Modeling Individual Differences Using Dirichlet Processes

Daniel J. Navarro[a], Thomas L. Griffiths[b], Mark Steyvers[c], and Michael D. Lee[a]

[a] *Department of Psychology, University of Adelaide*
[b] *Department of Cognitive and Linguistic Sciences, Brown University*
[c] *Department of Cognitive Sciences, University of California, Irvine*

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 explains the variation between individuals, but rather as representatives of a latent, arbitrarily rich structure. As more people are seen, and more details about the individual differences are revealed, the number of inferred groups is allowed to grow. 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 three applications that analyze the individual differences in category learning, choice of publication outlets, and web browsing behavior.

Key Words: individual differences, Dirichlet processes, Bayesian nonparametrics

"I am surprised that the author has used this data set. In my lab, when we collect data with such large individual differences, we refer to the data as "junk". We then re-design our stimuli and/or experimental procedures, and run a new experiment. The junk data never appear in publications"

– An anonymous reviewer in 2005, commenting on research that sought to model individual differences in cognition.

1 Introduction

Suppose we asked one hundred people which number was the most unlucky. Of those people, fifty said '13', forty said '4', and ten said '87'. This variation is unlikely to be due to noise in the cognitive process by which people make unluckiness judgments: If we replicated the experiment with the same people, the *same* fifty people would probably say 13 again. It seems much more likely that most of the observed variation arises from genuine differences in what those people believe. A complete explanation of people's answers would have to account for this variation.

Often, cognitive modeling ignores individual variation, because it uses data that have been averaged or aggregated across subjects. The potential benefit of averaging data is that, if the performance of subjects really is the same except for noise, the averaging process will tend to remove the effects of the noise, and the resultant data will more accurately reflect the underlying psychological phenomenon. When the performance of subjects has genuine differences, however, it is well known (e.g., Ashby, Maddox, & Lee, 1994; Estes, 1956; Myung, Kim, & Pitt, 2000) that averaging produces data that do not accurately represent the behavior of individuals, and provide a misleading basis for modeling. In our unlucky numbers experiment, the average unlucky number is approximately 17, which was not given as an answer by any participant. More fundamentally, the practice of averaging data restricts the focus of cognitive modeling to issues of how people are the same. While modeling invariants is fundamental, it is also important to ask how people are different. If experimental data reveal individual differences in cognitive processes, we should seek to model this variation rather than ignore it. From the unlucky number data, we might discover that, while fifty people were drawing on an Anglo-Saxon tradition in which 13 is considered unlucky, forty were drawing on a corresponding Chinese tradition in which 4 is considered unlucky. Moreover, the remaining ten participants might turn out to be Australian cricket fans (87 is considered an unlucky number for Australian batsmen).

Cognitive modeling that attempts to accommodate individual differences usually assumes that each subject behaves in accordance with a unique parameterization of a model, and so evaluation is undertaken against the data from each subject independently (e.g., Ashby, Maddox & Lee, 1994; Nosofsky, 1986; Wixted & Ebbesen, 1997). Although this avoids the problem of corrupting the underlying pattern of the data, it also foregoes the

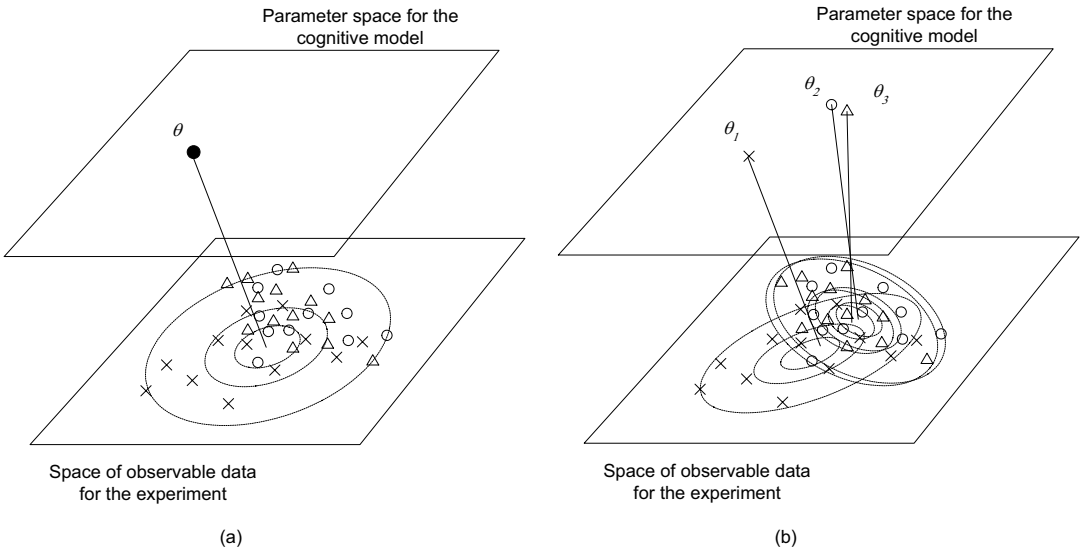


Fig. 1. Standard modeling approaches for data from many subjects, involving the aggregation of data (panel a), and the modeling of each individual independently (panel b). The data are plotted in the lower data space, with different symbols for each participant. The upper parameter space shows the parameter distributions inferred under each modeling approach.

potential benefits of averaging, and guarantees that modeling is affected by all of the noise in the data. In our hypothetical unlucky numbers experiment, it seems unlikely to be a coincidence that fully half of the participants said exactly the same thing. A more parsimonious account is that the fifty people who said 13 are in some way related to one another, but are not related to the forty people who said 4 or the ten people who said 87. Moreover, suppose we discovered an Australian cricket fan with a bad memory, and this person accidentally says 86. Individual subject analysis does not allow us to "share statistical strength" between the cricket fans, in the sense that having seen many 87 answers could be used to correct the 'noisy' 86 answer. In general, modeling everybody independently increases the risk of overfitting, and hence reduces the ability to make accurate predictions or to generalize to new contexts.

Notwithstanding the ongoing debate about the relative merits of fitting aggregated versus individual data (e.g., Maddox & Estes, 2005), the previous discussion suggests that *both* viewpoints are unsatisfying. To provide a visual illustration of this point, consider the hypothetical data shown in Figure 1. The figure depicts the outcome of a simple experiment in which we collect noisy data from three participants. The three participants' data are indicated with crosses, circles, and triangles. The crosses form a roughly elliptical shape from the lower left to the upper right of the data space, whereas the circles and triangles form ellipses that slant from the upper left to the lower right. On the left hand side (panel a), we aggregate across participants, and estimate a single parameter value θ that produces a distribution that is roughly circular, indicated by the contour plot.

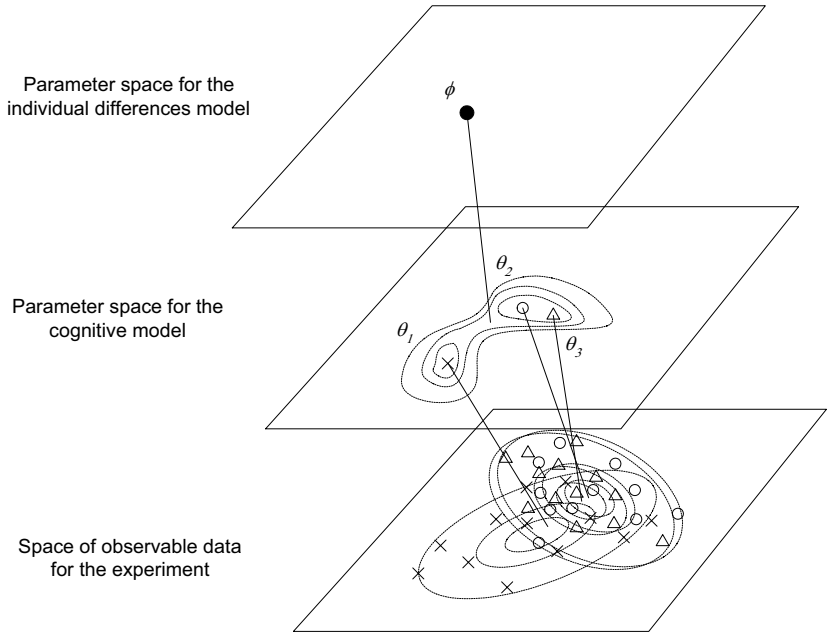


Fig. 2. The model-based view of individual differences. The data are plotted in the lower data space, with different symbols for each participant. The middle parameter space shows the parameter values inferred for each participant based on their data *and* an individual differences model that describes how these parameters can vary between people. The upper parameter space shows the inferred parameter values for this individual differences model.

The aggregate looks nothing like the individuals. On the right hand side (panel b), we estimate a parameter value independently for each participant. The inferred parameter values θ_1, θ_2 and θ_3 and their associated contour plots now do capture the basic aspects of everyone's performance. However, this accuracy has come at the cost of losing sight of the similarity between two of the participants. Using the individual fitting approach, this relationship $\theta_2 \approx \theta_3$ is not represented. Even if we observed a large number of people with very similar parameter values, we could make no formal inference about the relationship between them. Ultimately, neither the aggregate nor the individual view captures the pattern of similarities and differences apparent in the data. Aggregated models can learn similarities and individual models can learn differences, but modeling individual variation in cognition requires being able to learn both simultaneously.

Because of these difficulties, a number of authors have considered more sophisticated ways of expressing individual differences within models of cognitive processes (e.g., Lee & Webb, in press; Peruggia, Van Zandt & Chen, 2002; Rouder, Sun, Speckman, Lu & Zhou, 2003; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003; Webb & Lee, 2004). The central innovation is to provide an explicit model for the kinds of individual differences that might appear in the data, in much the same way as established methods in psychometric models like Item Response Theory (e.g., Lord, 1980; Hoskens & de Boeck, 2001;

Junker & Sijtsma, 2001). The general approach, illustrated schematically in Figure 2, is to supplement the cognitive model that describes variation within a single participant's data with an individual differences model that describes *how* cognitive parameters can vary across people. Using sufficiently flexible individual differences models, it is possible to learn both the similarities and differences between people.

Model-based approaches to individual differences vary in terms of the class of distributions that are allowed to describe variation in parameter values, reflecting different assumptions about which aspects of individual differences are the most important to capture. In this paper we introduce a new model-based framework for understanding 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 in the observed sample to grow naturally as more data appear, and avoids the complex computations that are often required when one chooses an individual differences model by standard model selection methods. We illustrate the infinite groups model by considering simple multinomial models that predict the frequencies of responses across a set of categories. However, the idea generalizes to more general classes of probabilistic models.

The structure of the paper is as follows: We begin with an overview of existing frameworks for modeling individual differences, and their interpretations as Bayesian hierarchical models. We then introduce the infinite groups approach as a principled way to address some of the problems associated with these frameworks, including model selection problems. Next, we provide a brief tutorial on the Dirichlet process, which forms the basis of our approach, and discuss how model selection proceeds when working with the infinite groups framework. We then derive the infinite groups model for discrete data and present illustrative simulation studies. Finally, we present three applications that analyze the individual differences in categorization performance, choice of publication outlets, and web browsing behavior.

2 Hierarchical Bayesian Models for Individual Differences

Two dominant model-based approaches have emerged in the literature on individual differences. In a *stochastic parameters model* (e.g., Peruggia et al., 2002; Rouder et al., 2003), every participant is assumed to have a unique parameter value θ that is sampled from a parametric distribution, as illustrated in Figure 3a. The intuition behind the approach is that, while every person is unique, the variation between people is not arbitrary, and can be described by a distribution over the parameters. These distributions are generally smooth and unimodal, reflecting a general tendency at the mode, and a noise model de-

scribing the variations that exist across individuals' parameters. In contrast, the idea that underlies the *groups model* is that there exist a number of distinct types of qualitatively different performance. Accordingly, this approach assumes that people fall into one of a number of fundamentally distinct groups. Within a group, people are assumed to behave in the same way, but the groups themselves can vary in all kinds of ways. 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 parameterization θ, as illustrated in Figure 3b.

In order to understand the assumptions that underlie these two frameworks, it is helpful to view them as *hierarchical Bayesian models* (e.g., Lindley & Smith, 1972). Suppose we have data from an experiment that involves n participants. If the ith individual participant provides m_i observations, we can denote these observations by the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i\,m_i})$. By specifying a cognitive model, we assume that these data can be described as *i.i.d.* samples from the distribution $x_{ij} \sim F(\cdot \,|\, \theta_i)$. Additionally, by specifying an individual differences model, we assume that there is a distribution $\theta_i \sim G(\cdot \,|\, \boldsymbol{\phi})$ that we can use to describe the parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ for each of the participants. Since we now have two distinct levels at which we wish to construct models, we can write a model of this form as the two-level hierarchical model,

$$\begin{aligned} x_{ij} \,|\, \theta_i &\sim F(\cdot \,|\, \theta_i) \\ \theta_i \,|\, \boldsymbol{\phi} &\sim G(\cdot \,|\, \boldsymbol{\phi}). \end{aligned} \tag{1}$$

In this expression $\boldsymbol{\phi}$ denotes the parameters used to describe the individual differences model $G(\cdot \,|\, \boldsymbol{\phi})$. Letting $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ refer to the complete data set, we can write the likelihood function for this hierarchical Bayesian model as,

$$\begin{aligned} p(\boldsymbol{x} \,|\, \boldsymbol{\phi}) &= \prod_i p(\boldsymbol{x}_i \,|\, \boldsymbol{\phi}) \\ &= \prod_i \int \left(\prod_j F(x_{ij} \,|\, \theta_i) \right) G(\theta_i \,|\, \boldsymbol{\phi}) \, \mathrm{d}\theta_i. \end{aligned} \tag{2}$$

To apply Bayesian inference to this model, we also need to define a prior on $\boldsymbol{\phi}$. We will assume that $\boldsymbol{\phi} \sim \pi(\cdot)$ for some appropriate distribution $\pi(\cdot)$. Statistical inference in this model is achieved by finding $p(\boldsymbol{\theta}, \boldsymbol{\phi} \,|\, \boldsymbol{x})$, the joint posterior distribution over parameter values and individual difference models. However, we are often only interested in some aspects of this joint posterior, so only some parts are reported. Two cases of particular interest are,

(1) *Posterior over parameters for the cognitive model.* One role of $G(\cdot \,|\, \boldsymbol{\phi})$ is to induce dependencies between the parameters θ_i. In some contexts this is all that the researcher requires, so it is natural in these situations to consider the marginal distri-

bution $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. The idea in this case is that we would use the dependencies induced via our individual differences model to produce better parameter estimates.

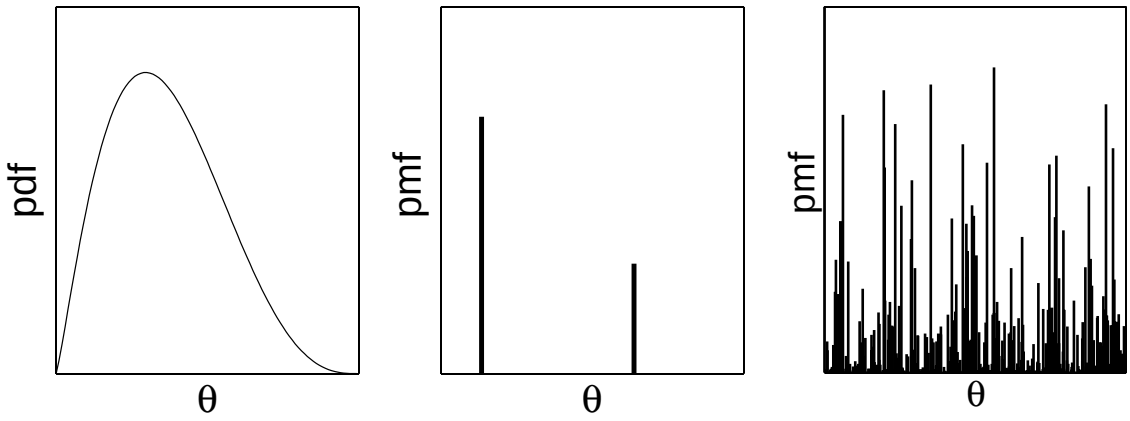
(2) *Posterior over the parameters for the individual differences model.* A second role for $G(\cdot \mid \boldsymbol{\phi})$ is to provide a theoretical account of the variation across the parameters θ_i. In those contexts, the researcher may wish to report the marginal distribution $p(\boldsymbol{\phi} \mid \boldsymbol{x})$. The idea in this case is to learn the structure of individual variation from the data.

In this paper we are interested more in the second case than the first, and it is necessary to distinguish between the two. This is particularly important since stochastic parameter models are generally motivated by the first case, while group models are often applied in the second. This difference in focus is reflected in the fact that, while both stochastic parameter models and group models can be viewed as hierarchical models, they differ in the form of the distribution $G(\cdot \mid \boldsymbol{\phi})$ that describes individual variation. In the stochastic parameters model, $G(\cdot \mid \boldsymbol{\phi})$ is usually a tractable distribution such as a Gaussian, with $\boldsymbol{\phi}$ corresponding to the parameters of that distribution, as in Figure 3a. In contrast, if we have a model with k groups, the individual differences model $G(\cdot \mid \boldsymbol{\phi})$ is a weighted collection of k point masses, as depicted in Figure 3b. That is,

$$G(\cdot \mid \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{z=1}^{k} w_z \, \delta(\cdot \mid \theta_z), \tag{3}$$

where $\delta(\cdot \mid \theta_z)$ denotes a point mass distribution located at θ_z and where $\sum_{z=1}^{k} w_z = 1$. In the groups model, $\boldsymbol{\phi} = (\boldsymbol{w}, \boldsymbol{\theta})$. It is important to notice that in this expression, $\boldsymbol{\theta}$ refers to the locations of the k spikes that make up the distribution $G(\cdot \mid \boldsymbol{w}, \boldsymbol{\theta})$ and are thus parameters of the individual differences model. The parameter values for the individual subjects are then sampled from this distribution, and are all equal to one of these k values. Notationally, we will distinguish between these two uses through the subscripts: θ_i will denote the parameters for the ith participant, while θ_z will denote parameter for group z. If the subscript is ambiguous, we will make it clear in each context.

The hierarchical Bayesian model perspective reveals some of the strengths and weaknesses of the two approaches. Assuming that individual parameters θ_i follow a simple parametric distribution, as in the stochastic parameters model, simplifies the problem of learning an individual differences model from 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, perhaps arising from different interpretations of a task. In this case the model cannot capture the most important aspect of the variation between people. Unimodal distributions naturally suggest an interpretation in terms of variation away from a single prototypical parameter value at the mode, which is misleading in many situations. To return to the unlucky numbers experiment, we might end up estimating a distribution with a mean of about 17, and a large enough variance to capture the performance of the



(a) (b) (c)

Fig. 3. 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). The continuous measure shown in panel a is a probability density function (pdf) while the discrete measures in panels b and c are probability mass functions (pmf).

individual subjects. While this account may provide reasonable estimates of the individual parameters (case 1), it is unsatisfactory as an explanation of these parameters (case 2). We would prefer to recognize that the data here have three distinct modes, located at 4, 13, and 87.

Unlike the stochastic parameters approach, the parameter distributions allowed by group models can naturally account for multimodality in individual differences. By postulating two groups, for instance, we arrive at a bimodal distribution. While this is desirable, given our goal of learning group structure from data, it introduces the problem of how many groups we should include in our individual differences model. In finite group models, this is viewed as a model selection problem. The fixed number of groups k is taken to define a family of individual differences distributions $\mathcal{M}_k$, and we are required to determine which of these families is best for our data. As a result, model selection issues are central to the application of group models to psychological data, and often make statistical inference very difficult computationally. In this paper we explore an *infinite groups models*, which retains the flexibility of the finite groups model but allows straightforward inference. Questions of model selection will still arise, but in a different and more theoretically satisfying fashion.

3 The Infinite Groups Model

Although the infinite groups model has implications for the model selection problem, it is motivated by a more psychological concern with finite group models. The statistical model described in Equation 3 assumes that k is a fixed value, independent of sample size. Such a model requires, rather implausibly, that future subjects will belong to one of the same set of k groups that were observed previously. No provision is made in this model for the idea that, should more data be observed, more groups could be observed. In contrast, we start with the assumption that there are an *infinite* number of latent groups, only some number of which are observed in any finite sample. The consequence is that k is now variable, and can grow with the data.

To build the infinite groups model, we adopt a distribution on individual parameters $\boldsymbol{\theta}$ 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 $G(\cdot \,|\, \boldsymbol{\phi})$ to be a weighted combination of an infinite number of point masses, as in Figure 3c. That is, the individual differences model is assumed to be of the form,

$$G(\cdot \,|\, \boldsymbol{w}, \boldsymbol{\theta}) = \sum_{z=1}^{\infty} w_z \, \delta(\cdot \,|\, \theta_z). \tag{4}$$

Once again, $\delta(\cdot \,|\, \theta_z)$ denotes a point mass distribution located at θ_z, and since the w_z values denote mixture weights, they must sum to 1. 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. This model is psychologically plausible: 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 behavior that has never been seen before.

3.1 Finite-Dimensional Priors

In order to apply Bayesian inference in the hierarchical model defined by Equations 1 and 4, we need to define a prior $\pi(\cdot)$ over the possible individual differences models $G(\cdot)$. A specific individual differences model is defined by the the countably infinite number of elements of $\boldsymbol{w}$ and $\boldsymbol{\theta}$ in Equation 4, where w_z denotes the probability that $G(\cdot)$ assigns to the zth point mass, and θ_z denotes the location of that point mass. In other words, we need a prior over the infinite dimensional space $\mathcal{W} \times \Theta$ that covers the possible values for the parameter vectors $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{\theta} \in \Theta$. To see how we might place a sensible prior on this infinite dimensional space, it is helpful to consider how we might proceed in the

finite case when $G(\cdot)$ consists of only k point masses, and then take the limit as $k \to \infty$. This approach is a standard way of eliciting infinite-dimensional priors (e.g., Neal, 2000; Rasmussen, 2000; Green & Richardson, 2001; Griffiths & Ghahramani, 2005). Note that this procedure does not explicitly derive the prior distribution itself. Rather, it provides a principled motivation for a particular choice of prior.

In a finite groups model with k groups (i.e., Equation 3), a standard prior is

$$\begin{aligned} \theta_z &\sim G_0(\cdot) \\ \boldsymbol{w} \,|\, \alpha, k &\sim \text{Dirichlet}(\cdot \,|\, \boldsymbol{\zeta}). \end{aligned} \tag{5}$$

In this prior, each of the k location parameters θ_z is sampled independently from the *base distribution* $G_0(\cdot)$. This base distribution provides a prior over the kinds of parameter values that are likely to capture human performance in a particular task. Choosing the base distribution is no different to setting a prior in any other Bayesian context, and so this prior should be chosen in the usual way. That said, there are differing views as to what ought to be the 'usual way' (e.g., de Finetti, 1974; DeGroot; 1970; Kass & Wasserman, 1996; Jaynes, 2003), but it is outside the scope of this paper to discuss this debate between subjective and objective views of Bayesian inference. Whatever approach is adopted, the base distribution $G_0(\cdot)$ is not affected when we make the transition from finite models to infinite models.

For our purposes, the relevant part of this prior is the distribution over the weights. Placing a prior over the weights is made difficult by the constraint that they need to sum to 1. Typically, in the finite case, we would use a k-dimensional Dirichlet distribution as a prior over these weights. The general form for a k-dimensional Dirichlet with parameters $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_k)$ is given by,

$$p(\boldsymbol{w} \,|\, \boldsymbol{\zeta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\zeta})} \left(\prod_{z=1}^{k} w_z^{\zeta_z - 1} \right) I(\boldsymbol{w}), \tag{6}$$

where $I(\boldsymbol{w}) = 1$ if the weights $\boldsymbol{w}$ sum to 1, and $I(\boldsymbol{w}) = 0$ otherwise. The Dirichlet distribution is a higher-dimensional version of the Beta distribution, with the Beta distribution corresponding to the case where $k = 2$. The normalizing function $\mathcal{Z}(\boldsymbol{\zeta})$ for the Dirichlet distribution is given by

$$\begin{aligned} \mathcal{Z}(\boldsymbol{\zeta}) &= \int \left(\prod_{z=1}^{k} w_z^{\zeta_z - 1} \right) I(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} \\ &= \frac{\prod_{z=1}^{k} \Gamma(\zeta_z)}{\Gamma\left(\sum_{z=1}^{k} \zeta_z \right)}. \end{aligned} \tag{7}$$

In this expression $\Gamma(y) = \int_0^\infty u^{y-1} e^{-u}\, \mathrm{d}u$ is the standard Gamma function, which generalizes the factorial function: If y is a non-negative integer, then $\Gamma(y+1) = y!$ When the Dirichlet distribution is used as a prior in a finite groups model, it is typical to use a symmetric Dirichlet distribution in which all parameters are equal. The reason for using a symmetric distribution stems from the prior being insensitive to the ordering of the location parameters $\boldsymbol{\theta}$. Since the location parameters are independent of one another, their order (i.e., the value of the index z) is irrelevant. This exchangeability requires that the prior over $\boldsymbol{w}$ be set so that the index is also irrelevant, which is achieved by setting a symmetric prior.

For the purposes of deriving a prior over infinite groups, we will assume that all parameter values ζ_z are set to α/k. The choice of α/k as the parameter value follows from recognizing that the sum of the parameters of a Dirichlet distribution can be interpreted as indicating how heavily to weight the prior. To understand this property of the Dirichlet parameters it may help to consider an example using an idealized bent coin. Suppose that data are produced by n independent flips of a bent coin. We might propose a simple model, in which these are *i.i.d.* Bernoulli trials with an unknown probability p of obtaining a head. The prior we set over this unknown p could be Dirichlet with only $k = 2$ possible outcomes, corresponding to a Beta distribution. Since we do not know which way the coin is bent, the distribution over p should be symmetric. We will set the parameters to $\alpha/2$. If we then observe h heads and $t = n - h$ tails in our data, our posterior distribution is still a Beta, since the Beta family is conjugate [1] to the Binomial likelihood. The parameters of the posterior Beta are $h + \alpha/2$ and $t + \alpha/2$. As a result, the expected posterior value of p is $\bar{p} = (h + \alpha/2)/(n + \alpha)$. From the denominator, it is evident that α is commensurate with n, in terms of its influence on this estimate. This property generalizes to larger k. Our goal here is to specify a prior over an infinite-dimensional outcome space $\mathcal{W}$ that embodies only a limited amount of information, so it is helpful to choose the prior in a way that keeps the amount of information independent of the dimensionality k. The α/k prior achieves this by ensuring that the sum of the parameter vector is always α. For more details on the α/k prior, see Neal (2000) and Ishwaran and Zarepour (2002).

To find the limiting prior as $k \to \infty$, it is helpful to to rewrite the finite-dimensional model in a way that lets us integrate out $\boldsymbol{w}$. To do this, we introduce the group membership variable g_i, indicating the group to which the ith observation belongs. Since w_z gives the probability that the ith observation belongs to the zth group, we can say that $p(g_i = z \,|\, \boldsymbol{w}) = w_z$. With this membership variable introduced, the finite-dimensional model with

[1] A family of prior distributions is conjugate to a particular likelihood function if the posterior distribution belongs to the same family as the prior (e.g., Bernardo & Smith, 2000, pp. 265–285).

this prior becomes,

$$\begin{aligned} x_{ij} \,|\, \boldsymbol{\theta}, g_i = z &\sim F(\cdot \,|\, \theta_z) \\ g_i \,|\, \boldsymbol{w} &\sim \text{Multinomial}(\cdot \,|\, \boldsymbol{w}) \\ \boldsymbol{w} \,|\, \alpha, k &\sim \text{Dirichlet}(\cdot \,|\, \tfrac{\alpha}{k}) \\ \theta_z \,|\, G_0 &\sim G_0(\cdot), \end{aligned} \tag{8}$$

where the multinomial is of sample size one. Since the group assignment variables g_i are conditionally independent given the weights $\boldsymbol{w}$, when we integrate out the weights we induce a conditional dependence between the group assignments. If we have observed the first $i-1$ group assignments $\boldsymbol{g}_{-i} = (g_1, \ldots, g_{i-1})$, we want the conditional probability $p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha, k)$ that is obtained by integrating out $\boldsymbol{w}$. This distribution is given by,

$$p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha, k) = \int p(g_i = z \,|\, \boldsymbol{w}) p(\boldsymbol{w} \,|\, \boldsymbol{g}_{-i}, \alpha, k) \,\mathrm{d}\boldsymbol{w}.$$

We have already seen that $p(g_i = z \,|\, \boldsymbol{w}) = w_z$. By applying Bayes' theorem we observe that the second term is the posterior probability,

$$p(\boldsymbol{w} \,|\, \boldsymbol{g}_{-i}, \alpha, k) \propto p(\boldsymbol{g}_{-i} \,|\, \boldsymbol{w}) p(\boldsymbol{w} \,|\, \alpha, k).$$

Since the first term is a multinomial probability and the second term is Dirichlet, conjugacy implies that the posterior is also Dirichlet. If we let s_z denote the number of previous observations that were assigned to group z, we can use the size vector $\boldsymbol{s} = (s_1, \ldots, s_k)$ to indicate how many observations fall in each group. The posterior probability $p(\boldsymbol{w} \,|\, \boldsymbol{g}_{-i}, \alpha, k)$ is a non-symmetric Dirichlet with the parameter vector $\boldsymbol{s} + \alpha/k$. We can now solve the integral.

$$\begin{aligned} p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha, k) &= \frac{1}{\mathcal{Z}(\boldsymbol{s} + \frac{\alpha}{k})} \int w_z \left(\prod_u w_u^{s_u + \frac{\alpha}{k} - 1} \right) I(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w} \\ &= \frac{\mathcal{Z}\left(\boldsymbol{s} + \frac{\alpha}{k} + \mathbf{1}^{(z)}\right)}{\mathcal{Z}\left(\boldsymbol{s} + \frac{\alpha}{k}\right)} \\ &= \frac{s_z + \frac{\alpha}{k}}{i - 1 + \alpha}. \end{aligned} \tag{9}$$

In this expression, $\mathbf{1}^{(z)}$ is a k-length vector of zeros with a 1 in position z. The last line follows from Equation 7 and the fact that $\Gamma(y+1) = y\Gamma(y)$.

3.2 Extension to Infinite-Dimensional Priors

The finite-dimensional prior can now be extended to the infinite case by letting $k \to \infty$ (see Neal, 2000; Ishwaran & Zarepour, 2002). Consider first the probability that the ith observation falls in a group z that already contains at least one member (i.e. $s_z > 0$). In this case, the limiting probability is

$$\begin{aligned} p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha) &= \lim_{k\to\infty} \left(\frac{s_z + \frac{\alpha}{k}}{i - 1 + \alpha} \right) \\ &= \frac{s_z}{i - 1 + \alpha}. \end{aligned} \tag{10}$$

We now consider the probability that the ith observation falls in one of the infinitely many groups that as yet contain no observations. If there are k_{-i} groups observed among the first $i-1$ observations, and letting $\mathcal{Q}$ denote the set of $k - k_{-i}$ currently empty groups, then the probability that the ith observation belongs to one of them is,

$$\begin{aligned} p(g_i \in \mathcal{Q} \,|\, \boldsymbol{g}_{-i}, \alpha) &= \lim_{k\to\infty} \left(\sum_{u\in\mathcal{Q}} \frac{s_u + \frac{\alpha}{k}}{i - 1 + \alpha} \right) \\ &= \frac{\alpha}{i - 1 + \alpha} \lim_{k\to\infty} \left(\frac{k - k_{-i}}{k} \right) \\ &= \frac{\alpha}{i - 1 + \alpha}. \end{aligned} \tag{11}$$

Notice that α remains commensurate with sample size in the limiting prior (in the derivation above the sample size is $i-1$) and so can be interpreted as a measure of prior information. In the bent coin example, α acted to drag the estimator toward the prior, thereby shaping predictions about future data. In the infinite groups model, large α increases the probability that future data will be drawn from a previously unobserved group. Since new groups have parameter values drawn from the prior $G_0(\cdot)$, larger α increases the influence of the prior. Moreover, since large α values tend to introduce more groups, it can be thought of as a *dispersion* parameter.

The group assignments g_i define a partition of the subjects, with each subject being assigned to a single group. The distribution over partitions induced by taking the limit of a Dirichlet-multinomial model, as we did in Equations 10 and 11, is the same as that induced by a stochastic process called the *Chinese Restaurant Process* (CRP: e.g., Aldous, 1995; Pitman, 1996) with dispersion α. The CRP gets its name from a metaphor based on Chinese restaurants in San Francisco that seem to have limitless seating capacity. In this metaphor, every possible group corresponds to a table in an infinitely large Chinese restaurant. Each observation corresponds to a customer entering the restaurant and sit-

ting at a table. People are assumed to prefer sitting at popular tables (with probability proportional to the number of people already sitting at the table), but it is always possible for them to choose a new table (with probability proportional to α). This gives exactly the conditional distribution over group assignments obtained in Equations 10 and 11, with the joint distribution over group assignments written $\boldsymbol{g} \,|\, \alpha \sim CRP(\cdot \,|\, \alpha)$. The resulting model becomes

$$\begin{aligned} x_{ij} \,|\, \theta_z, g_i = z &\sim F(\cdot \,|\, \theta_z) \\ \boldsymbol{g} \,|\, \alpha &\sim CRP(\cdot \,|\, \alpha) \\ \theta_z \,|\, G_0 &\sim G_0(\cdot). \end{aligned} \tag{12}$$

To complete the motivation of our prior, it is helpful to find the prior distribution over parameter values θ_i, by integrating out the group assignment variables g_u. Since these are just indicator variables, this is straightforward:

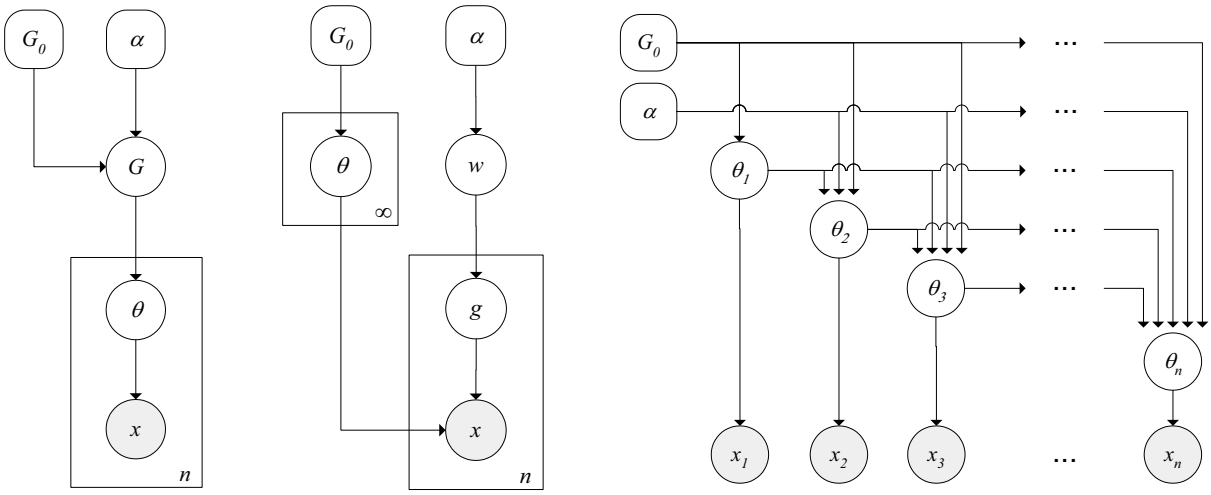
$$\theta_i \,|\, \boldsymbol{\theta}_{-i}, \alpha, G_0 \sim \frac{\alpha}{i-1+\alpha} G_0(\cdot) + \sum_{z=1}^{k_{-i}} \frac{s_z}{i-1+\alpha} \delta(\cdot \,|\, \theta_z). \tag{13}$$

To avoid confusion, it is important to recognize that θ_z denotes the parameter value associated with all the members of group z, whereas $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1})$ denotes parameters assigned to particular observations, as does θ_i. The conditional probability described in Equation 13 is a mixture between the empirical distribution of the $i-1$ previously observed parameters and the base distribution $G_0(\cdot)$.

A sequence of parameter values sampled from Equation 13 is sometimes said to be sampled from a *Pólya urn* (PU: e.g., Blackwell & MacQueen, 1973) parameterized by $G_0(\cdot)$ and α. In a Pólya urn scheme, we imagine an urn full of α colored balls, such that the proportion of balls with color θ is equal to $G_0(\theta)$. We sample θ_1 by drawing a ball at random from the urn and recording its color. We then return the ball to the urn and drop in another ball of the same color, effectively "updating" the urn. Using this Pólya urn formulation to express the induced prior on θ, our model may be written,

$$\begin{aligned} x_{ij} \,|\, \theta_i &\sim F(\cdot \,|\, \theta_i) \\ \theta_1, \ldots, \theta_\infty \,|\, G_0, \alpha &\sim PU(\cdot \,|\, G_0, \alpha). \end{aligned} \tag{14}$$

This description now allows us to select an appropriate infinite-dimensional prior: We want to choose a prior over the individual differences distribution $G(\cdot)$, subject to the constraint that the marginal prior over the set of individual parameters $\theta_1, \ldots, \theta_\infty$ is a Pólya urn scheme. As noted by Blackwell and MacQueen (1973), one prior that meets this requirement is the *Dirichlet process* (DP: Ferguson, 1973).



(a) (b) (c)

Fig. 4. Three graphical representations of a model employing a Dirichlet process prior. Each panel depicts the same model, but shown from a different perspective. Panel a depicts the standard view (Equation 15), panel b shows the Sethuraman construction (Equation 16) and panel c displays the construction via the Pólya urn scheme (Equation 14).

The Dirichlet process is widely used in nonparametric Bayesian statistics as a method for placing priors on infinite mixture models (e.g., Lo, 1984; Escobar & West, 1995; Rasmussen, 2000; Neal, 1996, 2000; Blei, Griffiths, Jordan & Tenenbaum, 2004), and has sometimes been applied in psychometrics as a generic prior over probability distributions in Bayesian Item Response Theory models (e.g., Duncan, 2004; Qin, 1998). This connection between the Dirichlet process and the Pólya urn scheme suggests that our prior on the individual differences distribution should be a Dirichlet process. Having elicited a principled prior, we may now formally specify the infinite groups model in the following way:

$$\begin{aligned} x_{ij} \,|\, \theta_i &\sim F(\cdot \,|\, \theta_i) \\ \theta_i \,|\, G &\sim G(\cdot) \\ G \,|\, G_0, \alpha &\sim DP(\cdot \,|\, G_0, \alpha). \end{aligned} \tag{15}$$

This model is illustrated graphically in Figure 4a. Parameters arise from an unknown distribution $G(\cdot)$, and our uncertainty about this distribution is reflected through the Dirichlet process prior. Grey circles denote observed variables, white circles denote latent variables, and the rounded squares denote parameters whose values are assumed to be known. Plates indicate a set of independent replications of the processes inside them (Buntine, 1994). For comparison, the Pólya urn formulation in Equation 14 is shown in panel c.

4 The Dirichlet Process

In nonparametric problems, the goal is to learn from data without making any strong assumptions about the class of parametric distributions (e.g., Gaussian) that might describe the data. The rationale for the approach is that the generative process for a particular data set is unlikely to belong to any finite-dimensional parametric family, so it would be preferable to avoid making this false assumption at the outset. From a Bayesian perspective, nonparametric assumptions require us to place a prior distribution that has broad support across the space of probability distributions. However, Bayesian nonparametrics are not widely known in psychology (but see Karabatsos, in press), so a brief discussion may be helpful.

The Dirichlet process, now a standard prior in Bayesian nonparametrics, was constructed by Freedman (1963) during a discussion of *tail-free processes*, and the associated statistical theory was developed by Ferguson (1973, 1974). The Dirichlet process represents a partial solution to the problem of Bayesian nonparametric inference, in the sense that it does have broad support, but the sampled distributions are discrete with probability 1 (e.g., Ferguson, 1973; Blackwell, 1973; Sethuraman, 1994; Ghosh & Ramamoorthi, 2003, pp. 102-103). As a result it is often used as a prior over discrete distributions [2] and it is in this capacity that we have used the Dirichlet process in this paper. Since we require an infinite number of variables (the countably infinite elements of $\boldsymbol{w}$ and $\boldsymbol{\theta}$) to describe a sample from the Dirichlet process, it is often referred to as an *infinite-dimensional model*.

4.1 Stick-Breaking Priors

The simplest description of the Dirichlet process is as an example of a *stick-breaking prior* (e.g., Ishwaran & James, 2001; Ishwaran & Zarepour, 2002). This construction was first discussed by McCloskey (1965), and formalized by Sethuraman (1994). In this formulation, we start by noting that discrete distributions can be written

$$G(\cdot) = \sum_{z=1}^{\infty} w_z \, \delta(\cdot \,|\, \theta_z),$$

[2] One reason for the popularity of the Dirichlet process is tractability, since the Dirichlet process is conjugate to *i.i.d.* sampling (Ferguson, 1973). If the prior over $G(\cdot)$ is a Dirichlet process with parameters α and $G_0(\cdot)$, and we observe *i.i.d.* data with empirical distribution $G_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta(\cdot \,|\, \theta_i)$, then the posterior distribution over $G(\cdot)$ is a Dirichlet process with dispersion $\alpha + n$ and base distribution $\frac{\alpha}{\alpha+n}G_0(\cdot) + \frac{n}{\alpha+n}G_n(\cdot)$. However, it is important to note that since the Dirichlet process concentrates on discrete distributions, it can be unsuitable as a prior over densities. For instance, Diaconis and Freedman (1986) provide an example of pathological behavior when the Dirichlet process is used in this way.

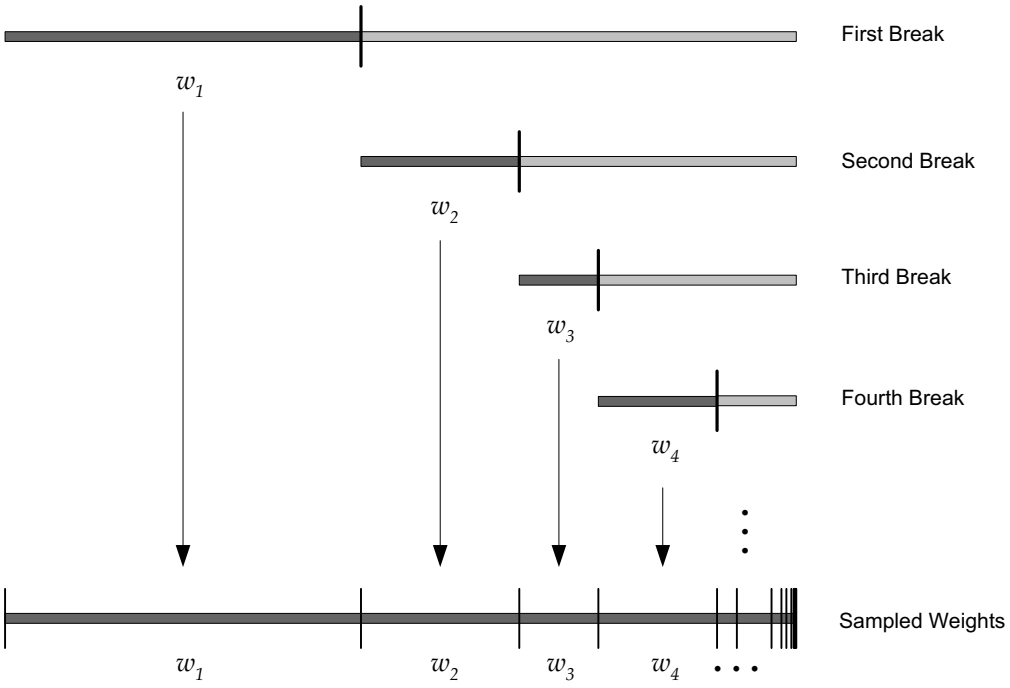


Fig. 5. A graphical depiction of the stick-breaking process, showing successive breaks of a stick with starting length one, and how the lengths of the pieces correspond to sampled weights.

as per Equation 4. Since the distribution can be described by the infinite set of point masses θ_z and the infinite set of weights w_z, this construction specifies two separate priors. As illustrated in Figure 4b, the base distribution places a prior over the locations of the point masses, while the dispersion parameter can be used to place a stick-breaking prior over their weights, denoted $\text{Stick}(1, \alpha)$. In other words, an infinite mixture model that uses a Dirichlet process prior (i.e. Equation 15) can be rewritten,

$$\begin{aligned} x_{ij} \,|\, \theta_1, \ldots, \theta_\infty, g_i = z &\sim F(\cdot \,|\, \theta_z) \\ g_i \,|\, w_1, \ldots, w_\infty &\sim \text{Multinomial}(\cdot \,|\, w_1, \ldots, w_\infty) \\ w_1, \ldots, w_\infty \,|\, \alpha &\sim \text{Stick}(\cdot \,|\, 1, \alpha) \\ \theta_z \,|\, G_0 &\sim G_0(\cdot), \end{aligned} \tag{16}$$

where the multinomial distribution in the second line is of sample size 1. The stick-breaking process can be illustrated in the following way. Imagine we started with a stick of length 1, broke it in two, and took the length of one of the pieces to be the first weight. We then broke the remaining piece in two, using one of the resulting pieces as the second weight. This process continues for a countably infinite number of breaks, as illustrated in Figure 5, and results in an infinite set of stick-lengths that sum to 1 with probability 1. More formally, at each step of the process the proportion of the stick w'_z that is broken

off follows a Beta distribution [3], so that

$$w'_z \mid \alpha \sim \text{Beta}(\cdot \mid 1, \alpha).$$

Thus, the length of the zth stick fragment is given by

$$w_z = w'_z \prod_{u=1}^{z-1} (1 - w'_u).$$

A nice property of the stick-breaking construction is that it allows us to draw approximate samples from the Dirichlet process, by sampling the values of w_z from the stick-breaking process until the sum of the observed values is sufficiently close to 1. Having done so, we then sample the corresponding θ_z values independently from $G_0(\cdot)$, and treat the resulting (sub)probability distribution as an approximation to $G(\cdot)$. By doing this, we can get a sense of what these distributions look like. Figure 6 shows three distributions sampled from three different choices of $G_0(\cdot)$, and a dispersion parameter value of $\alpha = 100$ in each case. As is immediately apparent, the base distribution places a prior on the shape of $G(\cdot)$. By way of comparison, Figure 7 shows a number of distributions sampled from a Dirichlet process with a uniform distribution over $[0, 1]$ as the base distribution $G_0(\cdot)$, and dispersion parameters of $\alpha = 100$ (top row), $\alpha = 20$ (middle row), and $\alpha = 5$ (bottom row). It illustrates the manner in which smaller values of α tend to concentrate the distribution on fewer values of θ_z.

4.2 Learning the Dispersion of Data

A difficulty with Dirichlet process models, noted by Antoniak (1974), is that it is usually too restrictive to specify a value of α *a priori*. The dispersion parameter reflects the degree of variability in the parameter values, and is something we would prefer to learn from data. In order to do so, we first need to understand the relationship between the dispersion α and the number of groups k that will manifest among n subjects. Note that k now refers not to the 'true' number of groups, but to the number of manifest groups. Antoniak (1974) shows that the probability $p(k \mid \alpha, n)$ that k groups will be observed in n samples from a model with a Dirichlet process prior is

$$\begin{aligned} p(k \mid \alpha, n) &= \frac{n! \Gamma(\alpha)}{\Gamma(\alpha + n)} z_{nk} \alpha^k \\ &= n \text{B}(\alpha, n) z_{nk} \alpha^k \end{aligned} \tag{17}$$

[3] In the more general class of stick-breaking priors the parameters of the Beta variate can vary across breaks, with the zth Beta distribution having parameters a_z, b_z.

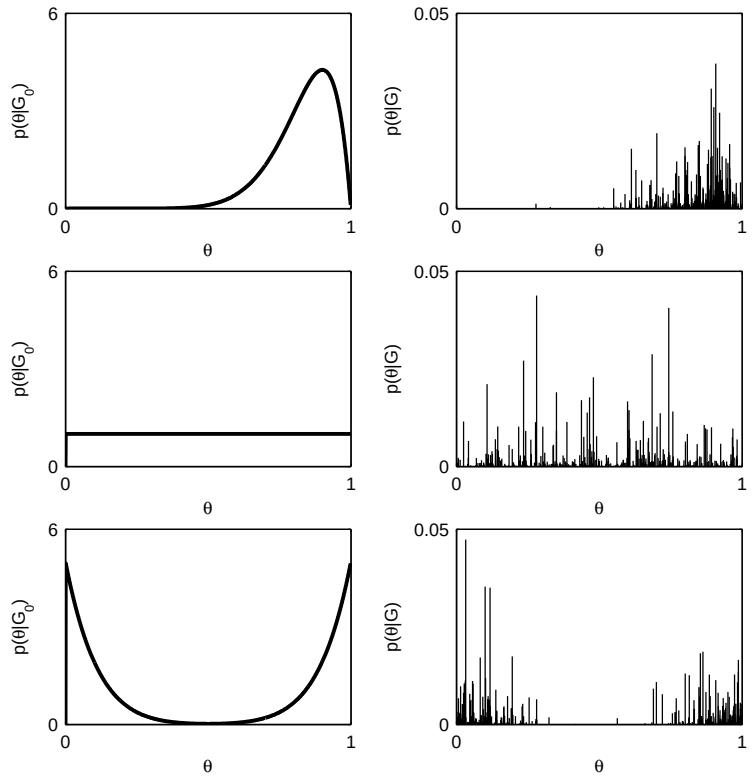


Fig. 6. Distributions sampled from a Dirichlet process with $\alpha = 100$, and three different base distributions $G_0(\cdot)$. Base distributions are shown on the left, and sampled distributions are shown on the right. In the top line, the base distribution is $\text{Beta}(\cdot \mid 10, 2)$, while in the middle row it is a uniform $\text{Beta}(\cdot \mid 1, 1)$, while in the bottom row it is an equal mixture of a $\text{Beta}(\cdot \mid 10, 1)$ and a $\text{Beta}(\cdot \mid 1, 10)$.

$$\propto z_{nk} \alpha^k,$$

where $\text{B}(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)} = \int_0^1 \eta^{u-1}(1-\eta)^{v-1} d\eta$ is a standard Beta function and z_{nk} is an unsigned Stirling number of the first kind. The unsigned Stirling numbers count the number of permutations of n objects having k permutation cycles (Abramowitz & Stegun, 1972, pp. 824), and are found by taking the absolute value of the corresponding signed Stirling numbers $z_{nk} = |s_{nk}|$. There is no analytic expression for s_{nk}, but it is easily calculated using the recurrence formula $s_{nk} = s_{n-1,k-1} - (n-1)s_{n-1,k}$, and the special cases $s_{nn} = 1$ for all n and $s_{n0} = 0$ for $n > 0$. Note that the use of s and z in this notation is unrelated to the previous use as the group sizes and indices (the two uses will not come into conflict). Antoniak (1974) also observes that the expected number of components sampled from a Dirichlet process is given by,

$$E[k \mid \alpha, n] = \sum_{k=1}^{n} k \, p(k \mid \alpha, n)$$

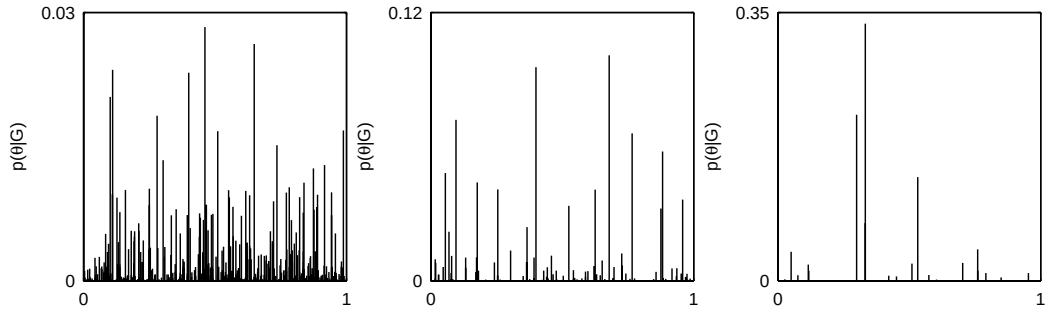


Fig. 7. Distributions sampled from a Dirichlet process with a uniform distribution over $[0, 1]$ as the base distribution $G_0(\cdot)$, and dispersion parameters of $\alpha = 100$ (left), $\alpha = 20$ (middle), and $\alpha = 5$ (right). In all cases there are a countably infinite number of components (most of which are too small to see), but the distributions vary in the extent to which the probability mass is concentrated on a few points.

$$
\begin{aligned}
&= \alpha \sum_{k-1}^{n} \frac{1}{\alpha + k - 1} \\
&\approx \alpha \ln \left(\frac{n + \alpha}{\alpha} \right).
\end{aligned} \tag{18}
$$

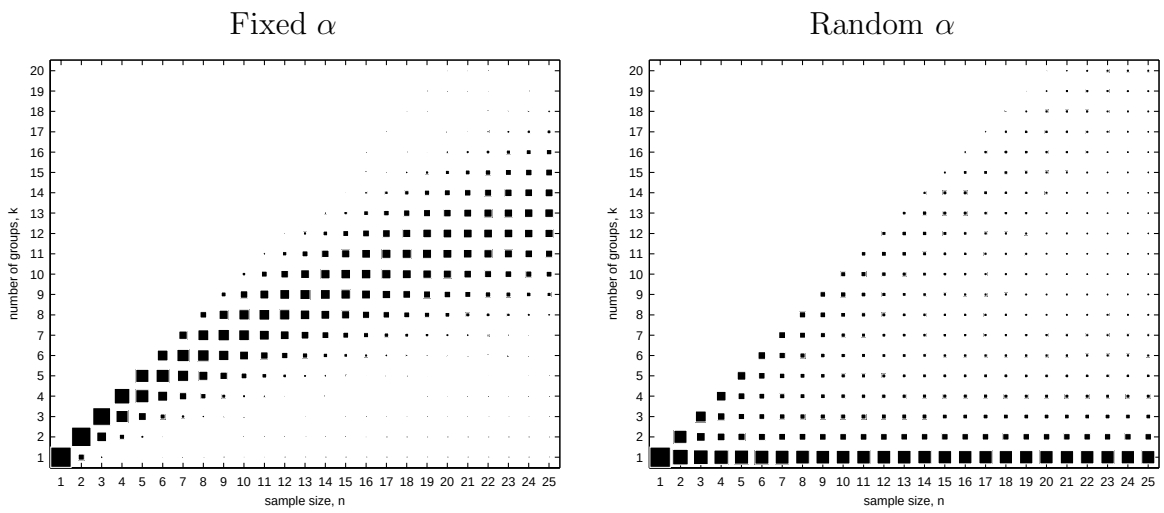
Thus, although $k \to \infty$ with probability 1 as $n \to \infty$ (Korwar & Hollander, 1973), the number of components increases in approximately logarithmically with the number of observations. This is illustrated in Figure 8a, which shows how the prior over the number of components grows changes as a function of n, for a Dirichlet process with $\alpha = 10$.

In many contexts the dispersion α is unknown, so we specify a prior distribution $p(\alpha)$, allowing us to learn α from data. The resulting model is known as a *Dirichlet process mixture*. Antoniak (1974) notes that the the posterior distribution for α is influenced only by the number of distinct groups k, and not by the details of the allocation of observations to those groups. Therefore, since $p(k \mid \alpha, n)$ provides the likelihood function for k, we can apply Equation 17 to find the posterior distribution over α given some observed data containing k groups. Since the prior on α is not dependent on the sample size n, we may write,

$$
\begin{aligned}
p(\alpha \mid k, n) &\propto p(k \mid \alpha, n)\, p(\alpha \mid n) \\
&= p(k \mid \alpha, n)\, p(\alpha) \\
&\propto \mathrm{B}(\alpha, n)\, \alpha^k\, p(\alpha).
\end{aligned} \tag{19}
$$

A common choice for $p(\alpha)$ is the (inverse) Gamma distribution $\alpha \mid a, b \sim \mathrm{Gamma}(\cdot \mid a, b)$ in which $p(\alpha) \propto \alpha^{a-1} e^{-b\alpha}$ (Escobar & West, 1995). If so, the posterior distribution becomes,

$$
p(\alpha \mid k, n) \propto \alpha^{a+k-1} e^{-b\alpha}\, \mathrm{B}(\alpha, n). \tag{20}
$$



(a) (b)

Fig. 8. Prior distributions over the number of components k for sample sizes n ranging from 1 to 25. The panel on the left shows the prior for a Dirichlet process with dispersion $\alpha = 10$, where the area of each rectangle is proportional to the probability associated with the corresponding value of k for a given n. The panel on the right shows the marginal prior over k for a Dirichlet process mixture in which the prior over α is an inverse Gamma$(\cdot \,|\, 10^{-10}, 10^{-10})$ distribution.

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 $p(\alpha) \propto 1/\alpha$ (e.g., Jeffreys, 1961; Kass & Wasserman, 1996). However, since this Gamma$(\cdot \,|\, 0, 0)$ prior is improper, we have chosen to approximate it with the proper but extremely similar prior Gamma$(\cdot \,|\, 10^{-10}, 10^{-10})$. Figure 8b shows the marginal prior over k resulting from this choice of prior.

5 Model Selection With Infinite Groups

One benefit to the infinite groups model is the principled perspective that it provides on the model order selection problem. Since model order selection problems are commonplace in psychological modeling (e.g., Landauer & Dumais, 1997; Griffiths & Steyvers, 2004; Lee, 2001; Lee & Navarro, 2005) it is worth discussing this point in a little more detail.

When working with finite models, it is natural to think of k as the intrinsic model order. Every value of k describes a different family of distributions in Equation 3, and so it is easy to think of k as defining a model $\mathcal{M}_k$ consisting of all discrete distributions that consist of exactly k point masses. This means that, when inferring a finite group model to account for individual differences, we need to address the model selection question of choosing a model $\mathcal{M}_k$, and a parameter estimation problem in which we pick a distribution

$G(\cdot) \in \mathcal{M}_k$. From a Bayesian standpoint (e.g., Wasserman, 2000) we would find a posterior distribution over the models $p(\mathcal{M}_k \mid \boldsymbol{x})$ and use this to draw our inference about k. In order to find this posterior, we need a prior distribution $p(\mathcal{M}_k)$. However, since it is not easy to see how this prior might be chosen, it is quite common to use Bayes factors (e.g., Kass & Raftery, 1995). This corresponds implicitly to the use of a uniform prior over model orders, which may not be appropriate. [4] It seems unlikely, for example, that experimental data from 40 subjects—thus requiring the consideration of model orders 1, 2, ..., 40—is equally as likely to contain 23 different groups of subjects as it is to contain two different groups of subjects.

The infinite groups model takes a different view. By assuming that the distribution $G(\cdot)$ has an infinite number of groups, we no longer have any model classes to select between. In this framework, we view k as the *variable* expression of $G(\cdot)$ through finite data. When we set a prior in this approach, it is over the distributions themselves: A prior that we have derived from basic considerations about the structure of the model. This, in turn, *implies* a prior over k that reflects the rate at which new groups are expected to appear when sampling from $G(\cdot)$. At no point do we need to specify artificial model classes. Moreover, since the natural way to think about inference is to do posterior sampling over $G(\cdot)$, the number of observed groups k will emerge in inferring $G(\cdot)$, rather than via a dedicated model selection procedure.

6 Modeling Discrete Data With Infinite Groups

We now turn to the specification and application of the infinite groups model to situations in which subjects provide discrete data. Suppose that n people perform some task in which m possible responses can be made on each trial, and the ith person experiences r_i trials. We will specify a simple cognitive model in which there is a multinomial distribution with parameter vector $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{im})$ for the behavior of participant i. In this situation, the natural way to describe data from the ith participant is with the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$, in which x_{ih} counts the number of times that participant i made response h. Note that this is a slight change from the previous notation, since $\boldsymbol{x}_i$ is now a vector of counts rather than a list of the outcomes for every trial. Since our cognitive model is multinomial, it is natural to use a Dirichlet distribution as the prior over $\boldsymbol{\theta}_i$. Specifically, we will assume a symmetric Dirichlet base distribution with parameter β. This cognitive model, including

[4] Note that Lee and Webb's (in press) approach to finite group selection is a little different to standard model order selection. Rather than placing an implicit uniform prior over k, they use an implicit uniform prior over the possible partitions of n subjects.

the prior, is written

$$\begin{aligned}\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_i &\sim \text{Multinomial}(\cdot \,|\, \boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i \,|\, \beta &\sim \text{Dirichlet}(\cdot \,|\, \beta).\end{aligned}$$

If we now assume that each person belongs to one of an infinite number of latent groups, we would incorporate an individual differences model by assuming that each parameter value $\boldsymbol{\theta}_i$ is drawn from some discrete distribution $G(\cdot)$, and use the Dirichlet process to place a prior over these distributions. However, since we do not wish to make strong assumptions about the dispersion parameter α, we use the Dirichlet process mixture model in which we assume that α follows an inverse Gamma distribution. If we write this model using the stick-breaking notation (as in Equation 16), we obtain the model

$$\begin{aligned}\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\infty, g_i = z &\sim \text{Multinomial}(\cdot \,|\, \boldsymbol{\theta}_z) \\ g_i \,|\, w_1, \ldots, w_\infty &\sim \text{Multinomial}(\cdot \,|\, w_1, \ldots, w_\infty) \\ w_1, \ldots, w_\infty \,|\, \alpha &\sim \text{Stick}(\cdot \,|\, 1, \alpha) \\ \alpha \,|\, a, b &\sim \text{Gamma}(\cdot \,|\, a, b) \\ \boldsymbol{\theta}_z \,|\, \beta &\sim \text{Dirichlet}(\cdot \,|\, \beta),\end{aligned} \tag{21}$$

where the multinomial in the second line is of sample size 1. The model is illustrated in Figure 9. Performing inference in this infinite groups model using the mixture of Dirichlet processes prior means being able to estimate the joint posterior distribution $p(\boldsymbol{g}, \boldsymbol{\theta}, \alpha \,|\, \boldsymbol{x}, a, b, \beta)$. A straightforward Gibbs sampling scheme for drawing samples from this posterior distribution is presented in the Appendix. From these posterior samples we can construct estimates of the posterior distribution itself using some density estimation technique (e.g., Hastie, Tibshirani & Friedman, 2001, pp. 182–190).

Using this model to make inferences from data there are several marginal posteriors that are of particular interest, corresponding to different theoretical questions. Some examples include:

(1) *How many groups?* This question corresponds to the model order selection problem, by asking how many groups are manifest in the data. To answer this, we want to know $p(k \,|\, \boldsymbol{x})$, the posterior probability that there are k distinct groups in the sample $\boldsymbol{x}$. Notice that this is a property of the observed data, not an inference about a population parameter.

(2) *How dispersed is the population?* The complementary question to (1) is to ask how groups might be distributed in the population. Of course, in an infinite population it is not sensible to ask how many groups exist. The relevant distribution is $p(\alpha \,|\, \boldsymbol{x})$,

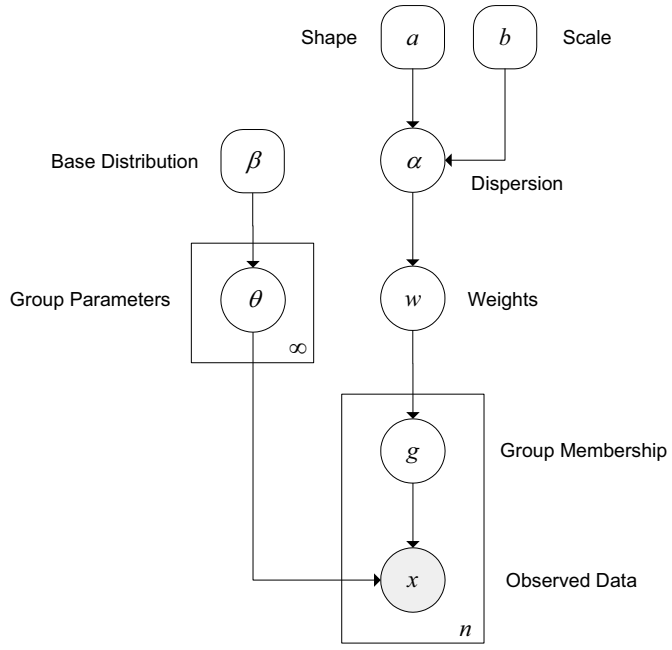


Fig. 9. Dependencies in the infinite groups model for discrete data as it is used here. Shaded circles denote observed variables, white circles are latent variables, rounded squares denote known parameter values, and plates indicate a set of independent replications of the processes shown inside them.

the posterior distribution over the dispersion parameter. If most subjects fall into a single group, then the posterior over α will place most mass on small values, since the population is unlikely to be highly dispersed.

(3) *What are the groups?* Clearly, in drawing inferences about the sample we want to know not just how many groups there are, but also which people tend to belong to the same groups. In this case, we want to know $p(\boldsymbol{g} \mid \boldsymbol{x})$. In some cases, we might want to find the *maximum a posteriori* (MAP) estimate for the group structure, namely $\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} p(\boldsymbol{g} \mid \boldsymbol{x})$. Alternatively, we might aim to get a sense for the full distribution $p(\boldsymbol{g} \mid \boldsymbol{x})$ by finding specific groupings that consistently appear in the posterior.

(4) *What performance characterizes a group?* The original motivation for proposing a groups model was to learn which subjects could be characterized in the same way. Having inferred that some subjects belong to the same group, we would like to know what parameter values of the cognitive model describe their performance. In this case, we want to know $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{g})$, or some other summary measure for this distribution such as $E[p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{g})]$.

To provide a simple illustration of the performance of the model in the context of the first question "*how many groups?*", we created random data sets with $n = 100$ people and $r = 100$ discrete observations per person, where each observation denotes a choice of

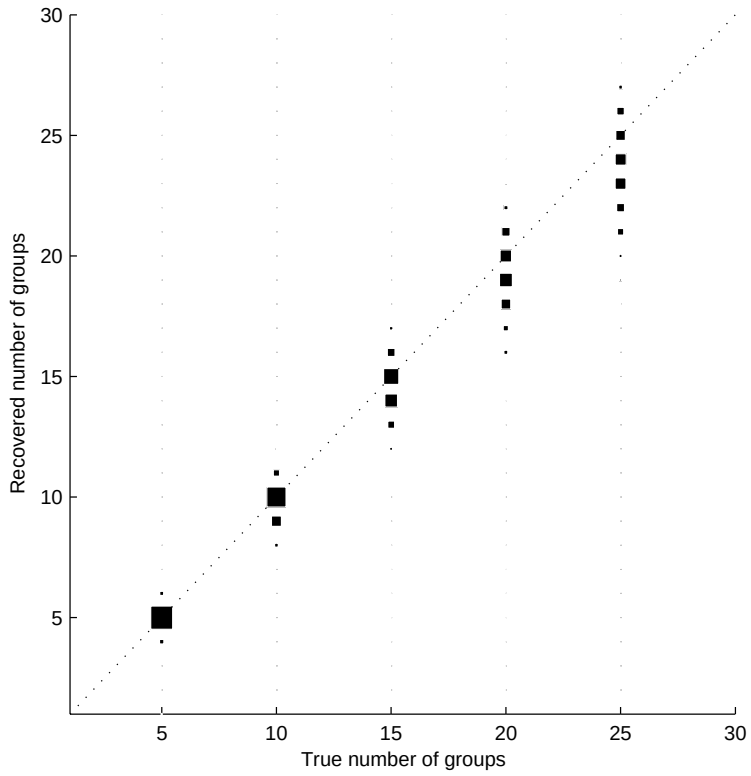


Fig. 10. Simulations in which $n = 100$ people provide $s = 100$ observations each, and $m = 20$ response options are possible on every trial. The true number of groups varies from 5 to 25. After a burn-in of only 500 samples, and using only a single draw from the posterior distribution, the Gibbs sampler performs reasonably well.

one of $m = 20$ 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, subject to the constraint that each group contained at least one member. The number of groups in the data varied from 5 to 25, with 500 data sets generated for each. For each data set, we ran the Gibbs sampler for 500 iterations and then drew a single sample from the posterior distribution. Figure 10 plots the distribution over the recovered number of groups as a function of the true number of groups represented in the data. Inspection of this figure shows that, for the most part, the Gibbs sampler recovers the appropriate number of groups in the data. There is a slight tendency to underestimate the number of groups in some cases, but as argued by Kontkanen et al. (2005), this is not undesirable behavior when extracting a partition, since it generally reflects "different" groups with parameter values so similar that they cannot be distinguished without much more data. Erring on the side of simpler models would appear to be the right thing to do in this situation.

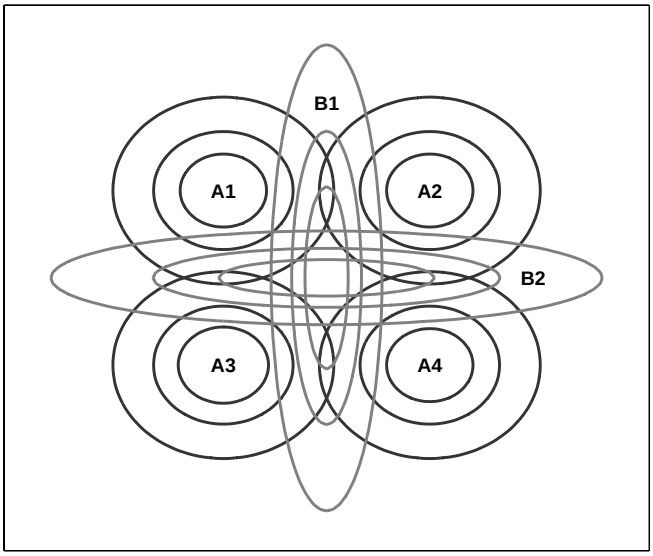


Fig. 11. The category densities used in McKinley and Nosofsky's (1995) experiment 2. Category A (dark grey) is a mixture of four Gaussians, while category B (light grey) is a mixture of two Gaussians. The 30%, 60% and 90% confidence ellipses are shown for each of the six densities.

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[5] ability to discriminate between the two probabilistic categories shown in Figure 11. The stimuli were circles with a radial line running through them, and so the two dimensions depicted in Figure 11 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 (light grey) is a mixture of two Gaussians. On any given trial in the experiment, a stimulus was sampled from one of the six Gaussian distributions. Subjects 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 subjects, we applied the infinite groups model to the data from this experiment. In doing so, we were interested in how the sub-

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

jects' classification performance varied as a function of the *source*. For the ith participant we obtain the data vector $\boldsymbol{x}_i = \left(x_i^{(A_1)}, x_i^{(A_2)}, x_i^{(A_3)}, x_i^{(A_4)}, x_i^{(B_1)}, x_i^{(B_2)}\right)$ in which $x_i^{(l)}$ records the number of correct responses to stimuli generated from distribution l. We are also given a vector of sample sizes, $\boldsymbol{r}_i = \left(r_i^{(A_1)}, r_i^{(A_2)}, r_i^{(A_3)}, r_i^{(A_4)}, r_i^{(B_1)}, r_i^{(B_2)}\right)$ indicating how many trials of each type appeared in each subjects' data. The natural thing to model is the probability of making the correct response to stimuli sampled from each of the six components. So the model would model would predict that for the ith participant, $p(\text{Correct} \,|\, \text{Sample from } A_1) = \theta_i^{(A_1)}$. The cognitive model therefore describes binomial distributions, so that if the ith participant belongs to group z,

$$x_i^{(l)} \,|\, r_i^{(l)}, \theta_z^{(l)}, g_i = z \sim \text{Binomial}\left(\cdot \,|\, \theta_z^{(l)}\right)$$

for all $l \in (A_1, A_2, A_3, A_4, B_1, B_2)$, and where the binomial is of sample size $r_i^{(l)}$. Group z would therefore have the parameter vector $\boldsymbol{\theta}_z = \left(\theta_z^{(A_1)}, \theta_z^{(A_2)}, \theta_z^{(A_3)}, \theta_z^{(A_4)}, \theta_z^{(B_1)}, \theta_z^{(B_2)}\right)$, where each element of this vector is a binomial rate. The fact that we have specified a multidimensional space for $\boldsymbol{x}$ and $\boldsymbol{\theta}$ has no bearing on the stick-breaking prior over $\boldsymbol{w}$, so it is still appropriate to write the infinite discrete groups model as

$$\begin{aligned}
g_i \,|\, w_1, \ldots, w_\infty &\sim \text{Multinomial}(\cdot \,|\, w_1, \ldots, w_\infty) \\
w_1, \ldots, w_\infty \,|\, \alpha &\sim \text{Stick}(\cdot \,|\, 1, \alpha) \\
\alpha \,|\, a, b &\sim \text{Gamma}(\cdot \,|\, a, b).
\end{aligned}$$

The only modification that we need to make is to specify a multidimensional base distribution $G_0(\cdot)$. To do so, we assume that each of the binomials has the same symmetric Beta prior, implying that

$$\theta_z^{(l)} \sim \text{Beta}(\cdot \,|\, \beta),$$

for all $l \in (A_1, A_2, A_3, A_4, B_1, B_2)$.

For each of the 10 subjects 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 (see Jaynes, 2003, pp. 382–386), which is obtained by setting $\beta = 1$, and since we have no strong beliefs about α we would like a scale-invariant prior (see Jeffreys, 1961) in which $a \to 0$, $b \to 0$. Once again, in order to ensure a proper prior, we chose $a = b = 10^{-10}$ as a compromise between ignorance and propriety. To approximate the posterior distribution over α, k and other

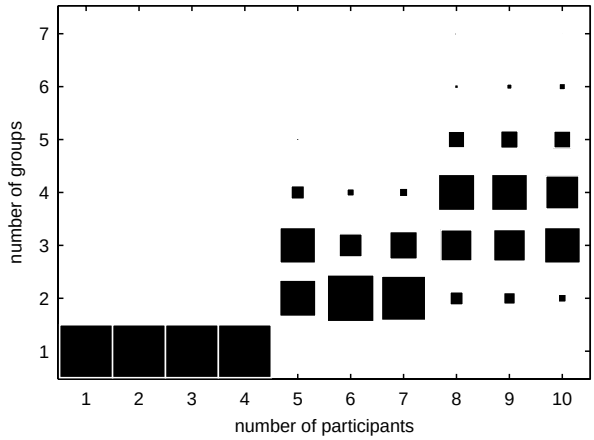


Fig. 12. Estimated posterior over k as a function of n. Assuming subjects arrive in a fixed order, from the first to the tenth person, we can see that the number of inferred groups changes as more people are observed. The area of the squares is proportional to the posterior probability of k given n.

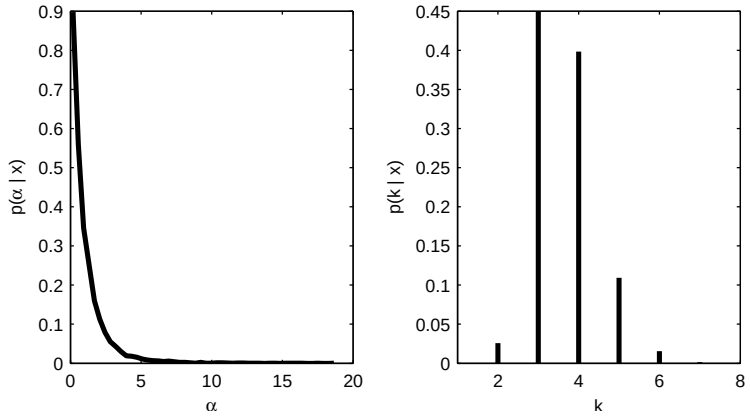


Fig. 13. Estimated posterior distributions over α and k when the infinite groups model is applied to McKinley and Nosofsky's (1995) experiment 2.

relevant parameters, we used the Gibbs sampler to draw 10,000 samples from the joint posterior distribution (after an initial burn-in period of 1,000 iterations), with a lag of 5 iterations between samples to minimize autocorrelation between samples.

We first consider the question of selecting the model order (*"how many groups?"*) by examining how the distribution $p(k \mid \boldsymbol{x})$ changes as a function of n. To do this, we imagine that the 10 subjects entered the lab in order of participant ID. Figure 12 shows how the posterior distribution over k changes as more subjects are observed: the model grows with the data. 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. The last of these posterior

Table 1
Estimated probability with which subjects in McKinley and Nosofsky's (1995) experiment 2 belong to the same group. For visual clarity, the probabilities are given as percentages.

	1	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

distributions (in the case when $n = 10$) is illustrated in Figure 13b. We can also use our samples to ask about the amount of variability that we believe exists in the population (*"how much dispersion?"*). Figure 13a shows the estimated posterior density $p(\alpha \mid \boldsymbol{x})$, which indicates a strong preference for smaller values of α. However, with such a small sample size, this distribution still reflects the near-ignorance prior that we chose for this analysis.

Turning now to the third question (*"what are the groups?"*), the small number of subjects allows us to present a nice summary of the behavior of the posterior distribution $p(\boldsymbol{g} \mid \boldsymbol{x})$. To do so, Table 1 shows the estimated (marginal) posterior probability that any two subjects belong to the same group. This table reveals a rich pattern of similarities and differences, indicating that the relationships between subjects is not arbitrary. To illustrate this, we turn to a characterization of the groups themselves (*"what performance?"*). For these data, it is most informative to plot some of the raw data rather than report parameter values, because the data have a natural two-dimensional graphical structure while the parameters are naturally six-dimensional. Figure 14 plots the last 300 stimuli observed by subjects 5, 7, 8 and 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 subjects 8 and 9 do a good job of learning the category structures on both dimensions. As a result, subjects 5 and 7 rarely appear in the same group as one another or with subjects 8 and 9 (with probabilities ranging from 0% to 7%), while subjects 8 and 9 almost always (91%) co-occur. In other words, the relational structure implied by Table 1 reflects the qualitative individual differences that are apparent from visual inspection of Figure 14.

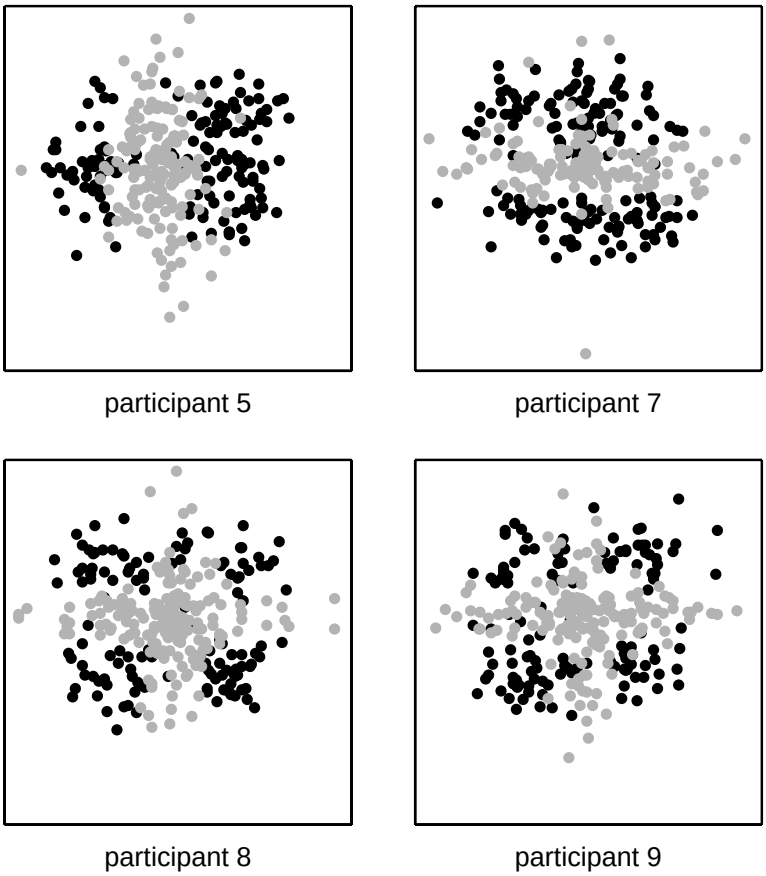


Fig. 14. Last 300 trials for subjects 5, 7, 8 and 9 in McKinley and Nosofsky's (1995) experiment 2. Black dots denote "A" responses, and grey dots denote "B" responses.

8 Individual Differences Among Psychologists

Another application of the infinite groups model regards the publication habits of psychologists. As an initial investigation, we took the publication lists posted on the websites of staff in psychology departments at the following six institutions: Boston College, Cardiff University, Johns Hopkins University, The University of Edinburgh, Florida Atlantic University and Colorado State University. This yielded a total of 125 academics publishing in 254 outlets[6]. Since most academics list only recent or selected publications, the data represent a subset of publication behavior that people prefer to announce. The distribution of the number of listed publications per academic was highly skewed (the skewness was 5.25), with a median value of 7 and an interquartile range of 10.5.

[6] The original data set contained 508 outlets, but half of them were missing from the analyzed data set due to a corrupted file.

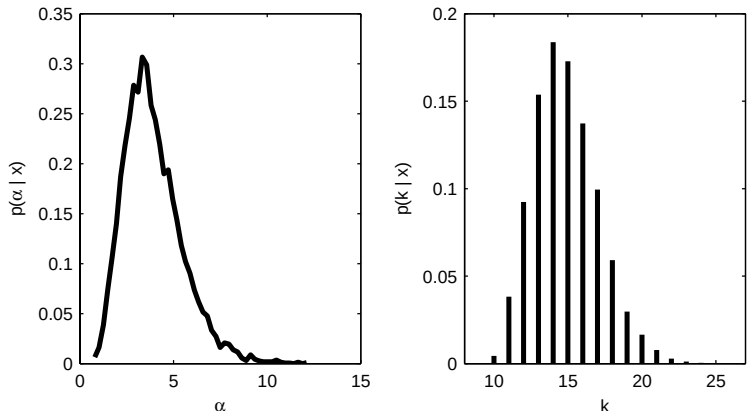


Fig. 15. Estimated posterior distributions over α and k when the infinite groups model is applied to the publications data.

Using the infinite groups model, we would like to learn the patterns of similarity and difference in declared publication preferences among these authors. The model in this case is straightforward version of the usual infinite groups model, with a single group corresponding to a multinomial distribution over the 254 outlets. Again, we start only with the belief that any author is able to publish in any outlet, implying that $\beta = 1$. Not knowing anything *a priori* about the dispersion, we set the prior by setting $a = b = 10^{-10}$. After an initial burn-in period of 1000 samples to ensure that the Gibbs sampler had converged, we drew 10,000 samples from the posterior distribution over groups, with a lag of 5 iterations between samples. The resulting posterior distributions over α and k are shown in Figure 15, and suggest that there are most likely between 12 and 17 groups in these data.

One difficulty with the analysis of these data is that the full posterior distribution over group assignments cannot be displayed easily. In order to provide insight into the structure that the infinite groups model extracts from these data, we undertook the following analysis. We took a set of ten successive samples (again, with a lag of five) from the Markov chain used to produce Figure 15, and averaged across the corresponding ten partitions to find an estimate for the expected probability with which each academic belongs to each group. In order to interpret the groups, we can list the names of the people that are expected to belong to them. Alternatively, we can find the "prototypical performance" associated with each group. In this case, we can calculate the expected probability with which a group member publishes in a particular journal. A simple way of interpreting the groups is to provide a list of typical journals for each group, since journal names are highly informative, whereas the author names often are not.

Note that this is a rather different analysis to the one we would obtain if we partitioned the journals themselves. In this case we are interested in groups of people, and measure their common behavior in terms of the journals they publish in. This does not necessarily

Table 2
Prototypical journals for the five most prominant author-clusters. The rankings are based on data that are normalized for the base rates of both journals and authors. All five represent structures that are found across most of the posterior distribution.

1. Cognitive Psychology
Journal of Experimental Psychology: Learning, Memory & Cognition
Journal of Experimental Psychology: Human Perception & Performance
Brain & Language
Perception & Psychophysics
Quarterly Journal of Experimental Psychology

2. Behavioral Psychology
Journal of Experimental Psychology: Animal Behavior Processes
Quarterly Journal of Experimental Psychology
Journal of the Experimental Analysis of Behavior
Behavioral Neuroscience
Applied Ergonomics

3. Social Psychology
Journal of Personality & Social Psychology
Personality & Social Psychology Bulletin
Cognition & Emotion
Social Cognition
The Behavioral & Brain Sciences

4. Developmental Psychology
Developmental Psychology
Journal of Experimental Child Psychology
Developmental Review
Infant Behavior and Development
Learning & Individual Differences

5. Medicine & Differential Psychology
Personality & Individual Differences
Intelligence
Diabetic Medicine
British Medical Journal
Diabetes Care

Table 3
Categories for the MSNBC.com web pages.

1. Front Page	7. Miscellaneous	13. Summary
2. News	8. Weather	14. Bulletin Board Service
3. Technology	9. Health	15. Travel
4. Local	10. Living	16. MSN-News
5. Opinion	11. Business	17. MSN-Sport
6. On-Air	12. Sports	

produce partitions of journals, however, since multiple groups of people may use the same journal. In short, the idea is we want groups of authors because we are interested in their individual differences: in this analysis, journal usage is the "parameter" for a group of authors, rather than the other way around. We should also mention the reason for using a small number of nearby samples. In this analysis, we want to partially preserve the autocorrelation between samples. This is because the full posterior distribution is likely to be multimodal, and while "local averaging" across samples from the same peak is likely to be beneficial, averaging across samples from different peaks would likely corrupt the analysis. We repeated this procedure across a large number of randomly chosen locations in the Markov chain, and looked for stable clusters of authors, defined as those that produced strong agreement in the rank ordering of journals.

Table 2 shows the top five journals for the five most prominant groups found in the posterior distribution. As indicated by the labels, four of the five groups have a very natural interpretation in terms of sub-fields within the discipline, namely cognitive, behavioral, social and developmental psychology. The fifth group contains journals that are representative of both medical research (e.g., *Diabetic Medicine* and *British Medical Journal*) and differential psychology (e.g., *Intelligence* and *Personality & Individual Differences*). While there is a possibility that this reflects a broader correlation in the interests of psychologists, it seems more likely that this cluster results from the multiple interests of some members of our sample.

9 Individual Differences in Web Browsing

The final application considers the behavior of 1000 people browsing on MSNBC.com and news-related portions of MSN.com on September 28, 1999. Rather than record every webpage viewed, each page is classified using one of the 17 categories listed in Table 3, such as "news", "technology" and "health". For every user the data count the number of times they visited pages belonging to each of the categories. The number of webpages that belonged to each category varied from 10 to 5000. This data set is taken from a much larger public database that records the behavior of all 989,818 (anonymous) users

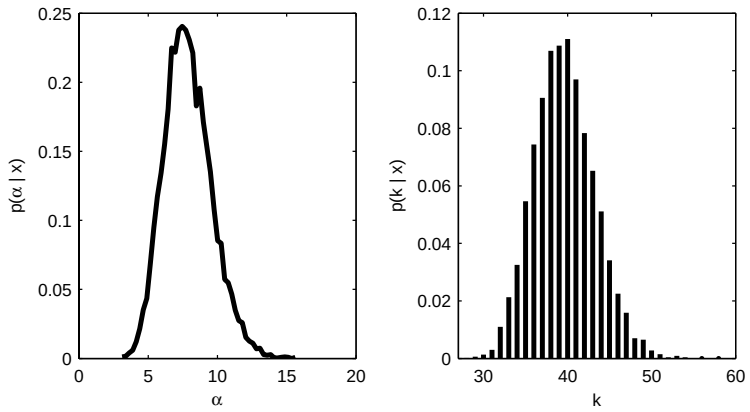


Fig. 16. Estimated posterior distributions over α and k when the infinite groups model is applied to the web data.

that visited MSNBC.com on that day, previously analyzed in some detail by Cadez et al. (2003).

One reason for considering these data is that they represent the unconstrained behavior of people engaged in a natural task. The analysis of large, natural data sets is not a standard approach in cognitive psychology which has traditionally been dominated by the experimental method. Although generally effective, this approach tends to restrict the domain of psychology to simplistic, often implausible contexts. As a complementary approach, analyzing large data sets collected in rich environments provide a reflection of real-world behavior and decision-making. By applying cognitive models in such contexts, we may obtain insights that are not easily obtained in the laboratory.

To analyze these data using the infinite groups model, we group visitors to the site by the frequencies with which they visit each of the 17 categories of websites. Once again, the cognitive model is a multinomial over the 17 categories, and we want to find groups of people who have the same multinomial. To do so, we again assume that $\beta = 1$ and $a = b = 10^{-10}$. After a burn-in of 1000 samples, we again drew 10,000 samples from the posterior $p(\boldsymbol{g} \mid \boldsymbol{x})$ with a lag of 5 iterations between samples. The posterior distributions over α and k are shown in Figure 16, and suggest that there are approximately 40 different groups represented among these 1000 people. In order to provide an interpretable summary of the full posterior over groups, we repeated the analysis used in the last section, in which we associate groups with an "expected performance profile". In this case, we find the expected distribution over the 17 categories for each different group. However, since there is a great deal of uncertainty about the make-up of many of the groups, we restrict the analysis to a few of the prominent and consistent groups.

As with the publication data, there is evidence for stable groupings of people. Across most of the posterior distribution we observe the three groups illustrated on the left hand side of Figure 17. In each case, there is a group of people who visit only one type of

web page, either "front page", "summary" or "weather". Given the extremely tight focus of these distributions, we might safely conclude that these people were engaged in very specific searches. Their interactions with the web environment were presumably oriented towards a very specific objective (e.g., find a weather report). On the other hand, there is some evidence for groups such as those illustrated on the right hand side of Figure 17. In these cases, people visited a range of different pages, particularly "front page", "news", "technology" and "on-air" pages. Distributed patterns of hits such as these suggest a different interpretation of user behavior. In these cases, people appear to be engaged in exploratory search through the web environment (i.e., genuinely "browsing" rather than simply "looking-up").

There is a great deal of variety in the kinds of exploratory browsing patterns observed across the posterior distribution. An exploratory analysis suggests that the clustering of "front page", "news" and "technology" pages is highly stable, in the sense that across most of the posterior there exist large groups that assign high probability to all three categories. However, there is a considerable degree of (apparently smooth) variation in the relative interest in these three topics. This is illustrated in the comparison between panels d and e in Figure 17, which show the same qualitative pattern of preferences, but display subtle differences in the various probabilities. Moreover, when we consider panel f, the same "clumping" of "front page", "news", "technology", is observed, but with the addition of "local" and "bulletin board service" instead of "on-air". Finally, there is some variation across the full posterior distribution in terms of the kinds of patterns it identifies.

Taken together, these results suggest that, while the infinite groups model is highly successful at identifying the focused search behavior illustrated on the left hand side of Figure 17, the more complex variation in exploratory browsing behavior is only captured in part. The apparently smooth variation from panel d to panel e suggests that a more complete account of individual differences in web browsing may require multimodal and continuous parameter distributions. The fact that there are similarities between panel d and panel f, for instance, suggests that we may need to explore models that allow structured relationships between groups.

10 General Discussion

Cognitive models aim to describe and predict how people think and act. Since different people think and act in different ways, we require models that allow us to learn complicated patterns of variation. The individual differences framework outlined in this paper provides a powerful method of representing the similarities and differences between people. By using a group model we can capture multimodality in individual differences, thereby remaining sensitive to the possibility of qualitative differences in performance. By adopting the Dirichlet process prior, we are able to view observed groups not as a fixed structure that

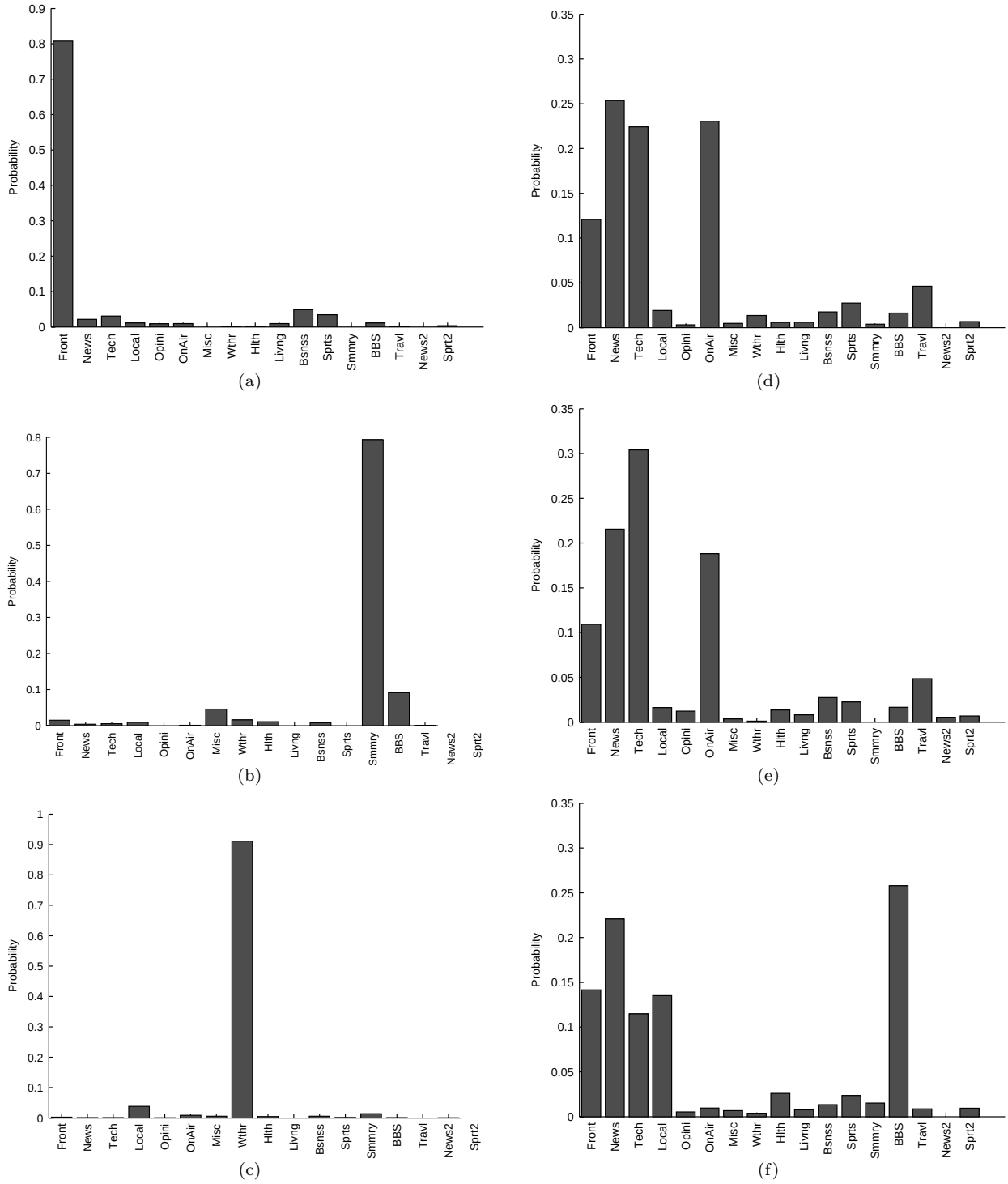


Fig. 17. Six different groups observed in the web data. In the three groups shown on the left, people visited only one type of page, either "front page", "summary" or "weather". All three groups on the right show a consistent tendency to visit "front page", "news", "technology" and "on-air" pages, but with different relative frequencies in each case. In addition, group (f) also shows interest in "health" and "bulletin board" pages.

fully explains the variation between individuals, but rather as representatives of a latent, arbitrarily rich structure. Additionally, by placing a prior over the dispersion we are able to learn about the extent of the the variability itself.

Our approach could be extended in a number of ways, enabling us to capture a greater range of individual differences phenomena. One very simple extension would be to generalize the Dirichlet process prior to other stick-breaking priors. As Ishwaran and James (2001) note, this family is quite general, incorporating the Poisson-Dirichlet process (Pitman & Yor, 1997) and Dirichlet-Multinomial processes (Muliere & Secchi, 1995) among others. Alternatively, one might choose to move beyond priors over the discrete distributions, instead using a different class of nonparametric priors, one that covers continuous distributions such as Pólya trees (Kraft, 1964; Ferguson, 1974) or Dirichlet diffusion trees (Neal, 2003).

A different extension to the framework can be motivated by returning to the unlucky numbers experiment. In this example there is an issue regarding how to treat the one person who responds 86. Does this person belong with the ten people who said 87? It may be the case that this person is not a cricket fan, and is a representative of a genuinely new group (fans of Agent 86 in the TV show *Get Smart*, perhaps). It is difficult to distinguish these cases, particularly since group models are rather unforgiving in their requirement that all group members share the same parameter value. One of the merits of the stochastic parameters approach is that it allows some smooth variation. If our data consisted only of cricket fans, a stochastic parameters model would learn an individual differences distribution centered on 87, since this is the typical behavior, but allow some variability to be expressed. However, once we reintroduce the 13 group and the 4 group, a unimodal stochastic parameter model will be inadequate.

A natural solution to this problem would be to build individual differences models that capture the strengths of both frameworks. One approach would be to adopt an *infinite stochastic groups model*, which would produce multimodal continuous distributions by convolving each point mass with a continuous distribution. In this approach, we would assume that there are distinct groups of subjects in our data, as with the infinite groups approach. However, within a group we would allow there to be continuous, unimodal variation, as with the stochastic parameters approach. Indeed, one of the reasons that we have avoided conducting some sort of competition between the different frameworks is that they are designed to address different phenomena. Accordingly, we feel that the better approach is to pursue more powerful modeling frameworks that integrate the best features of each.

Another direction for future work would be to allow structured relationships between groups. One possibility would be to postulate a separate Dirichlet process prior over each parameter. Alternatively, we could use a hierarchical Dirichlet process (Teh, Jordan, Beal & Blei, 2004), in which the distribution sampled from a Dirichlet process is itself a Dirichlet process. Finally, we may wish to consider an *idiosyncratic strategies model*, in which it is

assumed that all subjects draw on a common set of strategies but combine them in an unique way (e.g., Girolami & Kabán, 2004). In short, the infinite groups model is not by itself an authoritative account of individual differences. Rather, it is a representative of a large class of flexible models, each suggesting a fruitful approach for the development of powerful new cognitive models of individual differences.

Acknowledgements

Correspondence address: Daniel Navarro, Department of Psychology, University of Adelaide, SA 5005, Australia. E-mail: daniel.navarro@adelaide.edu.au, Tel.: +61 8 8303 5265, Fax.: +61 8 8303 3770. This research was supported by Australian Research Council grant DP-0451793. We thank Yves Rosseel for providing a copy of the categorization data, Victoria Dennington for collecting the publication data, as well as MSNBC and the UCI KDD archive (http://kdd.ics.uci.edu/) for making the web data available. We would also like to thank Jeff Rouder, E. J. Wagenmakers and an anonymous reviewer for helpful comments, and Hemant Ishwaran for providing some useful pointers.

References

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

Aldous, D. J. (1985). Exchangeability and related topics. In P. L. Hennequin (Ed.) *École d'Été Probabilités de Saint-Flour XII*. Springer-Verlag.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

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.

Bernardo, J. M. & Smith, A. F. (2000). *Bayesian Theory*. New York: Wiley.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Annals of Statistics, 1*, 356-358.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics, 1*, 353-355.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Blei, D. M., Griffiths, T. L., Jordan, M. I. & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. Saul, and B.

Schölkopf (eds), *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Buntine, W. L. (1994). Operations for learning with graphical models, *Journal of Artificial Intelligence Research, 2*, 159-225.

Cadez, I. V., Heckerman, D., Meek, C., Smyth, P. & White, S. (2003). Model-based clustering and visualization of navigation patterns on a Web site, *Journal of Data Mining and Knowledge Discovery, 7*, 399-424.

Chen, M., Shao, Q. & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Cowles, M. & Carlin, B. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review.*Journal of the American Statistical Association, 91*, 833-904.

Creutz, M., Jacobs, L. & Rebbi, C. (1979). Monte Carlo study of Abelian lattice gauge theories. *Physical Review D, 20*, 1915-1922.

de Finetti, B. (1974). *Theory of Probability* (vols 1 & 2). New York: Wiley.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: Wiley.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics, 14*, 1-26.

Duncan, K. A. (2004). Case and covariate influence: Implications for model assessment. *Ph.D. Thesis*. Columbus, OH: Ohio State University.

Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

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

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin 53*, 134140.

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

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics, 2*, 615-629.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Annals of Mathematical Statistics, 34*, 1386-1403.

Geman, S & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Recognition and Machine Intelligence, 6*, 721-741.

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.

Girolami, M. & Kabán, A. (2004). Simplicial mixtures of Markov chains: Distributed modeling of dynamic user profiles. In S. Thrun, L. Saul & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems, 16* (pp. 9-16). Cambridge, MA: MIT Press.

Green, P., & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics, 28*, 355-377.

Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian

buffet process. *Technical report no. GCNU TR 2005-001*, Gatsby Institute for Computational Neuroscience, University College London.

Griffiths, T. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228-5235.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

Hoskens, M. & de Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25*, 19-37.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96*, 161-173.

Ishwaran, H. & Zarepour, M. (2002). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics, 30*, 269-283.

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.

Junker, B. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Karabatsos, G. (in press, this issue). Bayesian nonparametric model selection and model selection. *Journal of Mathematical Psychology*.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Society, 90*, 773-795.

Kass, R. E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343-1370.

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J. & Tirri, H. (2005). An MDL framework for data clustering. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications* (pp. 323–354). Cambridge, MA: MIT Press.

Korwar, R. & Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability, 1*, 705-711.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability, 1*, 385-388.

Landauer T. K. & Dumais, S. (1997). A Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45*, 149-166.

Lee, M. D. & Navarro, D. J. (2005). Minimum description length and psychological clustering models. In P. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications* (pp. 355-384). Cambridge, MA: MIT 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.

Lo, A. (1984). On a class of Bayesian nonparametric estimates. *Annals of Statistics, 12*, 351-357.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum.

McCloskey, J. W. (1965). A model for the distribution of individuals by species in an environment. Unpublished Ph.D. thesis, Michigan State University.

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.

Maddox, W. T., & Estes, W. K. (2005). Predicting true patterns of cognitive performance from noisy data. *Psychonomic Bulletin & Review 11* , 1129–1135.

Muliere, P & Secchi, P. (1995). A note on a proper Bayesian bootstrap. *Technical Report 18*, Universitá degli Sudi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition, 28*, 832840.

Myung, I. J. & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

Neal, R. M. (1996). Bayesian learning for neural networks. *Lecture Notes in Statistics, 118*. New York: Springer-Verlag.

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

Neal, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics 7*, 619-629.

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

Oaksford, M. & Chater, N. (1998). *Rational Models of Cognition*. Oxford University Press.

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.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In T. S. Ferguson et al. (Eds) *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* (pp. 245-267). Hayward, CA: Institute of Mathematical Studies.

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability, 25*, 855-900.

Qin, A. L. (1998). Nonparametric Bayesian models for item response data. *Ph.D. Thesis*. Columbus, OH: Ohio State University.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In S.A. Solla, T.K. Leen and K.-R. Müller (eds.). *Advances in Neural Information Processing Systems 12* , pp. 554-560. Cambridge, MA: MIT Press.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for skewed variables with an application to response time distributions. *Psychometrika, 68*, 587-604.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4*, 639-650.

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

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2004). Hierarchical Dirichlet processes. *Technical Report 653*. Department of Statistics, University of California, Berkeley.

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian Inference. *Behavioral and Brain Sciences, 24*, 629-641.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*, 92-107.

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.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition 25*, 731-739.

Appendix: A Gibbs Sampler for Infinite Discrete Groups

Statistical inference in the infinite discrete groups model (Equation 21) can be achieved using Gibbs sampling, a Markov chain Monte Carlo (MCMC) method for sampling from the posterior distribution over the variables in a Bayesian model. The Gibbs sampler was introduced to statistics by Geman and Geman (1984), although it was already well-known in physics under the name of the "heat-bath algorithm" (Creutz, Jacobs & Rebbi, 1979). Gilks et al. (1995) and Chen et al. (2000) provide good discussions of MCMC methods, while Neal (2000) provides a detailed discussion of Gibbs sampling in Dirichlet process models.

To build a Gibbs sampler for a model with a Dirichlet process prior, we find an expression for the conditional distribution $p(g_i \,|\, \boldsymbol{g}_{-i}, \boldsymbol{x})$. This allows us to specify a Gibbs sampler in which we repeatedly sweep through all of the observations, reassigning the group variable g_i by sampling from this distribution. This results in a sequence of sampled assignment vectors $\boldsymbol{g}$ that form an Markov chain that converges to samples from $p(\boldsymbol{g} \,|\, \boldsymbol{x})$. In this approach, we have integrated out the $\boldsymbol{\theta}$ variables, so the Gibbs sampler does not provide samples from the joint posterior $p(\boldsymbol{g}, \boldsymbol{\theta} \,|\, \boldsymbol{x})$. However, since the joint distribution can be factorized into

$$p(\boldsymbol{g}, \boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\boldsymbol{\theta} \,|\, \boldsymbol{g}, \boldsymbol{x}) p(\boldsymbol{g} \,|\, \boldsymbol{x}),$$

it is simple enough to generate samples from the joint distribution by also drawing samples from $p(\boldsymbol{\theta} \,|\, \boldsymbol{g}, \boldsymbol{x})$. This distribution is straightforward to specify, firstly by noting that since the draws from the base distribution are independent. Therefore,

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{g}, \boldsymbol{x}) = \prod_{z \in \mathcal{Q}^*} p(\theta_z \,|\, \boldsymbol{g}, \boldsymbol{x}),$$

where $\mathcal{Q}^*$ refers to the set of k_{-i} currently non-empty groups. Secondly, we can write $p(\theta_z \,|\, \boldsymbol{g}, \boldsymbol{x})$ as the posterior distribution,

$$\begin{aligned} p(\theta_z \,|\, \boldsymbol{g}, \boldsymbol{x}) &\propto p(\boldsymbol{g}, \boldsymbol{x} \,|\, \theta_z) p(\theta_z \,|\, \beta) \\ &\propto \left(\prod_{i \,|\, g_i = z} p(\boldsymbol{x}_i \,|\, \theta_z) \right) p(\theta_z \,|\, \beta), \end{aligned}$$

where we have now reintroduced the dependence on β, the parameter value that describes our base distribution. Noting that the first term is a multinomial probability and the second term is Dirichlet, we can use conjugacy to infer that

$$\theta_z \,|\, \boldsymbol{g}, \boldsymbol{x}, \beta \sim \text{Dirichlet}(\cdot \,|\, \beta_z^*), \tag{22}$$

where $\beta_z^* = \beta + \sum_{i \,|\, g_i = z} \boldsymbol{x}_i$.

We now turn to the derivation for the conditional distribution over the group assignments. However, it is important to note that our model employs a mixture of Dirichlet processes, in which a prior over α is employed. Accordingly, our Gibbs sampler needs to sweep through the group assignment variables and the dispersion variable. We will begin by finding an expression for $p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha, \boldsymbol{x})$, the posterior probability that the ith participant is assigned to the group z, given some values for the other group assignment variables and a value for the dispersion (we will come back to the question of resampling the dispersion in a moment). Using Bayes' rule, we can write

$$p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha, \boldsymbol{x}) \propto p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha) p(x_i \,|\, g_i = z, \boldsymbol{g}_{-i}, \boldsymbol{x}_{-i})$$

The first term gives the prior probability that a new sample g_i from the Dirichlet process belongs to group z, where z may refer to a member of the set $\mathcal{Q}^*$ of k_{-i} currently non-empty groups, or it may refer to one of the infinite set $\mathcal{Q}$ of currently-empty groups. Using the conditional distributions described in Equations 10 and 11,

$$p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha) \propto \begin{cases} \frac{s_{-i,z}}{n-1+\alpha} \text{ if } g_i \in \mathcal{Q}^* \\ \frac{\alpha}{n-1+\alpha} \text{ otherwise} \end{cases} \tag{23}$$

where $s_{-i,z}$ counts the number of subjects (not including the ith) that are currently assigned to group z. This is a legitimate approach since samples from the CRP distribution are exchangeable; so for the purposes of the Gibbs sampling procedure we can always treat the ith observation as if it were in fact the last (or nth) one.

The second term $p(x_i \,|\, g_i = z, \boldsymbol{g}_{-i}, \boldsymbol{x}_{-i})$ is the likelihood of the ith participant's data, assuming they belong to group z. This can be written,

$$\begin{aligned} p(x_i \,|\, g_i = z, \boldsymbol{g}_{-i}, \boldsymbol{x}_{-i}) &= \int p(x_i \,|\, \theta_z) p(\theta_z \,|\, \boldsymbol{g}_{-i}, \boldsymbol{x}_{-i}, g_i = z) \, d\theta_z \\ &\propto \int \prod_{h=1}^{m} (\theta_{zh})^{x_{ih}} \left(\frac{\Gamma(m\beta + q_{-i,z})}{\prod_{h=1}^{m} \Gamma(\beta + q_{-i,z,h})} \prod_{h=1}^{m} (\theta_{zh})^{(\beta - 1 + q_{-i,z,h})} \right) d\theta_z \\ &= \frac{\Gamma(m\beta + q_{-i,z})}{\prod_{h=1}^{m} \Gamma(\beta + q_{-i,z,h})} \frac{\prod_{h=1}^{m} \Gamma(\beta + q_{\cdot,z,h})}{\Gamma(m\beta + q_{\cdot,z})} \end{aligned}$$

where the second line uses Equation 22. In this expression $q_{-i,z,h}$ denotes the number of times that a participant (not including the ith) currently assigned to group j made response h, and $q_{-i,z}$ denotes the total number of responses made by these subjects. The terms $q_{\cdot,z,h}$ and $q_{\cdot,z}$ are defined similarly, except that the ith subject's data are not excluded. Taking these results together, the required conditional posterior probability is given by,

$$p(g_i = z \,|\, \boldsymbol{g}_{-i}, \alpha, \boldsymbol{x}) \propto \begin{cases} \frac{\Gamma(m\beta + q_{-i,z})}{\prod_{h=1}^{m} \Gamma(\beta + q_{-i,z,h})} \frac{\prod_{h=1}^{m} \Gamma(\beta + q_{\cdot,z,h})}{\Gamma(m\beta + q_{\cdot,z})} \frac{s_{-i,z}}{n-1+\alpha} \text{ if } g_i \in \mathcal{Q}^* \\ \frac{\Gamma(m\beta)}{\prod_{h=1}^{m} \Gamma(\beta)} \frac{\prod_{h=1}^{m} \Gamma(\beta + q_{\cdot,z,h})}{\Gamma(m\beta + q_{\cdot,z})} \frac{\alpha}{n-1+\alpha} \text{ otherwise} \end{cases} \tag{24}$$

We can use Equation 24 in order to draw Gibbs samples for the group assignment variables. However, since we are using a Dirichlet process mixture, we also need to resample α. Throughout this paper, we treat the prior over α as an inverse Gamma$(\cdot \,|\, a, b)$ distribution. Using Antoniak's (1974) results, the conditional posterior over α depends only on the number of observed groups k and the sample size n, not the specific data or the group assignments. Thus, by expanding the Beta function $\mathrm{B}(\alpha, n)$ in Equation 20 we observe that

$$p(\alpha \,|\, \boldsymbol{g}, \boldsymbol{x}) = p(\alpha \,|\, k, n)$$

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

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

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

This approach comes from Escobar and West (1995). Using this joint distribution, we can find $p(\alpha \,|\, \eta, k, n)$ and $p(\eta \,|\, \alpha, k, n)$. These distributions are simply,

$$\begin{aligned} \alpha \,|\, \eta, k, n &\sim \text{Gamma}(\cdot \,|\, a + k - 1, b - \ln \eta) \\ \eta \,|\, \alpha, k, n &\sim \text{Beta}(\cdot \,|\, \alpha, n). \end{aligned} \tag{25}$$

Equations 24 and 25 define the Gibbs sampler. On every iteration of the Gibbs sampler, we sweep through all the group assignments g_i, sampling them from their conditional distributions, as well as the dispersion α and the dummy variable η. Over time, these converge to samples from the full posterior distribution $p(\boldsymbol{g}, \alpha \,|\, \boldsymbol{x})$, where convergence can be measured in a number of ways (see Cowles & Carlin, 1996). Given this distribution, it is straightforward to make other inferences such as $p(k \,|\, \boldsymbol{x})$.