Testing the psychological validity of cluster construction biases

Joseph L. Austerweil (Joseph_Austerweil@Brown.edu)
Brown University, Department of Cognitive, Linguistic, and Psychological Sciences, Providence, RI 02912

Abstract

To generalize from one experience to the next in a world where the underlying structures are ever-changing, people construct clusters that group their observations and enable information to be pooled within a cluster in an efficient and effective manner. Despite substantial computational work describing potential domain-general processes for how people construct these clusters, there has been little empirical progress comparing different proposals to each other and to human performance. In this article, I empirically test some popular computational proposals against each other and against human behavior using the Markov chain Monte Carlo with People methodology. The results support two popular Bayesian nonparametric processes, the Chinese Restaurant Process and the related Dirichlet Process Mixture Model.
Keywords: Clustering, Bayesian modeling, Connectionist modeling, Categorization

Introduction

A child observing a 3-tined fork for the first time easily infers its function. She infers that it is more similar to the 4-tined forks she typically uses than to other eating utensils such as spoons and knives. To generalize information from one observation (e.g., a 4-tined fork) to another observation (e.g., a 3-tined fork), people group their observations into clusters (e.g. *forks*). A cluster summarizes the information common to the stimuli within it, which enables efficient generalization and learning. In the example above, the child represents the category of *utensils* as a combination of three clusters: *forks*, *spoons*, and *knives*. A novel property learned about one fork should be generalized more to other forks (within-cluster) than to other utensils that are not forks. Given that underlying structures of the world are ever-changing and individuals only observe limited information, the clusters that encode these structures need to be flexible to accommodate new types of items. However, a new item can be too different to fit in any existing cluster, and a new cluster must be constructed. For example, when the child sees her first pair of chopsticks, she knows they are part of the category *utensils*, but they do not fit within her current set of clusters (*forks*, *spoons*, and *knives*). So she must create a new cluster to represent *chopsticks*. How does the mind construct clusters to represent observations?

Cognitive scientists have proposed several domain-general cluster construction processes with in a variety of domains, such as categorization (Anderson, 1991; Love, Medin, & Gureckis, 2004), associative learning (Gershman, Blei, & Niv, 2010), causal inference (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010), perceived orientation (Austerweil, Friesen, & Griffiths, 2011), and word segmentation (Goldwater, Griffiths, & Johnson, 2009). Despite the proliferation of different proposals, there have been few attempts to compare the proposals and evaluate their psychological validity. This is not due to lack of interest, but rather the difficulty of empirically comparing these models. Some challenges include the combinatorial explosion in the number of possible clusterings for even a small number of items (e.g., there are more than 100,000 possible clusterings for 10 items; Pitman, 2002), the indirect and at times weak effect of cluster construction processes, and the sensitivity of the clusters reported by participants to instructions (e.g., providing a number of clusters affects participants' responses; Pothos et al., 2011).

In this article, I present one of the first empirical experiments that directly tests the psychological validity of computational models that construct clusters to group observations. This new approach avoids the combinatorial explosion of having to test an insurmountably large number of stimuli on some measure. Instead, I test people's prior expectations for constructing clusters and compare them to model predictions using an adaptive experimental technique, Markov chain Monte Carlo with People (MCMCP, which is also called iterated learning; Griffiths & Kalish, 2007; Sanborn, Griffiths, & Shiffrin, 2010). In this method, participant responses are used to guide the experiment towards testing clusters that are consistent with their prior expectations. Previous work has shown that MCMCP is especially effective for eliciting peoples prior expectations over a large set of possibilities despite the combinatorial explosion (e.g., happy and sad faces; Martin, Griffiths, & Sanborn, 2012).

In this article I first describe previous work testing models that construct clusters. Next, I explain MCMCP and present the experimental methods. Then, I describe the models tested in this paper (Bayesian nonparametric models, SUSTAIN, and a few alternatives). Last, I discuss the results and conclude with implications and directions for future work.

Previous Empirical Work

There are few empirical tests comparing human behavior to different computational models that construct clusters. One exception is Pothos et al. (2011), who tested how people construct clusters in the domain of unsupervised categorization without instructing participants to use a certain number of clusters. Participants were asked to sort nine two-dimensional stimuli (spider-like images varying in the length of their "body" and "legs") into groups that felt intuitive. The authors compared participant sortings to their "goodness" according to several unsupervised category learning models. Their results were equivocal, but generally, they supported SUSTAIN and a few other models better than the Rational Model of Categorization (a Dirichlet process mixture model using a particular approximation method; Anderson, 1991; Neal, 1998). Their results are an informative first comparison of human and model clustering in unsupervised categoriza-

tion. However, they only tested nine of the potentially infinite number of stimulus sets (i.e., body-length $\times$ leg-length combinations) within their domain. This limits the conclusions one can draw from their results because it is unclear whether their results reflect an experimental bias due to the stimulus sets picked by the experimenters or actual properties of human cluster construction. For example, one might suspect that the most preferred clustering would be a single cluster with all the items in it, but none of their stimulus sets afforded a single cluster solution. Thus, based on their results alone, we might suspect that people are biased against adopting a single cluster to represent stimuli, which seems unintuitive and is not necessarily true.

Markov Chain Monte Carlo With People (MCMCP)

In the studies presented here I use MCMCP to circumvent the problem of how to pre-select stimulus sets that are most informative about how people construct clusters. MCMCP is an experimental methodology that adapts MCMC algorithms to construct experiment stimuli online, based on the participants previous response. MCMC algorithms are a class of methods that approximate a probability distribution by constructing a Markov chain that converges to that distribution (Gilks, Richardson, & Spiegelhalter, 1996). A familiar MCMC procedure is shuffling a deck of cards (Aldous & Diaconis, 1986). The state of the Markov chain at each step is the order of cards in a deck. A "shuffle" transitions the deck from one order to another order. After sufficiently many shuffles, the deck will reach "stationarity," meaning that if you stopped shuffling at any point, the probability of any order of cards would be the same. The distribution of states that the chain visits (after sufficiently many transitions) is known as a *stationary distribution*.

MCMC algorithms are methods for defining a Markov chain whose stationary distribution is any complex distribution of interest (e.g., expectations over clusterings). They do so by defining the transition procedure to be a simplified form of the complex distribution. In a Gibbs sampling within-subjects MCMCP experiment, a Markov chain is a sequence of trials within the experiment and the state of the Markov chain is a sample from the desired probability distribution (e.g., a clustering). The chain is initialized to a possible sample (e.g., 7 balls are clustered into one cluster of size 4 and a second cluster of size 3). On each trial, participants observe the sample except for one of its items, which is hidden (e.g., given that 6 balls are clustered into one cluster of size 4 and a second cluster of size 2, do you think a seventh ball will part of cluster 1, 2, or a new cluster?). Their choice becomes the value for the hidden part of the sample (e.g., if the participant chose cluster 1, there would now be 5 balls in cluster 1 and 2 balls in cluster 2). On the next trial of the chain, participants replace a different part of the sample. This is repeated many times. In these types of experiments, trials for multiple chains are interleaved so that participants do not realize their choices determine future trials.

Experimental methods

There were two conditions in the within-subjects MCMCP experiment, each of which elicited a different type of expectation related to clustering. The first condition (*Balls*) explored peoples expectations as to the size of clusters. The state of each Markov chain in this condition was the assignment of items to clusters. At the beginning of the experiment, each chain was initialized to a different possible clustering. Transitions then reassigned one of the items to a new cluster (based on the other cluster assignments). Because it is a MCMCP experiment, after sufficiently long, the Markov chains in the *Balls* condition can be treated as samples from peoples prior expectations of the size of clusters. The second condition (*Sticks*) explored peoples expectations over categories varying on a single continuous dimension (vertical lines varying in height). On each trial of this condition, participants observed seven vertical lines of different heights and produced the height of an eighth line. Biases due to clustering expectations should influence participant judgments. For example, if a participant observed four short and three tall lines, they should produce either a short or tall line, but not a line of medium height.

There were 24 participants (Amazon Mechanical Turk) paid \$7.50 for completing the $\approx$45 minute experiment. The order of the two conditions was counterbalanced across participants. The within-subject procedure allows for the optimal cluster construction parameter for each participant to be compared across conditions. The parameters are not necessarily related (e.g., their values could be domain-dependent), but if they are related, it would provide strong empirical support that the models are capturing some characteristic of a domain-general process.

Balls: MCMCP Over Clusterings

In each trial the colors of six balls were described, and a seventh ball was not described (hidden). Participants were always given verbal descriptions and never observed the image of any colored balls. There were seven possible ball colors: red, orange, yellow, green, blue, purple, and brown. Both the subjects and models were given the task of predicting the color of the hidden seventh ball in the set, given the six observed colors. The choices on a trial included any observed color name, plus one additional color name. For example, if they were told there are 4 red balls and 2 green balls, their options for the seventh ball were red, green and one other randomly chosen color (e.g., blue). The *Balls* condition differs from standard probability matching experiments because the trials and number of choices change depending on participant responses.

This procedure was conducted for 15 MCMCP chains, which all started at a different initial state. The state were defined by all possible combinations of cluster sizes, regardless of color (e.g., $(7), (6,1), (5,2), (5,1,1)$ and so forth un-

til $(1,1,1,1,1,1,1)$, where $(a,b,\ldots,)$ means a cluster of a balls of color A, a cluster of b balls of color B, and so on). The color names assigned to each cluster were randomly determined for each participant but were consistent during that participants session.

After the first iteration of each chain, the ball that the participant predicted replaced the hidden ball. On the next iteration, their response became part of the observed set of balls, and a ball becomes hidden (different from the one of the previous iteration). Which ball becomes hidden was determined randomly, with the constraint that every ball gets replaced once before it is replaced again. By the end of the experiment, all balls within a chain were replaced twice. Also, a ball was replaced in each chain before the next ball in a chain was replaced. This procedure resulted in 210 trials.

Participants were told that they would be presented with a series of urns filled with balls of different colors. On each trial they would be told the colors of 6 balls from an urn and would be asked "what you thought was most likely to be the color of the next ball drawn from the urn?". To mitigate any inter-trial dependencies (and to stay faithful to the assumptions of the cluster construction models), participants were told that "the urns are unrelated and so balls from one urn provide no information about the balls in another urn" and "although you get to see 6 balls from each urn, there are many balls in each urn." Finally, each urn was labeled with a unique number.

Sticks: MCMCP Over 1-D Categories

In each trial, the heights of seven sticks were visually presented to participants. Analogous to the *Balls* condition, there is an eighth hidden stick height. On each trial, the participants and models predicted the height of the eighth stick, given the observed seven sticks. Participants controlled the length of a stick on the screen by moving their mouse vertically and clicked to submit their response.

This procedure was conducted for 25 MCMCP chains, all initialized to different states. To capture a diverse range of stimulus distributions, their heights were initialized by sampling from the following Beta distributions: four were *Uniform* (Beta(1,1)), four were centered on the *Middle* stick height (Beta(5,5)), five were *Bimodal* at the extremes (Beta(0.2,0.2)), four were *Very Small* (Beta(0.2,1)), four were *Very Large* (Beta(1,0.2)), one was *Small* (Beta(5,1)), and one was *Large* (Beta(1,5)). The value was rescaled to 0.07 to 1.11 inches. The width of a stick was 0.02 inches. Otherwise the procedure was identical to the *Balls* condition. This resulted in 368 trials.

Constructivist Models

Bayesian Nonparametric Models

Bayesian models posit a set of possible structures, formulate prior expectations over these structures as probability distributions, and then integrate observed information into the distribution over structures via Bayes' rule. In Bayesian nonparametric models (see Austerweil, Gershman, Tenenbaum, & Griffiths, in press, for a review), the set of possible structures is infinite, which allows them to capture a wide array of structures while maintaining explicit prior expectations over the structures. In this subsection, I discuss the Chinese Restaurant Process (**CRP**; Aldous, 1985), the Pólya Urn (Blackwell & MacQueen, 1973), the two-parameter generalization of the CRP called the Pitman-Yor Process (**PYP**; Pitman, 2002), the Dirichlet process mixture model (**DPMM**; Antoniak, 1974; Ferguson, 1973), and the Pitman-Yor process mixture model (**PYPMM**; Pitman, 2002).

A commonly used Bayesian nonparametric process is the CRP with parameter $\alpha > 0$ that governs the propensity for constructing clusters. It is a culinary metaphor that defines a probability distribution directly over clusterings. According to it, customers (observations) $\mathbf{z}_N = (z_1,\ldots,z_N)$ enter a restaurant with infinite tables of infinite capacity. The first customer starts the process by sitting at the first table. Customer N sits at occupied table k with probability $m_k/(N+\alpha-1)$, where m_k is the number of customers at the table, or an unoccupied table with probability $\alpha/(N+\alpha-1)$. The PYP is equivalent to the CRP except that a small "discount" (parameter $0 \leq d \leq 1$) is taken whenever a customer sits at a new table and given back to the probability of a future new table. So, for the PYP, customer N sits at occupied table k with probability $(m_k-d)/(N+\alpha-1)$ or an unoccupied table with probability $(\alpha+Kd)/(N+\alpha-1)$, where K is the number of occupied tables. This defines a clustering of items where two items are in the same cluster when their corresponding customers are sitting at the same table. Note that these processes implement forms of probability matching (Shanks, Tunney, & McCarthy, 2002), where the probability of choosing a previously unobserved value decreases in the number of observed items. When each table k of the CRP is associated with a parameter θ_k from an arbitrary distribution $G(\cdot)$, the process on $(\mathbf{z},\theta) = (z_1,\ldots,z_N,\theta_1,\ldots)$ is called the Pólya Urn (Blackwell & MacQueen, 1973) due to its equivalence to a generative process where colored balls are sequentially drawn from an urn, and each time a ball is drawn, it is put back in the urn with an additional ball of the same color (which inspired the *Balls* condition). The urn, balls, and colors are analogous to the restaurant, customers, and parameter associated with each table, respectively.

To connect clusters to observations, Bayesian nonparametric models typically assume each cluster is associated with a parameter that determines a distribution over the observations. I.e., observations in a cluster are generated from the distribution determined by the cluster's parameter. I assume that items given their cluster membership are normally distributed ($x_n|z_n = k \sim N(\theta_k,\sigma_x^2)$), where the cluster parameter θ_k defines the mean of the observations and is generated from a Normal distribution with known mean μ_0 and variance σ_0^2 ($G = N(\mu_0,\sigma_0^2)$). This mixture model is a DPMM or PYPMM when the CRP and PYP are used to generate cluster membership, respectively.

Connectionist Model

Although many connectionist models use a fixed architecture, one of the most popular connectionist models of category learning, **SUSTAIN** (Love et al., 2004), changes its architecture by constructing new nodes when it cannot explain its current observation. There are a few variants of SUSTAIN, which are used depending on whether categorization information is given. Because participants do not get category information in the Experiment, I focus on the purely unsupervised variant of the model, where a layer of clusters compete to encode observations. SUSTAIN starts small (with one cluster centered on the first observation) and constructs new clusters whenever the activation of the most activated cluster is below a threshold τ. The activation of a cluster k, h_k, for an input x_n decays exponentially in the distance between the position of cluster k, θ_k, and the input $h_k = e^{-\lambda d(x,\theta_k)}$, where λ is the tuning of the input dimension and $d(x,\theta)$ is the distance between x and θ, which is the Hamming distance (0 if equal, 1 otherwise) for discrete stimuli and $\frac{1}{2}|x-y|$ for continuous stimuli (following Love et al., 2004). This activation rule is equivalent to the activation rule used by Love et al. (2004) when inputs are one-dimensional. When a new cluster is created for an item, the cluster's parameter is set to the current item's value. Otherwise, the "winning" cluster (the one with largest activation) is updated according to $\Delta\theta_k = \eta(x-\theta_k)$, where η is the learning rate. When a new cluster was not created, the dimensional tuning was also updated via $\Delta\lambda = \eta e^{-\lambda d(x,\theta_k)}(1-\lambda d(x,\theta_k))$ where k is the index of the winning cluster. Following Love et al. (2004), the output activation for item x was given by $o = h_k^{\beta+1}/\sum_j h_j^{\beta}$, where j ranges over the network's clusters and β is a nonnegative lateral inhibition parameter. To convert the output activations to a probability distribution over items, I used the exponentiated Luce choice rule with parameter w over the output activation o for a given item (as compared to the activation of other possible items).

Alternative Models

There were two sets of alternative models depending on the observable property of the given items. When the cluster assignments were directly observed (the *Balls* condition of the Experiment), I used two alternative models: **Max + Noise** and **Random**.[1] With probability $1-\varepsilon$, the Max + Noise model generated the modal cluster (the cluster with the most items) and with probability ε it generated a random cluster from the remaining options. Importantly, the sizes of the previously observed clusters only matter for determining the modal cluster. The Random model simply chooses uniformly at random from the possible choices.

When the observable property of a given item was a dimension, there were two alternative models: a **prototype** model (Reed, 1972) and an **exemplar** model (Medin & Schaffer, 1978; Nosofsky, 1986). For both models, the same distance function was used $d(x,y) = \lambda|x-y|$, where λ is the dimensional tuning parameter. The prototype model assumes that participants represent the given items as the average of their values. For the prototype model, the activation of a new item y to the given items $\mathbf{x}$ is $h = \exp\{-d(\theta,y)\}$, where θ is the average of the given items $\mathbf{x}$. Conversely, the exemplar model assumes that participants represent the given items explicitly. For the exemplar model, the activation of a new item y to the given items $\mathbf{x}$ is $h = \sum_{n=1}^{N}\exp\{-d(x_n,y)\}$. For both models, the probability of a new item is given by the exponentiated Luce choice rule of the new item's activation (as compared to the other possible items) with parameter w.

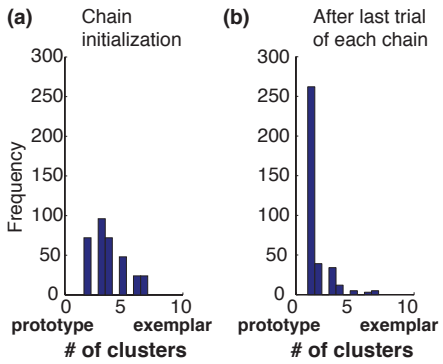

Figure 1: Results of *Balls* condition. (a) The number of clusters after the first response. The distribution is roughly uniform over possible clusterings of the balls. (b) The number of clusters at the end of the experiment. The distribution is tightly peaked at one cluster, which provides support to people being initially biased towards prototype representations.

Experimental Results

Balls: MCMCP over Clusterings

Figure 1a shows the number of clusters at the start of the experiment and Figure 1b shows the numbers of clusters after the last block of this condition of the experiment. At the end of the experiment, the distribution of the number of clusters per trial is peaked at one, which supports the hypothesis that given only a small number of items from a category, people are biased to represent the category using a prototype (Smith & Minda, 1998). Although the majority of chains converged to one cluster, some chains still contained more than one cluster at the end of the experiment.[2] Thus, the bias towards prototype representations is not as strong as it could be.

Figure 2 presents the Akaike Information Criterion (AIC; Claeskens & Hjort, 2008), a measure of model fit that penalizes models with larger numbers of parameters, for the PYP (black), CRP (red), SUSTAIN (purple), Max+Noise (blue),

[1] For the *Balls* condition, the exemplar and prototype (defined by the mode) models with an exponentiated Luce choice rule are equivalent to the CRP and Max + Noise models, respectively.

[2] It is possible that every chain would converge to a single cluster with further testing. This is unlikely because the distribution over the number of clusters barely changed over the last few trials.

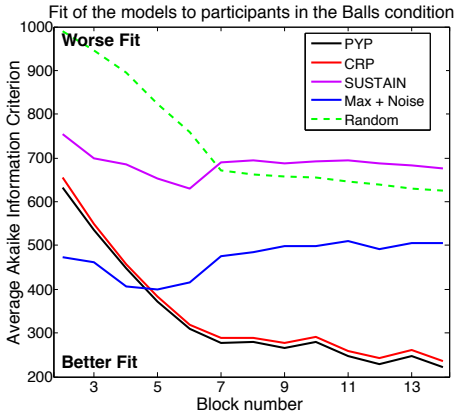

Figure 2: Average Akaike Information Criterion (AIC) fits of the PYP (black), CRP (red), SUSTAIN (purple), Max+Noise (blue), and Random (green) models for participants in the *Balls* condition over the course of the experiment (Note: Better fits have smaller values). Although PY and CRP are barely distinguishable in aggregate (with PY having slightly better CRP), CRP has significantly better AIC when compared to the other models at the level of individual participants.

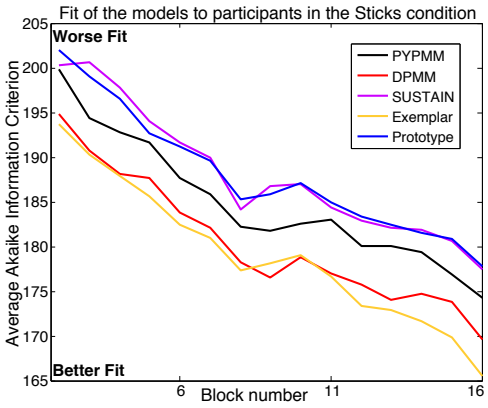

Figure 3: Average AIC fits of the PYPMM (black), DPMM (red), SUSTAIN (purple), Exemplar (yellow), and Prototype (blue) models for participants in the *stick* condition over the course of the Experiment (Note: Better fits have smaller values). Although PYPMM and DPMM are barely distinguishable in aggregate (with perhaps the Exemplar model being slightly better near the end of the experiment), DPMM has significantly better AIC when compared to the other models at the level of individual participants.

and Random (green) models over the 14 blocks of the experiment (generating a clustering for each item twice).[3] In this condition of the experiment, SUSTAIN is equivalent to a two-step generative process: first, decide to make a new cluster or use an old cluster with some probability, and if an old cluster is used, pick an old cluster uniformly at random. Unlike participants, it is not biased towards clusters of larger sizes, and thus has poor fit to the results. Although it is tempting to conclude that the PY model captures the results better than the CRP model from Figure 2, this would be premature because Figure 2 reports aggregate fits, rather than the results of individual participants. Fitting each model to individual participants provides a more appropriate analysis and different results: the CRP provides significantly better fit to the results of individual participants in the experiment (CRP had the best AIC for 16 of 24 participants; $p < 0.05$ for a Binomial sign test). The PY and random models fit five and three subjects best, respectively. SUSTAIN fits the results poorly because, unlike participants, it has no bias towards larger clusters. The Max + Noise and Random models also fit the data poorly.

Sticks: MCMCP Over 1-D Categories

Figure 3 presents the AIC of the PYMM (black), DPMM (red), SUSTAIN (purple), Exemplar (yellow), and Prototype (blue) models over the 16 blocks of the experiment (replacing the length of each stick twice). Although it might be tempting to conclude that the Exemplar model performs similar to if not better than the DPMM (which are better than the rest), again analyzing the individual participants is more appropriate and tells a different result: The DPMM provides significantly better fit to the results of individual participants in the experiment (DPMM had the best AIC for 16 of 24 participants; $p < 0.05$ for a Binomial sign test). The PYPMM and the Exemplar models provided the best fit for one and seven participants, respectively. Thus, in corroboration with the results of the *Balls* condition, the DPMM seems to provide the best description to people's expectations over a one-dimensional stimulus and the implicit bias of the clusters provides a benefit over the simpler Exemplar model (though it is too weak to show in the aggregate results).

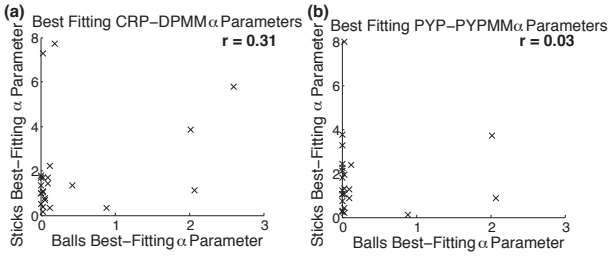

Figure 4: Relation between the cluster construction parameters of the (a) CRP-DPMM and (b) PYP-PYPMM fit to individual participants. Although neither correlation is significant ($p = .14$ and $p = .89$ two-tailed, respectively), the CRP-DPMM correlation is suggestive.

[3] Using the Bayesian Information Criterion (Claeskens & Hjort, 2008) yields the same statistical results as AIC for both conditions.

Balls-Sticks Comparisons

Given the within-subjects design, it is possible to explore whether the parameters of the Bayesian nonparametric models reflected an aspect of a domain-general cluster construction process. Figures 4 (a) and (b) present the relation between the cluster construction parameters of the CRP-DPMM ($r = .31, p = .14$ two-tailed) PYP-PYPMM ($r = .03, p = .89$ two-tailed) fit to individual participants, respectively. Although neither correlation is significant, the CRP-DPMM correlation is suggestive (especially because there are outliers and a one-tailed test is justified). Future work should test this possibility further.

Concluding Remarks

This article describes results about how people and computational models construct clusters by comparing different models to human performance. First, popular culinary metaphors from Bayesian nonparametrics (the CRP and PYP) implement forms of probability matching and the CRP is equivalent to an Exemplar model over cluster membership. Second, I used the MCMCP methodology to compare how people construct clusters to different computational proposals. The CRP and DPMM best captured the expectations of individual participants over clusterings and the stimulus distribution. Further work is needed to reconcile these results with those of Pothos et al. (2011) and to follow up on the intriguing possibility that the cluster construction parameter in the CRP and DPMM captures an important aspect of an underlying process used by people to construct clusters across domains.

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de probabilités de Saint-Flour xiii* (pp. 1–198). Berlin: Springer.

Aldous, D., & Diaconis, P. (1986). Shuffling cards and stopping times. *The American Mathematical Monthly*, *93*(5), 333-348.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

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

Austerweil, J. L., Friesen, A. L., & Griffiths, T. L. (2011). An ideal observer model for identifying the reference frames of objects. In *Advances in NIPS 24*.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (in press). Structure and flexibility in Bayesian models of cognition. In J. R. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press.

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

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge Univ. Press.

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

Gershman, S., Blei, D., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441-480.

Kemp, C., Tenenbaum, J., Niyogi, S., & Griffiths, T. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories. *Cognitive Science*, *36*, 150-162.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models* (Tech. Rep. No. 9815). Department of Statistics, University of Toronto.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *JEP: General*, *115*(1), 39–57.

Pitman, J. (2002). *Combinatorial stochastic processes.* (Notes for Saint Flour Summer School)

Pothos, E. M., Perlmann, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*, 83-100.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*, 63-106.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233-250.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *JEP:LMC*, *24*(6), 1411-1436.