UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

A Rational Model of Dimension-reduced Human Categorization

Permalink

https://escholarship.org/uc/item/17z1w2g9

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Hong, Yifan

Wang, Chen

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

A Rational Model of Dimension-reduced Human Categorization

Yifan Hong (hongyf23@mails.tsinghua.edu.cn)
Department of Industrial Engineering
Tsinghua University

Chen Wang (chenwang@tsinghua.edu.cn)
Department of Industrial Engineering
Tsinghua University

Abstract

Humans can categorize with only a few samples despite the numerous features. To mimic this ability, we propose a novel mixture of probabilistic principal component analyzers (mPPCA) model with dimension-reduced category representations, along with a theoretical analysis of rational dimensionality choices in categorization. Tests on the `CIFAR-10H` natural image categorization dataset show that introducing a single principal component for each category effectively improves predictions of human categorization patterns. We further use mPPCA to account for human category generalization with very few samples. In our experiments with visual patterns of varying size and color, combining principal components and the hierarchical prior leads to significantly better predictions of human generalization within and beyond previously learned categories.

Introduction

Human categorization grasps commonalities across items despite their differences. Although natural stimuli have numerous features, people can learn new categories with just a few instances (Lake et al., 2015) and generalize to novel observations (Salakhutdinov et al., 2012; Tiedemann et al., 2022). For example, a child can recognize a giraffe with only verbal descriptions. Theories suggest that people group instances with similar features together, and categories can be represented with past exemplars (Nosofsky, 1986) or abstract prototypes (Reed, 1972). Rational models (Anderson, 1991; Griffiths et al., 2007) provide a unifying perspective, casting categorization as optimal (Bayesian) inference. These models are insightful but struggle to explain categorization in *few-shot* settings where dimensions outnumber samples. For example, the prototype model with full-rank covariance cannot obtain a reliable estimate directly, and exemplar-based approaches are naturally biased on unbalanced categories.

This paper proposes a novel dimension-reduced category representation under the rational framework. Each category is described by a prototype and a set of principal components (PCs), characterizing the location and within-category variations, respectively. Additionally, we provide a theoretical rationale for when dimension-reduced representations reflect adaptation to the environment's information structure. A rational agent should remove a dimension from the category representation *if and only if* the dimension provides relatively more information about category differences than about within-category variation. On the natural image dataset with human labels `CIFAR-10H` (Peterson et al., 2019), representation with merely a single principal component proves highly effective in predicting human categorization patterns.

The dimension-reduced representation is compatible with a hierarchical prior over principal components. The resulting model, *mixture of probabilistic principal component analyzers* (mPPCA), suggests a principled way of generalization in the few-shot setting. Within an existing category, a principal component can serve as a low-dimensional local feature system to locate subcategories. For a new category, mPPCA prefers generalizing along principal components of existing categories. Behavioral experiments with simple visual patterns confirmed the anticipated generalization patterns. Compared with classical models, mPPCA provides significantly better predictions of how humans generalize categories.

Background

Models of human category learning

Categorization groups instances with similar features. Category representations enable accurate predictions and consistent generalizations. Cognitive models of categorization make various assumptions about category representations. For example, the prototype model (Reed, 1972) assumes that categories can be represented as abstract prototypes. People assign an instance to the category with probability proportional to the similarity to the prototype. The exemplar model (Nosofsky, 1986) considers a representation with all known category members. The rational model of categorization (RMC) (Anderson, 1991) offers a different perspective. It postulates that human categorization results from adapting to the optimal prediction of unobserved features. RMC models categories as probability distributions. Denote $x_n \in \mathbb{R}^d$ and c_n the new observation and its category assignment, respectively, and $\mathbf{x}_{n-1}$ and $\mathbf{c}_{n-1}$ the set of previous observations and their category memberships, respectively. The (posterior) predictive distribution of the features for a new observation is given by

$$P(x_n|\mathbf{x}_{n-1},\mathbf{c}_{n-1}) = \sum_{k=1}^{K} P(c_n = k|\mathbf{c}_{n-1}) \cdot P(x_n|c_n = k, \mathbf{x}_{n-1}, \mathbf{c}_{n-1}). \tag{1}$$

The formulation decomposes the prediction task the prior bias towards a particular category and the likelihood of an observation from that category. The prior can take the form of a Chinese restaurant process (CRP) (Blackwell & MacQueen,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

1973), a sequential process that over $\mathbf{c}_{n-1}$ that allows for infinite many categories. A sample is assigned to an existing category k with probability proportional to the number of existing samples M_k. Meanwhile, a new category emerges with probability proportional to a concentration parameter $\gamma > 0$:

$$P(c_n = k|\mathbf{c}_{n-1}) \propto \begin{cases} M_k & \text{if } M_k > 0 \ (k \text{ is old}) \\ \gamma & \text{if } M_k = 0 \ (k \text{ is new}) \end{cases} \quad (2)$$

The CRP is the marginal distribution of category assignment corresponding to a Dirichlet Process (DP), which governs the joint distribution of category assignments and parameters for each category (Teh et al., 2010). DP has the constructive process known as the stick-breaking construction (Blei & Jordan, 2006). For the prior probability measure G of category parameter θ (without specifying the category), we have $G \sim DP(\gamma, H)$, and G can be constructed as follows.

$$\begin{aligned} \beta_k \sim \text{Beta}(1,\gamma), \quad \theta_k \sim H, \\ \pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l), \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}. \end{aligned} \quad (3)$$

where H is the base measure in the DP. The intuition is to sequentially sample for each category parameter θ_k a proportion β_k from the remaining part of a stick (with $\sum_{k=1}^{\infty} \pi_k = 1$).

The likelihood in Equation (1) $P(x_n|c_n = k, \mathbf{x}_{n-1}, \mathbf{c}_{n-1})$ can be a multivariate normal distribution for continuous variables with parameters $\theta_k = (\mu_k, \Sigma_k)$ for category k. The distribution specifies the mean parameters μ_k representing category prototypes and the covariance parameters Σ_k defining dimensional variations.

RMC, with the CRP prior, enjoys the flexibility to learn an indefinite number of categories. It can also be used in supervised and unsupervised settings and allows subcategory modeling (Griffiths et al., 2007). However, RMC models dimensional variations with full-rank covariance, and it is generally difficult to discern the similarity between covariances.

Models of human generalization patterns

Humans exhibit consistent generalization patterns in the feature space, e.g., isotropic or dimension-aligned (Smith, 1989). Through category learning, they gradually exhibit preferences to generalize along some meaningful axis, such as size or color. Shepard (1987) uses L_1 metric and L_2 metric to describe these generalization patterns.

In rational models, the covariance matrix reflects graded generalization that rotates and scales the feature space. It implies a direction of strong generalization through its first principal component. Researchers have imposed a mixture prior on the covariance matrix (Heller et al., 2009) to highlight a preference to reuse dimensions for strong generalization. Consider a mixture of inverse Wishart distributions with J components. Denote Φ_j as the parameters for the j-th component. The prior for the covariance matrix of category k is given by $P(\Sigma_k|\Phi_1, \ldots, \Phi_J) = \sum_{j=1}^{J} P(u_k = j) P(\Sigma_k|\Phi_j)$ where u_k indicates which component to take effect. This model can also include infinitely many components using the CRP prior (Sanborn et al., 2021).

Notice that the covariance Σ_k holds full-rank information about rotation and scaling. However, humans tend to focus only on a subset of dimensions for categorization (Aha & Goldstone, 1992). Besides, the implicity of the generalization direction implied by the covariance makes it challenging to identify local structures. Therefore, a model incorporating dimension-reduction for each category can be favorable.

Dimension-reduced category representation

To properly characterize human categorization, we consider a combination of two elements: a flexible local dimension-reduced representation and a hierarchical structure for generalization. We start with dimension reduction for each category and then move on to a hierarchical model in the next section.

We propose a low-dimensional representation of categories based on probabilistic principal component analysis (PPCA, Tipping & Bishop (1999)). For a category $c \in C$, PPCA assumes that an observation $x_n \in \mathbb{R}^d$ is generated from a low-dimensional latent variable $z_n \in \mathbb{R}^q$ $(q < d)$ with transformation

$$x_n = W_c z_n + \mu_c + \varepsilon_n. \quad (4)$$

The columns of the *loading* matrix $W_c \in \mathbb{R}^{d \times q}$ suggest the directions of strong generalization. The latent variable $z_n \in \mathbb{R}^q$ indicates variations in these directions. $\mu_c \in \mathbb{R}^d$ is the category prototype. We assume normal priors for the latent variables $z_n \sim N(0, I_q)$ and noises $\varepsilon_n \sim N(0, \sigma^2 I_d)$ so as to marginalize out z_n analytically.

Assume the observations are from a DP mixture of categories. Denote $\{\theta_c\} = \{(\mu_c, W_c, \sigma_c^2)\}$ the parameters of all categories $c \in C$. With $n-1$ observations $\mathbf{x}_{n-1}$ and their category assignments $\mathbf{c}_{n-1}$, the DP updates the joint posterior

$$\begin{aligned} &P(\{\theta_c, \beta_c\}|\mathbf{c}_{n-1}, \mathbf{x}_{n-1}) \propto \\ &\quad P(\{\theta_c\}) P(\{\beta_c\}) P(\mathbf{c}_{n-1}|\{\beta_c\}) P(\mathbf{x}_{n-1}|\mathbf{c}_{n-1}, \{\theta_c\}) \end{aligned} \quad (5)$$

where $\{\beta_c\}$ come from the stick-breaking process, and $\{\theta_c\}$ are sampled from the base measure H of the DP. The marginal posterior distribution of the category parameter θ is derived as

$$\begin{aligned} &P(\{\theta_c\}|\mathbf{c}_{n-1}, \mathbf{x}_{n-1}) \propto P(\{\theta_c\}) P(\mathbf{x}_{n-1}|\mathbf{c}_{n-1}, \{\theta_c\}), \\ &P(\{\beta_c\}|\mathbf{c}_{n-1}, \mathbf{x}_{n-1}) \propto P(\{\beta_c\}) P(\mathbf{c}_{n-1}|\{\beta_c\}). \end{aligned} \quad (6)$$

To formulate H, we assume independent normal for μ_c and multivariate normal for W_c. Later we will modify the prior for W to incorporate shared principal component dimensions across categories. The features of an observation x_n given its category assignment c_n follow the multivariate normal distribution $x_n|c_n, \theta_{c_n} \sim N(\mu_{c_n}, W_{c_n} W_{c_n}^T + \sigma_{c_n}^2 I_d)$. We then introduce the PPCA *classifier* as the predictive distribution of category assignment c_n given observation x_n

$$P(c_n|x_n, \mathbf{c}_{n-1}, \mathbf{x}_{n-1}) = P(c_n|\mathbf{c}_{n-1}) P(x_n|c_n, \mathbf{c}_{n-1}, \mathbf{x}_{n-1}), \quad (7)$$

where $P(c_n|\mathbf{c}_{n-1})$ is obtained from the CRP, and the latter involves simulating the posterior of θ_{c_n},

$$P(x_n|c_n,\mathbf{c}_{n-1},\mathbf{x}_{n-1}) = \int_{\theta_{c_n}} P(x_n|c_n,\theta_{c_n})dP(\theta_{c_n}|\mathbf{c}_{n-1},\mathbf{x}_{n-1}). \tag{8}$$

Meanwhile, given category assignment c_n, the latent variable z_n has the posterior $z_n|x_n,c_n,\theta_{c_n} \sim N\big((W_{c_n}^T W_{c_n} + \sigma_{c_n}^2 I_q)^{-1} W_{c_n}^T (x_n - \mu_{c_n}), \sigma_{c_n}^2 (W_{c_n}^T W_{c_n} + \sigma_{c_n}^2 I_q)^{-1}\big)$. The PCs for each category span a low-dimensional feature system, with the latent variables explicitly capturing within-category variations.

Theoretical analysis of dimension reduction

When is a low-dimensional representation better than a full-rank representation? We explore this question by considering the limiting case of PPCA when $\sigma^2 \to 0$, so that it reduces to PCA, and by focusing on two categories $C = \{a,b\}$. Observations from each category c follow $x|c \sim N(\mu_c, \Sigma_c), \forall c \in C$. For simplicity, we assume equal covariance $\Sigma_a = \Sigma_b = \Sigma$. The probability of assigning observation x to the correct category (set to be a without loss of generality) is a sigmoid function

$$p(a|x) = \frac{e^{-\tau_q(x,\mu_a)}}{e^{-\tau_q(x,\mu_a)} + e^{-\tau_q(x,\mu_b)}} = \frac{1}{1 + e^{-\{\tau_q(x,\mu_b) - \tau_q(x,\mu_a)\}}}, \tag{9}$$

where τ_q is the projected distance of X to the subspace spanned by the first q PC dimensions ($q < d$), specified by eigenvectors $u_i, i = 1,\ldots,q$ of the covariance Σ, with decreasing eigenvalues $\lambda_1 \geq \ldots \geq \lambda_q$. The squared distance between category prototypes in the full-dimension space $r_{ab} = ||\mu_a - \mu_b||^2$ implies the amount of *total information*, while $r_i = ||(\mu_a - \mu_b)^T u_i||^2$ describes the proportion of total information explained by the i-th PC (with $\sum_{i=1}^d r_i = r_{ab}$). We call $\alpha_q \triangleq \tau_q(x,\mu_b) - \tau_q(x,\mu_a)$ the *sample discrimination index* for the q-dimensional PC subspace, reflecting how far the observation is from the wrong category relative to the correct one.

To investigate when should the $(q+1)$-th PC dimension be removed, and use the representation with the first q PCs instead, we study the signal-to-noise ratio (SNR) of the sample discrimination index

$$\text{SNR}_q = \frac{\mathbb{E}_x[\alpha_q]^2}{Var_x[\alpha_q]}, \quad q = 0,1,...,d-2.$$

Proposition 1 presents the necessary and sufficient condition for dimension-reduction in category representations to increase SNR. Proofs are presented in the appendix.

Proposition 1. *For given category prototypes μ_a, μ_b, discarding the $(q+1)$-th PC dimension ($q = 0,1,\ldots,q-2$) from the category representation increases the signal-to-noise ratio of α_q ($SNR_q > SNR_{q+1}$) if and only if*

$$\lambda_{q+1} < \left(\frac{r_{q+1}}{\sum_{i=q+2}^d r_i} + 2\right)\left(\frac{\sum_{i=q+2}^d r_i \lambda_i}{\sum_{i=q+2}^d r_i}\right). \tag{10}$$

The left-hand side characterizes within-category variation on the $(q+1)$-th PC. The first term on the right-hand side reflects the information provided by the $(q+1)$-th PC dimension for differentiating categories. The second term is a weighted average amount of within-category variation on the remaining PCs. Equation (10) suggests excluding a PC if it provides more information about category differences than within-category variation. It further implies an improved accuracy bound in categorization, which is a monotone function of SNR.

Corollary 1. *If* (10) *holds, dimension-reduction improves the lower bound for the classification accuracy of the PCA classifier.*

Hierarchical prior on feature dimensions

Mixture of PPCA (mPPCA) model

Now we present the mixture of PPCA (mPPCA), a nonparametric Bayesian hierarchical model. It introduces dependencies between categories by sharing PCs among them.

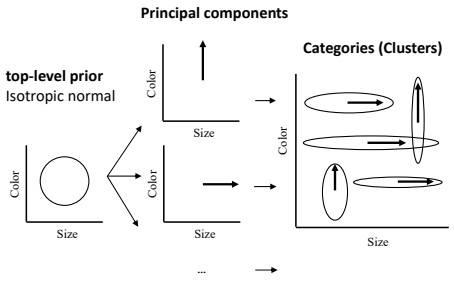


Figure 1: Conceptual schematic of mPPCA, adapted from (Sanborn et al., 2021).

There are two infinite mixtures in the prior of mPPCA. On the lower level, mPPCA describes observations as an infinite mixture of categories. The CRP prior (2) over category assignment allows infinitely many categories, but materializes only a finite set given the observations. Each category is represented by PPCA with its own parameters μ_c, W_c. In this section, we assume each category has only one direction of strong generalization, represented by a *single* PC dimension $w_c \in \mathbb{R}^d$. On the higher level, we introduce another infinite mixture to share PC dimensions among categories. For each category c, an ownership indicator u_c indexes a *component* ν_j in the top-level CRP, which determines the category PC w_c. Observations are drawn from the generative process presented in Figure 1 (See the appendix for a formal description).

Adding a hierarchical prior changes the inference process. The posterior can be decomposed as the product of the conditional distribution of component-level parameters $\{\beta'_j, \nu_j\}$ and the marginal of category-level parameters $\{\theta_c, \beta_c\}$.

$$p(\{\beta'_j,\nu_j\},\{\theta_c,\beta_c\}|\mathbf{x}_{n-1},\mathbf{c}_{n-1}) = p(\beta'_j,\nu_j|\{\theta_c\})p(\{\theta_c,\beta_c\}|\mathbf{x}_{n-1},\mathbf{c}_{n-1}) \tag{11}$$

where the second term is Equation (5). The first term is the posterior of the CRP mixture with concentration parameter γ'

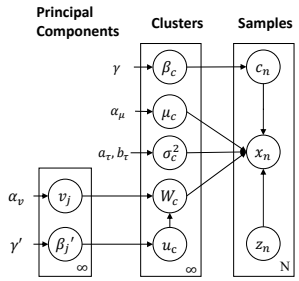


Figure 2: Graphical representation of mPPCA.

and normal base measure with mean $\mathbf{0}$ and covariance $\frac{1}{\alpha_v}I$,

$$p(\{\beta'_j, \nu_j\}|\{\theta_c\}) \propto p(\{\beta'_j\}|\gamma')p(\nu_j|\alpha_\nu) \cdot p(\{u_c\}|\{\beta'_j\})p(\{w_c\}|\{u_c\}, \{\nu_j\}). \quad (12)$$

Equation (11) implies that the full posterior can be derived by the marginal posterior of category-level parameters and the conditional probability of component-level parameters.

Explaining few-shot generalization with mPPCA

mPPCA suggests a principled approach to generalize in the few-shot setting, operating across two distinct aspects. Within existing categories, local PC helps locate prototypes for subcategories. Beyond an existing category, shared PC components facilitate the learning of new categories.

Learning sub-categories The principal components guide generalization within categories. Within a category c with learned posterior over prototype μ_c and the category PC w_c, specifying a sub-category with latent variables requires only a low-dimensional latent variable z_{sub} as input. The category PC w_c serves as a local feature system to locate subcategory prototype

$$\mathbb{E}[\mu_{sub}] = \mathbb{E}[\mu_c] + \mathbb{E}[w_c]z_{sub}. \quad (13)$$

Using point estimates $\hat{\mu_c}, \hat{w_c}$ for convenience, the subcategory can be learned with no observation but only an instruction specifying z_{sub}.

$$p(x|z_{sub}, \hat{\mu_c}, \hat{w_c}, \sigma^2) = N(x|\hat{\mu_c} + \hat{w_c}z_{sub}, \sigma^2 I) \quad (14)$$

We assume the covariance of the subcategory does not involve the category PC, which follows the intuition given by the theoretical analysis.

Learning new categories Hierarchical prior over PCs guides generalization of a new category. Given only one sample x_{new} from a new category, a covariance cannot be estimated directly. By contrast, hierarchical prior in mPPCA allows the new category to inherit generalization patterns from the existing ones. mPPCA suggests a category with mean $\mu_{new} = x_{new}$ and a PC w_{new} sampled from the CRP posterior

$$P(w_{new}|x_{new}) \propto \sum_{u_{new}} P(w_{new}|u_{new}, \{\nu_j\})P(u_{new}|\{\beta'_j\}) \quad (15)$$

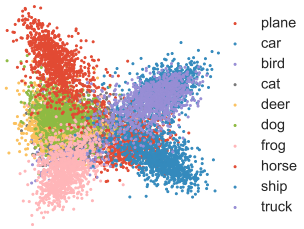


Figure 3: Two-dimensional linear-discriminant-analysis projections of the `ResNet18` stimulus representations.

Table 1: Performance of categorization models on `CIFAR-10H` dataset, using pre-trained `ResNet` for feature representation.

	Accuracy	SBA	Rank Correlation
Exemplar	0.6037	0.3270	0.4007
Prototype	**0.9102**	0.2737	0.4182
PPCA (dim 1)	0.8784	**0.4947**	**0.7665**
PPCA (dim 2)	0.8554	0.4885	0.7592

with strong generalization along existing PCs. Categorization then follows Equation (7) and Equation (8).

Experiments

We carry out two sets of experiments to test the model predictions. We first explore human categorization of natural images. Then we turn to artificial categories to examine human few-shot generalization behavior.

Natural image categorization

We explore human categorization of natural images using `CIFAR-10H` (Peterson et al., 2019), which includes 50 human labels for each of the 10000 images in `CIFAR-10` test set.

Procedure We use pre-trained convolutional network `ResNet18` (Figure 3) as the feature map [1], as it can predict human categorization decisions(Singh et al., 2020). We train the categorization models on `CIFAR-10` training set, and then test their performance and resemblance to human categorization on `CIFAR-10H`. For PPCA, we use maximum likelihood estimation of model parameters since there are no new categories or subcategories.

Metrics Besides accuracy, we record second best accuracy (SBA) and rank correlation with human data. SBA is the proportion of images on which the model predicts the second common human choice correctly. Rank correlation evaluates the ordinal associations between the model predictions and human choices. When models are comparable on accuracy, SBA and rank correlation evaluates the prediction of graded human categorization patterns.

[1] Adapted under the MIT licence from https://github.com/huyvnphan/PyTorch_CIFAR10.

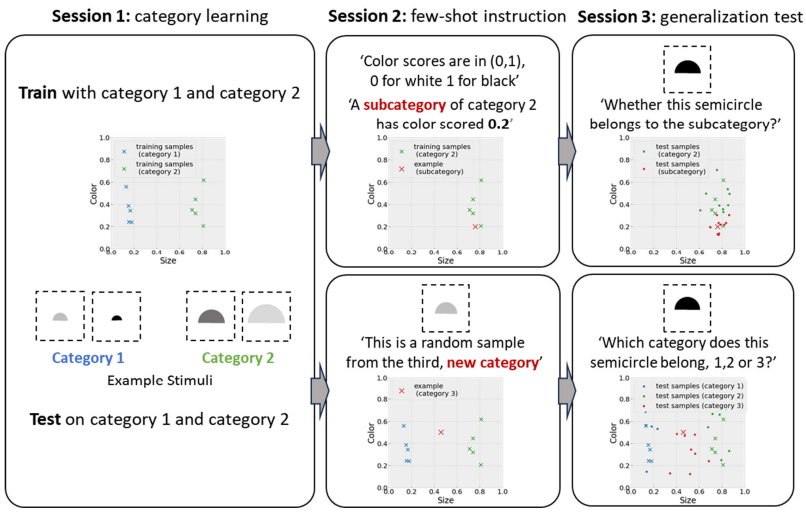


Figure 4: Procedure of few-shot generalization experiment. Category 1 or 2 contains semicircles of regular size but varying colors. After category learning in Session 1, Session 2 provides either a sample or an instruction. The new category is similar to Category 1 and 2 but with different sizes. The subcategory is generated from an isotropic Gaussian distribution, aligned with Category 2 on the size dimension. Generalization patterns are tested in Session 3.

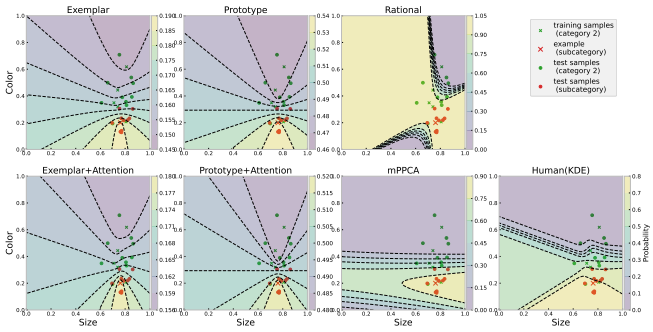


Figure 5: Predicted patterns of subcategory. Dashed lines represent equal generalization probability and dots are the training and test exemplars.

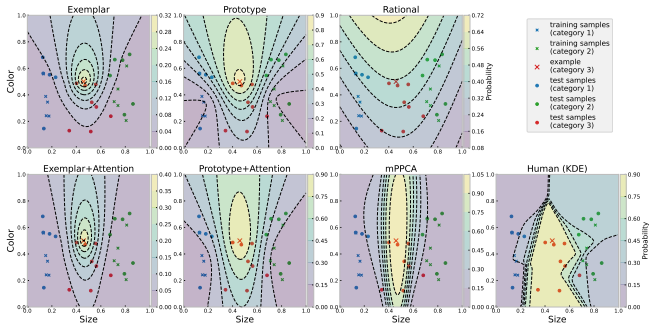


Figure 6: Predicted patterns of new category. Dashed lines represent equal generalization probability and dots are the training and test exemplars.

Results One-dimensional category representation can effectively capture human categorization of natural images. mPPCA model with a single PC in each category representation achieves better SBA and rank correlation (Table 1), surpassing both exemplar and prototype models. Meanwhile, increasing dimensionality does not further improve performance.

Category few-shot generalization

In this study, we conducted two behavioral experiments to study human few-shot generalization of a new subcategory or category. We design artificial categories to mitigate the effect of human priors on natural images (Ma et al., 2023).

Stimuli The stimuli are semicircles with varying *color* and *size*, two commonly used separable dimensions in previous studies (Smith, 1989; Heller et al., 2009). We used dimension rating data from a pilot study to scale stimulus parameters based on perceived similarity. Each category corresponds to a bi-variate normal distribution in the size-color space, having small variance on the size dimension and large variance on the color dimension. The subcategory has small variance on both dimensions. Stimuli are independent and identically sampled from the (sub)categories.

Procedure Participants go through 3 sessions for each experiment: a category learning session containing the training and test phases, a few-shot instruction session, and a generalization test session (Figure 4). First, in the *category learning session*, participants get familiar with the categories and their variations. They undergo training and test phases, with 20 samples in each phase from Category 1, 2 or neither. Training lasts until participants correctly categorize all the training samples. No feedback is available during the test. Second, in the *few-shot instruction session*, participants learn about a

Table 2: Performance in predicting human few-shot generalization of subcategories and new categories

	subcategory learning		new category learning	
Model	expected accuracy	correlation	expected accuracy	correlation
Exemplar	0.517±0.043	-0.102±0.118	0.594±0.063	0.372±0.128
Exemplar+Attention	0.498±0.041	-0.102±0.118	0.620±0.064	0.407±0.117
Prototype	0.599±0.020	0.351±0.091	0.562±0.030	0.607±0.051
Prototype+Attention	0.555±0.012	0.351±0.091	0.638±0.030	**0.688±0.044**
Rational model	**0.668±0.039**	0.374±0.068	0.467±0.019	0.570±0.044
mPPCA (**Ours**)	**0.662±0.033**	**0.451±0.065**	**0.705±0.028**	**0.696±0.040**

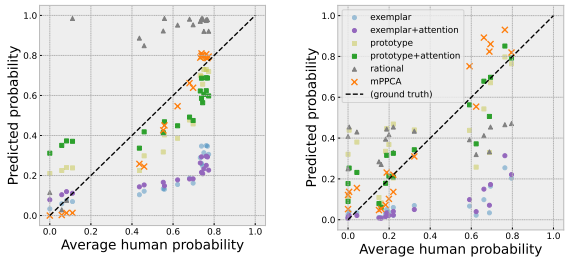
new (sub)category. In the *subcategory* experiment, verbal description of a subcategory is provided, describing its category PC (color) score. In the *new category* experiment, one sample from the new category is provided. Third, in the *generalization test session*, participants categorize 20 test stimuli. The *subcategory* experiment consists of samples from Category 2, some of which come from the specified subcategory. The participants judge whether the test stimuli come from the subcategory. In the *new category* experiment, participants classify the samples into Category 1, 2, or the new one.

Participants We recruited 200 participants online for each experiment on *Credamo*, with 172 and 186 passing the attention tests (participants are required to choose a specific category for that test), respectively. Participants were compensated for participating. The median time for both experiments was 9 minutes. We obtained IRB approval at our institute (IRB ref number is omitted for anonymity.). Participants undergo informed consent before they took the experiment.

Results After training, most participants effectively learned the new (sub)category (appendix, Figure 8). We compare mPPCA to prototype and exemplar models, with or without attention mechanism, and the rational model. The attention mechanism scales the original space with dimensional weights optimized based on cluster variations. The rational model is adapted for the subcategory learning by treating the subcategory as a cluster. The model setup is detailed in the Appendix. Overall, exemplar models fail to capture the learning of a new category. Prototype models cannot generalize well. The rational model suffers from identifying a cluster as the subcategory and covariance estimation for the new category. mPPCA provides a better account of human few-shot categorization (Tukey's HSD, $\alpha < 0.05$; see Table 2).

In the subcategory experiment, mPPCA produces predictions with significantly higher accuracy and correlation with human choices. Its generalization pattern matches human behavior (Figure 5). Both exemplar models and prototype models, with or without attention mechanism, underestimate the probability of the subcategory (Figure 7a). Attention mechanism harms categorization performance within a category. Instead, humans adopt flexible feature weighting within or beyond categories, as is captured by mPPCA.

In the new category experiment, exemplar models, affected by unbalanced categories, underestimate the probability of the new category (Figure 7b). Rational model uses an uninformative prior for the covariance of the new category, deviating from human behavior (Figure 6). Prototype models (with or without attention) produce similar generalization patterns as mPPCA (Figure 9), both providing good predictions. mPPCA predicts human category assignments more accurately.



(a) subcategory (b) new category

Figure 7: Model prediction probabilities of assigning each stimulus to the new (sub)category.

Discussion

In this work, we propose a rational model of categorization with dimension-reduced category representation. Such a low-dimensional representation benefits categorization in an environment where certain dimensions provide more information about category differences than about internal variations. The model captures human categorization patterns on `CIFAR-10H`, and can effectively predict human few-shot generalization within or beyond categories.

PPCA has interesting properties worth exploring. It does not impose orderings among PCs, enabling context-dependent ordering of features, as in cross-categorization (Shafto et al., 2011). Besides, PPCA does not assume orthogonality and can learn correlated features, which is similar to human feature learning (Austerweil & Griffiths, 2010).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) 72192824.

References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In *Proceedings of the fourteenth annual conference of the cognitive science society* (pp. 534–539).

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

Austerweil, J., & Griffiths, T. (2010). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for dirichlet process mixtures.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Heller, K. A., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. *Advances in Neural Information Processing Systems*, *22*.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Ma, W. J., Kording, K. P., & Goldreich, D. (2023). *Bayesian models of perception and action: An introduction*. MIT press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9617–9626).

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2012). One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of icml workshop on unsupervised and transfer learning* (pp. 195–206).

Sanborn, A. N., Heller, K., Austerweil, J. L., & Chater, N. (2021). Refresh: A new approach to modeling dimensional biases in perceptual similarity and categorization. *Psychological Review*, *128*(6), 1145.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Singh, P., Peterson, J., Battleday, R., & Griffiths, T. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. In *Proceedings of the 42nd annual conference of the cognitive science society.*

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, *96*(1), 125.

Teh, Y. W., et al. (2010). Dirichlet process. *Encyclopedia of machine learning*, *1063*, 280–287.

Tiedemann, H., Morgenstern, Y., Schmidt, F., & Fleming, R. W. (2022). One-shot generalization in humans revealed through a drawing task. *Elife*, *11*, e75485.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *61*(3), 611–622.

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the annual conference of the cognitive science society* (Vol. 27, pp. 2277–2282).

Appendix: model and theory

Generative process of mPPCA

(1) For each component in the higher-level mixture,

(a) Draw probabilistic PC $\nu_j \sim N(0, \frac{1}{\alpha_\nu} I_d)$.

(b) Draw stick-breaking weight $\beta_j^* \sim \text{Beta}(1, \gamma^*), \pi_j^* = \beta_j^* \prod_{i=1}^{j-1}(1-\beta_j)$.

(2) For each category in the lower-level mixture,

(a) Draw component assignment $u_c \sim \text{Mult}(\{\pi_j^*\})$. $w_c = \nu_{u_c} + \xi_c$, where ξ_c is a normal noise term.

(b) Draw category prototype $\mu_c \sim N(0, \frac{1}{\alpha_\mu} I_d)$.

(c) Draw stick-breaking weight $\beta_c \sim \text{Beta}(1, \gamma), \pi_c = \beta_c \prod_{l=1}^{c-1}(1-\beta_c)$.

(d) Draw noise variance $\sigma_c^2 \sim \text{Inv-Gamma}(a_\tau, b_\tau)$.

(3) For each sample x_n, $n = 1, ..., N$,

(a) Draw category assignment $c_n \sim \text{Mult}(\{\pi_c\})$

(b) Draw latent variable $z_n \sim N(0, 1)$.

(c) Draw observation

$$x_n | z_n, c_n \sim N(\mu_{c_n} + w_{c_n} z_n, \sigma_c^2 I_d)$$

Proofs for the theoretical analysis

Proof for Proposition 1

Proof. We first formalize some useful notation. The covariance matrix of the categories has eigen-decomposition $\Sigma_c = U \Lambda U^T$, where the diagonal matrix $\Lambda = \text{diag}(\lambda_1, ..., \lambda_d)$ consists of the eigenvalues, and columns of U are corresponding eigenvectors. The truncated matrix containing first q columns of U is denoted as U_q, with corresponding $\Lambda_q \text{diag}(\lambda_1, ..., \lambda_q)$, $q < d$. Linear projection into the subspace is represented as $P = W(W^T W)^{-1} W^T = U_q U_q^T$.

According to the definition, the sample distinction index $\alpha \triangleq ||(I-P)(x-\mu_b)||^2 - ||(I-P)(x-\mu_a)||^2$. For any given category prototypes, μ_a, μ_b, and projection matrix $P = U_q U_q^T$, the expectation and variance of α can be derived as

$$\mathbb{E}_x[\alpha] = ||(I-P)(\mu_a - \mu_b)||^2 = r_{ab} - \sum_{i=1}^{q} r_i, \tag{16}$$

$$\begin{aligned} \text{Var}_x[\alpha] &= 4(\mu_a - \mu_b)^T \Sigma_c (\mu_a - \mu_b) - \\ 4((\mu_a - \mu_b)^T U_q \Lambda_q U_q^T (\mu_a - \mu_b)) &= 4 \sum_{i=q+1}^{d} \lambda_i r_i. \end{aligned} \tag{17}$$

When considering distance to the principal subspaces spanned by the first q eigenvectors, the signal-to-noise ratio of α

$$\text{SNR}_q = \frac{\mathbb{E}_x[\alpha]^2}{var_x(\alpha)} = \frac{1}{4} \frac{(r_{ab} - \sum_{i=1}^{q} r_i)^2}{\sum_{i=q+1}^{d} \lambda_i r_i} \tag{18}$$

Hence, the decision to exclude dimension $q+1$ will increase signal-to-noise ratio ($\text{SNR}_{q+1} > \text{SNR}_q$) if and only if $\lambda_{q+1} < \frac{2r_{ab} - 2\sum_{i=1}^{q+1} r_i + r_{q+1}}{(r_{ab} - \sum_{i=1}^{q+1} r_i)^2} \sum_{i=q+2}^{d} r_i \lambda_i$, which leads to inequality 10 with direct transformation. □

Proof for Corollary 1

Proof. PCA corresponds to the limit of PPCA as $\sigma^2 \to 0$. Hence, the classifier chooses with probability 1 the category whose principal subspace is the closest. This leads to $p(\hat{y} = a | a, b) = p(\alpha > 0 | a, b)$. From the one-sided Chebyshev's inequality, we derive the lower bound of classification accuracy

$$P(\alpha > 0 | a, b) \geq \frac{E_x[\alpha]^2}{Var(\alpha | a, b) + E_x[\alpha]^2} = \frac{\text{SNR}}{1 + \text{SNR}} \tag{19}$$

Since it is a monotonic function of signal-to-noise ratio, we immediately arrives at the corollary. □

Appendix: experiment details

Details of category few-shot generalization experiment

Here, we present some detailed results that are not included in the main body due to space constraints.

Model setup To provide predictions of human categorization, our models experience the same set of data. Given both stimuli x_n and labels c_n, the higher-level mixture is disentangled from the lower level. We first obtain the posterior of the lower-level mixtures. Then we use variational inference for the high-level mixture, i.e. the global PCs, according to Equation (12). Because of the task context, participants treat the mentioned categories with equal expectation. We set the base rate to equal values, which leads to better prediction for all models. In the new category experiment, all stimuli in the train and test phase of Session 1 is used to get a more reliable estimate of global PCs. This will not be necessary for the subcategory experiment, since only the local PC is needed.

We compare our model with the exemplar model (Nosofsky, 1986), prototype model (Reed, 1972), with and without attention mechanism, and the rational model (Anderson, 1991). All models provide predictions without access to human choice. The attention mechanism scales the original space with a set of dimensional weights, optimized based on cluster variations. Prototype model with attention mechanism generates generalization pattern similar to that of hierarchical models (Salakhutdinov et al., 2012; Sanborn et al., 2021), since the categories are dimension-aligned in the experiment.

Rational model represents a category as a infinite mixture of clusters. For subcategory prediction, we assume rational model treats subcategory as one of its clusters. The model first use the instruction to identify the subcategory as one of the clusters. Then for each new sample y, we estimate the probability of it belonging to each cluster

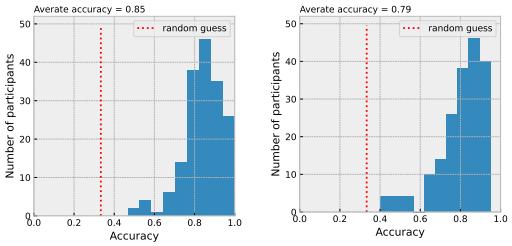
$$P(y \in \text{Subcategory} | x_{\text{sub}}) = \sum_k P(k | x_{\text{sub}}) P(y | k),$$

where k indicates clusters within the category.

Meanwhile, we cannot estimate the new category's covariance with one sample. Using a prior on cluster covariance is

not fair since category and cluster belong to different levels. As a result, we consider similarity by calculating the sum of similarity to clusters of that category. This is similar to the varying abstraction model (Vanpaemel et al., 2005).

Session 1: learning The training and testing phases helps the participants familiarize the stimuli, and learn the category structure in this artificial environment. Figure 8 shows that the subjects have indeed learned the categories, with accuracy significantly surpassing random guess.



(a) subcategory (b) new category

Figure 8: Subject categorization test accuracy in Session 1. The majority of participants learn the new subcategory or the new category effectively.

Session 3: generalization Test on generalization of categories is the main part of the experiment. Here we present the model predictions on the test stimuli, given the same training experience as human participants.

Figure 7 plots the predicted probability of assigning test stimuli to the subcategory (Figure 7a) and new category (Figure 7b) against human assignment probability. For the **subcategory experiment**, exemplar models systematically underestimate the probability of the subcategory. Notice that even for the quantitatively best-performing PPCA, there is some underestimate of assignment probability, especially on those stimuli with human assignment probability around 0.5. We consider this may be an effect of task context. Given specific instruction in **Session 2** about the existence of a subcategory, participants may naturally tend to choose the subcategory, when they are actually uncertain about the category membership. Prototype models exhibit complex nonlinear patterns. They cannot capture human generalization with flexible switching between contexts. For the **new category** experiment, the exemplar models again underestimates the probability of the new category. Prototype models provide similar predictions, but generally deviates more from the "ground truth".

Figure 5 and Figure 6 provides the generalization gradients of the subcategory and new category, respectively. In the subcategory experiment, rational model fails to identify the subcategory. This is because learned clusters are not aligned with the subcategory. Exemplar and prototype-based models cannot adjust to the category context flexibly. mPPCA matches human behavior quite well. In the new category setting, mPPCA is similar to prototype with attention. However, we point out that a fixed set of attention cannot account for human categorization. Therefore, mPPCA stands out in explaining human categorization patterns in our experiment.

For the new category experiment, we also use heatmap in Figure 9 to illustrate categorization patterns. We can see that mPPCA and prototype model (with attention) provide predictions similar to human categorization probability. Without attention mechanism, prototype model fails to focus on important dimensions for the current task. Exemplar models, on the other hand, underestimates the probability of the new category.

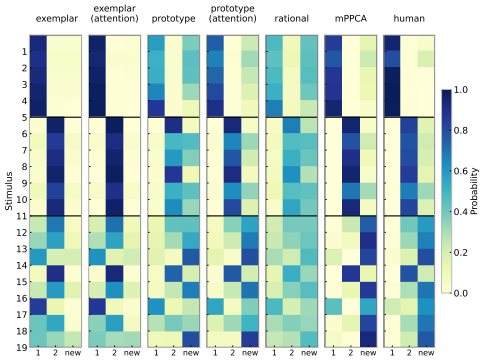


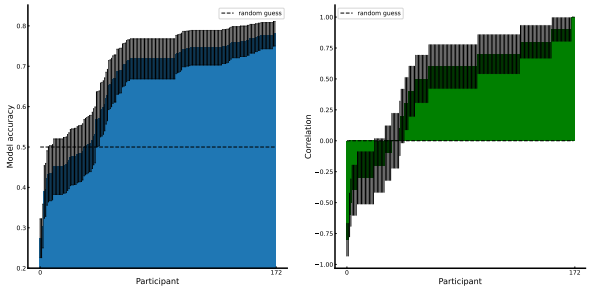
Figure 9: Stimulus-category similarity heatmaps of models in the new category experiment. On the right most is human choice probability.

For more detailed analysis, we show prediction performance for each participant, and each randomly generated stimuli. Figure 10a and Figure 11a illustrate the expected accuracy of mPPCA when predicting human choice probability on the subcategory and new category experiment, respectively. Figure 10b and Figure 11b show the correlation with human categorization on the subcategory and new category experiment, respectively. mPPCA provides a good estimation in these two experiments.

Figure 10c and Figure 11c show the expected accuracy on for each randomly generated stimulus in the subcategory and new category experiments. In the subcategory experiment, mPPCA performs at least comparably with other models, and is significantly better on some of them. In other words, mPPCA dominates the baseline models, both in terms of accuracy and correlation. In the new category experiment, mPPCA is outperformed by the exemplar models on simuli from category 1 and category 2. This is caused by the bias of exemplar models towards these categories, which have more training samples. In general, mPPCA provides a better account of human categorization pattern.

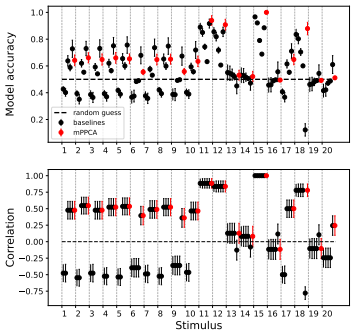
Impact statements

This study shares many of the potential societal impacts as other computational cognitive science research. This study focuses on the human behavior of categorization. The major goal of this study is to better the understanding of human mind



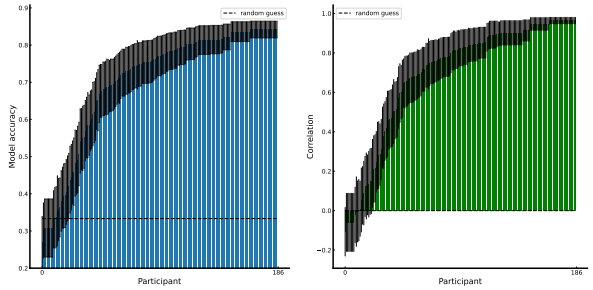
(a) Model accuracy for each individual participant in the subcategory experiment.

(b) Model correlation to each individual participant in the subcategory experiment.



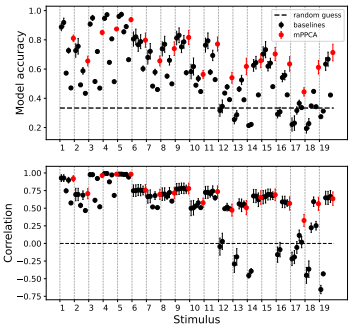
(c) Model prediction performance on each stimulus in the subcategory experiment.

Figure 10: Participant and stimulus-level analysis of the subcategory experiment.



(a) Model accuracy for each individual participant in the subcategory experiment.

(b) Model correlation to each individual participant in the new category experiment.



(c) Model prediction performance on each stimulus in the new category experiment.

Figure 11: Participant and stimulus-level analysis of the new category experiment

using computational models. It is necessary to guard against intentional manipulation of humans with the insight provided by cognitive science studies. The gravity of this issue may not be obvious for the current study, but because categorization is a fundamental cognitive activity, we believe it is critical to be cautious about the abuse of scientific discoveries.

During our behavior experiments, human participants were recruited online. The experiments have minimal risk. We followed existing protocols and went through the informed consent procedure. The participants were aware of the procedure and can withdraw at any time. They received fair compensation for participation. They allowed the data to be used for the present study. Private information is not used or disclosed. We obtained the IRB approval at our institute. The IRB ref is not disclosed in the current version for anonymity.