Modeling Individual Differences with Dirichlet Processes

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
Department of Psychology, University of Adelaide, SA 5005, Australia

Thomas L. Griffiths (thomas_griffiths@brown.edu)
Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912, USA

Mark Steyvers (msteyver@uci.edu)
Department of Cognitive Sciences, University of California, Irvine, Irvine CA 92697, USA

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide, SA 5005, Australia

Abstract

We introduce a Bayesian framework for modeling individual differences, in which subjects are assumed to belong to one of a potentially infinite number of groups. In this model, the groups observed in any particular data set are not viewed as a fixed set that fully explain the variation between individuals, but rather as representatives of a latent, arbitrarily rich structure. As more people are seen, the number of observed groups is allowed to grow, as more details about the individual differences are revealed. We use the Dirichlet process – a distribution widely used in nonparametric Bayesian statistics – to define a prior for the model, allowing us to learn flexible parameter distributions without overfitting the data, or requiring the complex computations typically required for determining the dimensionality of a model. As an initial demonstration of the approach, we present an application of the method to categorization data.

Much of cognitive science involves the development and evaluation of models. Models formalize theoretical predictions and have been successfully applied to a range of phenomena. A recurrent problem, however, is that individual differences between people are often overlooked. This occurs because, most often, models are evaluated against data that have been averaged or aggregated across subjects, and so assume that there are no individual differences between them. In this paper we introduce a new framework for modeling individual differences. Informed by recent insights in statistics and machine learning (e.g., Escobar & West, 1995; Neal, 2000), our *infinite groups model* makes it possible to divide subjects who behave similarly into groups, without assuming an upper bound on the number of groups. This model is sufficiently flexible to capture the heterogeneous structure produced by different subjects pursuing different strategies, allows the number of groups to grow naturally as we observe more data, and avoids the complex computations often required for determining the dimensionality of an individual differences model.

Modeling Individual Differences

In those cases of cognitive modeling that recognize individual differences, it is usually assumed that each subject behaves in accordance with a different parameterization, θ, of a single model, and that model is evaluated against the data from each subject separately (e.g., Ashby, Maddox & Lee, 1994; Nosofsky, 1986). Although this avoids the problem of corrupting the underlying pattern of the data, it also foregoes the potential benefits of averaging, and guarantees that models are fit to all of the noise in the data. Clearly, individual subject analysis increases the risk of overfitting, and hence reduces the ability to make accurate predictions or to generalize to new contexts. As a result, a number of authors have considered more economical ways of expressing individual differences, which seek to describe the ways in which people are the same as well as the ways in which they are different (e.g., Peruggia, Van Zandt & Chen, 2002; Rouder, Sun, Speckman, Lu & Zhou, in press; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003; Webb & Lee, 2004).

Two dominant approaches have emerged in the literature on modeling individual differences. In the *stochastic parameters model* (e.g., Peruggia et al., 2002; Rouder et al., in press), every participant is assumed to have a unique parameter value θ that is randomly sampled from a parametric distribution, as illustrated in Figure 1a. In contrast, the *groups model* assumes that people fall into one of a number of different clusters. Within a group, people are assumed to behave in essentially the same way, but each group is qualitatively different. Under this approach to individual differences modeling (e.g., Lee & Webb, in press; Steyvers et al., 2003; Webb & Lee, 2004), the goal is to partition subjects into a number of groups and associate each group with a parameter set θ, as illustrated by the parameter distribution shown in Figure 1b.

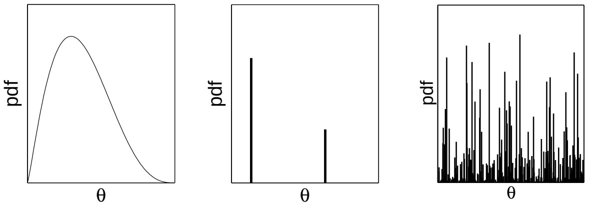
Hierarchical Bayesian Models

The assumptions underlying these two approaches to modeling individual differences can be understood by viewing both as *hierarchical Bayesian models* (e.g., Lindley & Smith, 1972). If data arise from a parametric distribution $x \sim F(\cdot \,|\, \theta)$ described by some cognitive model, then there is assumed to exist a higher-order process $G(\cdot \,|\, \phi)$ that generates the values of θ. A two level hierarchical model with parameters ϕ is written,

$$\begin{array}{rcl} \theta \,|\, \phi & \sim & G(\cdot \,|\, \phi) \\ x \,|\, \theta & \sim & F(\cdot \,|\, \theta). \end{array}$$

To apply Bayesian inference to such a model, we also need to define a prior on ϕ. We will assume that $\phi \sim \pi(\cdot)$ for an appropriate distribution $\pi(\cdot)$.

In the stochastic parameters model $G(\cdot \,|\, \phi)$ is usually a tractable distribution such as a Gaussian, with ϕ corre-



(a) (b) (c)

Figure 1: Parameter distributions associated with stochastic parameters approach to individual differences (panel a), the original groups approach (panel b), and the infinite groups approach (panel c).

sponding to the parameters of that distribution. In the groups model $G(\cdot \,|\, \phi)$ is a weighted collection of k point masses, as depicted in Figure 1b. That is,

$$G(\cdot \,|\, w, \xi) = \sum_{j=1}^{k} w_j \delta(\xi_j), \tag{1}$$

where $\delta(\xi_j)$ denotes a point mass located at $\theta = \xi_j$ and where $\sum_{j=1}^{k} w_j = 1$. In the groups model, ϕ corresponds to the parameters (w, ξ).

This perspective reveals some of the strengths and weaknesses of these two models. Assuming that θ follows a parametric distribution, as in the stochastic parameters model, simplifies the problem of fitting individual differences models to data, but places strong constraints on the kind of variation that can manifest across subjects. A particularly severe problem arises when we specify a unimodal distribution to capture individual differences that are inherently multimodal. In this case the model cannot capture the most important aspect of the variation between people.

Unlike the stochastic parameters approach, the parameter distributions postulated by group models naturally account for multimodality in individual differences. By postulating two groups, for instance, we arrive at a bimodal distribution. However, there is a lack of flexibility in parameter distributions consisting only of a few point masses. Moreover, a computational problem faced by the groups model is the difficulty in choosing the number of groups. This is typically treated as a *model selection* problem, addressed by evaluating a series of models which assume different numbers of groups. Such a strategy is time-consuming, and makes the false assumption that there really is a fixed number of groups, with future subjects belonging to the same set of groups. In the remainder of the paper we will explore a model that combines the strengths of these two approaches, having the flexibility of the groups model, but a simple inference algorithm.

The Infinite Groups Model

In the infinite groups model, we adopt a distribution on θ that is more flexible than the parametric distribution assumed by the stochastic parameters model, but still allows efficient inference. We assume that subjects are drawn from an infinite number of groups, taking $G(\cdot \,|\, \phi)$ to be a weighted combination of an infinite number of point masses, as in Figure 1c. That is,

$$G(\cdot \,|\, w, \xi) = \sum_{j=1}^{\infty} w_j \delta(\xi_j). \tag{2}$$

While we assume that the number of groups is unbounded, any finite set of subjects will contain representatives from a finite subset of these groups. By avoiding setting an upper bound on the number of groups, we no longer need to perform explicit model selection to identify the number of groups. This model has an inherent psychological plausibility: people can vary in any number of ways, only some of which will be observed in a finite sample. With infinitely many groups, there is always the possibility that a new subject can display a pattern of behavior that has never been seen before. Moreover, the approach requires us to make our assumptions explicit, in the form of a well-defined prior distribution over the number of observed groups. In contrast, in most finite-order model selection scenarios these assumptions are usually swept up in an implicit and often inappropriate uniform prior over model order (for a notable exception, see Courville, Daw, Gordon & Touretzky, 2004).

In order to apply Bayesian inference in the hierarchical model defined by Equation 2, we need to define a prior $\pi(\cdot)$ on the parameters of G, in this case w and ξ. In a finite groups model with k groups (i.e., Equation 1), a standard prior is

$$\begin{array}{rcc} \xi & \sim & G_0(\cdot) \\ w \,|\, \alpha, k & \sim & \text{Dirichlet}(\alpha/k, \ldots, \alpha/k). \end{array} \tag{3}$$

In this prior, the locations of the k point masses ξ_j are sampled from the *base distribution* denoted $G_0(\cdot)$. The Dirichlet distribution over w gives us a prior over the different ways in which k groups could be weighted, in which $p(w \,|\, \alpha, k) \propto \prod_{j=1}^{k} w_j^{(\alpha/k)-1}$. Specifying the base distribution, G_0 and the dispersion parameter of the Dirichlet distribution, α, defines a prior over distributions $G(\cdot)$ for any finite groups model.

If we take the limit of the prior $\pi(\cdot)$ defined by Equation 3 as $k \to \infty$, we obtain the distribution known as the *Dirichlet process* (Ferguson, 1973). This distribution takes its name from the fact that is very much like an infinite dimensional Dirichlet distribution (see Schervish, 1995, pp. 52-60; Ghosh & Ramamoorthi, 2003). The Dirichlet process provides a prior for infinite models, and inference in models using this prior is generally straightforward. If the distribution $G(\cdot)$ is sampled from a Dirichlet process, we write $G \,|\, G_0, \alpha \sim DP(\cdot \,|\, G_0, \alpha)$, so the infinite groups model can be written

$$\begin{array}{rcl} G \,|\, G_0, \alpha & \sim & DP(\cdot \,|\, G_0, \alpha) \\ \theta \,|\, G & \sim & G(\cdot) \\ x \,|\, \theta & \sim & F(\cdot \,|\, \theta), \end{array}$$

where α is the dispersion parameter (setting the prior on w) and $G_0(\cdot)$ is the base distribution on ξ. In this model, the base distribution $G_0(\cdot)$ represents our prior beliefs about the kinds of parameter values θ that are likely to capture human performance in a particular task, while the dispersion parameter α represents the amount of variation that we expect to see in a finite sample. If α is low, then most people will be expected to behave similarly to one another, and the distribution $G(\cdot)$ will concentrate most of its mass on a few points. However, if α is large, then people will be expected to be very different to one another, and $G(\cdot)$ will spread its mass over a large number of points.

The Dirichlet process defines a distribution over the assignment of subjects to groups. Since w_j gives the probability that the ith observation belongs to the jth group, it is convenient to introduce the group membership variable g_i, such that $p(g_i = j) = w_j$. Given the group assignments of $i-1$ subjects, it is straightforward to compute the probability distribution over g_i given $g_{-i} = \{g_1, \ldots, g_{i-1}\}$. In a finite groups model with the prior given in Equation 3, we can integrate over $w_1, \ldots, w_k$, and obtain

$$p(g_i = j \mid g_{-i}, \alpha, k) = (n_j + \alpha/k)/(i - 1 + \alpha),$$

where n_j denotes the number of elements in g_{-i} that are equal to j. If we let $k \to \infty$, the limiting probabilities are

$$p(g_i = j \mid g_{-i}, \alpha) \quad \propto \quad \begin{cases} \frac{n_j}{i-1+\alpha} & \text{if } j \leq k_{-i} \\ \frac{\alpha}{i-1+\alpha} & \text{otherwise,} \end{cases}$$

where k_{-i} denotes the number of distinct groups present in g_{-i}. This gives the appropriate conditional probability under the Dirichlet process (Neal, 2000).

Through the distribution over group assignments, the Dirichlet process induces a prior $p(k \mid \alpha, n)$ over the number of unique groups k that will manifest among n subjects. Antoniak (1974) shows that $p(k \mid \alpha, n) \propto z_{nk}\alpha^k$, where z_{nk} is an unsigned Stirling number of the first kind (see Abramowitz & Stegun, 1972). He also observes that the expected number of components sampled from a Dirichlet process is approximately given by,

$$E[k \mid \alpha, n] \approx \alpha \ln \left(\frac{n + \alpha}{\alpha} \right).$$

Thus, although $k \to \infty$ with probability 1 as $n \to \infty$, the number of components increases approximately logarithmically with the number of observations.

In most contexts the dispersion α is unknown, so we specify a prior distribution $p(\alpha)$, allowing us to learn α from data. The posterior distribution over α is given by

$$p(\alpha \mid k, n) \propto \mathrm{B}(\alpha, n)\alpha^k p(\alpha),$$

where $\mathrm{B}(\alpha, n)$ is a standard Beta function. A common choice for $p(\alpha)$ is the Gamma distribution $\alpha \mid a, b \sim \text{Gamma}(\cdot \mid a, b)$ in which $p(\alpha) \propto \alpha^{a-1}e^{-b\alpha}$ (Escobar & West, 1995). If so,

$$p(\alpha \mid k, n) \propto \alpha^{a+k-1}e^{-b\alpha}\mathrm{B}(\alpha, n). \qquad (4)$$

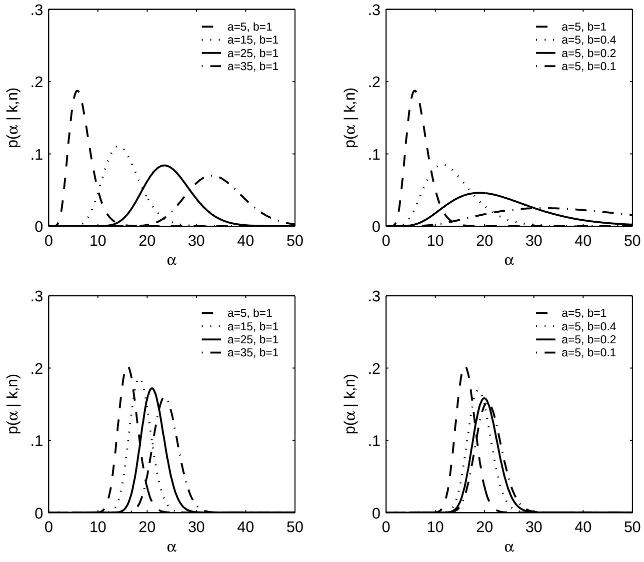


Figure 2: Posterior distributions over α given k and n. In the top row $k = 8$ and $n = 10$, while in the bottom row $k = 79$ and $n = 1000$. In both cases, the expected value of α is approximately 20. On the left hand side, the Gamma priors over α have the same scale parameter but vary in shape. On the right hand side,the priors have the same shape parameter but vary in scale.

In particular, Escobar and West (1995) note that if we let $a \to 0$ and $b \to 0$ we obtain a so-called scale-invariant prior in which $p(\alpha) \propto 1/\alpha$ (see Jeffreys, 1961). Posterior distributions for α are shown in Figure 2. The influence of the prior is shown by varying both the shape (left hand side) and the scale (right hand side) parameters.

Modeling Discrete Data

We now turn to the derivation and application of the infinite groups model to situations in which participants provide discrete data. Suppose that n people perform some task in which m possible responses can be made on each trial, and each person experiences s trials. We can describe the ith participant's responses with the vector $x_i = (x_{i1}, \ldots, x_{im})$, where x_{ih} counts the number of times that they made the hth response. Using the Dirichlet process, we assume that each person belongs to one of an infinite number of groups, and that the parameters for the jth group describe a multinomial rate $\theta_j = (\theta_{j1}, \ldots, \theta_{jm})$ such that θ_{jh} denotes the probability with which a member of group j makes response h on any given trial. Since this likelihood function is multinomial, it is convenient to assume that the base distribution $G_0(\cdot)$ is a Dirichlet distribution with symmetric parameter β. We will assume that the dispersion parameter α is unknown, and follows a Gamma distribution with parameters a and b. The model is illustrated in Figure 3.

This model lends itself to inference by Gibbs sampling, a Monte Carlo method for sampling from the posterior distribution over the variables in a Bayesian model (see Neal, 2000; Gilks, Richardson & Spiegelhalter, 1995). In

$$p(g_i = j \mid g_{-i}, \alpha, x) \quad \propto \quad \begin{cases} \dfrac{\Gamma(m\beta + q_{-i,j})}{\prod_{h=1}^{m} \Gamma(\beta + q_{-i,j,h})} \dfrac{\prod_{h=1}^{m} \Gamma(\beta + q_{\cdot,j,h})}{\Gamma(m\beta + q_{\cdot,j})} \dfrac{n_{-i,j}}{n-1+\alpha} & \text{if } j \leq k_{-i} \\ \dfrac{\Gamma(m\beta)}{\prod_{h=1}^{m} \Gamma(\beta)} \dfrac{\prod_{h=1}^{m} \Gamma(\beta + q_{\cdot,j,h})}{\Gamma(m\beta + q_{\cdot,j})} \dfrac{\alpha}{n-1+\alpha} & \text{otherwise} \end{cases} \tag{5}$$

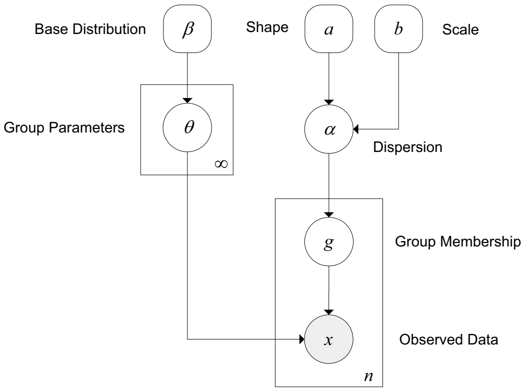


Figure 3: Dependencies in the infinite groups model for discrete data as it is used here. Shaded circles denote observed variables, white circles are latent variables, rounded squares denote known parameter values, and plates indicate a set of independent replications of the processes shown inside them.

Gibbs sampling, we fix all assignments except one and sample that assignment from the conditional posterior $p(g_i \mid g_{-i}, x)$, a procedure which eventually converges to samples from the complete posterior $p(g \mid x)$. Since the Dirichlet is conjugate to the multinomial, it is straightforward to show that the conditional posterior distribution over the ith group assignment is given by Equation 5. In this expression, $q_{-i,j,h}$ denotes the number of times that a participant (not including the ith) currently assigned to group j made response h, and $q_{-i,j}$ denotes the total number of responses made by these participants. The terms $q_{\cdot,j,h}$ and $q_{\cdot,j}$ are defined similarly, except that the ith participant's data are not excluded.

For the dispersion parameter, we treat the prior over α as a Gamma$(\cdot \mid a, b)$ distribution. Using Antoniak's (1974) results, the conditional posterior over α depends only on the number of observed groups k, not the specific assignments. Thus, by expanding the Beta function $\mathrm{B}(\alpha, n)$ in Equation 4 we observe that

$$p(\alpha \mid g, x) \propto \alpha^{a+k-1} e^{-b\alpha} \int_0^1 \eta^{\alpha-1}(1-\eta)^{n-1} d\eta.$$

Since this conditional distribution is difficult to directly sample from, it is convenient to employ a "data augmentation", in which we view $p(\alpha \mid g, x)$ as the marginalization over η of the joint distribution,

$$p(\alpha, \eta \mid k, n) \propto \alpha^{a+k-1} e^{-b\alpha} \eta^{\alpha-1}(1-\eta)^{n-1}.$$

Using this joint distribution, we can find $p(\alpha \mid \eta, k, n)$ and $p(\eta \mid \alpha, k, n)$. These distributions are simply,

$$\begin{array}{lcl} \alpha \mid \eta, k, n & \sim & \mathrm{Gamma}(\cdot \mid a+k-1, b - \ln \eta) \\ \eta \mid \alpha, k, n & \sim & \mathrm{Beta}(\cdot \mid \alpha, n). \end{array} \tag{6}$$

Equations 5 and 6 define the Gibbs sampler.

As a simple illustration, we created random data sets with n people and s discrete observations per person, where each observation denotes a choice of one of m response options. The sample was divided into k groups, and each group associated with a multinomial rate θ sampled from a uniform distribution. People were allocated randomly to groups using a uniform distribution, subject to the constraint that each group contained at least one member. Note that this allocation scheme means that the Dirichlet process model is misspecified (which is as it should be in any worthwhile simulation). Results are shown in Figure 4. The posterior distributions over α are shown on the left and the posterior distributions over k are shown on the right. On the whole, the distributions converge on sensible answers, though in both cases they reflect a fair degree of uncertainty about the number of groups present in the sample.

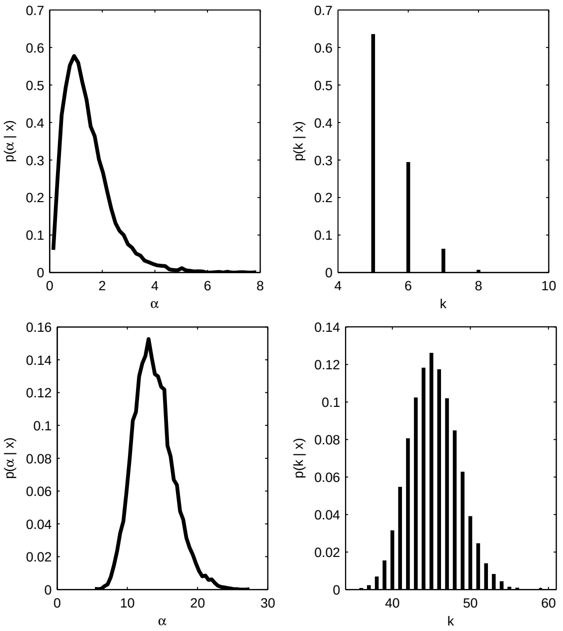


Figure 4: Simulations in which people provide $s = 50$ observations each, and $m = 20$ response options are possible on every trial. In the upper panels there are $n = 20$ people and $k = 5$ groups. In the lower panels there are $n = 200$ people and $k = 50$ groups.

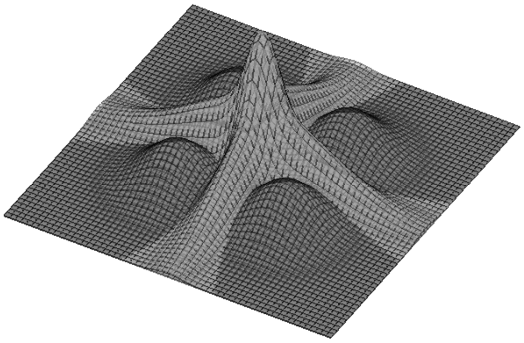


Figure 5: The category densities used in McKinley and Nosofsky's (1995) experiment 2. Category A (dark grey) is a mixture of four Gaussians, while category B (light grey) is a mixture of two Gaussians.

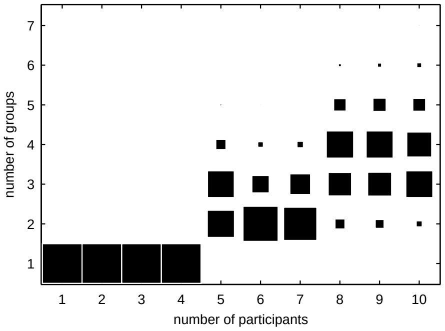


Figure 6: Posterior over k as a function of n. The area of the squares is proportional to the posterior probability of k given n.

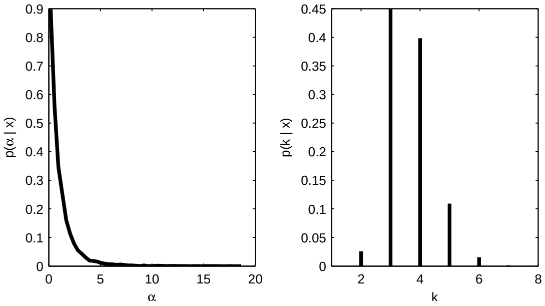


Figure 7: Posterior distributions over α and k when the infinite groups model is applied to McKinley and Nosofsky's (1995) experiment 2.

Table 1: Percentage of occasions on which participants in McKinley and Nosofsky's (1995) experiment 2 appear in the same group.

	2	3	4	5	6	7	8	9	10
1	37	0	73	0	58	68	36	43	55
2		22	34	1	68	56	3	5	67
3			1	57	1	0	0	0	2
4				0	44	52	53	60	43
5					0	0	0	0	0
6						86	4	7	93
7							11	16	86
8								91	4
9									7

Individual Differences in Categorization

We now present an application of the infinite groups model. An elegant category learning experiment by McKinley and Nosofsky (1995) investigated 10 people's[1] ability to discriminate between the two probabilistic categories shown in Figure 5. The stimuli were circles with a radial line running through them, and so the two dimensions depicted in Figure 5 correspond to the radius of the circle, and the angle of the line. Category A (dark grey) is a mixture of four Gaussian distributions, while category B is a mixture of two Gaussians. On any given trial in the experiment, a stimulus was sampled from one of the six Gaussian distributions. Participants were asked whether it came from category A or category B, and provided feedback as to the accuracy of their response. Because the categories are inherently probabilistic and the category densities are quite complicated, this task is very difficult, and shows evidence of differences not only during the course of category learning, but in the final structures learned.

In order to learn about the variation between participants, we applied the infinite groups model to the data from this experiment. Since each trial is labeled by the Gaussian distribution (i.e., A1, A2, A3, A4, B1, or B2) from which it was sampled, a natural way of viewing each participant's data is in terms of the probability of making the correct response to stimuli sampled from each of the six components. For each of the 10 participants we used only the last 300 trials of the experiment, in order to look for differences in the learned category structure, rather than differences in the learning process itself. In order to conduct a Bayesian analysis, we set principled *a priori* parameter values rather than fitting the model to the data. Since we know that both responses (i.e., "A" and "B") are possible but are otherwise "ignorant", the natural choice for the base distribution is the uniform distribution $\beta = 1$ (see Jaynes, 2003, pp. 382–386), and since we have no strong beliefs about α we set the scale-invariant prior (see Jeffreys, 1961) in which $a \rightarrow 0$, $b \rightarrow 0$.

To illustrate the manner in which the model grows with the data, imagine that the 10 participants entered the lab in order of participant ID. Figure 6 shows how the posterior distribution over k changes as more participants are observed. Initially there is evidence for only a single group, but once the 10th participant is observed, there is strong evidence for about 3 or 4 groups, as illustrated in Figure 7. A more detailed description of the relationships between participants is presented in Table 1, which shows the (marginal) probability that any two participants belong to the same group, and reveals a rich pattern of similarities and differences. A subset of this interaction is illustrated in Figure 8, which plots the last 300 stimuli observed by participants 5, 7, 8 and

[1] McKinley and Nosofsky (1995) actually report data for 11 participants. However, the data currently available to us include only 10 of these.

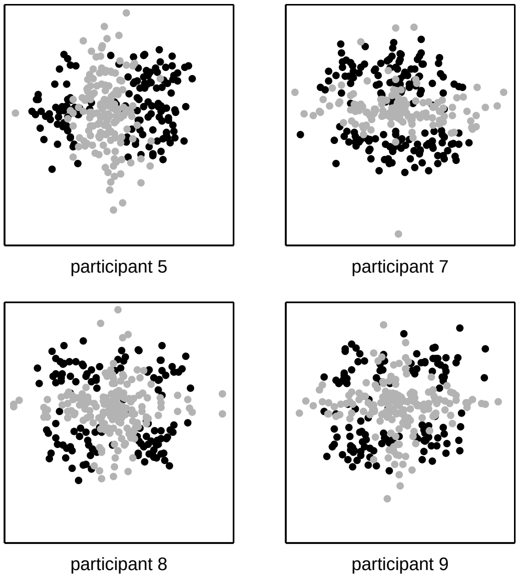


Figure 8: Last 300 trials for participants 5, 7, 8 and 9 in McKinley and Nosofsky's (1995) experiment 2. Black dots denote "A" responses, and grey dots denote "B" responses.

9, and the decisions that they made. Broadly speaking, participant 5 is sensitive only to variation along the x-axis, participant 7 is sensitive only to variation on the y-axis, while participants 8 and 9 do a good job of learning the category structures on both dimensions. As a result, participants 5 and 7 rarely appear in the same group as one another or with participants 8 or 9 (with probabilities ranging from 0% to 7%), while participants 8 and 9 almost always (91%) co-occur.

General Discussion

The individual differences framework outlined in this paper provides a natural method of representing the similarities and differences between people. Moreover the groups that are seen in any particular sample are not viewed as a fixed structure that fully explains the variation between individuals, but rather as representatives of a latent, arbitrarily rich structure. This means that, had we subsequently observed another 100 participants in the McKinley and Nososky (1995) experiment, the number and variety of observed groups would grow as more detail about individual differences are revealed.

The approach can be extended in a number of ways. Firstly, in many situations we may want continuous multimodal parameter distributions. An "infinite stochastic groups" model would convolve each of the point masses in the Dirichlet process with a continous distribution, giving an infinite model that subsumes both the groups model and the stochastic parameters model. A second direction in which the framework could be extended would be to allow structured relationships between groups. Finally, we may wish to consider an "idiosyncratic strategies model", in which it is assumed that all participants draw on a common set of strategies but combine them in an unique way.

Acknowledgements

Research supported by Australian Research Council grant DP-0451793. We thank Yves Rosseel for providing a copy of the categorization data.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics, 2*, 1152-1174.

Ashby, F. G., Maddox, W. T. & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science 5*, 144-151.

Courville, A. C., Daw, N. D., Gordon, G. J. & Touretzky, D. S. (2004). Model uncertainty in classical conditioning. In S. Thrun, L. Saul & B. Schölkopf (Eds) *Advances in Neural Information Processing Systems, 16* (pp. 977–984). Cambridge, MA: MIT Press.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90*, 577-588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1*, 209-230.

Ghosh, J. K. & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* New York: Springer.

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

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). London: Oxford University Press.

Lee, M. D. & Webb, M. R. (in press). Modeling individual differences in cognition. *Psychonomic Bulletin & Review.*

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society (Series B), 34*, 1-41.

McKinley, S. C. & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance, 21*, 128–148.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Peruggia, M., Van Zandt, T., & Chen, M. (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. In C. Gatsonis, A. Kass, R. E. Carriquiry, A. Gelman, D. Higdon, D. K. Pauler, & I. Verdinelli (Eds.), *Case Studies in Bayesian Statistics 6* (pp. 319334). New York: Springer-Verlag.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J. & Zhou, D. (in press). A parametric hierarchical framework for inference with response time distributions. *Psychometrika.*

Schervish, M. J. (1995). *Theory of Statistics.* New York: Springer.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J. & Blum, B. (2003). Inferring causal networks from observations and intervations. *Cognitive Science, 27*, 453-487.

Webb, M. R., & Lee, M. D. (2004). Modeling individual differences in category learning. In K. Forbus, D. Gentner & T. Regier, (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pp. 1440-1445. Mahwah, NJ: Erlbaum.