this sum is taken over the location parameters for regions that belong to all categories, and n_m denotes the number of terms in this sum. Along the same lines, define

$$s^2 = (n-1)^{-1} \sum_{q \neq k} (m_{jq} - \bar{m})^2$$

to be the sample variance of these terms. Then if we let $\nu_0 = 2\phi_1$, and $\sigma_0^2 = (\phi_1\phi_2)^{-1}$, the updated parameters are described by defining

$$\begin{aligned}
\tau_n &= \tau_0 + n_m \\
\nu_n &= \nu_0 + n_m \\
\mu_n &= (\tau_0\mu_0 + n_m\bar{m})/\tau_n \\
\sigma_n^2 &= (\nu_0\sigma_0^2 + (n_m\tau_0/\tau_n)(\bar{m} - \mu_0)^2 + (n_m - 1)s^2)/\nu_n
\end{aligned}$$

Given these expressions, the updated parameters are $\phi_{1n} = \nu_n/2$ and $\phi_{2n} = 1/(\phi_{1n}\sigma_n^2)$. The posterior predictive distribution $p(m_{kj} \mid \boldsymbol{m}_{-kj}, \boldsymbol{\theta})$ is found by marginalizing out the posterior distribution. After a scale and location shift, the predictive distribution follows a Student's t distribution with ν_n degrees of freedom:

$$m_{kj} \mid \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2 \;\sim\; \sqrt{(\tau_n + 1)\sigma_n^2/\tau_n}\,\text{Student}(\nu_n) + \mu_n.$$

This gives us an expression for $p(m_{kj} \mid \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2)$, which after a little algebra reduces to,

$$p(m_{kj} \mid \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2) = \frac{\Gamma\left(\phi_{1n} + \frac{1}{2}\right)}{\sqrt{2\pi\phi_{1n}}\Gamma\left(\phi_{1n}\right)}\left(1 + \frac{\tau_n\left(m_k - \mu_n\right)^2}{2\phi_{1n}\sigma_n^2(\tau_n + 1)}\right)^{-\phi_{1n}-\frac{1}{2}}$$

Since we now have expressions for $p(\boldsymbol{x}_k \mid m_k, s_k)$ and $p(m_{kj} \mid \boldsymbol{m}_{-kj}, \mu_0, \tau_0, \phi_1, \phi_2)$, we can now propose a posterior sampling scheme for m_{kj}. Although there are a great many Metropolis-Hastings possibilities, it is straightforward to construct a pseudo-Gibbs sampler by developing a sufficiently-accurate numerical approximation to the conditional posterior. Since this distribution for m_{kj} is univariate and the probabilities are easily calculated (up to the normalizing constant), a simple approach is to break the continuous distribution into a large number of discrete intervals and assume the posterior likelihood is constant over an interval. An interval is sampled from the discrete distribution, and then a point in the interval randomly sampled.

Resampling Sizes

The second part of the sampler involves drawing from the conditional posterior over the size variables, $p(s_{kj}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}_{-kj}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{\theta})$. As before, we apply Bayes' theorem and drop irrelevant variables:

$$\begin{aligned}
&p(s_{kj} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}_{-kj}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{\theta}) \\
&\quad\propto\; p(\boldsymbol{x}_k \mid \boldsymbol{x}_{-k}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{\theta})\; p(s_{kj} \mid \boldsymbol{x}_{-k}, \boldsymbol{z}, \boldsymbol{s}_{-kj}, \boldsymbol{m}, \boldsymbol{r}, \boldsymbol{\theta}) \\
&\quad=\; p(\boldsymbol{x}_k \mid m_k, s_k)\; p(s_{kj} \mid \boldsymbol{s}_{-kj}, \beta_1, \beta_2, \xi).
\end{aligned}$$

The likelihood function remains the same as described for the location variables. The second term involves integrating over λ_j, and is given

$$p(s_{kj} \mid \boldsymbol{s}_{-kj}, \beta_1, \beta_2, \xi) = \int p(s_{kj}|\lambda_j, \xi)\; p(\lambda_j|\boldsymbol{s}_{-kj}, \beta_1, \beta_2)\; d\lambda_j$$

As before, note that $\boldsymbol{s}_{-kj}$ includes size variables for all categories, since observations from all categories can influence the posterior distribution over λ_j. The integral

can be solved by noting that the gamma prior on λ_j is conjugate to the gamma likelihood on s_{kj}. This implies that the posterior distribution also follows a gamma distribution, with updated parameters $\beta_{1n} = \beta_1 + n_s\xi$ and $\beta_{2n} = \beta_2 + \sum_{q \neq k} s_{qj}$, and where $n_s = n_m$, discussed in the last section. The posterior distribution can be written,

$$\lambda_j \quad | \quad \beta_1, \beta_2, \boldsymbol{s}_{-kj} \quad \sim \quad \text{Gamma}(\beta_{1n}, \beta_{2n}).$$

The resulting probability for s_{kj} is found by marginalizing over this posterior, which gives

$$p(s_{kj} \mid \boldsymbol{s}_{-kj}, \beta_1, \beta_2, \xi) = \frac{\Gamma(\xi + \beta_{1n})}{\Gamma(\xi)\Gamma(\beta_{1n})} \frac{{\beta_{2n}}^{\beta_{1n}} {s_t}^{\xi-1}}{(\beta_{2n} + s_t)^{\beta_{1n}+\xi}}.$$

In this case, we adopt a Metropolis sampler with a normal distribution for the proposal, centred on the existing value and with a fixed variance of 0.1. A new candidate value s^*_{kj} is sampled from this distribution, and is accepted if a standard uniform random variate u on $[0, 1]$ is less than

$$\frac{p(\boldsymbol{x}_k|m_k, s^*_k)}{p(\boldsymbol{x}_k|m_k, s_k)} \frac{p(s^*_{kj} \mid \boldsymbol{s}_{-kj}, \boldsymbol{\theta})}{p(s_{kj} \mid \boldsymbol{s}_{-kj}, \boldsymbol{\theta})}$$

where s^*_k has the same values as s_k for all dimensions except the jth.

Resampling Regions When the Category is Known

In the third part of the sampler, we look to sample the region assignment for the ith stimulus. For the moment, we assume that $i \neq t$, so the value of z_i is known to the learner, and is not a latent variable. Since the observations are exchangeable, we may treat the ith observation as if it were the last, and sample from the distribution $p(r_i|\boldsymbol{r}_{-i}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta})$. For convenience, I introduce the notation $m_i = (m_{i1}, \ldots, m_{iv})$ in this section, to refer to the center of the region to which the ith stimulus belongs. Strictly, this should be written m_{r_i}, but this simplifies matters and no confusion is introduced. The same applies to the size of the region, s_i. By applying Bayes' theorem and simplifying, we see that

$$\begin{aligned} & p(r_i|\boldsymbol{r}_{-i}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) \\ & \quad \propto \quad p(x_i|\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}_{-i}, \boldsymbol{z}, \boldsymbol{\theta})\ p(r_i|\boldsymbol{r}_{-i}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}_{-i}, \boldsymbol{z}, \boldsymbol{\theta}) \\ & \quad = \quad p(x_i \mid m_i, s_i)\ p(r_i \mid \boldsymbol{r}^{(z_i)}_{-i}, z_i, \alpha), \end{aligned}$$

where $\boldsymbol{r}^{(z_i)}_{-i}$ refers to the collection of regions to which the other members of category z_i are assigned. Now consider each of these terms in turn. The prior probability term is given by integrating out the stick-breaking weights w,

$$\begin{aligned} p(r_i \mid \boldsymbol{r}^{(z_i)}_{-i}, z_i, \alpha) \quad &= \quad \int p(r_i|w) p(w \mid \boldsymbol{r}^{(z_i)}_{-i}, z_i, \alpha) dw \\ &= \quad \int w_i\ p(w|\boldsymbol{r}^{(z_i)}_{-i}, z_i, \alpha)\ dw \end{aligned}$$

The marginal distribution for a stick-breaking prior with discrete likelihood is called a Chinese restaurant process (CRP: see Navarro, Griffiths, Steyvers & Lee, in press, for a simple discussion). We can write the conditional CRP probability very simply:

$$p(r_i = k \,|\, \boldsymbol{r}_{-i}^{(z_i)}, z_i, \alpha) = \begin{cases} \frac{n_k}{n_z+\alpha} & \text{if } k \in \boldsymbol{r}_{-i}^{(z_i)} \\ \frac{\alpha}{n_z+\alpha} & \text{otherwise} \end{cases}$$

where n_k denotes the number of items currently assigned to region k, and n_z denotes the total number of items (in category z). Strictly, $\alpha/(\alpha + n_z)$ is the probability of sampling a hitherto unseen consequence, not the probability of any specific member of the set of unseen consequences.

Turning to the likelihood, $p(x_i \,|\, m_i, s_i)$, if $r_i \in \boldsymbol{r}_{-i}^{(z_i)}$ we can use the uniform likelihood function described earlier. In the case where the item is assigned to an entirely new region (i.e., when $r_i \notin \boldsymbol{r}_{-i}^{(z_i)}$), m_i and s_i are random samples from the "base distribution" over possible regions. The likelihood of x_i given that m_i and s_i are random samples from the base distribution is given by:

$$p(x_i \,|\, r_i \notin R_+) \;\; = \;\; \int p(x_i \,|\, m_i, s_i)\; p(m_i \,|\, \boldsymbol{m}_{-i}, \boldsymbol{\theta})\; p(s_i \,|\, \boldsymbol{s}_{-i}, \boldsymbol{\theta})\; d(m_i, s_i)$$

This integral is intractable for the current model, but we can address this by a data augmentation technique in which values for m_i and s_i are actually sampled from the base distribution. However, rather than sampling a single value to deal with this integral, we adopt the more efficient auxiliary variable approach for non-conjugate priors proposed by by Neal (2000: algorithm 8), in which multiple values are sampled at every stage. If we sample g auxiliary variables, the probability that r_i comes from the base distribution is evenly spread across g samples. Thus, each variable is assigned probability $\alpha/(g(n_z + \alpha))$. In the current implementation, the samples are drawn by first sampling values for λ_j, τ_j and μ_j from their conditional posteriors distributions (discussed earlier), and then sampling the locations and sizes from the distributions described by λ_j, τ_j and μ_j. The values of λ_j, τ_j and μ_j are discarded as soon as the assignment r_i is made, as are the g auxiliary regions, unless of course the stimulus is assigned to the region. Note that if the region to which x_i had previously been assigned was a singleton (i.e., if x_i was the only stimulus assigned to the region), then one of the auxiliary regions must correspond to the region to which x_i had been previously assigned. See Neal (2000) for details regarding this sampler.

Resampling Regions When the Category is Unknown

Extending the sampler to the case when the category assignment z_t is unknown is straightforward[1]. In this situation, it is most convenient to sample z_t and r_t from

[1] In supervised learning tasks, this case only arises for the current stimulus, but in unsupervised or semi-supervised tasks it can apply for other items. However, because the current model provides a single set of parameters μ, τ and λ shared by all categories, it is unlikely to constitute a good model for unsupervised tasks: as noted by Navarro (submitted), stronger priors for within-category homogeneity and between-category heterogeneity are probably required.

their joint posterior,

$$\begin{aligned} & p(r_t, z_t|\boldsymbol{r}_{-t}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}, \boldsymbol{z}_{-t}, \boldsymbol{\theta}) \\ & \quad \propto \; p(x_t|\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}_{-t}, \boldsymbol{z}, \boldsymbol{\theta}) \; p(r_t, z_t|\boldsymbol{r}_{-t}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{x}_{-t}, \boldsymbol{z}_{-t}, \boldsymbol{\theta}) \\ & \quad = \; p(x_t \,|\, m_t, s_t) \; p(r_t, z_t \,|\, \boldsymbol{r}_{-t}, \boldsymbol{z}_{-t}, \alpha, \eta), \end{aligned}$$

The likelihood term remains the same as in the previous section, and may be dealt with using the auxiliary variable approach suggested by Neal (2000). The joint prior over the region and category is simple,

$$p(r_t, z_t \,|\, \boldsymbol{r}_{-t}, \boldsymbol{z}_{-t}, \alpha, \eta) = p(r_t \,|\, \boldsymbol{r}_{-t}^{(z_t)}, z_t, \alpha) p(z_t \,|\, \boldsymbol{z}_{-t}, \eta)$$

The prior over the region assignment follows the same CRP probability described in the last section. Since we have a symmetric Dirichlet(η) prior on the category base rates ζ, it is simple to observe that

$$p(z_t = q \,|\, \boldsymbol{z}_{-t}, \eta) = \frac{n_q + \eta}{n + v\eta}$$

where n_q denotes the number of previously observed stimuli that belonged to category q, and n is the total number of stimuli observed so far excluding the item t, so $n = t - 1$. These expressions are sufficient to produce a sampler for (z_t, r_t).

References

Gilks, W. R. , Richardson, S., & Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Navarro, D. J. (submitted). From natural kinds to complex categories. Submitted to *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Navarro, D. J., Griffiths, T. L., Steyvers, M, & Lee, M. D. (in press). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9*, 249-265.