Similarity as Inference

Samuel J. Gershman (gershman@fas.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA 02138, USA

December 16, 2017

Abstract

The perceived similarity between two objects varies with the distribution of other objects in the same context. The most famous theory of similarity judgment, Tversky's contrast model, can capture some of these contextual effects by assuming that the context changes the weighting of shared and distinctive object features in determining similarity judgment. However, the contrast model does not by itself explain how the weights are determined by context. This paper shows how a Bayesian theory of similarity judgment, first proposed by Kemp, Bernstein, and Tenenbaum (2005), can fill the gap. The key ideas are that (a) perceived similarity reflects the posterior probability that two objects arose from the same generative process; and (b) the parameters of the generative process are estimated based on all the objects in the context. The Bayesian theory is mathematically equivalent to Tversky's contrast model, but now the weights are directly determined by the estimated parameters. Several context effects can be captured using this theory, supporting the claim that similarity judgment is a form of inference.

Keywords: similarity; Bayesian inference; feature representation

Introduction

A familiar experience to travelers is the surprising affinity one feels for a compatriot encountered abroad, while the same compatriot might not even be noticed in one's home country. Likewise, two apples might not seem particularly different until one visits an apple orchard laden with apples of different sizes, shapes and tastes. These examples illustrate the fundamentally context-sensitive nature of similarity: the same two objects can be perceived as more or less similar depending on the distribution of other objects in the same context (Goldstone, Medin, & Halberstadt, 1997; Goodman, 1972; Medin, Goldstone, & Gentner, 1993; Tversky, 1977). Despite its ubiquity, the origin of context-sensitivity is still poorly understood—we lack a computational account of when and why context effects occur.

The "contrast model" developed by Tversky (1977) is the most influential and comprehensive attempt to grapple with context-sensitivity. In its simplest form, the contrast model

assumes that the judged similarity between objects a and b is given by a linear combination of three components:

$$s(a,b) = w_1 F(a \cap b) - w_2 F(a - b) - w_3 F(b - a), \tag{1}$$

where $a \cap b$ denotes the set of features shared by a and b, $a - b$ denotes the set of features possessed by a but not b, and $F(\cdot)$ is a function that measures the salience of a feature set. The weights $\{w_1, w_2, w_3\}$ determine the relative importance of shared and distinctive components. In some cases, Tversky assumed an additive decomposition whereby $F(a) = \sum_{i \in a} F(i)$. We can then think of $F(i)$ as the salience of feature i. Because feature salience interacts multiplicatively with the component weights, we will elide the two concepts and simply talk about "feature weights."

The contrast model captures certain kinds of context effects by assuming that context influences feature weights. Being from the US is not particularly salient when you are in the US, but when you meet another US citizen in a foreign country, this feature has diagnostic value in distinguishing individuals. Tversky canonized this idea as the *diagnosticity principle*: features that can be used to discriminate between categories have greater weight. Although category-based diagnosticity has received checkered support (Evers & Lakens, 2014; Goldstone et al., 1997), the notion that the distribution of features influences feature weights is manifested in a number of experimental results, some of which are discussed below.

The key question addressed here is how to determine the weights. Tversky's diagnosticity principle was essentially a heuristic; our goal is to show how the weights emerge from more fundamental computational principles. We describe a Bayesian derivation of Tversky's contrast model (Kemp et al., 2005), which we term the *Bayesian contrast model*. Crucially, the Bayesian contrast model grounds feature weights in subjective beliefs about distributional properties of the environment. The novel contribution of this paper is to demonstrate that the model can predict how weights change with manipulations of the feature distribution.

The Bayesian contrast model

In an elegant but under-appreciated paper, Kemp et al. (2005) showed how Tversky's contrast model could be derived from Bayesian principles. The advantage of adopting a Bayesian viewpoint is that it explains not just *how* similarity is computed, by *what* similarity is—it provides an analysis of the computational problem being solved by the information processing system (cf. Marr, 1982). According to Kemp and colleagues, the computational problem of similarity is to determine whether two objects arose from the same generative process or two independent generative processes. Formally, let H_1 denote the hypothesis that the two objects (a and b) were sampled from the same generative process, and let H_2 denote the hypothesis that the two objects were sampled from independent processes. The posterior probability of H_1 is given by Bayes' rule:

$$P(H_1|a,b) = \frac{P(a,b|H_1)P(H_1)}{P(a,b|H_1)P(H_1) + P(a,b|H_2)P(H_2)}. \tag{2}$$

Following Kemp and colleagues, we use the log odds form of Bayes' rule:

$$\log\left[\frac{P(H_1|a,b)}{P(H_2|a,b)}\right] = \log\left[\frac{P(a,b|H_1)}{P(a,b|H_2)}\right] + \log\left[\frac{P(H_1)}{P(H_2)}\right]. \tag{3}$$

Under the assumption that the two hypotheses have equal prior probability, we arrive at the formulation of Kemp and colleagues:

$$s(a,b) = \log\left[\frac{P(a,b|H_1)}{P(a,b|H_2)}\right]. \tag{4}$$

Because each generative process is governed by some parameters (θ), we need to marginalize over these unknown parameters to determine the log odds:

$$s(a,b) = \log\left[\frac{\int_\theta P(a,b|\theta)d\theta}{\int_\theta P(a|\theta)d\theta \int_\theta P(b|\theta)d\theta}\right]. \tag{5}$$

In the simulations reported below, we measure similarity on the more intuitive probability scale rather than the log odds scale, via the logistic sigmoid transform:

$$P(H_1|a,b) = \frac{1}{1+e^{-s(a,b)}}. \tag{6}$$

As a side note, this formulation is similar to the ratio formulation of the contrast model (Tversky, 1977), and appears in Tenenbaum and Griffith's (2001) alternative Bayesian derivation of the contrast model.

Kemp and colleagues showed that this general formalism can subsume several different notions of similarity, including transformational models (Hahn, Chater, & Richardson, 2003; Imai, 1977). We focus on their derivation of the contrast model, which assumes the following generative process:

$$\theta^i \sim \text{Beta}(\alpha,\beta), \tag{7}$$

$$a^i \sim \text{Bernoulli}(\theta^i), \tag{8}$$

where θ^i is the probability that the ith feature is "active" ($a^i = 1$). The hyperparameters α and β specify the prior expectations about feature occurrence. We use a somewhat more intuitive parametrization:

$$m = \frac{\alpha}{\alpha+\beta}, \qquad v = \frac{1}{\alpha+\beta}, \tag{9}$$

where m denotes the mean and v denotes the "variability" (which is monotonically related to variance, but algebraically simpler). Kemp and colleagues showed that the log odds for this generative model can be expressed in the following form (using our reparametrization):

$$s(a,b) = k_1|a \cap b| - k_2|a-b| - k_2|b-a|, \tag{10}$$

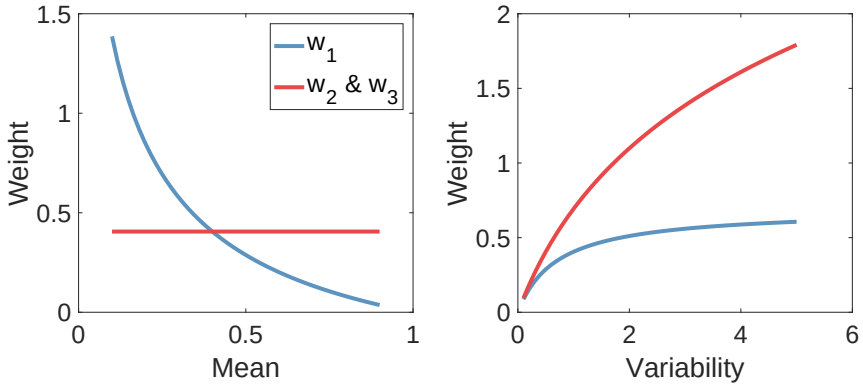

Figure 1: **Parametric effects of hyperparameters on component weights**. (Left) Mean feature activation, $\alpha/(\alpha+\beta)$. (Right) Feature variability, $1/(\alpha+\beta)$.

where $|\cdot|$ denotes set cardinality, and the weights are given by:

$$k_1 = \log\left(\frac{m+v}{m}\right) - \log(1+v), \tag{11}$$

$$k_2 = \log(1+v). \tag{12}$$

The Bayesian similarity measure is thus mathematically equivalent to the contrast model when $w_1 = k_1$, $w_2 = w_3 = k_2$, and $F(\cdot) = |\cdot|$.

The model can be generalized straightforwardly to the case where each feature is associated with its own hyperparameters:

$$\theta^i \sim \text{Beta}(\alpha^i, \beta^i), \tag{13}$$

leading to:

$$s(a,b) = \sum_{i \in a \cap b} k_1^i - \sum_{i \in a-b} k_2^i - \sum_{i \in b-a} k_d^i, \tag{14}$$

where k_1^i and k_2^i are defined with respect to the feature-specific mean m^i and variability v^i. This generalization allows us to connect the Bayesian contrast model to the additive decomposition used by Tversky to model differential feature salience. Specifically, a feature will be salient to the extent that its corresponding weight is large. Figure 1 illustrates how the weights change with the mean and variability: the shared component weight ($w_1 = k_1$) decreases with the mean and increases with the variability, whereas the distinctive component weight ($w_2 = w_3 = k_2$) does not vary with the mean and increases with the variability (with a steeper slope than the common component weight).

We have so far treated m and v as known, but in the real world (and in experimental tasks) they may need to be estimated from experience. The simplest assumption is that the

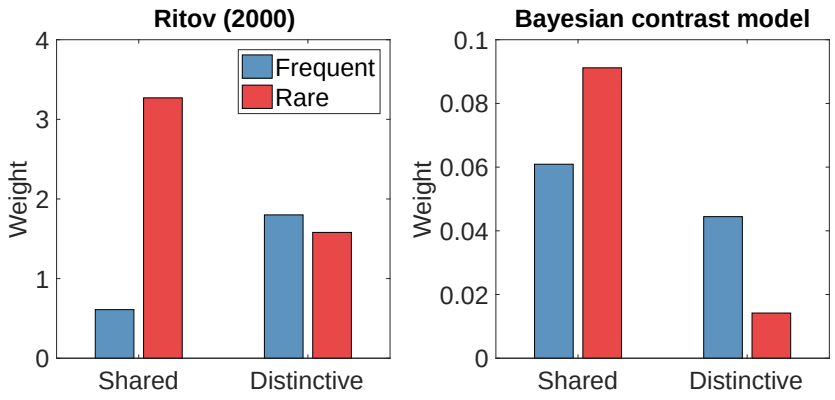

Figure 2: **Changing feature probability has opposite effects on weights for shared and distinctive components**. (Left) Data from Ritov (2000). (Right) Weights derived from the Bayesian contrast model.

distribution of object features provides the data used to estimate the hyperparameters. In the non-additive case, suppose we have N features, M of which are active. The posterior sufficient statistics are $\alpha = \alpha_0 + M$ and $\beta = \beta_0 + N - M$, where α_0 and β_0 represent the prior beliefs about the feature distribution before observing any data. We will take this to be a uniform distribution, corresponding to $\alpha_0 = \beta_0 = 1$, though the results reported below are robust to changes in the prior. The results are analogous for the additive case: $\alpha^i = \alpha_0^i + M^i$ and $\beta^i = \beta_0^i + N^i - M^i$, where now the sufficient statistics re feature-specific.

In the next section, we explore the implications of this analysis, showing how several context effects are natural consequences of the Bayesian model. The central theme running through these examples is that even though similarity is a relation between two objects, the salience of shared and distinctive features depends on the entire distribution of object features.

Simulations

One of the key predictions of the Bayesian contrast model is a *rare feature match effect*: when two objects match on a rare feature, they should be perceived as more similar compared to when the two objects match on a frequent feature. This prediction has been confirmed in a number of studies, which we weave together here.

Ritov (2000) manipulated the frequency of features. In the "frequent" condition, there were 5 possible features (ice cream flavors), only two of which were ever active, whereas in the "rare" condition, there were 17 possible features. Thus, any given feature was less likely in the rare condition relative to the frequent condition. Ritov found that the weight attached to shared features was greater when features were rare compared to when they were frequent; this pattern flipped for the weight attached to distinctive features (Figure 2).

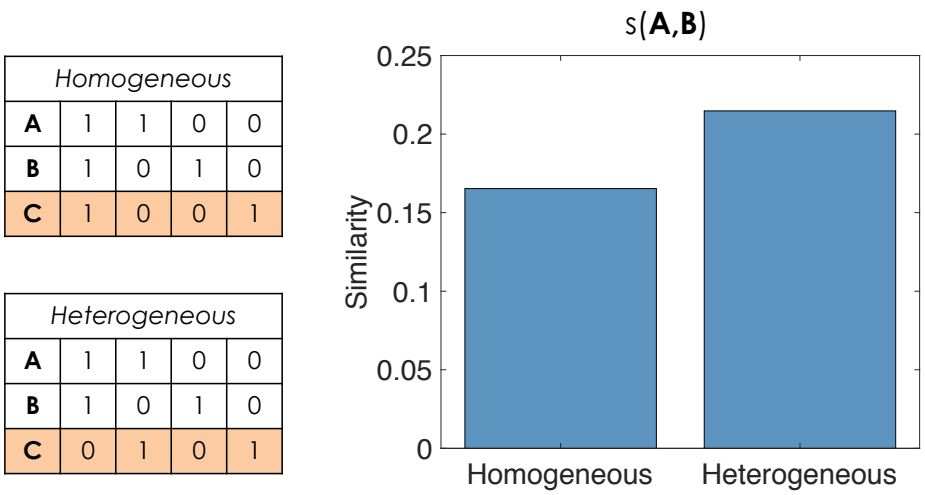

Figure 3: **Simulation of the extension effect**. Adding an object (C) that extends the variability of the first feature (heterogeneous condition) increases the similarity between objects A and B compared to when the added object does not extend the variability (homogeneous condition).

The Bayesian contrast model (using the non-additive version for simplicity) captures this finding because the shared feature weight decreases with feature probability while the distinctive feature weight increases with feature probability (Figure 1). The model does not seem to capture the fact that in Ritov's data, the distinctive weight is larger than the shared weight in the frequent condition. However, that difference was much smaller (0.87 for frequent vs. 0.83 for rare) when Ritov restricted the analysis to trials on which the feature sets were kept constant across conditions, suggesting that this difference is not a robust property of the data.

Another manifestation of the rare feature match effect is what Tversky (1977) referred to as the *extension effect*: adding objects to a set such that the set becomes heterogeneous with respect to a particular feature causes that feature to be salient relative to a condition in which the objects are homogeneous with respect to the feature. Heterogeneity has the effect of decreasing feature frequency, and thereby increasing the weight of matches. As shown in Figure 3, The additive version of the Bayesian contrast model reproduces the extension effect.

Finally, we consider the *variability-based* diagnosticity principle proposed by Goldstone et al. (1997): high-variability features will attain higher salience. Goldstone and colleagues found evidence for this principle in an experiment that asked participants to choose one of three alternative objects that best matched a standard (S). Two of the objects (A and C) were fixed across conditions (within a set), and one object (B) was manipulated across conditions. In the "shared match" condition, A and B share a feature that matches the

Shared match			
S	1	1	1
A	1	1	0
B	1	0	0
C	0	1	1

Shared mismatch			
S	1	1	1
A	1	1	0
B	0	0	0
C	0	1	1

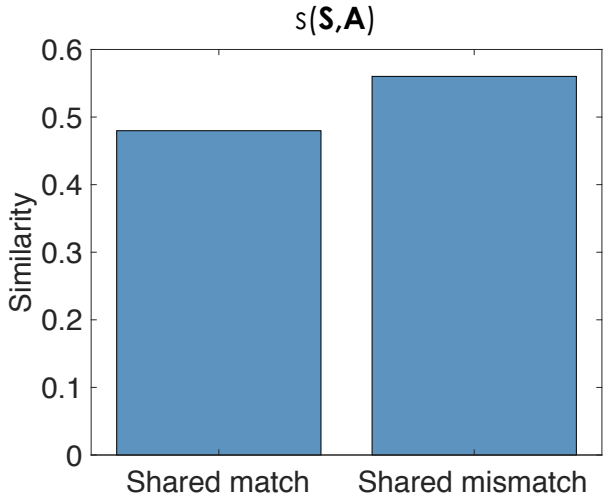

Figure 4: **Variability-based diagnosticity**. When objects A and B share a feature that mismatches the standard (S), the similarity between the standard and A is greater compared to when A and B share a match.

standard. In the "shared mismatch" condition, A and B share a feature that mismatches the standard. Because the shared mismatch condition has greater variability on the critical feature, Goldstone and colleagues predicted (and confirmed experimentally) that the similarity between A and S would be greater in the shared mismatch condition. The Bayesian contrast model explicitly motivates variability-based diagnosticity, since variability increases the weight for both shared and distinctive components Figure 1. Accordingly, simulation of the Goldstone and colleagues experiment shows that shared mismatches induce greater similarity than shared matches (Figure 4). It is important to note, however, that this effect could potentially be explained entirely by the fact that in the shared mismatch condition, the first feature shared by S and A has lower frequency, and thus should have higher weight (without considering the weights of distinctive features). Thus, more experiments are necessary to determine the relative contributions of shared and distinctive features.

Re-analysis of existing similarity data

Although most studies of similarity do not explicitly manipulate feature frequency, there are endogenous variations across domains. To examine this, we re-analyzed 49 publicly available data sets.[1] For the purposes of this analysis, we interpret 'similarity' broadly as 'proximity,' since some of these data sets correspond to confusions in identification paradigms.

In the Appendix, we describe a novel procedure for simultaneously estimating the features

[1]Available courtesy of Michael Lee: `http://faculty.sites.uci.edu/mdlee/similarity-data/`.

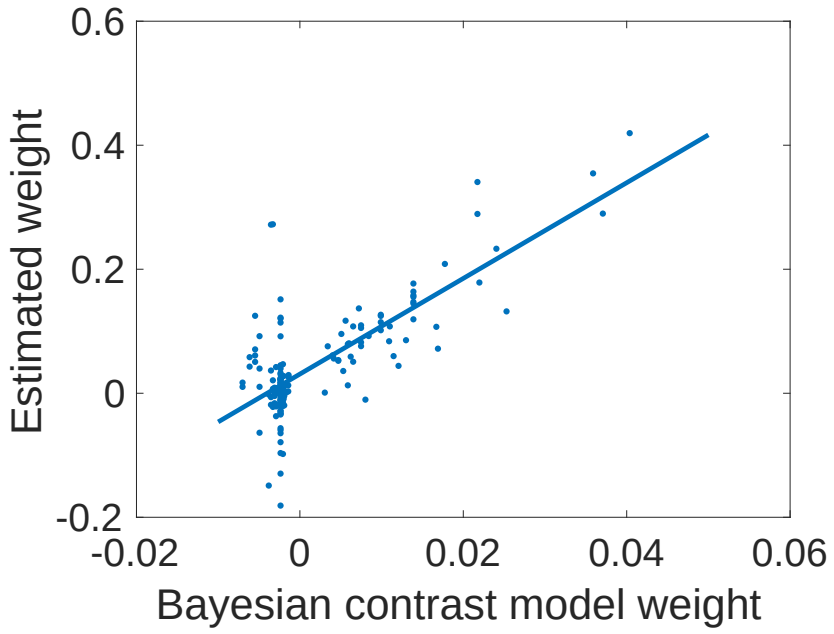

Figure 5: **Comparison of empirical weight estimates and the predictions of the Bayesian contrast model**. Each point represents a single weight (3 weights for each data set, 49 data sets total).

and the weights of the contrast model. Using this method, we estimated the weights and features separately for each data set and found an aggregate correlation of $r = 0.75$ ($p < 0.00001$) between the estimated weights and the Bayesian contrast model's weights (Figure 5). Thus, endogenous variations in weights is well-predicted by properties of the empirical feature distribution. This correlation is particularly notable given that the Bayesian contrast model has no free parameters.

This re-analysis gives us the opportunity to test the quantitative predictions of the Bayesian contrast model, as shown in Figure 1. Inspection of the weights (Figure 6) confirm the prediction that the shared component weight (w_1) should decrease with the mean and increase with the variability, while the distinctive component weight (average of w_2 and w_3) stays mostly flat as a function of the mean and increases with variability. Although in theory the two curves should cross when plotted as a function of the mean, the empirical values of the mean in these data sets are sufficiently low that the cross-over regime is not encountered here. Similarly, although the weights should increase more dramatically as a function of variability, the variability is sufficiently low that the weights stay close to 0, as predicted by the model. The only major discrepancy is that theoretically the distinctive component weight should increase more rapidly than the common component weight as a function of variability, whereas empirically the opposite is the case.

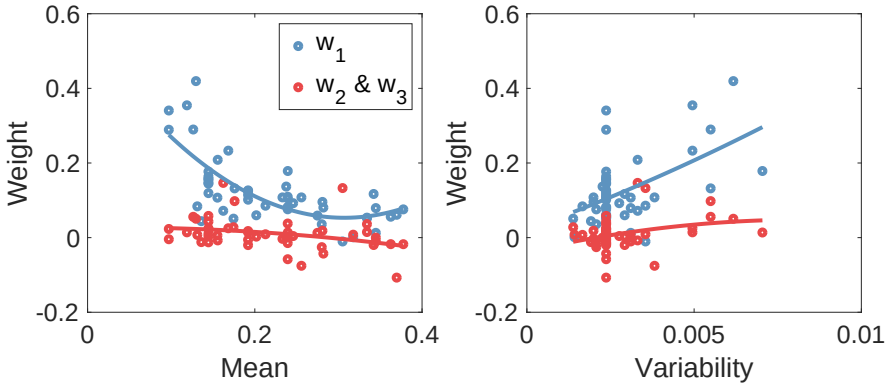

Figure 6: **Parametric effects of hyperparameters on estimated component weights**. (Left) Mean feature activation. (Right) Feature variability. Lines show second-order polynomial fits. The weights for the distinctive component is expressed as the average of w_2 and w_3.

The average distinctive weight is not significantly different from 0 [$t(48) = 1.63, p = 0.11$]. This means that the estimated contrast model tends to only use common features to determine similarity, consistent with the classic additive clustering model (Shepard & Arabie, 1979) and its descendants (e.g., Navarro & Griffiths, 2008; Navarro & Perfors, 2010). Nonetheless, there is structure in the distinctive weights: they correlate positively with v ($r = 0.3, p < 0.05$), as predicted by the Bayesian contrast model. This suggests that distinctive weights must be taken into account in order to understand context effects on similarity judgment.

Discussion

This paper addresses a long-standing challenge in the study of similarity: how does context affect similarity judgment? The answer is provided, at least in part, by a Bayesian model of similarity (Kemp et al., 2005), according to which similarity judgments reflect the posterior probability that two objects were sampled from the same generative process (as opposed to two independent processes). Under a certain set of assumptions about the generative process (namely, that it is a Beta-Bernoulli process), Kemp and colleagues showed that the model is equivalent to the seminal contrast model of Tversky (1977). The contribution of the present work is to show how the hyperparameters of the generative process, estimated from the distribution of objects, allows the Bayesian contrast model to capture a range of contrast effects. In particular, the model correctly predicts that increasing the mean feature activation reduces the weight of shared features, while increasing the variability of feature activation strengthens the weight of both shared and distinctive features. Moreover, the model can capture variation in weights across a large range of experiments, without any

parameter tuning.

Several other Bayesian models also predict a *rare feature match effect* (rare feature matches influence similarity more than frequent feature matches). Tenenbaum and Griffiths (2001), deriving a version of the contrast model from a theory of concept generalization, predicted this effect as a consequence of the *size principle*: hypotheses with smaller sizes have higher likelihoods. In this case, hypotheses correspond to features and the size corresponds to the number of objects possessing a feature. Navarro and Perfors (2010) offered another variation on this theme, and demonstrated the effect empirically for a wide range of features in the Leuven concept database (De Deyne et al., 2008). One limitation of these models is that they make somewhat restrictive assumptions on the feature weights, requiring one or more weights in the contrast model to be 0, whereas the model presented here allows the common and distinctive weights to vary with context. Our analysis of 49 data sets suggests that this variation is linked to systematic structure in the similarity data.

There are a number of aspects of similarity judgment that are not captured by the Bayesian contrast model. First, it cannot handle asymmetry: it requires that the weights of the two distinctive components always be equal. However, Tversky (1977) showed that these weights are asymmetric in some cases. For example, when one object is more prototypical, its distinctive features tend to receive greater weight. Thus, North Korea is judged as more similar to China than China is to North Korea. There are a number of potential ways to model this asymmetry, which go beyond the constraints of the Bayesian contrast model. One is to adopt a transformational view of similarity (Hahn et al., 2003; Imai, 1977), according to which two objects are judged to be similar to the extent that the transformation distance between them is smaller. If transformation probability is biased in one direction (see Hahn, Close, & Graf, 2009), then the resulting similarities can be asymmetric. Kemp et al. (2005) showed that transformational models of similarity can also be derived from a Bayesian analysis of generative processes. A related approach is to assume that similarities reflect conditional probabilities (Griffiths, Steyvers, & Tenenbaum, 2007), which would allow $P(a|b)$ to differ from $P(b|a)$. Finally, some asymmetries may arise from stimulus bias (Holman, 1979; Nosofsky, 1991), which would obviate the need for an asymmetric similarity model.

Another limitation of the Bayesian contrast model is that it does not capture the role of structured relational knowledge (Goldstone, Medin, & Gentner, 1991; Markman & Gentner, 1993; Medin et al., 1993). For example, judged similarity can be increased by changing features such that two objects become superficially more different but relationally more similar (Goldstone et al., 1991). An important direction for future research will be to contemplate versions of the Bayesian framework for similarity built upon generative processes defined over relational structures (e.g., Kemp, Tenenbaum, Niyogi, & Griffiths, 2010; Piantadosi, Tenenbaum, & Goodman, 2016). Such models could provide a bridge to alignment-based theories of similarity (Forbus, Gentner, & Law, 1995; Goldstone, 1994). Some progress in this direction has been made by Lu, Chen, and Holyoak (2012).

An interesting extension of the Bayesian framework would be to incorporate a structure learning mechanism for clustering objects into categories (Gershman & Niv, 2010). The Beta-Bernoulli generative process and its generalization to categorical variables (the Dirichlet-

Multinomial generative process) have been at the heart of unsupervised category learning models (Anderson, 1991; Gershman, Blei, & Niv, 2010; Sanborn, Griffiths, & Navarro, 2010). These models could explain why implicit groups appears to alter similarity judgment (Tversky, 1977), though an important caveat is that the evidence for such grouping effects is still rather weak (Evers & Lakens, 2014; Goldstone et al., 1997). Thus, more experimental work is needed to pin down the empirical foundation of this idea.

Bayesian models have been powerful tools for articulating computational-level hypotheses about cognition. The work of Kemp et al. (2005) is notable in that it connects an abstract Bayesian model directly to a well-known descriptive model of similarity. This connection allows us to understand *why* similarity judgments appear to be well-described in many cases by a weighting of shared and distinctive components. Of course, similarity is multifaceted, as recognized by many authors (e.g., Medin et al., 1993; Sloman & Rips, 1998), and thus likely to resist any single unifying explanation. But by occupying a sufficiently abstract computational perspective, we may begin to understand how the diverse functions of similarity fit together—we may begin to recognize the similarities between similarities.

Acknowledgments

This work was supported by the Office of Naval Research Science of Autonomy program (N00014-17-1-2984) and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

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

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030–1048.

Evers, E. R., & Lakens, D. (2014). Revisiting tversky's diagnosticity principle. *Frontiers in Psychology*, *5*.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

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

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*, 251–256.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 3–28.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*, 222–262.

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition*, *25*, 237–255.

Goodman, N. (1972). *Problems and projects*. Bobbs-Merrill.

Griffiths, T. L., & Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*, 1–32.

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape-similarity judgments. *Psychological Science*, *20*, 447–454.

Holman, E. W. (1979). Monotonic models for asymmetric proximities. *Journal of Mathematical Psychology*, *20*, 1–15.

Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica*, *41*, 433–447.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).

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

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617–648.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467.

Marr, D. (1982). *Vision.* W.H. Freeman.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, *20*, 2597–2628.

Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*, 256–268.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*, 94–140.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*, 392–424.

Ritov, I. (2000). The role of expectations in comparisons. *Psychological Review*, *107*, 345–357.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as

combinations of discrete overlapping properties. *Psychological Review*, *86*, 87–123.
Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, *65*, 87–101.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Appendix: nonparametric estimation of the contrast model

Here we develop a novel Bayesian nonparametric estimation procedure for Tversky's contrast model (Tversky, 1977). Note that this is conceptually distinct from the Bayesian contrast model: it represents the *experimenter's* uncertainty about the weights and features used by subjects in an experiment. As such, it is not a normative model but rather a data analysis tool that allows us to simultaneously estimate the weights and features of the contrast model.

The estimation procedure rests upon the following generative model:

$$\mathbf{Z} \sim \text{IBP}(\alpha) \tag{15}$$

$$s(a,b) \sim \mathcal{N}(\hat{s}(a,b), \sigma^2) \tag{16}$$

$$w_i \sim \mathcal{N}(0, \tau^2). \tag{17}$$

where

$$\hat{s}(a,b) = w_1 \mathbf{z}_a \mathbf{z}_b^\top - w_2 \mathbf{z}_a (1 - \mathbf{z}_b^\top) - w_3 (1 - \mathbf{z}_a) \mathbf{z}_b^\top \tag{18}$$

is the output of Tversky's contrast model for the non-additive case with $F(\cdot) = |\cdot|$ (the additive case can be handled in a straightforward manner but we do not address it here). The matrix $\mathbf{Z}$ is an objects-by-features binary matrix indicating whether object n possesses a particular feature k ($Z_{nk} = 1$). The vectors $\mathbf{z}_n$ are rows of $\mathbf{Z}$. We assume that $\mathbf{Z}$ is drawn from an *Indian buffet process* (IBP; Griffiths & Ghahramani, 2011) with concentration parameter α. The nonparametric specification allows the model to infer the number of features automatically, but α induces a bias towards fewer active features. The features are combined to produce similarity judgments according to Tversky's contrast model with weights $\{w_1, w_2, w_3\}$ sampled from a zero-mean Gaussian. This generative model is similar to the one developed by Navarro and Griffiths (2008), with the key difference that the features produce similarity judgments using the contrast equation rather than an additive clustering equation (which assumes that only common features determine similarity).

To fit the model, we used annealed Gibbs sampling, iteratively sampling each element of $\mathbf{Z}$ from its conditional distribution holding the others fixed, raised to the power of t (the iteration number). This produces an estimate of the maximum *a posteriori* value of $\mathbf{Z}$:

$$\hat{\mathbf{Z}} = \operatorname*{argmax}_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{S}), \tag{19}$$

where $\mathbf{S}$ is the matrix of observed similarity judgments. Typically 50 iterations were sufficient to reach a local maximum. We analytically marginalized out the weights (see Griffiths & Ghahramani, 2011) during sampling and then computed the posterior mean conditional on the point estimate of $\mathbf{Z}$. The hyperparameters were fixed to $\sigma^2 = 1, \tau^2 = 1, \alpha = 1$.