Rational model of categorization

CS786

5th October 2024

Note for class

- I will introduce RMC more gently in the live class
- For the assignment, since you already have a coded-up version of the model,
  - these slides
  - plus this paper
  - should see you through

Setup

- People behave as if they are storing prototypes sometimes
  - Category judgments evolve over multiple presentations
  - Sensible thing to do in situations where the category is not competing with others for membership
- People behave as if they are storing exemplars sometimes
  - Probability matching behavior in describing category membership
  - Sensible thing to do when discriminability at category boundaries becomes important
- Shouldn’t we have models that can do both?

A Bayesian observer model of categorization

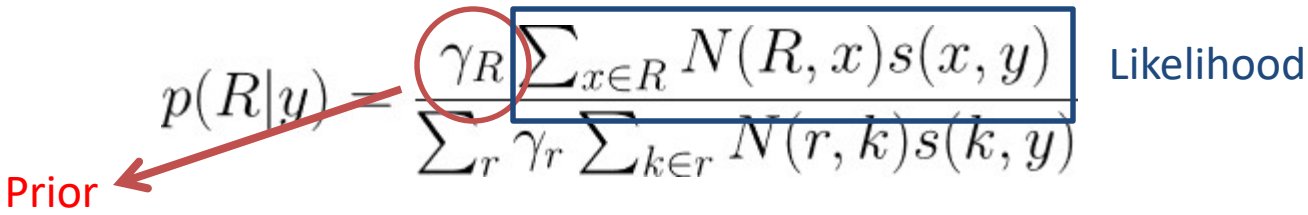
- Want to predict category label **c** of the **N**th object, given
  - All objects seen before and
  - Their category labels
- Sequential Bayesian update

$$p(c_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{p(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) p(c_N = j | \mathbf{c}_{N-1})}{\sum_c p(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) p(c_N = j | \mathbf{c}_{N-1})}$$

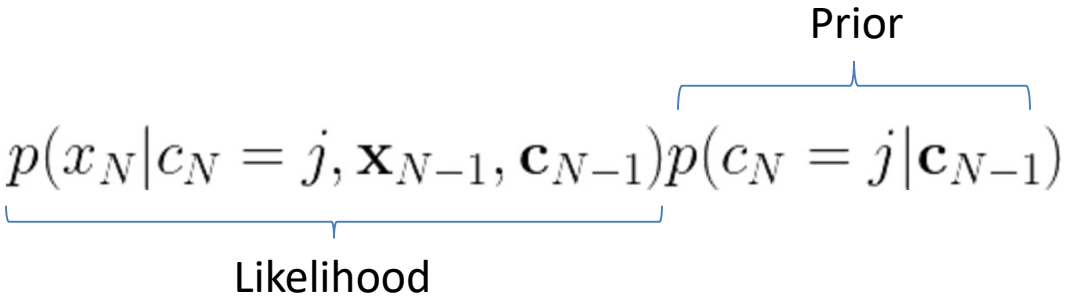
Assumption: only category labels influence the prior. Can you think of situations when this would be broken?

Connection with classic models

- Remember the GCM category response calculation?



- Look at the numerator of the Bayesian model



Let's call this likelihood $L_{N,j}$

A unifying view

- In an exemplar model

$$L_{N,j} = \sum_{i|c_i=j} L_{N,i}$$

- In a prototype model

$$L_{N,j} = L_{N,p_j}$$

- Crucial insight
  - Prototype model is a clustering model with one cluster
  - Exemplar model is a clustering model with N clusters

https://hekyll.services.adelaide.edu.au/dspace/bitstream/2440/46850/1/hdl_46850.pdf

The clustering view of categorization

- Stimuli are grouped in clusters
- Clusters are associated with categories
  - Either non-exclusively (Anderson's RMC, 1992)
  - Or exclusively (Griffiths' HDP, 2006)
- Now the likelihood can look like

$$L_{N,j} = \sum_{k}^{K_j} L_{N,p_{j,k}}$$

- First proposed as the varying abstraction model (Vanpaemal et al., 2005)

How many clusters need we learn per category?

- Deeply related to the non-parametric Bayesian question of how many clusters we need to fit a dataset
- Problem addressed by Anderson's Rational Model of Categorization (RMC)
- Modeled category learning as a Dirichlet Process

RMC: big picture

- Treat categories as just another label for the data
- Category learning is equivalent to learning the joint distribution $p(\mathbf{x}_N, \mathbf{c}_N)$
- Learn this as a mixture model

$$p(\mathbf{x}_N, \mathbf{c}_N) = \sum_{\mathbf{z}_N} p(\mathbf{x}_N, \mathbf{c}_N | \mathbf{z}_N) p(\mathbf{z}_N)$$

- $p(\mathbf{z}_N)$ is a distribution over all possible clusterings of the N items

http://blog.datumbox.com/the-dirichlet-process-the-chinese-restaurant-process-and-other-representations/

Coin toss example

- Say you toss a coin N times
- You want to figure out its bias
- Bayesian approach
  - Find the generative model
  - Each toss ~ Bern(θ)
  - θ ~ Beta(α,β)
- Draw the generative model in plate notation

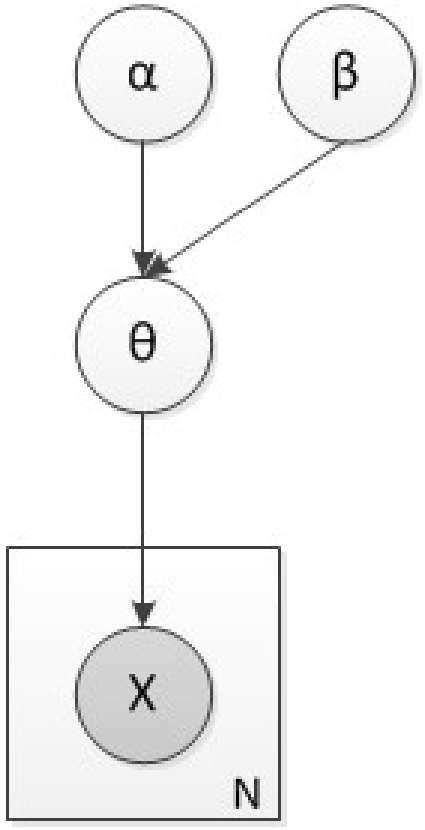


Plate notation

- Random variables as circles
- Parameters, fixed values as squares
- Repetitions of conditional probability structures as rectangular ‘plates’
- *Switch* conditioning as squiggles
- Random variables observed in practice are shaded

Conjugacy

- Algebraic convenience in Bayesian updating
- Posterior ← Prior x Likelihood
- We want the distributions to be parametric, the parameter is what is learned
  - we want the posterior to have the same parametric form as the prior
  - Conjugate prior = f(.) such that $f(\theta)g(x|\theta) \sim f(\theta^{new})$

Useful conjugate priors

likelihood	conjugate prior	posterior
$p(x \mid \theta)$	$p_0(\theta)$	$p(\theta \mid x)$
Normal (θ, σ)	Normal (μ_0, σ_0)	Normal (μ_1, σ_1)
Binomial (N, θ)	Beta (r, s)	Beta $(r + n, s + N - n)$
Poisson (θ)	Gamma (r, s)	Gamma $(r + n, s + 1)$
Multinomial $(\theta_1, \ldots, \theta_k)$	Dirichlet $(\alpha_1, \ldots, \alpha_k)$	Dirichlet $(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$

This one is important for us

The multinomial distribution

- For n independent trials that could yield one of k possible results, the multinomial distribution gives the probability of seeing any particular combination of outcomes

$$p(\mathbf{z}, \mathbf{p}) = \frac{n!}{z_1! z_2! \cdots z_k!} p_1^{z_1} p_2^{z_2} \cdots p_k^{z_k}$$

  - Each point can go into one of k clusters
  - z_i is the number of points in each cluster for the immediate observation
  - p_i is the fraction of points in each cluster in the long run
- Given 3 clusters A,B and C, with normalized empirical frequencies [0.3, 0.4, 0.3] seen over a large set, what is the probability of the partitioning AABB for a four data sample?
  - P(clustering) = 6 x 0.09 x 0.16 = 0.0864

The Dirichlet distribution

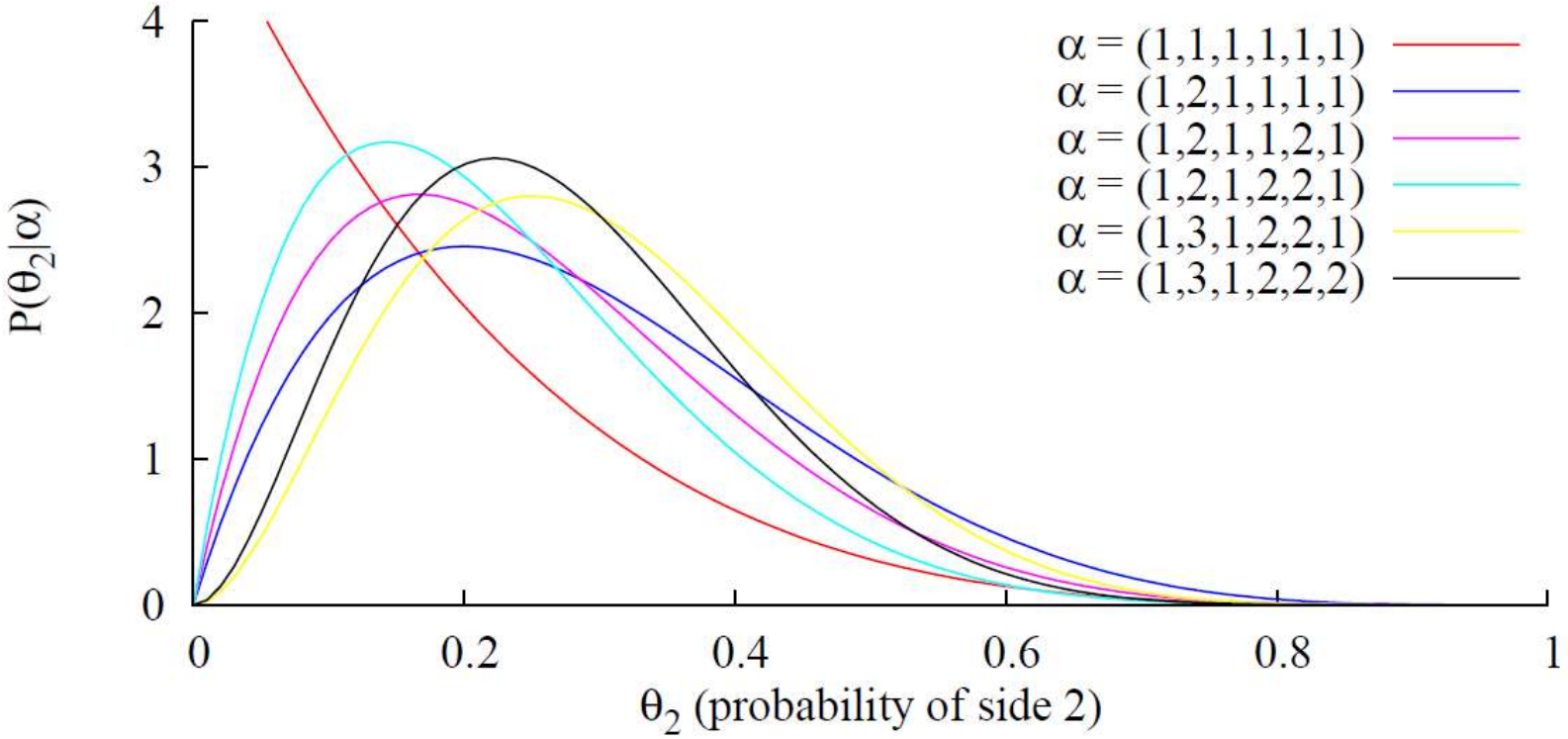
- A *k*-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex, and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

- Easier to understand
  - Prior Dir(α_1, α_2)
  - Likelihood Multi(θ_1 , θ_2)
  - Outcome {n_1, n_2}
  - Posterior Dir(α_1+ n_1, α_2 + n_2)
- Ignoring the normalization constant, what is the Dirichlet probability of a multinomial sample [0.1, 0.5, 0.4] with parameter 10
  - $(0.1)^9$ $(0.5)^9$ $(0.4)^9$ = 5e-16
- What would it be for parameter 0.2?
  - 22

Dirichlet distribution emits multinomial samples

- Data $d = (2, 5, 4, 2, 6)$



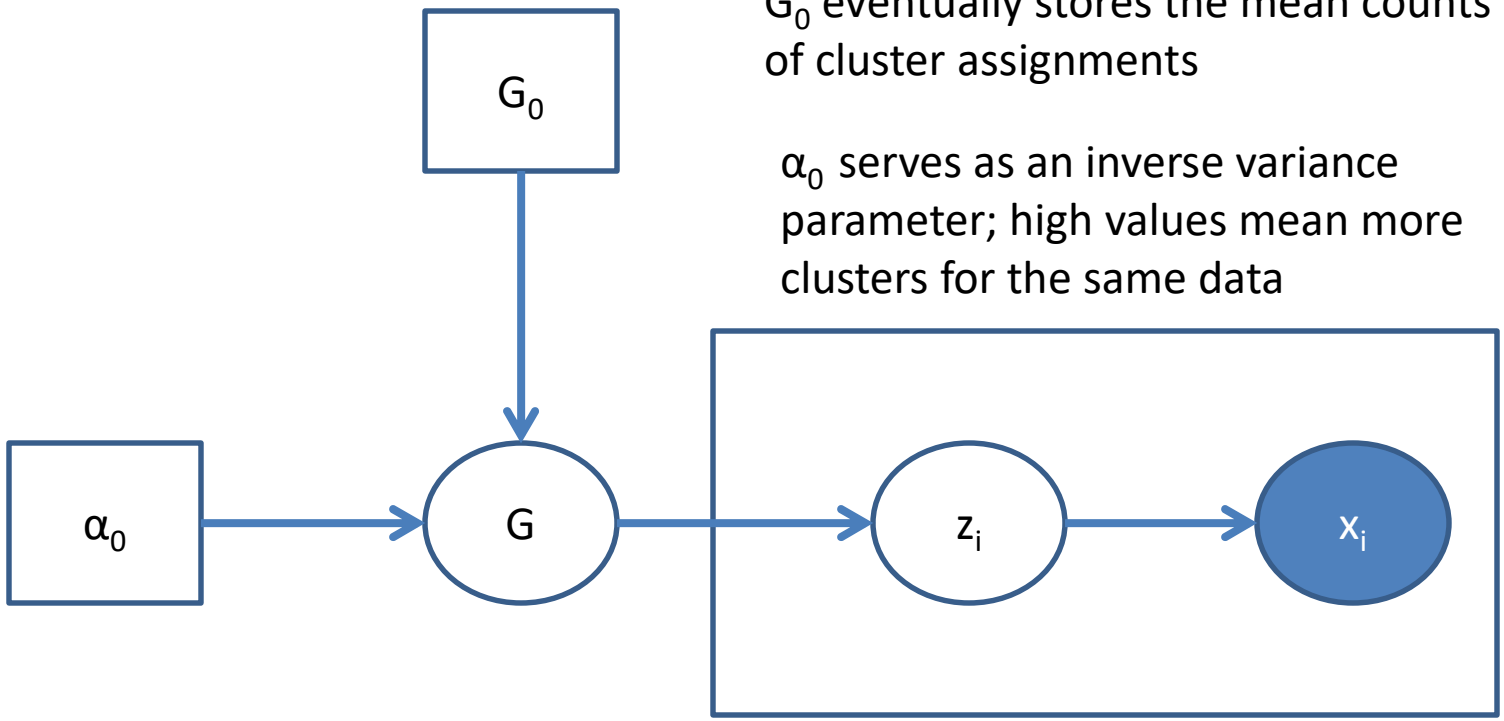
Dirichlet process

- Probability distribution over probability measures
  - A probability measure is a function that maps a probabilistic sample space to values in [0,1]
- G is a DP(α, G_0) distributed random probability measure if for any partition of the corresponding probability space **Θ** we have

$$G(A_1), \cdots, G(A_n) \sim \text{Dirichlet}(\alpha G_0(A_1), \cdots, \alpha G_0(A_n))$$

https://en.wikipedia.org/wiki/Probability_space

Dirichlet process mixture model



G_0 eventually stores the mean counts of cluster assignments

α_0 serves as an inverse variance parameter; high values mean more clusters for the same data

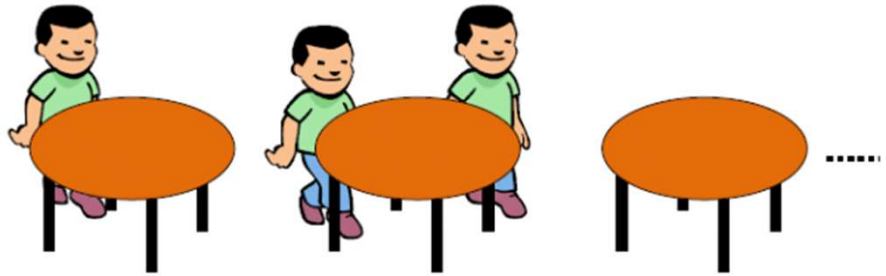
G is a Dirichlet distribution on possible partitions of the data
z is a sample from this distribution, a partition of the data
Learning the right parameter values ends up telling us which partitions are most likely

RMC prior is a Dirichlet process prior

- The prior reflects a generative process where

$$P(z_i = k|\mathbf{z}_{i-1}) = \begin{cases} \frac{M_k}{i-1+\alpha} & M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{i-1+\alpha} & M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases}$$

M is the count of cluster assignments to that point. Compare with the GCM frequency prior.



$P(z_i = c \mid c_{-i}) =$

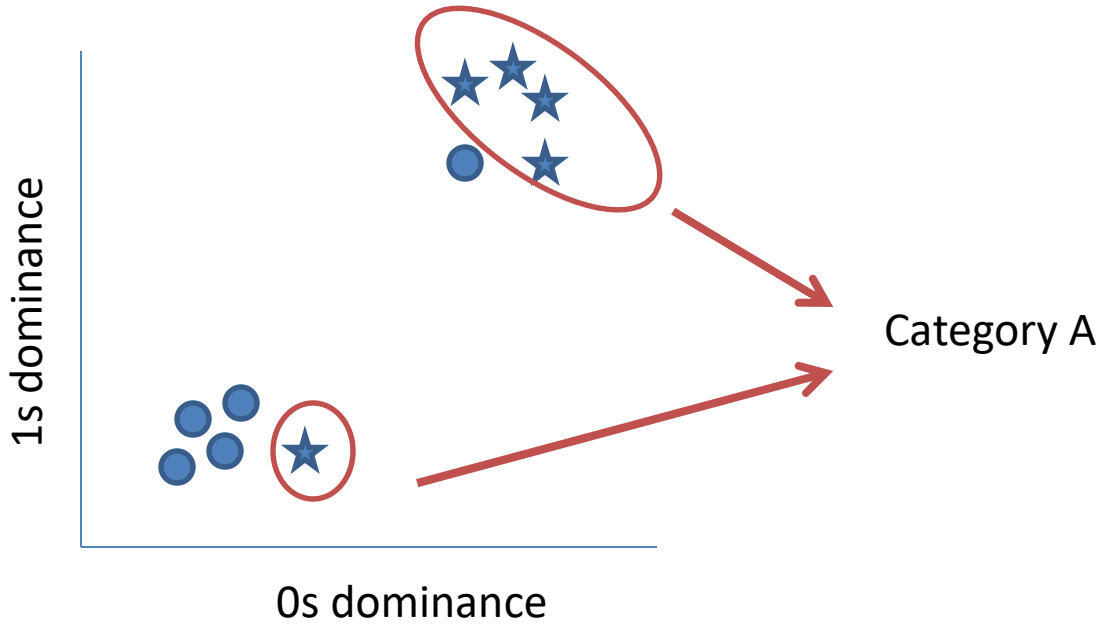
1	0	0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$	$\frac{\alpha}{3+\alpha}$

What does RMC do?

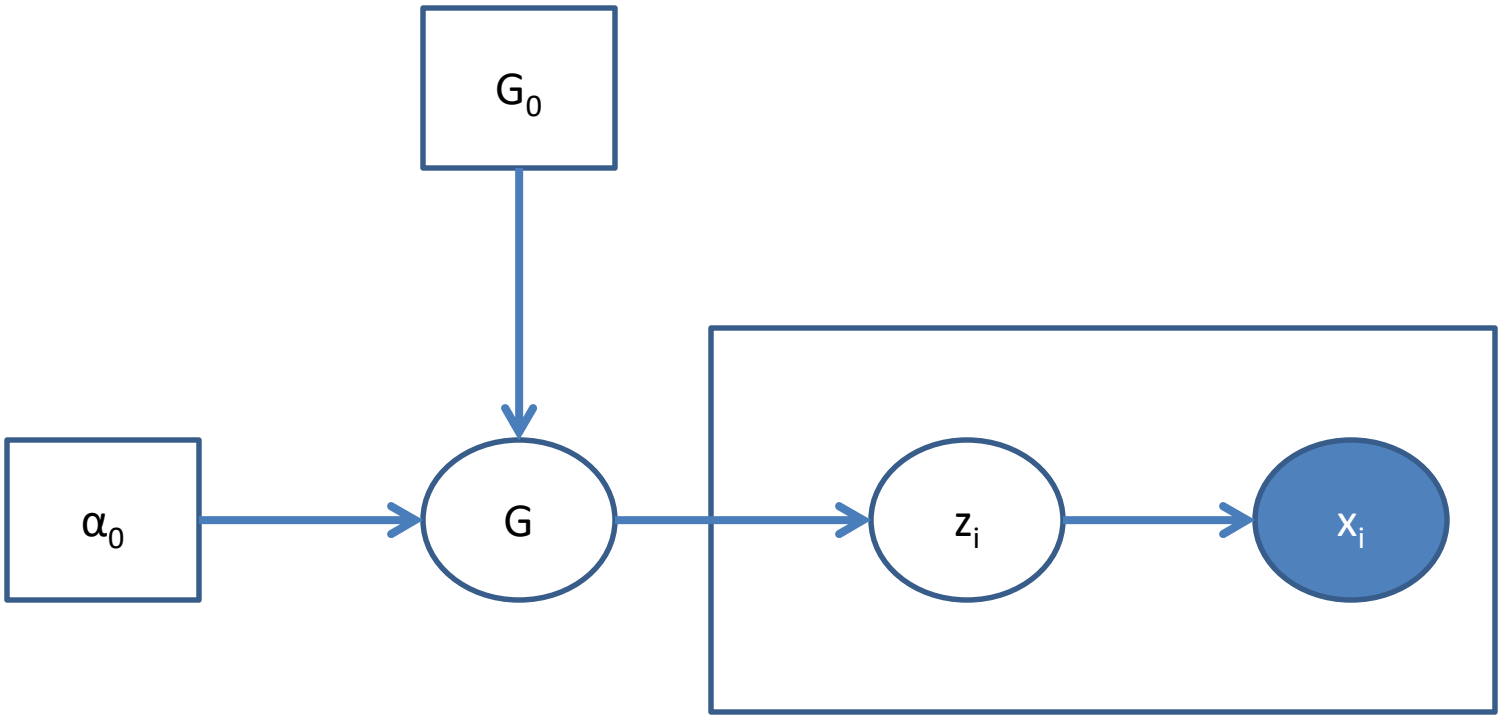
- It's a Dirichlet process mixture model that learns clusters in the data
- Each cluster is soft-assigned to any of the category labels through feedback
- How many clusters are learned across the entire dataset depends on the CRP prior
- Weaknesses
  - Same number of clusters across all categories
  - Order in which data points enter the model doesn't matter (exchangeability)

The value of RMC

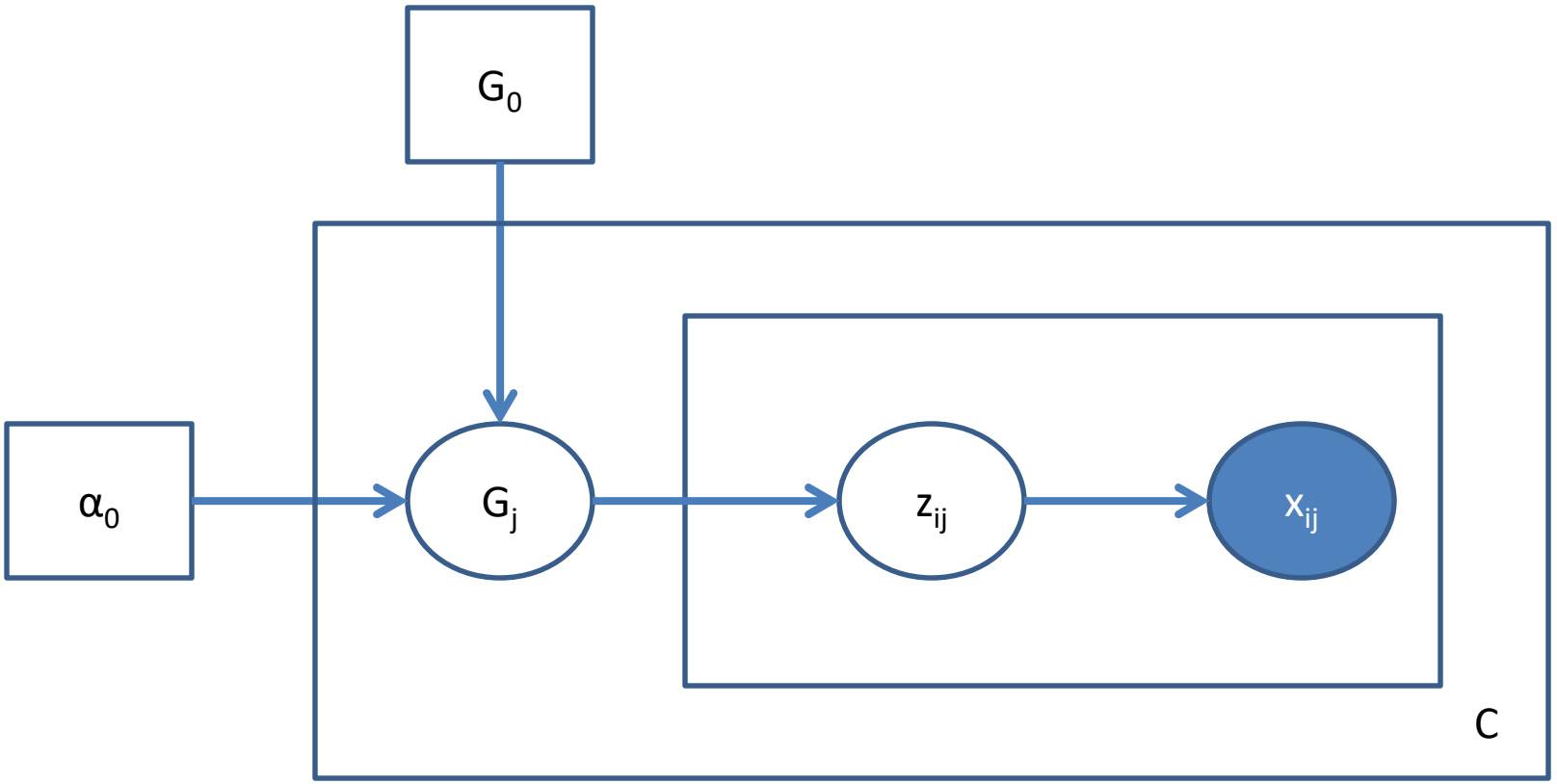
- Partially explains how humans might succeed in learning category labels for datasets that are not linearly separable in feature space



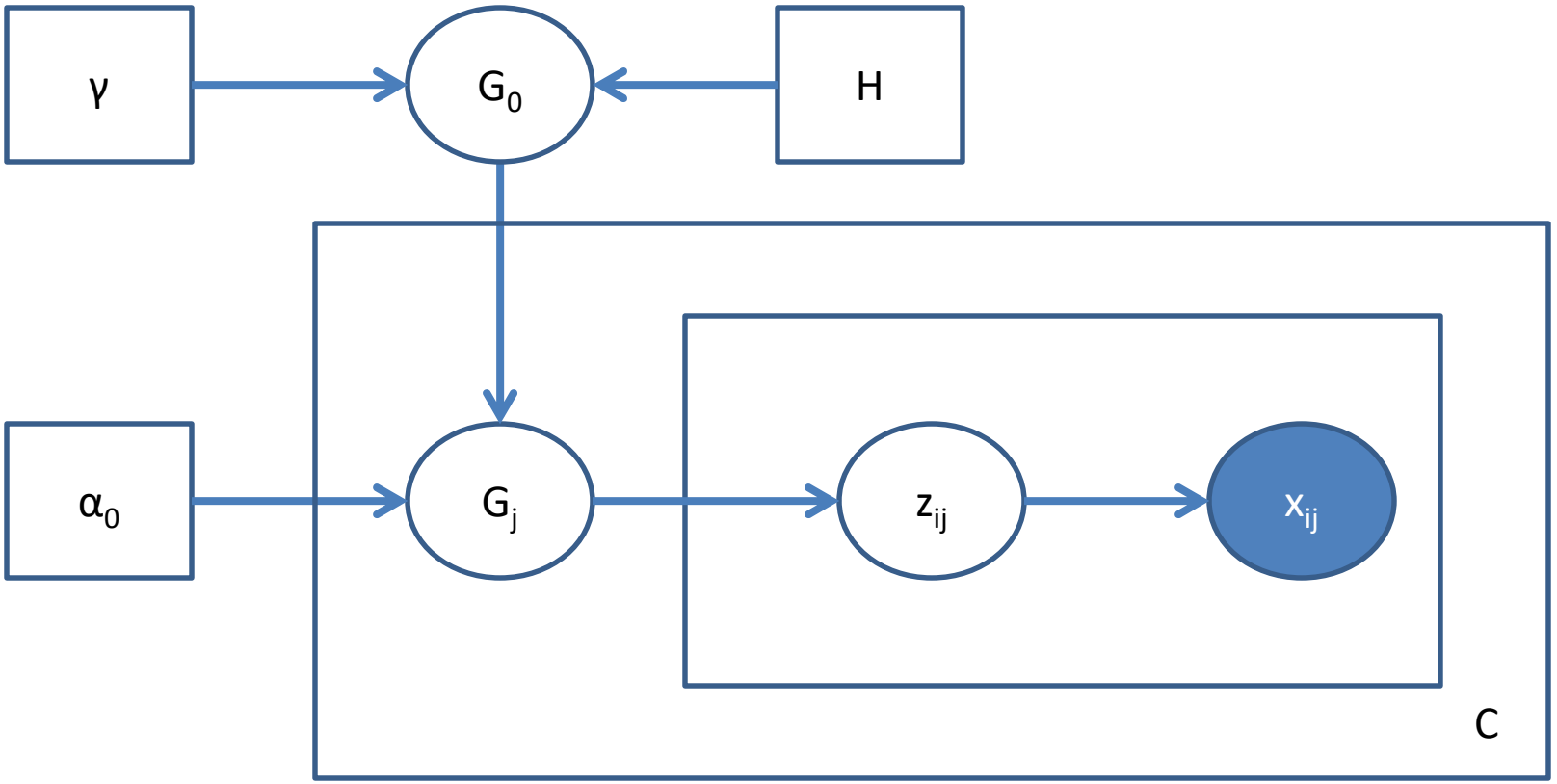
HDP model of categorization



HDP model of categorization



HDP model of categorization



C is the number of category labels
HDP learns clusters for each category label separately.
Can have varying numbers of clusters for each category now.

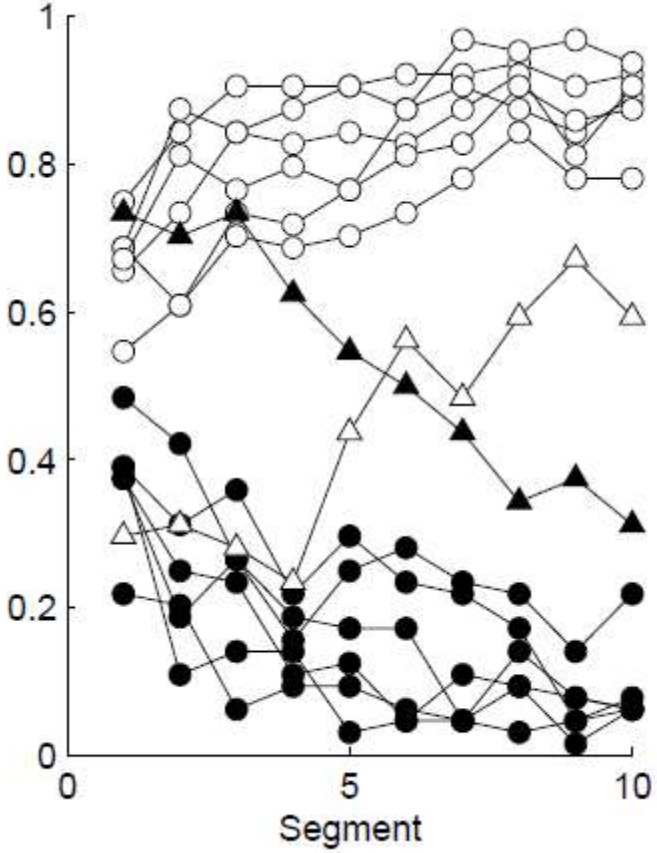
A unifying view of categorization models

- The HDP framework successfully integrates most past accounts of feature-based categorization

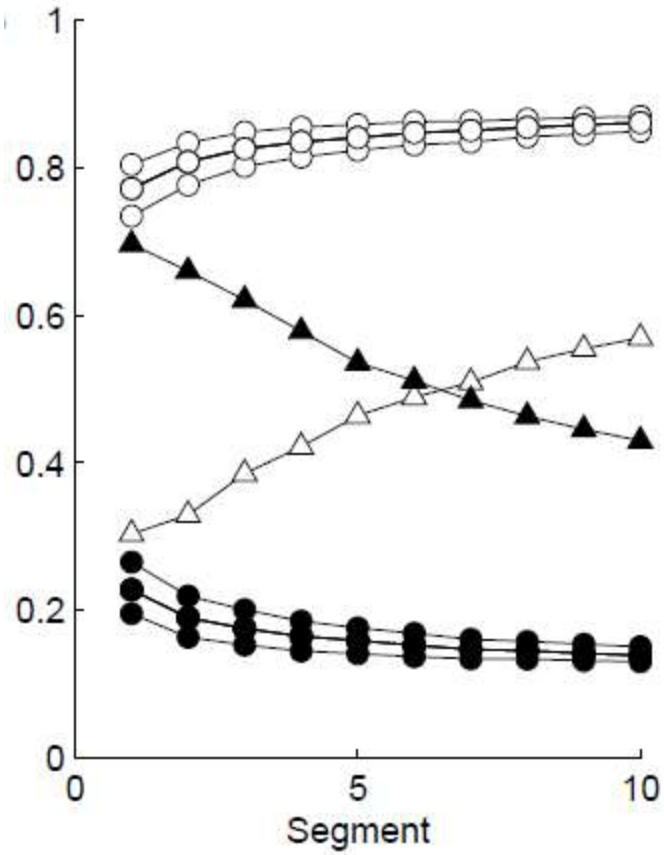
	$\gamma \in (0,\infty)$ categories share clusters	$\gamma \to \infty$ categories share no clusters
$\alpha \to 0$ one cluster per category	$HDP_{0,+}$	$HDP_{0,\infty}$ (prototype)
$\alpha \in (0,\infty)$ intermediate number of clusters	$HDP_{+,+}$	$HDP_{+,\infty}$
$\alpha \to \infty$ one stimulus per cluster	$HDP_{\infty,+}$ (RMC)	$HDP_{\infty,\infty}$ (exemplar)

Predicting human categorization dynamics

Human data



$HDP_{+,\infty}$



Insight: category-specific clustering seems to explain the data best

Open questions

- How to break order-independence assumptions in such models
  - Human categorization is order dependent
- The actual calculations in these models are formidable
  - What simplifications are humans using that let them do the same task using neuronal outputs?
- The likelihood function is just similarity based
  - Are similarity functions atomic entities in the brain, or are they subject to inferential binding like everything else