Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 12 (2011) 1–18

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Packing: A geometric analysis of feature selection and category formation

Action editor: Stephen Jose Hanson

Shohei Hidaka[a,*], Linda B. Smith[b]

[a] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[b] Department of Psychological and Brain Sciences, Indiana University, 1101 East Tenth Street, Bloomington, IN 47405-7007, United States

Received 3 June 2009; received in revised form 29 April 2010; accepted 6 July 2010
Available online 29 July 2010

Abstract

This paper presents a geometrical analysis of how local interactions in a large population of categories packed into a feature space create a global structure of feature relevance. The theory is a formal proof that the joint optimization of discrimination and inclusion creates a smooth space of categories such that near categories in the similarity space have similar generalization gradients. Packing theory offers a unified account of several phenomena in human categorization including the differential importance of different features for different kinds of categories, the dissociation between judgments of similarity and judgments of category membership, and children's ability to generalize a category from very few examples.

Keywords: Categorization; Cognitive development; Word learning; Feature selection

1. Introduction

There are an infinite number of objectively correct descriptions of the features characteristic of any thing. Thus, Murphy and Medin (1985) argue that the key problem for any theory of categories is feature selection: picking out the relevant set of features for forming a category and generalizing to new instances. The feature selection problem is particularly difficult because considerable research on human categories indicates that the features people think are relevant depend on the kind of category (Macario, 1991; Murphy & Medin, 1985; Samuelson & Smith, 1999). For example, color is relevant to food but not to artifacts; material (e.g., wood versus plastic) is relevant to substance categories but not typically to artifact categories. This leads to a circularity as pointed out by Murphy and Medin (1985); to know that something is a pea, for example, one needs to attend to its color, but to know that one should attend to its color, one has to know it is a potential pea. Children as young as 2 and 3 years of age seem to already know this and exploit these regularities when forming new categories (Colunga & Smith, 2005; Yoshida & Smith, 2003). This paper presents a new analysis of feature selection based on the idea that individual categories reside in a larger geometry of other categories. Nearby categories through processes of generalization and discrimination compete and these local interactions set up a gradient of feature relevance such that categories that are near to each other in the feature space have similar feature distributions over their instances. We call this proposal "Packing theory" because the joint optimization of generalization and discrimination yields a space of categories that is like a suitcase of well-packed clothes folded into the right shapes so that they fit tightly together. The proposal, in the form of a mathematical proof, draws on two empirical results: (1) experiments and theoretical analyses showing that the distribution of instances in a feature space is critical to the weighting of those features in adult category judgments and (2) evidence from young children's category judgments that near categories in the feature space are

* Corresponding author.
E-mail address: shhidaka@jaist.ac.jp (S. Hidaka).

1389-0417/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2010.07.004

generalized in similar ways. Distributions of instances in a seminal paper, Rips (1989) reported that judgments of instance similarity to a category and judgments of the likelihood that the instance is a member of the category did not align. The result, now replicated many times, shows that people take into account how broadly known instances vary on particular properties (Holland, Holyoak, Nisbett, & Thagard, 1986; Kloos & Sloutsky, 2008; Nisbett, Krantz, Jepson, & Kunda, 1983; Rips & Collins, 1993; Thibaut, Dupont, & Anselme, 2002; Zeigenfuse & Lee, 2009) with respect to judgments of the likelihood that an instance was a member of the category, people take into account the frequency of features across known instances and do not just judge the likelihood of membership in the category by similarity across all features. Importantly, however, similarity judgments in Rips' study were not influenced by the frequency distribution of features across category instances (see also, Holland et al., 1986; Nisbett et al., 1983; Rips & Collins, 1993; Stewart & Chater, 2002; Thibaut et al., 2002). Rather, the similarity relations of potential instances to each other and the importance of features to judgments of the likelihood of category membership appear to be separable sources of information about category structure with the distribution of features across known instances most critical in determining the importance of features in decisions about category membership. Fig. 1 provides an illustration. The figure shows the feature distribution on some continuous dimension for two categories, A and B. A feature that is highly frequent and varies little within a category is more defining of category membership than one that is less frequent and varies more broadly. Thus, a novel instance that falls just to the right side of the dotted line would be farther from the central tendency of B than A, but may be judged as a member of category B and not as a member of category A. This is because the likelihood of that feature given category B is greater than the likelihood of the feature given category A. This can also be conceptualized in terms of this feature having greater importance to category A than B.

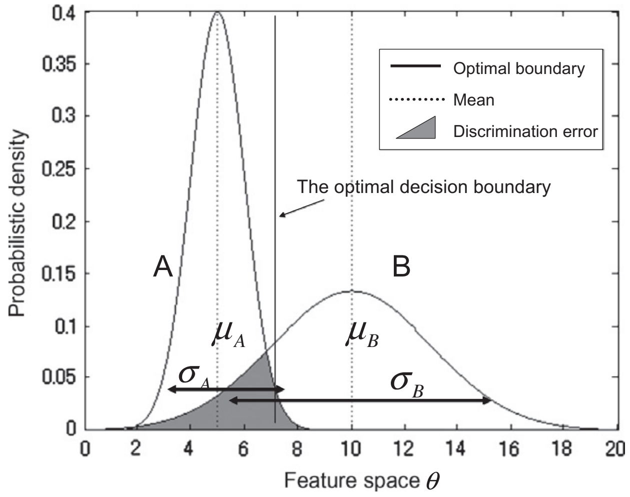


Fig. 1. Likelihoods of instances in two categories, A and B, in a hypothetical feature space. The solid line shows the optimal decision boundary between category A and B. The broken line shows mean value of the instances in the feature space for each category.

The potentially separate contributions of similarity and the category likelihoods of features provide a foundation for the present theory. Similarity is a reflection of the proximity of instances and categories in the feature space. The density of instances in this feature space result in local distortions of feature importance in the space, a result of competitions among nearby categories. The result is a patchwork of local distortions that set up a global gradient of feature importance that constrains and promotes certain category organizations over others as a function of location in the global geometry. Further, in this view, the weighting of features is *locally* distorted, but similarity is not.

1.1. Nearby categories

Studies of young children's novel noun generalizations also suggest that the proximity of categories to each other in a feature space influence category formation. These results derive from laboratory studies of how 2-and 3-year olds generalize a category to new instances given a single exemplar. In these experiments, children are given a novel never-seen-before thing, told its name ("This is a dax") and asked what other things have that name. The results show that children extend the names for things with features typical of animates (e.g., eyes) by multiple similarities, for things with features typical of artifacts by shape (e.g., solid and angular shapes), and for things with features typical of substances by material (e.g., nonsolid, rounded flat shape). The children systematically extend the name to new instances by different features for different kinds (Booth & Waxman, 2002; Gathercole & Min, 1997; Imai & Gentner, 1997; Jones et al., 1991; Jones & Smith, 2002; Kobayashi, 1998; Landau, Smith, & Jones, 1988, 1992, 1998; Markman, 1989; Soja et al., 1991; Yoshida & Smith, 2001; see also, Gelman & Coley, 1991; Keil, 1994).

Critically, the features that cue these categories having eyes or legs, being angular or rounded, being solid or nonsolid-may be treated as continuous rather than as discrete and categorical features. When the degree to which instances present these features is systematically varied so that named exemplars are more or less animal-like or more or less artifact-like (Colunga & Smith, 2005, 2008; Yoshida & Smith, 2003) children show graded generalization patterns: exemplars with similar features are generalized in similar ways and there is a graded, smooth, shift in the patterns of generalizations across the feature space. Based on an analysis of the structure of early-learned nouns, Colunga and Smith (Colunga & Smith, 2005, 2008; Samuelson & Smith, 1999) proposed that children's generalizations reflected the instance distributions of early-learned nouns that categories of the same general kind (artifacts versus animals versus substances) typically have many overlapping features and also have similar dimensions as the basis for including instances and discriminating membership in nearby categories. In brief, categories of the same kind will be similar to each other, located near each other in some larger feature geometry of categories, and have similar patterns of instance distributions. This has potentially powerful consequences: If feature importance is similar for nearby categories, then the location of categories or even one instance – within that larger space of categories could indicate the relevant features for determining category membership. Thus, children's systematic novel noun generalizations may reflect the distribution of features for nearby known categories.

The findings in the literature on children's generalizations from a single instance of a category may be summarized with respect to Fig. 2. The cube represents some large hyperspace of categories on many dimensions and features. Within that space we know from previous studies of adult judgments of category structure and from children's noun generalizations (Colunga & Smith, 2005; Samuelson & Smith, 1999; Soja et al., 1991) that solid, rigid and constructed things, things like chairs and tables and shovels are in categories in which instances tend to be similar in shape but different in other properties. This category generalization pattern is represented by the ellipses in the bottom left corner; these are narrow in one direction (constrained in their shape variability) but broad in other directions (varying more broadly in other properties such as color or texture). We also know from previous studies of adult judgments of category structure and from children's novel noun generalizations (Colunga & Smith, 2005; Samuelson & Smith, 1999; Soja et al., 1991), that nonsolid, nonrigid things with accidental shapes (things like sand, powder, and water) tend to be in categories well organized by material. This category generalization pattern is represented by the ellipses in the upper right corner of the hyperspace; these are broad in one direction (wide variation in shape) but narrow in other directions (constrained in material and texture). Finally, Colunga and Smith (2005, 2008) found evidence for a gradient of generalization patterns within one local region of feature space of early-learned noun categories. In sum, the evidence suggests that near categories (with similar instances) have similar generalization patterns and far categories (with dissimilar instances) have dissimilar generalization patterns.

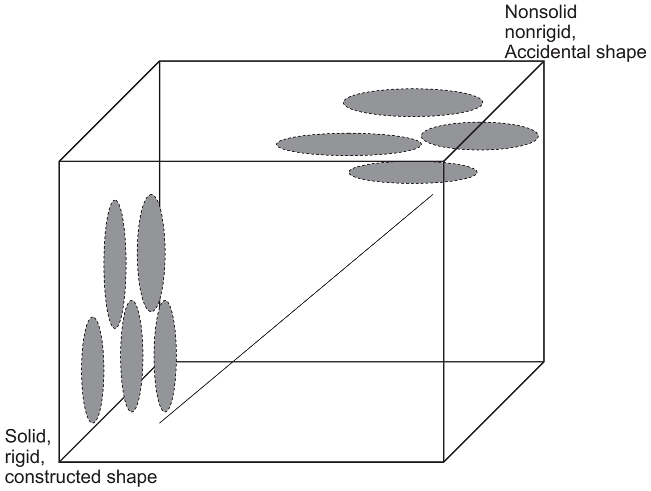


Fig. 2. A hyperspace of categories. The ellipses represent categories with particular generalization patterns (constrained in some directions but allowing variability in others). Packing theory predicts that near categories in the space will have similar generalization patterns and that there should be a smooth gradient of changing category generalizations as one moves in any direction in the space. Past research shows that categories of solid, rigid and constructed things are generalized by shape but categories of nonsolid, nonrigid, and accidentally shaped things are generalized by material. Packing theory predicts a graded transition in feature space between these two kinds of category organizations.

At present Fig. 2 represents a theoretical conjecture, one based on empirical evidence about one small region of the space of human categories and one that has not been theoretically analyzed. Still, it is a particularly interesting conjecture in the context of Rips' (1989) insight that the distributions of features are distinct from similarity and determine the relative importance of features in category judgment. Packing theory builds on these two ideas: distributions of instances and proximity of categories in the feature space to suggest how they jointly determine feature selection.

1.2. Starting assumptions

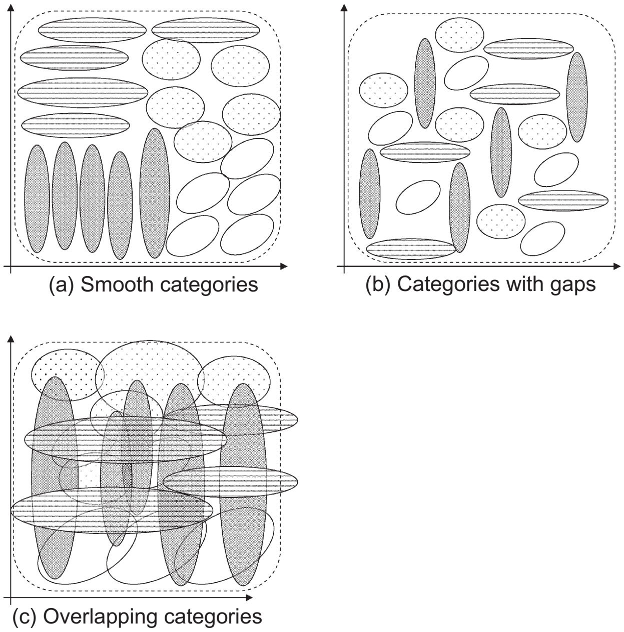
Three theoretical assumptions form the backdrop for Packing theory. First, as in many contemporary theories of categorization, we take an exemplar approach (Ashby & Townsend, 1986; Nosofsky, 1986; Nosofsky, Palmeri, & McKinley, 1994). We assume that noun categories begin with mappings between names and specific instances with generalization to new instances by some weighted function of feature similarity. Packing is about those weighted features. This means that categories do not have fixed or rule-like boundaries but rather probabilistic boundaries. Second, we assume that at the probabilistic edges of categories, there is competition among categories for instances. Competition characterizes representational processes at the cognitive, sensory, motor, cortical and subcortical levels. In general, activation of a representation of some property, event or object is at the expense of other complementary representations (Beck & Kastner, 2009; Duncan, 1996; Marslen-Wilson, 1987; Swingley & Aslin, 2007). Packing theory proposes that nearby categories have similarly shaped generalizations patterns because of the joint optimization of including nearby instances and discriminating instances associated with different categories. Third, the present approach assumes some feature-based representation of categories (McRae, Cree, Seidenberg, & McNorgan, 2005). However, we make no assumptions about the specific nature of these features or these origins. Although we will use perceptual features in ours discussions and simulations, the relevant features could be perceptual, functional or conceptual. Packing theory is a general theory, about any distribution of many instances in many categories across any set of features and dimensions. Moreover, the theory does not need the right pre-specification of the set of features and dimensions. Optimization within the theory depends only on distance relations in the space (and thus on the number of orthogonal, that is uncorrelated, dimensions but not on any assumptions about what orthogonal directions in that space constitute the dimensions). Further, the predictions are general; along any direction in that space (a direction that might consist of joint changes in two psychological dimensions, angularity and rigidity, for example), one should see near categories having more similar generalization patterns and far categories having more different generalization patterns. To present the theory, we will often use figures illustrating instance and category relations in a two-dimensional space; however, the formal theory as presented in the proof-and the conceptual assumptions behind it-assume a high dimensional space.

2. Packing theory

Geometry is principally about how one determines neighbors. If the structure of neighboring categories determines feature selection, then a geometrical analysis should enhance our understanding of why categories have the structure they do. Fig. 2 is the starting conjecture for the theory presented here and it suggests that near categories have similar instance distributions whereas as far categories have more dissimilar instance distributions. Thus, a geometry is needed when it represents both local distortions and the more global structure that emerges from a space of such local distortions. Many theorists of categorization have suggested that although Euclidean assumptions work well within small and local stimulus spaces, a Riemann (or non-Euclidian) geometry is better for characterizing the local and global structure of large systems of categories (Griffiths, Steyvers, & Tenenbaum, 2007; Steyvers & Tenenbaum, 2005; Tversky & Hutchinson, 1986). Packing theory follows this lead. We first present a conceptual understanding of the main idea that packing categories into a space creates a smooth structure and then the formal proof.

2.1. Well-packed categories

Fig. 3 shows three different sets of categories distributed uniformly within a (for exposition only 2-dimensional) feature space. Fig. 3a shows a geometry, like that of young children; near categories have similar patterns of feature distributions and far categories have different ones. Such a geometry is not logically necessary (though it may be psychologically likely). One could have a geometry of categories like that in Fig. 3b, where each category has its own organization unrelated to those of near neighbors and more specifically, a geometry in which near categories do not share similar feature importance. The two spaces of categories illustrated in Fig. 3a and b are alike in that in both of these spaces there is little category overlap. That is, in both of these spaces, the categories discriminate among instances. However, the categories in 3b are not smooth in that near categories have different shapes. Moreover, this structure leads to gaps in the space, possible instances (feature combinations) that do not belong to any known category. The categories in Fig. 3b could be pushed close together to lessen the gaps. However, given the non-smooth structure, there would always be some gaps, unless the categories are pushed so close that they overlap as in Fig. 3c. Fig. 3c then shows a space of categories with no gaps, but also one in which individual categories do not discriminate well among instances. The main point is that if neighboring categories have different shapes there will either be gaps in the space with no potential category corresponding to those instances or there will be overlapping instances. A smooth space of categories, a space in which nearby categories have similar shapes, can be packed in tighter. This is a geometry in which categories include and discriminate among all potential instances.


(a) Smooth categories

(b) Categories with gaps

(c) Overlapping categories

Fig. 3. A cartoon of populations of categories in a feature space illustrating three different ways those might categories might fit into the space. Each ellipsis indicates equal-likelihood contour of category. The broken enclosure indicates the space of instances to be categorized.

Packing theory proposes that a smooth space of categories results from the optimization with respect to two constraints: minimizing gaps and minimizing overlap. These constraints are understood in terms of the joint optimization of including all experienced and potential instances in a category and discriminating instances of nearby categories. The inclusion–discrimination problem is illustrated with respect to the simple case of two categories in Fig. 4. Each category has a distribution of experienced instances indicated by the diamonds and the crosses. We assume that the learner can be more certain about the category membership of some instances than others can; that is, the probability that each of these instances is in the category varies. If a category is considered alone, it might be

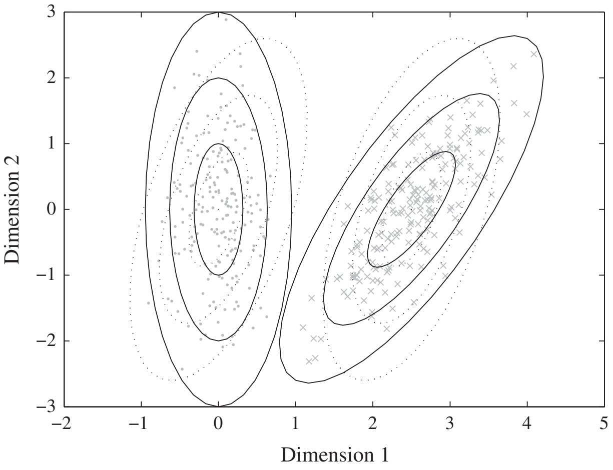


Fig. 4. Two categories and their instances on two-dimensional feature space. The dots and crosses show the respective instances of the two categories. The broken and solid ellipses indicate equal-likelihood contours with and without consideration to category discrimination respectively.

described in terms of its central tendency and estimated category distribution of instances (or covariance of the features over the instances). The solid lines that indicate the confidence intervals around each category illustrate this. However, the learner needs to consider instances not just with respect to a single category, but also with respect to nearby categories. Therefore, such a learner might decrease the weight of influence of any instance on the estimated category structure by a measure of its confusability between the categories. This is plausible because the shared instances in between these nearby categories are less informative as exemplars of categories than the other non-confusing instances. Thus, the estimation of category structure may "discount" the instances in between the two categories. Doing this results in an estimated category distribution, that is shifted such that the generalization patterns for the two categories are more aligned and more similar, which is shown with dotted lines. This is the core idea of packing theory.

2.2. Formulation of the theory

It is relatively difficult to describe the whole structure formed by a large number of categories when they locally interact across all categories at once. N categories have $\frac{N(N-1)}{2}$ possible pairs of categories. Moving one category to maximize its distance from some other category may influence the other $(N-1)$ categories, and those other categories' movements will have secondary effects (even on the first one), and so forth. Thus, two categories that compete with each other in a local region in a feature space *influence the whole structure by chains of category interactions*. The goal of the theory formulation is to describe the dynamics of category inclusion and discrimination in a general case, and to specify a stable optimal state for the N-categories case. Mathematically, the framework can be classified in one of variant of a broad sense of the Linear Discriminant Analysis (LDA, see Duda, Hart, & Stork, 2000), although we do not necessarily limit the theory to employ only linear transformation or assume homoscedastic category distributions (see also, Ashby & Townsend (1986) for the similar formulation using normal distribution in psychological field).

2.2.1. Inclusion

We begin with a standard formalization of the distribution of instances in a category as a multi-dimensional normal distribution; that is, the conditional probabilistic density of an instance having feature θ given category c_i is assumed to follow a multi-dimensional normal distribution:

$$P(\theta|c_i) = \left\{(2\pi)^2|\sigma_i|\right\}^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\theta-\mu_i)^T\sigma_i^{-1}(\theta-\mu_i)\right] \quad (1)$$

where μ_i and σ_i are respectively mean vector and covariance matrix of the features that characterize the instances of category c_i. The superscript "T" indicates transposition of the matrix. We identify the central tendency and distribution pattern of these features as the mean vector of the category, and the covariance matrix respectively. The motivation of this formulation of category generalization derives from Shepard (1958) characterization of the generalization gradient as an exponential decay function of given psychological distance.

When we have K_i instances $(k = 1, 2, \ldots, K_i)$ for category c_i, the log-likelihood G_i that these instances come from category c_i is the joint probability of all instances

$$G_i = \log\left\{\prod_{k=1}^{K} P(x_k|c_i)P(c_i)\right\} = \sum_{k=1}^{K_i} G_{ik} \quad (2)$$

where $G_{ik} = \log\{P(x_k|c_i)P(c_i)\}$.

Note that $\log(x)$ is a monotonic function for all $x > 0$. Thus we can identify a solution of x that maximizes the log-likelihood. For all categories $(i = 1, 2, \ldots, N)$, the log-likelihood is $G = \sum_{i=1}^{N} G_i$.

$$G = \sum_{i=1}^{N}\sum_{k=1}^{K_i} \log\{P(x_k|c_i)P(c_i)\} \quad (3)$$

This is the formal definition of the likelihoods of instances with respect to individual categories, which we call inclusion. In this formulation, the mean vectors $\hat{\mu}_i$ and covariance matrices $\hat{\sigma}_i$ $(i = 1, 2, \ldots, N)$, maximize the log-likelihood (likelihood) as follows:

$$\hat{\mu}_i = \frac{1}{K}\sum_{k=1}^{K_i} x_k \quad (4)$$

$$\hat{\sigma}_i = \frac{1}{K}\sum_{k=1}^{K_i}(x_k - \hat{\mu}_i)(x_k - \hat{\mu}_i)^T \quad (5)$$

These maximum likelihood estimates are simply the mean vectors and covariance matrices of all instances of a category.

2.2.2. Discrimination

Consider the simple case with two categories in a one-dimensional feature space as in the example from Rips (1989) in Fig. 1. Likelihoods of category A and B are shown as the solid lines: category A has the central tendency on the left side in which the instance is most likely, and category B has the central tendency on the right side. An optimal category judgment for a given instance is to judge it as belonging to the most likely category. Thus the optimal solution is to judge an instance on the left side of the dashed line in Fig. 1 to be in category A and otherwise to judge it to be in category B. Meanwhile, the error probability of discrimination in the optimal judgment is the sum of the non-maximum category likelihood, the shaded region in Fig. 1. Therefore, we formally define discriminability as the probability of discriminating error in this optimal category judgment.

Although the minimum or maximum function is difficult to solve, we can obtain the upper bound of the discriminating error (log-likelihood) F_{ij} for category c_i and c_j as follows:

$$F_{ij} = \log\left[\int_{\Omega}\left\{P(\theta|c_i)P(\theta|c_j)\right\}^{-\frac{1}{2}}\right] \tag{6}$$

In particular, when likelihoods of categories are normally distributed, it is called the Bhattecheryya bound (Duda et al., 2000). In fact, the non-maximum likelihood of a given pair is the classification (instance, category) error rate and the non-maximum likelihood has the following upper bound: $\min(P(\theta|c_i), P(\theta|c_j)) \leqslant P(\theta|c_i)^{\alpha}P(\theta|c_j)^{1-\alpha}$, where $0 \leqslant \alpha \leqslant 1$. Thus, Eq. (6) is the upper bound of error in the optimal classification with $\alpha = \frac{1}{2}$. The term $P(\theta|c_i)^{\alpha}P(\theta|c_j)^{1-\alpha}$ is the function of α, and a particular α may allow the tightest upper bound with a general case of α. (Eq. (6) is called the Chernoff bound). Here, we assume $\alpha = \frac{1}{2}$ for simplification of formulation.

$$F_{ij} = -\frac{1}{4}(\mu_i - \mu_j)^T(\sigma_i + \sigma_j)^{-1}(\mu_i - \mu_j) \tag{7}$$

$$-\frac{1}{2}\log\left|\frac{1}{2}(\sigma_i + \sigma_j)\right| + \frac{1}{4}\log\left(|\sigma_i||\sigma_j|\right) \tag{8}$$

The first term $(\mu_i - \mu_j)^T(\sigma_i + \sigma_j)^{-1}(\mu_i - \mu_j)$ indicates the distance between mean vectors of the categories weighted by their covariance matrices, which is zero when $\mu_i = \mu_j$. And the second and third terms, $-\frac{1}{2}\log|\frac{1}{2}(\sigma_i + \sigma_j)| + \frac{1}{4}\log(|\sigma_i||\sigma_j|)$, indicate the distance between the covariance matrices of categories which is zero when $\sigma_i = \sigma_j$. Thus, obviously the discrimination error is maximized when $\mu_i = \mu_j$ and $\sigma_i = \sigma_j$, when category c_i and c_j are identical. Meanwhile the discriminating error is minimized when the distance between two central tendencies or distributions goes to infinity $(\mu_i - \mu_j)^T(\mu_i - \mu_j) \rightarrow \infty$ or $\mathrm{tr}[(\mu_i - \mu_j)^T(\mu_i - \mu_j)] \rightarrow \infty$ where $\mathrm{tr}[X]$ is trace of the matrix X. The minimum and maximum concern is the one component of discrimination.

2.2.3. The packing metric

The joint optimization of discrimination by reducing the discriminating error and the optimization of the inclusion of instances with respect to the likelihood of known instances results in a solution that is constrained to an inter region between these two extremes. In the general case with N categories, we define the sum of all possible $\frac{N(N-1)}{2}$ pairs (including symmetric terms) of discriminating error as discriminability.

$$F = \log\left[\sum_{i=1}^{N}\sum_{j=1}^{N}P(c_i)^{\frac{1}{2}}P(c_j)^{\frac{1}{2}}\exp(F_{ij})\right] \tag{9}$$

More formally, we obtain a set of optimal solutions by deriving the differential of Eq. (3) (inclusion) and Eq. (8) (discrimination). Since the desired organization of categories should maximize both discriminability and inclusion simultaneously, we define the packing metric function as a weighted summation of Eqs. (3) and (8) with a multiplier:

$$L = G + \lambda(F - C) \tag{10}$$

where C is a particular constant, which indicates a level for discriminability F to satisfy. According to the Lagrange multiplier method, the differential of parameters $\frac{\partial L}{\partial X} = 0$ gives the necessary condition for the optimal solution of $X \subset \{\mu_i, \sigma_i, \lambda\}$ on condition that the discrimination error is a particular criterion $F = C$. Since the probability of discrimination error is calculated for all possible pairs (categories $i,j = 1, 2, \ldots, N$), this is the maximization of likelihood with latent variables of the discriminated pairs. The optimization is computed with the EM algorithm, in which the E-step computes the expectation with the unknown pairing probability, and the M-step maximizes the expectation of likelihood function (Dempster, Laird, & Rubin, 1977). Thus, the expected likelihood L, with respect to a variable for the ith category X_i, is calculated: $\frac{\partial L}{\partial X_i} = \sum_i^N\sum_j^N Q_{ij}\frac{\partial F_{ij}}{\partial X_i} + \sum_i^N\sum_k^{K_i}R_{ik}\frac{\partial F_{ik}}{\partial X_i}$, where the probability of category pairs are $Q_{ij} = \frac{\sqrt{P(c_i)P(c_j)}\exp(F_{ij})}{\sum_{i=1}^N\sum_{j=1}^N\sqrt{P(c_i)P(c_j)}\exp(F_{ij})}$, and the probability from categories to instances $R_{ik} = \frac{P(c_i)P(x_k|c_i)}{\sum_{i=1}^N\sum_{k=1}^{K_i}P(c_i)P(x_k|c_i)}$, calculated in the following derivation. Note that $P(x_k|c_i)$ is a given binary constant variable, either one or zero, in case of supervised learning, and it should be estimated from its likelihood in case of unsupervised learning.

2.2.4. Optimal solutions for covariance matrices

Consider the case in which the mean of a category, but not its covariance, is specified. Packing theory assumes a covariance for that category derives from the optimization of discrimination and inclusion with respect to the known categories. This is formally given as the solution of the optimal covariance of an unknown category when the mean of the category and the set of means and covariances of the other known categories are given. Thus, in that case, we solve the differential of the packing metric with respect to the covariance matrix. We next derive these differentials to show that these optimal solutions imply a smooth organization of categories in general. As a preview, the following derivations show that a solution that optimizes discrimination and inclusion implies a particular pattern of organization, that of smooth categories, that emerges out of chains of local category interactions. In addition, the form of the optimal solution also suggests how known categories constrain the formation of new categories.

The differential of the functions of likelihoods and discriminability with respect to covariance matrix σ_i is,

$$\frac{\partial F_{ij}}{\partial \sigma_{ij}} = -\frac{1}{4}\sigma_i^{-1}\left\{(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T + \hat{\sigma}_{ij} - \sigma_i\right\} \tag{11}$$

where $\hat{\sigma}_{ij} = 2\sigma_i(\sigma_i + \sigma_j)^{-1}\sigma_j$ and $\bar{\mu}_{ij} = \frac{1}{2}\hat{\sigma}_{ij}(\sigma_i^{-1}\mu_i + \sigma_j^{-1}\mu_j)^{-1}$ σ_j, and

$$\frac{\partial G_{ik}}{\partial \sigma_i} = -\frac{1}{2}\sigma_i^{-1}\left\{(\mu_i - x_{ik})(\mu_i - x_{ik})^T - \sigma_i\right\}\sigma_i^{-1} \tag{12}$$

See Appendix A for the detailed derivation of Eqs. (11) and (12). Since a covariance matrix must be a positive definite, we parameterize the covariance matrix σ_i using its lth eigenvector y_{il} and lth eigenvalue η_{il} ($l = 1, 2, \ldots, D$), that is $\sigma_i = \sum_{l=1}^{D}\eta_{il}y_{il}y_{il}^T$. Solving $\frac{\partial L}{\partial y_{il}} = \sum_{i=1}^{N}\sum_{k=1}^{K_i}\frac{\partial G_{ik}}{\partial y_{il}} + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial F_{ij}}{\partial y_{il}}$, we obtain the following generalized eigenvalue problem[1] (see also Appendix A):

$$\left[\sum_{k=1}^{K_i}R_{ik}(S_{ik} - \sigma_i) - \lambda\sum_{j=1}^{N}Q_{ij}(\widehat{S}_{ij} + \hat{\sigma}_{ij} - \sigma_i)\right]y_{il} = 0 \tag{13}$$

where $S_{ik} = (x_{ik} - \mu_i)(x_{ik} - \mu_i)^T$ and $\widehat{S}_{ij} = (\bar{\mu}_{ij} - \mu_i)(\bar{\mu}_{ij} - \mu_i)^T$. Because Eq. (13) is also an eigenvalue form with a particular constant λ, we identify $\lambda = \eta_{il}^{-1}$ and, using the relationship between the eigenvector and its matrix $\sigma_i y_{il} = \eta_{il}y_{il}$, we obtain the quadratic eigenvalue problem.

$$(\eta_{il}^2\Phi_{i2} + \eta_{il}\Phi_{i1} + \Phi_{i0})y_{il} = 0 \tag{14}$$

where $\Phi_{i0} = \sum_{j=1}^{N}Q_{ij}(\hat{\sigma}_{ij} + \widehat{S}_{ij})$, $\Phi_{i1} = -\sum_{k=1}^{K_i}R_{ik}\sigma_{ik} - \sum_{j=1}^{N}Q_{ij}$, $\Phi_{i2} = \sum_{k=1}^{K_i}R_{ik}I_D$, and I_D is Dth order identity matrix.

The Eq. (13) indicates that the specific structure of category distribution consists of three separate components. The first component in Φ_{i1} which includes S_{ik} represents the deviation of instances from the central tendency of category (scatter matrix of instances). The first term in Φ_{i0} represents the "harmonic mean" of nearby covariance matrices $\hat{\sigma}_{ij}$. The second term in Φ_{i0} represents the scatter matrix $\widehat{S}_{ij}$ that characterizes local distribution of the central tendencies to nearby categories.

To understand the meaning of these components, we draw the geometric interpretation of these three components, the scatter matrix of instances S_{ik}, the scatter matrix of categories $\widehat{S}_{ij}$, and the harmonic mean of covariance matrices $\hat{\sigma}_{ij}$ (Fig. 5). The probabilistic density Q_{ij} weighting to each component exponentially decays in proportion to the "weighted distance" F_{ij} between category c_i and c_j. This "weighted distance" is with respect to the distance between central tendencies and also the *distance between covariance matrices*. That means that interactions among categories are limited to a particular local region. In Fig. 5b, the likelihood contours of the closest categories to a target category are shown as ellipses, and the probabilistic weighting Q_{ij} between the target category (center) and others is indicated by the shading. With respect to this locality, the scatter matrix of categories $\widehat{S}_{ij}$ reflects the variance pattern from the center of category c_i to other relatively "close" category centers. The harmonic means of covariance matrices $\hat{\sigma}_{ij}$ indicate the averaged covariance matrices among the closest categories. Note that it is a "harmonic" average, not an "arithmetic" one, because the inverse (reciprocal number) of the covariance matrix (and not the covariance matrix per se) is appropriate for the probabilistic density function with respect to inclusion and discriminability. The point of all this is that categories with similar patterns of covariance matrices that surround another category, will influence the surrounded category, distorting the feature weighting at the edges so that the surrounded category is more similar to the surrounding categories in its instance distributions.

These effects depend on the proximity of the categories and the need to discriminate instances at the edge of the distributions of adjacent categories S_i is covariance of instances belonging to a category, which is the natural statistical property with respect to the likelihood of instances without discriminability. The magnitude of Q_{ij} and R_{ik} are quite influential in determining the weighting between the likelihood of the instances of a category and the discriminability between categories. If $Q_{ij} \to 0 (j = 1, 2, \ldots, N)$. Thus, if the distances of the central tendencies among categories increase (increasing gaps), the estimated category distribution will depend only on the covariance of instances S_i. Meanwhile, if the number of experiences instances of categories decreases (i.e., $K_i \to 0$ or $R_{ik} \to 0 (k = 1, 2, \ldots, K_i)$) but proximity to other categories remains the same, the estimated category distributions will depend more on discriminability, $(\widehat{S}_{ij} + \bar{\sigma}_{ij})$.[2] That is, when the number of known instances of some category is small, generalization to new instances will be more influenced by the known distributions of surrounding categories. However, when there are already many known instances of a category, the experiences instances and the known distribution of that

[1] Since the matrix in Eq. (13) includes $\hat{\sigma}_{ij}$ or $\widehat{S}_{ij}$ which has σ_i inside, it is not a typical eigenvalue problem, which has a fixed matrix. Eq. (13) can be considered an eigenvalue problem only when σ_i is given. Thus, iterative method for eigenvalue problems such as the power iteration method would be preferable to calculate numerical values.

[2] Although too few instances may cause a non-full-rank covariance matrix whose determinant is zero (i.e., $|\sigma_i| = 0$), in this special case, we still assume a particular variability $|\sigma_i| = C$. See also Method in Analysis 4.

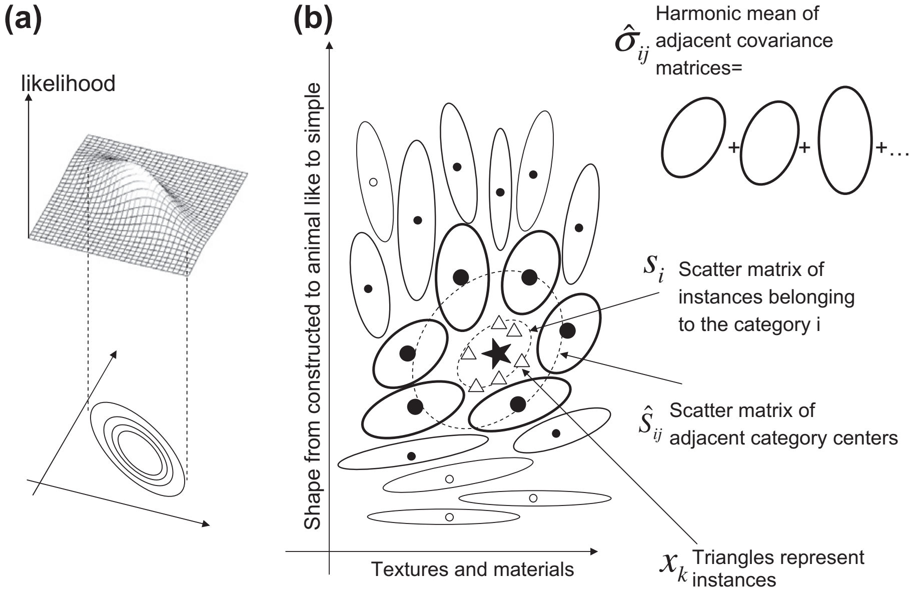

(a)
likelihood
(b)
Shape from constructed to animal like to simple
Textures and materials
$\hat{\sigma}_{ij}$ Harmonic mean of adjacent covariance matrices=
+
+
+…
S_i Scatter matrix of instances belonging to the category i
$\hat{S}_{ij}$ Scatter matrix of adjacent category centers
x_k Triangles represent instances

Fig. 5. An illustration of the local interactions among adjacent categories according to the packing theory. Each ellipsis indicates equal-likelihood contour of category. The star shows a central tendency of a focal category, and triangles show the few experienced instances of this focal category. The covariance matrix of the focal category is estimated with deviation of instances (triangles; S_{ik}), harmonic mean covariance matrix (ellipses) of adjacent categories ($\hat{\sigma}_{ij}$), and deviation of central tendencies (filled circles) of adjacent categories ($\widehat{S}_{ij}$).

category-will have a greater effect on judgments of membership in that category (and a greater effect on surrounding categories).

2.2.5. Optimal solutions for mean vectors

In the previous section, the optimal solution of covariance matrices was derived for a given set of fixed mean vectors. The optimal solutions for mean vectors may also be written as eigenvectors of a quadratic eigenvalue problem. The differential of discriminability and inclusion with respect to the mean vectors are:

$$\frac{\partial F_{ij}}{\partial \mu_i} = -\frac{1}{2}(\sigma_i + \sigma_j)^{-1}(\mu_i - \mu_j) \tag{15}$$

and

$$\frac{\partial G_{ik}}{\partial \mu_i} = -\sigma_i^{-1}(\mu_i - x_{ik}) \tag{16}$$

Then

$$\frac{\partial L_N}{\partial \mu_i} = -\sum_{k=1}^{K_i} R_{ik}\sigma_i^{-1}(\mu_i - x_{ik}) + \lambda \sum_{j=1}^{N} Q_{ij}(\sigma_i + \sigma_j)^{-1}(\mu_i - \mu_j) \tag{17}$$

The equation is rewritten with N times larger order of matrix as follows:

$$\Sigma^{-1}(\overline{R}\mu - \bar{x}) - \lambda\Phi\mu = 0 \tag{18}$$

In Eq. (18), each term is as follows: $\mu = (\mu_1^T, \mu_2^T, \ldots, \mu_N^T)^T$, $\bar{x}_i = \left(\sum_{k=1}^{K_i} R_{1k}x_{1k}^T, \sum_{k=1}^{K_i} R_{2k}x_{2k}^T, \ldots, \sum_{k=1}^{K_i} R_{Nk}x_{Nk}^T\right)^T$,

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_N \end{pmatrix} \tag{19}$$

$$\overline{R} = \begin{pmatrix} I_D \sum_{k=1}^{K_1} R_{1k} & 0 & \ldots & 0 \\ 0 & I_D \sum_{k=1}^{K_2} R_{2k} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & I_D \sum_{k=1}^{K_N} R_{Nk} \end{pmatrix}, \tag{20}$$

and

$$\Phi = \begin{pmatrix} \sum_{i=1}^{N} \overline{Q}_{1i} - \overline{Q}_{11} & \overline{Q}_{12} & \ldots & \overline{Q}_{1N} \\ \overline{Q}_{21} & \sum_{i=1}^{N} \overline{Q}_{2i} - \overline{Q}_{22} & \ldots & \overline{Q}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{Q}_{N1} & \overline{Q}_{N2} & \ldots & \sum_{i=1}^{N} \overline{Q}_{Ni} - \overline{Q}_{NN} \end{pmatrix} \tag{21}$$

where $\overline{Q}_{ij} = Q_{ij}(\sigma_i + \sigma_j)^{-1}$.

Since this equation indicates a typical form of the least square error with a constraint, it is also rewritten as a typical quadratic eigenvalue problem as follows (Tisseur & Meerbergen, 2001):

$$(\lambda^2 I_D - 2\lambda\overline{\Sigma}^2 - \alpha^{-2}\overline{\Sigma}\bar{x}\bar{x}^T\overline{\Sigma})\mu = 0 \tag{22}$$

where $\overline{\Sigma} = \overline{R}\Sigma^{-1}\overline{R}$ and $\alpha^2 = \mu^T\Phi\mu$. Thus, the optimal mean vector μ is one of the eigenvectors given by this eigenvalue problem. This also suggests a similar structure as in the optimal solution for the covariance. The magnitude of Q_{ij} and R_{ik} are quite influential in determining the weighting between the likelihood of the instances of a category and discriminability between categories. If the distances between all pairs of categories are infinite (i.e., $Q_{ij} \to 0(j = 1, 2, \ldots, N)$), the optimal mean vector mainly depends on mean vectors of instances (and the matrix assigning instances to categories) which is purely given by a set of instances (i.e., we obtain $\mu = \overline{R}^{-1}\bar{x}$ by assuming $\Phi = 0$ on Eq. (22)). On the other hand, if the number of instances of categories decreases (i.e., $K_i \to 0$ or $R_{ij} \to 0(k = 1, 2, \ldots, K_i)$), the optimal mean vector depends on both central tendency $\bar{x}$ and distribution Σ of instances (i.e., we obtain $\Phi\mu = \lambda^{-1}\Sigma^{-1}\bar{x}$ by assuming $\overline{R} = 0$). Thus in the latter extreme situation, the central tendencies μ strongly depends on the estimated distribution of instances of each category Σ. In other words, this optimization of central tendencies also indicates the emergence of *smooth* categories, that is, there is predicted correlation between distances in central tendencies and distributions. This correlation between the distance of two categories and their feature distributions is a solution to the feature selection problem. The learner can know the relevant features for any individual category from neighboring categories.

3. Analysis 1: Is the geometry of natural categories smooth?

If natural categories reside in a packed feature space in which both discrimination and inclusion are optimized, then they should show a smooth structure. That is, near by natural categories should not only have similar instances, but they should also have similar *frequency distributions* of features across those instances. Analysis 1 provides support for this prediction by examining the relation between the similarity of instances and the similarity of feature distributions for 48 basic level categories.

A central problem for this analysis is the choice of features across which to describe instances of these categories. One possibility that we considered and rejected was the use of features from feature generation studies (McRae et al., 2005; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Samuelson & Smith, 1999). In these studies, adults are given a category and asked to list the features characteristic of items in each category (e.g., has legs, made of wood, can be sat on). The problem with this approach is that the features listed by adults as important to those queried categories have (presumably) already been selected by whatever cognitive processes make categories coherent. Thus, there is the danger that the use of these generated features presupposes the very phenomenon one seeks to explain. Accordingly, we chose to examine a broad set of polar dimensions unlikely to be specifically offered as important to any of these categories. The specific features chosen do not need to be the exactly right features nor comprehensive. All they need to do is capture a portion of the similarity space in which instances and categories reside. If they do and if the packing analysis is right, these features should nonetheless define an *n*-dimensional space of categories, which shows some degree of smoothness: categories with instances similar to each other on these features should also show similar category likelihoods on these features.

To obtain this space, 16 polar opposites (e.g., wet–dry, noisy–quiet, weak–strong) were selected that broadly encompass a wide range of qualities (Hidaka & Saiki, 2004; Osgood, Suci, & Tannenbaum, 1957), that are also (by prior analyses) statistically uncorrelated (Hidaka & Saiki, 2004) but that neither by introspection nor by prior empirical studies seem to be specifically relevant to the particular categories examined in this study. In this way, we let the packing metric select the locally defined features.

The analysis is based on the assumption that categories with more variability in their feature distributions in the world will yield more variability in the subjects' judgments about the relevant features. Thus, the mean of the subjects' judgments for any category is used as an estimate of the mean of the feature distributions for the category and the covariance of the subjects' judgments is used as an estimate of covariance.

3.1. Method

3.1.1. Participants

The participants were 104 undergraduate and graduate students at Kyoto University and Kyoto Koka Women's University.

3.1.2. Stimuli

Participants were tested in Japanese. The English translations of the 16 adjective pairs in English were dynamic–static, wet–dry, light–heavy, large–small, complex–simple, slow–quick, quiet–noisy, stable–unstable, cool–warm, natural–artificial, round–square, weak–strong, rough hewn–finely crafted, straight–curved, smooth–bumpy, hard–soft. The 48 noun categories, in English, are butterfly, cat, fish, frog, horse, monkey, tiger, arm, eye, hand, knee, tongue, boots, gloves, jeans, shirt, banana, egg, ice cream, milk, pizza, salt, toast, bed, chair, door, refrigerator, table, rain, snow, stone, tree, water, camera, cup, keys, money, paper, scissors, plant, balloon, book, doll, glue, airplane, train, car, bicycle. These nouns were selected to be common with early ages of acquisition (Fenson et al., 1994).

3.1.3. Procedure

Participants were presented with one noun at a time and asked to judge the applicability of the 16 adjective pairs on a 5-point scale. For example, if the adjective pair was *small–big*, and the noun was *chair*, participants would be asked to rate the size of typical instances of a chair on the scale of 1 (indicating small) to 5 (indicating big). The presented order of the list of 48 nouns by 16 dimension-rating scale was randomly determined and differed across subjects.

3.1.4. The smoothness index

The adult judgments generate an initial space defined by the 48 noun categories and the mean and variance of the ratings of these nouns on the 16 polar dimensions. The mean μ_i and covariance σ_i of ith category over instances, is defined as the central tendencies and generalization patterns (Eq. (24)).

$$\mu_i = \frac{1}{M}\sum_{k=1}^{M} d_{ik} \tag{23}$$

$$\sigma_i = \frac{1}{M}\sum_{k=1}^{M}(d_{ik} - \mu_i)(d_{ik} - \mu_i)^T \tag{24}$$

where M is number of subjects (104) and d_{ik} is 16 dimensional column vector having kth subjects' adjective ratings of ith category. The smoothness index of the given categories is defined by the correlation of central tendencies and covariance as follows:

$$S = \frac{\sum_{i,j<i}(\|\mu_{ij}\| - \|\bar{\mu}\|)(\|\sigma_{ij}\| - \|\bar{\sigma}\|)}{\sqrt{\sum_{i,j<i}(\|\mu_{ij}\| - \|\bar{\mu}\|)^2 \sum_{i,j<i}(\|\sigma_{ij}\| - \|\bar{\sigma}\|)^2}} \tag{25}$$

where S is the smoothness index, a correlation coefficient between all possible paired distances of central tendencies $\|\mu_i\|$ and $\|\sigma_{ij}\|$. $\|\mu_{ij}\| = \{(\mu_i - \mu_j)^T(\mu_i - \mu_j)\}^{-\frac{1}{2}}$ is the Euclidian distance of the paired central tendencies of category i and j (μ_i is the mean vector given as dimensional column vector). $\|\sigma_{ij}\| = \mathrm{tr}\{(\mu_i - \mu_j)^T(\mu_i - \mu_j)\}^{-\frac{1}{2}}$ is the Euclidian distance of paired generalization patterns of category i and j (σ_i is covariance matrix given as dimensional square matrix). $\|\bar{\mu}\| = N^{-1}\sum_{i,j<i}\|\mu_{ij}\|$ and $\|\bar{\sigma}\| = N^{-1}\sum_{i,j<i}\|\sigma_{ij}\|$ with top bars indicates the mean of $\|\mu_{ij}\|$ and $\|\sigma_{ij}\|$ respectively, where N is number of possible combinations of pairs from n categories.

In sum, smoothness is measured as a correlation between the distance of categories, which is measured by the distances of the central tendencies, and the generalization pattern for each category, which is measured by the category's covariance matrix. Accordingly, we calculated the distances of the central tendencies for each of the 48 categories to each other and the distances of the generalization patterns (the covariance matrices) for each of the 48 categories to each other. If categories that are near in the feature space have similar generalization patterns, than the two sets of distances should be correlated with each other. Because distances between the means of categories A and B are dependent of the distances between the means of categories B and C,[3] we sampled independent paired distances in which no category appears in two different pairs. For 48 categories, the number of possible combinations of independent pairs is $\frac{48!}{2^{24}}$. We analyzed the median and the empirical distribution of 1000 such samplings. We also transformed the rating data using a logistic function, which corrects for the bounded rating scale. This corrected covariance $\hat{\sigma}_{ij}$ and mean $\hat{\mu}_i$ (having range $[-\infty, \infty]$) of dimensions i and j is defined by the following equation:

$$\hat{\sigma}_{ij} = \sigma_{ij}\{p_i(1-p_i)p_j(1-p_j)\}^{-\frac{1}{2}}$$

$$\hat{\mu}_i = \log p_i - \log(1-p_i)$$

$$p_i = \frac{\mu_i - 1}{4}$$

where μ_i is mean of ith dimension (range 1–5), and p_i is normalized mean having the range from zero to one. As the first differential with respect to corrected mean $\hat{\mu}_i$ of logistic function $p_i = (1 + \exp(\alpha\hat{\mu}_i))^{-1}$ is proportional to $p_i(1 - p_i)$, where α is a particular constant, we use this differential to transform mean and variance to theoretically homoscedastic space with mean $\hat{\mu}_i$ and covariance $\hat{\sigma}_{ij}$. The corrected mean $\hat{\mu}_i$ and covariance $\hat{\sigma}_{ij}$ are used for the smoothness index instead of the raw mean μ_i and covariance σ_{ij}. (See also Generalized Linear Model (McCullagh & Nelder, 1989) for the detail of logistic analysis.) As a supplemental measure, we also calculated smoothness by normalizing variance using a correlation matrix instead of a covariance matrix. The potential value of this approach is that it ignores artificial correlations between means and (the absolute value of) the variance.

3.2. Results and discussion

Fig. 6a shows a scatter plot of all possible pairs of categories; the x-axis is the Euclidian distance of the paired corrected mean vectors and the y-axis is the Euclidian distance of the paired corrected covariance matrices. The correlation between these two variables (with no dependence of paired distances) is the smoothness index (see Eq. (25) for its definition). The median correlation was 0.537 (95% confidence interval is from 0.223 to 0.756). Fig. 6b shows the same scatter plot using the correlation matrix instead of the covariance matrix as the measure of category likelihood; here the median correlation was 0.438 (95% confidence interval is from 0.137 to 0.699). These positive correlations between the distances of central tendencies and the distances of category likelihoods provide a first indication that natural categories may be smooth.

Fig. 6 raises an additional possible insight. Not only do categories near each other in feature space show similar patterns of feature distribution, but across categories the changes in the feature distributions appear to be continuous, both in terms of location in the feature space and in terms of the feature likelihoods. This seamlessness of transitions within the space of categories is suggested by the linear structure of the scatter plot itself. This can emerge only if there are no big jumps or gaps in feature space or in the category likelihoods.

[3] In fact, for arbitrary points A–C, the triangle inequality $|AB| + |BC| > |CA|$ is true, where $|AB|$ is a metric between point A and B.

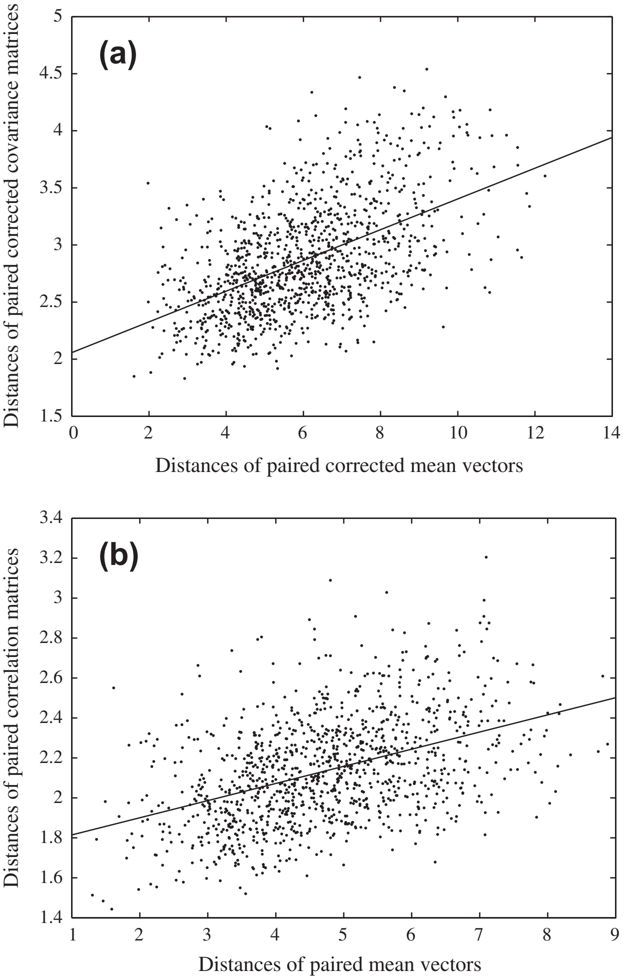


Fig. 6. If a space is smooth then the nearness of the categories (distances of the means) and similarity of the generalization patterns should be correlated. The two figures differ in their respective measures of the similarity of the generalization patterns of categories: (a) scatter plot of the Euclidean distances of the covariance matrices and the Euclidean distances of the means for pairs of categories. This correlation is the smoothness index, $S = 0.537$. (b) Scatter plot of the Euclidean distances of the correlation matrices and the Euclidean distances of the means for pairs of categories ($S = 0.438$).

Critically, the features analyzed in this study were not pre-selected to particularly fit the categories and thus the observed smoothness seems unlikely to have arrived from our choice of features or a priori notions about the kind of features that are relevant for different kinds of categories. Instead, the similarity of categories on any set of features (with sufficient variance across the category) may be related to the distribution of those features across instances.

Categories whose instances are generally similar in terms of their range of features also exhibit similar patterns of feature importance. The mathematical analysis of Packing theory indicates that this could be because of the optimization of discrimination and generalization in a geometry of crowded categories.

4. Analysis 2: Learning new categories

According to Packing theory, the generalization of a category to new instances depends not just on the instances that have been experienced for that category but also on the distributions of known instances for nearby categories. From one, or very few new instances, generalizations of a newly encountered category may be systematically aligned with the distributions of instances from surrounding categories. This is illustrated in Fig. 7: a learner who already knows some categories (shown as solid ellipses in Fig. 7a) and observes the first instance (a black star) of a novel category (a broken ellipsis) may predict the unknown generalization pattern shown by the broken ellipsis (Fig. 7b). Because nearby categories have similar patterns of likelihoods, the system (via competition among categories and the joint optimization of inclusion and discrimination) can predict the likelihood of the unknown category, a likelihood that would also be similar to other known and nearby categories in the feature space. If categories did not have this property of smoothness, if they were distributed like that in Fig. 7c, where each category has a variance pattern unrelated to those of nearby categories, the learner would have no basis on which to predict the generalization pattern.The goal of Analysis 2 is to show that the packing metric can predict the feature distribution patterns of categories unknown to the model. In the simulation, the model is given the mean and covariance of 47 categories (from Analysis 1) and then is given a single instance of the 48th category. The model's prediction of the novel probabilistic density is calculated by an optimal solution with respect to the configuration of surrounding known noun categories.

4.1. Method

On each trial of the simulation, one target category is assigned as unknown; the other 47categories serve as the background categories that are assumed to be already learned. Each of the background-knowledge categories is assumed to have a normal distribution, and the model predicts the covariance matrix of the target category of the base on the given mean vectors and covariance matrices of the categories that comprise the model's background knowledge. Because children are unlikely to have complete knowledge of any category, the mean and covariance for the background categories are estimated from a random sampling of 50% of the adult judgments. This is done 50 times with each of the 48 noun categories from Analysis1 serving as the target category.

4.1.1. Estimation of a novel category from the first instance

Within the packing model, the variance (and covariance) of the probabilistic density function is a critical determiner of the feature dimensions that are most important for a local region and category. Thus, to predict the distribution of instances for the target category, a category for which only one instance is given, we derive the covariance

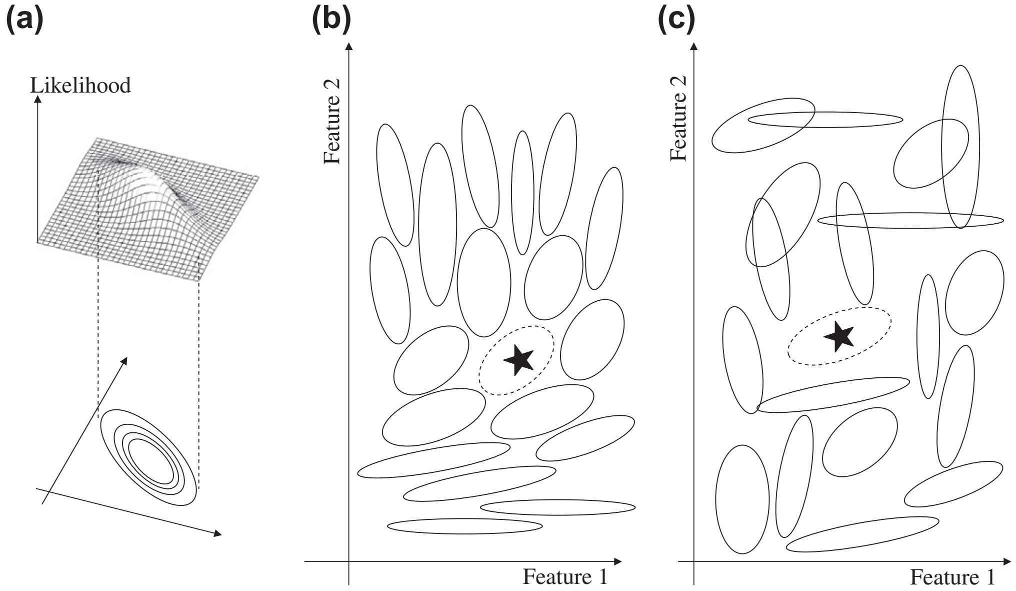


Fig. 7. (a) Each ellipsis indicates the equal-likelihood contour. Two schematic illustrations of a (b) smooth (c) and non-smooth space of categories. The broken ellipsis in each figure indicates the equal-likelihood contour of the unknown category, and the star indicates a given first instance of that category. The solid ellipses indicate equal-likelihood contour of known categories. A smooth space of categories provides more information for predicting the likelihood of the novel category contour.

estimation for the whole category. To do this, we let the scatter matrix of category i be zero (i.e., $S_i \approx 0$) by assuming the first instance is close to the true mean (i.e., $K_i = 1$ and $x_{il} = \mu_i$). In addition, we assume that the unknown likelihood of the novel category takes the form $G_i - \log(C)$, where C is a particular constant. In particular, in case. Based on this assumption, we can obtain the covariance matrix of category $C_i(\sigma_i)$ by solving the Eq. (13). Then the optimal covariance of the novel category is given as follows:

$$\sigma_i = \widehat{C} \sum_{j=1}^{N} Q_{ij}\left(\widehat{S}_{ij} + \hat{\sigma}_{ij}\right) \tag{26}$$

where $\widehat{C} = C|\sum_{j=1}^{N} Q_{ij}(\widehat{S}_{ij} + \hat{\sigma}_{ij})|$ is derived from the constraint $\frac{\partial L_N}{\partial \lambda} = G_i - C = 0$. Thus, estimated σ_i in Eq. (26) optimize the packing metric, and it is considered as a special case of the general optimal solution when the covariance matrix of instances is collapsed to be zero (because there is the only instance). This equation indicates that a novel category with only instance can be estimated with harmonic mean of nearer known covariance matrices $(\hat{\sigma}_{ij})$ and nearer scatter matrix of weighted means $(\widehat{S}_{ij})$. This directly means the covariance matrix of the novel category is estimated from the other covariance matrices of nearby categories.

We used Eq. (26) in order to calculate a covariance matrix of a novel category σ_i from an instance sampled from the category ($K_i = 1, x_k = \mu_i$) and other known categories (μ_j and σ_j, $j = 1, 2, \ldots, i-1, i+1, \ldots, 48$). The scaling constant in Eq. (26) is assumed to have the same determinant of covariance matrix as the target category as the adult judgment has (i.e, $C = |S_i|$ so as to have $|\sigma_i| = |S_i|$).

4.1.2. Control comparisons

The packing metric predicts the distribution of instances in the target category by taking into account its general location (indicated by the one given instance) and the distributions of known instances for nearby categories. It is thus a geometric solution that derives not from what is specifically known about the individual category but from its position in a geometry of many categories. Accordingly, we evaluate this central idea and the packing model's ability to predict the unknown distribution of instances by comparing the predictions of the packing model to two alternative measures of the distribution of instances in feature space for that target category that take into account only information about the target category and not information about neighboring categories. These two alternative measures are: (1) the actual distribution of all instances of the target category as given by the subjects in Analysis 1 and (2) three randomly selected instances from that subject generated distribution of instances. The comparison of the predictions of the packing metric to the actual distribution answers the question of how well the packing metric generates the full distribution given only a single instance but information about the distributions of neighboring categories. The second comparison answers the question of whether a single instance in the context of a whole geometry of categories provides better information about the shape of that category than more instances with no other information.

4.2. Results and discussion

The predicted covariance of the target category by the packing model correlates strongly with the actual distribution of instances as generated by the subjects in Analysis 1. Specifically, the correlations between the packing metric predictions for the target category and measures of the actual distributions from Analysis 1 were 0.876, 0.578 and 0.559 for the covariances and variances (136 dimensions), the variances considered alone (16 dimensions) and the covariances considered alone (120 dimensions). These are robust correlations overall; moreover, they are considerably greater than those derived from an estimation of the target category from three randomly chosen instances. For this "control" comparison, we analyzed the correlation of covariance matrix for each category calculated from randomly chosen three instances of adults' judgment with that calculated from the whole set of instances. Their average correlations of 50 different random set of samples were 0.2266 in variances ($\text{SD} = 0.2610$), 0.2273 in covariance ($\text{SD} = 0.1393$) and 0.4456 in both variance and covariance ($\text{SD} = 0.0655$). The packing metric – given one instance and *information about neighboring categories* – does a better job predicting category shape than a prediction from three instances. In sum, the packing metric can generate the distribution of instances in a category using its location in a system of known categories. This result suggests that a developing system of categories should, when enough categories and their instances are known, enable the learner to infer the distribution of newly encountered categories from very few instances. A geometry of categories – and the local interactions among them – creates knowledge of *possible* categories.

There are several open questions with respect to the joint optimization of inclusion and discrimination should influence category development in children who will have sparser instances and sparser categories than do adults. The processes presumed by Packing theory may be assumed to always be operation as they seem likely to reflect core operating characteristics (competition) of the cognitive system. But their effects will depend on the density of categories and instances in local regions of the space. An implication of the simulation is that the accuracy of novel word generalizations will be monotonic increasing function of number of categories. But here is what we do not know: as children learn categories, are some regions dense (e.g., dense animal categories) and other sparse (e.g., tools)? Are some regions of the space-even those with relatively many categories-sparse in the sense of relatively few experienced instances of any one category? Knowing just how young children's category knowledge "scales up" is critical to testing the role of the joint optimization proposed by Packing theory in children's category development.

The formal analyses show that for the bias inherent in the joint optimization of discrimination and inclusion require many categories (crowding) and relatively many instances in these categories. This crowding will also depend on the dimensionality of the space as crowding is more likely in a lower than in a higher dimensional space, and we do not know the dimensionality of the feature space for human category judgments. This limitation does not matter for testing general predictions since the optimization depends only on distance relations in the space (and thus on the number of orthogonal, that is uncorrelated, dimensions but not on any assumptions about what orthogonal directions in that space constitute the dimensions) and since the prediction of smoothness should hold in any lower-dimensional characterization of the space. The specification of the actual dimensionality of the space also may not matter for relative predictions about more and less crowded regions of people's space of categories. Still, insight into the dimensionality of the feature space of human categories would benefit an understanding of the development of the ability to generalize a new category from very few instances.

5. General discussion

A fundamental problem in category learning is knowing the relevant features for to-be-learned categories. Although this is a difficult problem for theories of categorization, people, including young children, seem to readily solve the problem. The packing model provides a unified account of feature selection and fast mapping that begins with the insight that the feature distributions across the known instances of a category play a strong role, one that trumps overall similarity, in judgments as to whether some instance is a member of that category. This fact is often discussed in the categorization literature in terms of the question of whether instance distributions or similarity matter to category formation (Holland et al., 1986; Nisbett et al., 1983; Rips, 1989; Rips & Collins, 1993; Thibaut et al., 2002). The packing model takes role of instance distributions and ties it to similarity in a geometry of category in which nearby categories having similar category-relevant features, showing how this structure may emerge and how it may be used to learn new categories from very few instances. The packing model thus provides a bridge that connects the roles of instance distributions and similarity. The fitting of categories into a feature space is construed as the joint optimization of including known and possible instances and discriminating the instances belonging to different categories. The joint optimization of inclusion and discrimination aligns nearby categories such that their distributions of instances in the feature space are more alike. The chain reaction of these local interactions across the population of categories creates a smooth space. Categories that are similar (near in the space) have similar distributions of instances; categories that are dissimilar (far in the space) have more dissimilar distributions of instances.

In this way, the packing model provides the missing link that connects similarity to the likelihood of instances. Both similarity and feature distributions are deeply relevant to understanding how and why human categories have the structures that they do. However, the relevance is not with respect to the structure of a single category, but with respect to the structure of a population of categories. Smoothness implies a higher order structure, a gradient of changing feature relevance across the whole that is made out of, but transcends, the specific instances and the specific features of individual categories. It is this higher order structure that may be useable by learners in forming new categories. This higher order structure in the feature space aligns with what are sometimes called "kinds" or "superordinate categories" in that similar categories (clothes versus food for example) will have similar features of importance and be near each other in the space. However, there are no hard and fast boundaries and the packing does not directly represent these higher order categories. Instead, they are emergent in the patchwork of generalization gradients across the feature space. How such a space of probabilistic likelihoods of instances as members of basic level categories relates to higher and lower levels of categories (and the words one learns to name those categories) is an important question to be pursued in future work.

5.1. Packing theory in relation to other topographic approaches

The packing model shares some core ideas with other topographic approaches such as self-organizing maps (SOM, see Kohonen, 1995; see also Tenenbaum, 2000, etc.). The central assumption underlying the algorithms used in SOM is that information coding is based on a continuous and smooth projection that preserves a particular topological structure. More particularly, within this framework, information coders (e.g., receptive fields, categories, memories) that are near each other code similar information whereas coders that are more distant code different types of information. Thus, SOM and other topographical representations posit a smooth representational space, just as the packing metric.

However, there are differences between the packing model and algorithms, such as SOM. In the packing metric, categories may be thought of as the information coders, but unlike the information coders in SOM, these categories *begin* with their own feature importance and their own location in the map, which is specified by the feature distributions of experienced instances. Within the packing model, local competition "tunes" feature importance across causes of the population of categories and creates a smooth space of feature relevance. SOM also posits a competition among information unites but of a fundamentally different kind. In the SOM algorithm, information coders do not explicitly have "their own type" of information. Rather it is the topological relation among information coders that implicitly specifies their gradients of data distribution. In the SOM learning process, the closest information units to an input is gradually moved to better fit the data point, and nearby points are moved to fit similar inputs. Thus nearby units end up coding similar inputs.

Topological algorithms such as SOM assume that a smooth structure is a good way to represent information and this assumption is well supported by the many successful applications of these algorithms (Kohonen, 1995). However, just why a smooth structure is "good" is not well specified. The packing metric might provide an answer from a psychological perspective. The packing neither assumes topological relations nor a smooth structure, but rather *produces* them through the joint optimization of discriminability and inclusion. Thus, a smooth space might be a good form of representation because of the trade off between discrimination and generalization.

5.2. Packing theory in relation to other accounts of fast mapping

Fast mapping is the term sometimes used to describe young children's ability to map a noun to a whole category given just one instance (Carey & Bartlett, 1978). Packing theory shares properties with two classes of current explanations of fast mapping in children: connectionist (Colunga & Smith, 2005; Rogers & McClelland, 2004, see also, Hanson & Negishi (2002) for a related model) and Bayesian approaches (Kemp, Perfors, & Tenenbaum, 2007; Xu & Tenenbaum, 2007). Like connectionist accounts, the packing model views knowledge about the different organization of different kinds as emergent and graded. Like rationalist accounts, the packing model is not a process model. Moreover, since the packing model is build upon a statistical optimality, it could be formally classified as a rationalist model (Anderson, 1990). Despite these differences, there are important similarities across all three approaches. Both the extant connectionist and Bayesian accounts of children's smart noun generalizations consider category learning and generalization as a form of statistical inference. Thus, all three classes of models are sensitive to the feature variability within a set of instances. All agree on the main idea behind the packing model that feature variability within categories determines biases in category generalization. All three also agree that the most important issue to be explained is higher order feature selection, called variously second order generalizations (Colunga & Smith, 2005; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002), over-hypotheses (Kemp et al., 2007), and smoothness (the packing model). Using the terms of Colunga and Smith (2005), the first-order of generalization is about individual categories and it is a generalization over instances. The second-order generalization is generalization of *distribution of categories* over categories. The central goal of all three approaches is to explain how people form higher-order generalizations.

There are also important and related differences among these approaches. The first set of differences concern whether or not the different levels are explicitly represented in the theory. Colunga and Smith's (2005) connectionist account represents only input and output associations, the higher order representations of kind – that shape is more relevant for solid things than for nonsolid things, for example – are *implicit* in the structure of the input–output associations. They are not explicitly represented and they do not pre-exist in the learner prior to learning. In contrast, the Bayesian approach for the novel word generalization (Kemp et al., 2007; Xu & Tenenbaum, 2007) has assumed categories structured as a hierarchical tree. The learner knows from the start that there are higher order and lower order categories *in a hierarchy*. Although the packing model is rationalist in approach, it is emergentist in spirit: Smoothness is not an a priori expectation and is not explicitly represented as higher order variable but is an emergent and graded property of the population as a whole. As it stands, the Packing model also makes no explicit distinction between learned categories at different levels such as the learning of categories of animal as dog, for example. The present model is considered only basic level categories and thus is moot on this point. However, one approach that the packing metric could take with respect to this issue is to represent all levels of categories in the same geometry, with many overlapping instances, letting the joint optimization of inclusion and discrimination find the stable solution given the distributional evidence on the inclusion and discrimination of instances in the overlapping categories. This approach might well capture some developmental phenomena. For example, children's tendency to agree that unknown animals are "animals" but that well known ones (e.g., dogs) are not. Within this extended framework, one might also want to include, outside of the packing model itself, real-time processes that perhaps activate a selected map of categories in working memory or that perhaps contextually shift local feature gradients, enabling classifiers to flexibly shift between levels and kinds of categories and to form ad hoc categories (Barsalou, 1985; Spencer, Perone, Smith, & Samuelson, in preparation).

The second and perhaps most crucial difference between packing theory and the other two accounts is the ultimate origin of the higher order knowledge about kinds. For connectionist accounts, the higher order regularities are latent structure in the input itself. If natural categories are smooth, by this view, it is solely because the structure of the categories in the world is smooth and the human learning system has the capability to discover that regularity. However, if this is so, one needs to ask (and answer) why the to-be-learned categories have the structure that they do. For the current Bayesian accounts, a hierarchical representational structure (with variabilized over-hypotheses) is assumed and fixed (but see the other approach that learns the structure Kemp & Tenenbaum, 2008). These over-hypotheses create a tree of categories in which categories near the tree will have similar structure. Again, why the system would have evolved to have such an innate structure is not at all clear. Moreover, the kind of mechanisms or neural substrates in which such hierarchical pre-ordained knowledge resides is also far from obvious.

The packing model provides answers and new insights to these issues that put smoothness neither in the data nor as a pre-specified outcome. Instead, smoothness is emergent in the local interactions of fundamental processes of categorization, inclusion, and discrimination. As the proof and analyses show, the joint optimization of discriminability and inclusion leads to smooth categories, regardless of the starting point. The packing model thus provides answer as to why categories are the way they are and why they are smooth. The answer is *not* that categories have the structure they do in order to help children learn them; the smoothness of categories in feature space is not a pre-specification of what the system has to learn as in the current Bayesian accounts of children's early word learning (although the smoothness of geometry of categories is clearly exploitable). Rather, according to Packing theory, the reason categories have the structure they do lies in local function of categories, in the first place: to include known and possible instances but to discriminate among instances falling in different categories. The probabilistic nature of inclusion and discrimination, the frequency distributions of individual categories, the joint optimization of discrimination, and inclusion in a connected geometry of many categories creates a gradient of feature relevance that is then useable by learners. For natural category learning, for categories that are passed on from one generation to the next, the optimization of inclusion and discrimination over these generations may make highly common and early-learned categories particularly smooth. Although the packing model is not a process model, processes of discrimination and inclusion and processes of competition in a topographical representation are well studied at a variety of levels of analysis and thus bridges between this analytic account and process accounts seem plausible.

5.3. Testable predictions

The specific contribution of this paper is a mathematical analysis that shows that the joint optimization of inclusion and discrimination yields a smooth space of categories and that given such a smooth space that optimization can also accurately predict the instance distributions of a new category specified only by the location of a single instance. What is needed beyond this mathematical proof is empirical evidence that shows that the category organizations and processes proposed by the packing model are actually observable in human behavior. The present paper provide a first step by indicating that the feature space of early-learned noun categories may be smooth (and smooth enough to support fast mapping). Huttenlocher, Hedges, Lourenco, Crawford, and Corrigan (2007) have reported empirical evidence that also provides support for local competitions among neighboring categories. Huttenlocher et al.'s (2007) method provides a possible way to test specific predictions from Packing theory in adults.

The local interactions that create smoothness also raise new and testable hypotheses about children's developing category knowledge. Because these local competitions depend on the frequency distributions over known instances and the local neighborhood of known categories, there should be observable and predictable changes as children's category knowledge "scales up". Several developmental predictions follow: (1) Learners who know (or are taught) a sufficiently large and dense set of categories, should form and generalize a geometry of categories that is smoother than that given by the known instances. (2) The generalization of any category trained with a specific set of instances should depend on the instance distributions of surrounding categories and be distorted in the direction of the surrounding categories; thus, children should show smoother category structure and smarter novel noun generalizations in denser category regions than sparser ones. (3) The effects of learning a new category on surrounding categories or surrounding on new categories should depend in formally predictable ways on the feature distributions of those categories.

6. Conclusion

Categories (and their instances) do not exist in isolation but reside in a space of many other categories. The local interactions of these categories create a gradient of higher order structure-different kinds with different feature distributions. This structure emergent from the interactions of many categories in a representational space constrains the possible structure of both known and unknown categories. Packing theory captures these ideas in the joint optimization of discrimination and generalization.

Acknowledgements

This study is based on in part on the PhD thesis of the first author. The authors thank Dr. Jun Saiki and Dr. Toshio Inui for their discussion on this work. This study was supported by grants from NIH MH60200.

Appendix A. Derivation of differential with respect to covariance matrix

We derive Eqs. (11) and (12) by expanding $\frac{\partial F_{ij}}{\partial \sigma_i}$ and $\frac{\partial G_{ik}}{\partial \sigma_i}$. For the derivation, we use the vectorizing operator v(X) which form a column vector from a given matrix X (see also Magnus and Neudecker (1988) and Turkington (2002) for the matrix algebra). A useful formula on vectorizing operator is as follows. For A: $m \times n$ matrix and B: $n \times p$ matrix,

$$\begin{aligned} v(AB) = v(I_m AB) &= \left(B^T \otimes I_m\right) v(A) \\ = v(AI_m B) &= \left(B^T \otimes A\right) v(I_m) \\ = v(ABI_p) &= \left(I_p \otimes A\right) v(B) \end{aligned}$$

where I_m is mth order identity matrix and $\otimes$ denotes Kronecker product. Moreover, we use the following formulae in order to expand the differential with respect to a matrix (see also Turkington, 2002). For $d \times d$ matrices X, Y, Z and a constant matrix A which is not function of X,

$$\frac{\partial |X|}{\partial v(X)} = |X| v(X^{-1T})$$
$$\frac{\partial v(X^{-1})}{\partial v(X)} = -\left(X^{-1} \otimes X^{-1T}\right)$$
$$\frac{\partial \mathrm{tr}(AX)}{\partial v(X)} = v(A^T)$$
$$\frac{\partial v(X)}{\partial x_k} = (\lambda_k I_d - X)\left(x_k^T \otimes I_d\right)$$
$$\frac{\partial v(Z)}{\partial v(X)} = \frac{\partial v(Y)}{\partial v(Z)} \frac{\partial v(X)}{\partial v(Y)}$$

where x_k and λ_k are respectively kth eigenvector and eigenvalue of a dth order real symmetric matrix X. We derive Eq. (11) from Eq. (8) using formulae above,

$$\begin{aligned} \frac{-4\partial F_{ij}}{\partial v(\sigma_i)} &= \frac{\partial v\left(\Delta\mu_{ij}\Delta\mu_{ij}^T \bar{\sigma}_{ij}^{-1}\right)}{\partial v(\sigma_i)} + 2\frac{\partial \log|2^{-1}\bar{\sigma}_{ij}|}{\partial v(\sigma_i)} - \frac{\partial \log|\sigma_i|}{\partial v(\sigma_i)} \\ &= \frac{\partial v\left(\bar{\sigma}_{ij}^{-1}\right)}{\partial v(\sigma_i)} v\left(\Delta\mu_{ij}\Delta\mu_{ij}^T\right) + 2\frac{\partial v(\sigma_i)}{\partial v(\sigma_i)} v\left(\bar{\sigma}_{ij}^{-1}\right) \\ &\quad - \frac{\partial v(\sigma_i)}{\partial v(\sigma_i)} v(\sigma_i^{-1T}) \\ &= -\left(\bar{\sigma}_{ij}^{-1} \otimes \bar{\sigma}_{ij}^{-1T}\right) v\left(\Delta\mu_{ij}\Delta\mu_{ij}^T\right) + 2v\left(\bar{\sigma}_{ij}^{-1}\right) - v(\sigma_i^{-1T}) \\ &= -v(\sigma_i^{-1}\Delta\mu_{ij}\Delta\mu_{ij}^T\sigma_i^{-1} - 2\bar{\sigma}_{ij}^{-1} - \sigma_i^{-1}) \\ &= -v(\sigma_i^{-1}(\mu_i - \bar{\mu}_{ij})(\mu_i - \bar{\mu}_{ij})^T\sigma_i^{-1} + \sigma_i^{-1}\hat{\sigma}_{ij}^{-1}\sigma_i^{-1} - \sigma_i^{-1}) \end{aligned}$$

where $\bar{\sigma}_{ij} = \sigma_i + \sigma_j$ and $\Delta\mu_{ij} = \mu_i - \mu_j$. And note that $\bar{\sigma}_{ij} = \sigma_i^{-1} - 2^{-1}\sigma_i^{-1}\hat{\sigma}_{ij}\sigma_i^{-1}$ and $\bar{\sigma}_{ij}\Delta\mu_{ij} = \sigma_i^{-1}\sigma_i^{-1}(\mu_i - \bar{\mu}_{ij})$ are used for the last line. Thus, we obtain Eq. (11). Likewise the derivation of Eq. (11), we derive Eq. (12) from Eq. (2) as follows:

$$\begin{aligned} \frac{-2\partial G_{ik}}{\partial v(\sigma_i)} &= \frac{\partial \mathrm{tr}(S_{ik}\sigma_i^{-1})}{\partial v(\sigma_i)} + \frac{\partial \log|\sigma_i|}{\partial v(\sigma_i)} \\ &= \frac{\partial \mathrm{tr}(\sigma_i^{-1})}{\partial v(\sigma_i)} v(S_{ik}) + \frac{\partial v(\sigma_i)}{\partial v(\sigma_i)} v(\sigma_i^{-1T}) \\ &= -\left(\bar{\sigma}_{ij}^{-1} \otimes \bar{\sigma}_{ij}^{-1T}\right) v(S_{ik}) + v(\sigma_i^{-1T}) \\ &= v\left(-\bar{\sigma}_{ij}^{-1} S_{ik} \bar{\sigma}_{ij}^{-1T} + \sigma_i^{-1T}\right) \end{aligned}$$

Next, we derive Eq. (13) using formula for the differential with respect to eigenvector as follows:

$$\begin{aligned} \frac{\partial L_N}{\partial s_{il}} &= \frac{\partial v(\sigma_i)}{\partial s_{il}} \frac{\partial L_N}{\partial v(\sigma_i)} \\ &= (\eta_{il} I_D - \sigma_i)(s_{il}^T \otimes I_D) v\left(\frac{\partial L_N}{\partial \sigma_i}\right) \\ &= (\eta_{il} I_D - \sigma_i) \frac{\partial L_N}{\partial v(\sigma_i)} s_{il}^T \end{aligned}$$

Since $|\eta_{il} I_D - \sigma_i| = 0$ is obvious by definition, $|\frac{\partial L_N}{\partial \sigma_i}| = 0$ is necessary in order to obtain non-obvious solution for $\frac{\partial L_N}{\partial s_{il}} = 0$. Therefore, we obtain Eq. (13).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Ashby, G. F., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154–179.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 11*, 629–654.

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research, 49*, 1154–1165.

Booth, A. E., & Waxman, S. (2002). Word learning is 'smart': Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition, 84*, B11–B22.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17–29.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review, 112*, 347–382.

Colunga, E., & Smith, L. B. (2008). Flexibility and variability: Essential to human cognition and the study of human cognition. *New Ideas in Psychology, 26*(2), 174–192.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society Series, B*(39), 1–38.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: John Wiley & Sons.

Duncan, J. (1996). Cooperating brain systems in selective perception and action. *Attention and Performance XVI: Information Integration, 18*, 193–222.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*(59), 1–173.

Gathercole, V. C. M., & Min, H. (1997). Word meaning biases or language-specific effects? Evidence from english, spanish, and korean. *First Language, 17*(49), 31–56.

Gelman, S. A., & Coley, J. D. (1991). Perspectives on language and thought: Interrelations in development. In S. A. German & J. P. Byrnes (Eds.), *Language and categorization: the acquisition of natural kind terms*. Cambridge: Cambridge University Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2).

Hanson, S. J., & Negishi, M. (2002). On the emergence of rules in neural networks. *Neural Computation, 14*, 2245–2268.

Hidaka, S., & Saiki, J. (2004). A mechanism of ontological boundary shifting. In *The twenty sixth annual meeting of the cognitive science society* (pp. 565–570).

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction*. Cambridge, MA: MIT Press.

Huttenlocher, J., Hedges, L. V., Lourenco, S. F., Crawford, L. E., & Corrigan, B. (2007). Estimating stimuli from contrasting categories: Truncation due to boundaries. *Journal of Experimental Psychology: General, 136*(3), 502–519.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition, 62*, 169–200.

Jones, S. S., & Smith, L. (2002). How children know the relevant properties for generalizing object names. *Developmental Science, 5*, 219–232.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62*, 499–516.

Keil, F. C. (1994). Mapping the mind: Domain specificity in cognition and culture. In L. A. Hirschfeld, S. A. Susan, & A. Gelman (Eds.), *The birth and nurturance of concepts by domains: The origins of concepts of living things*. MA: Cambridge University Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687–10692.

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General, 137*, 52–75.

Kobayashi, H. (1998). How 2-year-old children learn novel part names of unfamiliar objects. *Cognition, 68*, B41–B51.

Kohonen, T. (1995). *Self-organizing maps*. Heidelberg: Springer.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.

Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language, 31*(6), 807–825.

Landau, B., Smith, L. B., & Jones, S. S. (1998). Object shape, object function, and object name. *Journal of Memory and Language, 38*, 1–27.

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development, 6*, 17–46.

Magnus, J. R. (1988). *Linear structure*. Oxford: Oxford University Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71–102.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Reserch Methods, Instruments, & Computers, 37*, 547–559.

Murphy, G., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Nisbett, R., Krantz, D., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339–363.

Nosofsky, R. M. (1986). Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: Learning Memory, and Cognitition, 15*, 39–57.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, England: Cambridge University Press.

Rips, L. J., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General, 122*, 468–486.

Rogers, T. T., & McClelland, J. M. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: The MIT Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition, 73*, 1–33.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55*, 509–523.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*, 13–19.

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition, 38*, 179–211.

Spencer, J. P., Perone, S., Smith, L. B., & Samuelson, L. (in preparation). Non-Bayesian noun generalization from a capacity-limited system.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 893–907.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54*, 99–132.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*, 2319–2323.

Thibaut, J. P., Dupont, M., & Anselme, P. (2002). Dissociations between categorization and similarity judgments as a result of learning feature distributions. *Memory & Cognition, 30*(4), 647–656.

Tisseur, F., & Meerbergen, K. (2001). The quadratic eigenvalue problem. *SIAM Review, 43*(2), 235–286.

Turkington, D. A. (2002). *Matrix calculus and zero-one matrices: Statistical and econometric applications*. Cambridge, MA: Cambridge University Press.

Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review, 93*, 3–22.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245–272.

Yoshida, H., & Smith, L. B. (2001). Early noun lexicons in English and Japanese. *Cognition, 82*, 63–74.

Yoshida, H., & Smith, L. B. (2003). Shifting ontological boundaries: How Japanese- and English- speaking children generalize names for animals and artifacts. *Developmental Science, 6*, 1–34.

Zeigenfuse, M. D., & Lee, M. D. (2009). Finding the features that represent stimuli. *Acta Psychologica, 133*, 283–295.