International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:4, 2007

# Generalized Exploratory Model of Human Category Learning

Toshihiko Matsuka

***Abstract*—** One problem in evaluating recent computational models of human category learning is that there is no standardized method for systematically comparing the models' assumptions or hypotheses. In the present study, a flexible general model (called GECLE) is introduced that can be used as a framework to systematically manipulate and compare the effects and descriptive validities of a limited number of assumptions at a time. Two example simulation studies are presented to show how the GECLE framework can be useful in the field of human high-order cognition research.

***Keywords*—** artificial intelligence, category learning, cognitive modeling, radial basis functions.

## I. Introduction

THE past fifteen years have seen significant advances in adaptive network models of category learning. In particular, three models of human category learning have attracted much attention, namely ALCOVE [1], RASHNL [2], and SUSTAIN [3]. These models share several properties, because they can be considered special cases of Generalized Radial Basis Functions [4, 5], as discussed by several authors [6 – 8]. First, all three models are multilayer adaptive network models, with internally represented "reference points" ("basis units" in RBF terminology) in their memory (specifically, in the hidden layer). The models all use similarities between the internally represented reference points (RPs) and the input stimuli for calculating activations of RPs. Then, the weighted RP activations are fed forward to output nodes, whose activations are used to categorize the input stimuli. Next, all three models scale feature dimensions independently in calculating these input-to-RP similarities, and this scaling process is interpreted as reflecting dimensional attention processes [6]. In addition, all models incorporate as their basic learning method a form of gradient descent for incremental adjustments of both association weights and dimension-specific attention parameters. Finally, all three models may be considered as confirmatory models, because they are based on specific a priori assumptions about how humans process information in categorization (e.g. how stimuli are internally represented & how humans pay attention to stimuli's feature dimensions). These a priori assumptions are usually justified by the empirical results. There have been some modeling studies based on simulations trying to justify the validity of their assumptions by comparing several different cognitive models consisting of different model assumptions. In order to test these specific assumptions, the assumptions should be varied systematically, and tested by comparing the fit of the resulting competing models to the empirical data. However, the confirmatory nature of the recent computational models and the differences among these models, some being possibly crucial as described below, generally prevent us from making such systematic comparisons of competing model assumptions.

The significant differences among the three models are as follows. Firstly, the assumptions about how stimuli are internally represented are different. ALCOVE and RASHNL are exemplar models, in the sense that each stimulus in the training set is allocated as an RP in the "hidden" layer of the network, and the RPs reside in the fixed locations. In contrast, SUSTAIN is a prototype model that uses a reduced number of updateable or movable RPs in its hidden layer, corresponding to potential generalizations. In addition, SUSTAIN dynamically allocates new prototypes, thus it may use multiple prototype nodes for each category explicitly defined by the corrective feedback. Secondly, how RP activations are utilized in making category predictions and in adjusting parameter values during learning are different among the models. SUSTAIN utilizes only the single most activated RP for categorization and learning, whereas ALCOVE and RASHNL utilize the activations of all RPs. Thirdly, the assumptions about attention processes are different. RASHNL assumes limited attention capacity and rapid shifts in attention processes, whereas ALCOVE and SUSTAIN do not. Finally, the functions for computing similarity measures and RP activations are different.

There have been several studies comparing computational models of categorization, including but not limited to ALCOVE, RASHNL, and SUSTAIN (e.g., [9, 10]). Although these comparative studies provided information on the models' capabilities for reproducing human-like categorization learning, they did not necessarily provide information that can lead to specific understanding of the nature of human category learning. That is because model-to-model comparisons may not be informative for testing the

Manuscript received December 9, 2003. This work was supported in part by the Nation Science Foundation under Grant EIA-0205178 and the James McDonnel Foundation.

T. Matsuka is with Rutgers University Mind and Brain Analysis (RUMBA) Laboratory, Rutgers University. 101 Warren Street, Newark, NJ 07102 USA (phone: +1 973-353-5440 x239; fax: +1 973-353-1170; e-mail: matsuka@psychology.rutgers.edu).

plausibility of each specific assumption. Rather, such model comparisons are essentially omnibus tests collectively comparing all variations in assumptions at once. In other words, these studies involving model comparisons do not effectively point out which element, assumption, or structure of models was responsible for successful or unsuccessful replication of observed tendencies and phenomena in human category learning.

Since it has been difficult, if not impractical to use the results of these previous comparative studies to understand which specific assumptions are supported by the data, it seems desirable to develop and apply a general framework for modeling human category learning that allows us to manipulate and test one or a limited number of model assumptions at a time. The framework should be general and flexible, so that we can conduct standardized exploratory model comparisons of various types of human cognitive processes associated with categorization to better understand the nature of human category learning.

## II. New Model of Human Category Learning

### *A. Qualitative Descriptions*

GECLE (for Generalized Exploratory models of Category LEarning) is a framework for a general and flexible exploratory approach for modeling human category learning, that is capable of modeling human category learning with many variants using different model assumptions. This general framework allows model assumptions to be manipulated separately and independently. For example, one can manipulate assumptions about how stimuli are internally represented (e.g. exemplars vs. prototypes), or about how people selectively pay attention to input feature dimensions (e.g., paying attention to dimensions independently or not).

The GECLE model uses the Mahalanobis distances (in quadratic form) between the internally represented reference points (corresponding to either exemplars or prototypes) and the input stimuli as the measure of similarity between them. The entries in the covariance matrix, which are used for calculating the Mahalanobis distances, are considered as attention parameters that control a process called psychological or mental scaling which regulates perceived distances between an input stimulus and reference points. Thus, unlike other NN models of category learning, the GECLE does not necessarily assume that attention is allocated independently dimension-by-dimension. Rather, it assumes that humans in some cases might pay attention to correlations among feature dimensions. This allows the GECLE to model processes interpretable as dimensionality reduction or mental rotation in the perception and learning of stimuli. Such processes may increase the interpretability of stimuli in the categorization task and learning. Another motivation for the use of the Mahalanobis distance is that the capability of paying attention to correlations among feature dimensions may be needed for classification tasks defined on integral stimuli.

In the GECLE framework, the attention parameters (i.e., the diagonal and off-diagonal elements of the covariance matrices, see Equation 1) can be considered as *shape* and *orientation* parameters for receptive fields or attention coverage areas of the reference points. Note that virtually all neural network-based models of category learning incorporate the "dimensional attention processes" assumption (i.e., attention is allocated independently on a dimension-by-dimension basis), causing the models to stretch and shrink attention coverage areas orthogonal to feature dimensions. This type of attention process can be incorporated in GECLE models by constraining the off-diagonal entries in the covariance matrices to be equal to zero.

Another unique feature of GECLE's attention mechanism is that it allows each reference point to have uniquely shaped and oriented attention coverage area (Fig. 2D, 2E, and 2F). This type of attention coverage is denoted as "local attention coverage structure". Again, one can impose a restriction on the model's attention mechanism by fixing all covariance matrices to be the same, which may be referred to as "global attention coverage structure" (Fig. 2A, 2B, and 2C). Many NN models of category learning, ALCOVE, for example, incorporate the global attention coverage structure [1 –3].

The local attention coverage structure model is complex, but may plausibly model attention processes in human category learning. For example, it allows models to be sensitive to one particular feature dimension when the input stimulus is compared with a particular reference point that is highly associated with category X, while the same feature dimension receives little or no attention when compared with another reference point associated with category Y. Thus the local attention coverage structure causes models to learn and be sensitive to within-cluster or within-category feature configurations, while the global attention coverage structure essentially stretches or shrinks input feature dimensions in a consistent manner for all RP receptive fields and all categories.

Another way of interpreting GECLE's capabilities for paying attention to correlations among feature dimensions and having local attention coverage structures is that the model learns to define what the feature dimensions are for each RP and to allocate attention to those dimensions independently. In contrast, for almost all previous adaptive models of category learning, the definition of the feature dimensions is static and supplied by individuals who use the models.

The other notable characteristic of the GECLE's attention mechanism is that the user can manipulate characteristics of the distributions assumed for the activations of the reference points as described in next section. For example, one can have RP activation distributions with thicker tails to obtain more competition among the RPs.

To determine category membership of an input stimulus, GECLE use the similarities between the input stimulus and all reference points collectively as the evidence of category membership. The evidence of the input stimuli belonging to each category is represented with a numerical value, and categorization response is probabilistically determined based on relative strength of (transformed) evidence in the GECLE framework.

In its natural form, the GECLE may be considered as a model using prototype internal representation, because it tries to learn to locate its reference points at the centers of each category cluster. However, with a proper user-defined parameter setting, it can behave like a model with an exemplar-based internal representation.

### *B. Quantitative Descriptions (Algorithm)*

The feedforward and learning algorithms of the GECLE are typical for implementation of the Generalized Radial Basis Function [5, 6, 11]. GECLE uses the following function to calculate the distances or similarity between internally represented reference points (e.g., prototypes or exemplars) and input stimuli:

$$D_j(x,r) = (x - r_j)^T \Sigma_j^{-1} (x - r_j) \tag{1}$$

where $x$ is an $I$-tuple vector representing an input stimulus consisting of $I$ feature dimensions, $r_j$, also an $I$-tuple vector, that corresponds to the centroids of reference point $j$, expressing its characteristics, and $\Sigma_j^{-1}$ is the inverse of the covariance matrix, which defines the shape and orientation of the attention coverage area of reference point $j$. For a model with global attention coverage structure, there is only one global $\Sigma^{-1}$ for all reference points. The entries ($s_{im}$) in $\Sigma_j$ are assumed and constrained to satisfy the following conditions: $s_{ii} \geq 0$ & $|s_{im}| \leq \mathrm{MAX}(s_{ii}, s_{mm})$. That is attention strength is always non-negative and attention allocated to a covariation for any given pair of feature dimensions must be less than the maximum of the amount of attention allocated to either dimensions.

The psychological similarity measures $D_j(x,r)$ cause some activations in internal "hidden" units or reference points (i.e., exemplars or prototypes). The activation of "hidden" basis unit $j$, or $h_j$, is obtained by any differentiable nonlinear activation transfer function (ATF), or

$$h_j = G(D_j(x,r)) \tag{2}$$

given that its first derivative $G'(\cdot)$ exists. An exponential function, $\exp(-cD_j(x,r))$, is an example of an ATF. The ATF must be a differentiable function, because GECLE uses a gradient method for learning, where the partial derivatives are used for updating the learnable parameters. However, it is possible to eliminate this restriction by incorporating a form of derivative-free learning algorithm such as stochastic learning methods [12, 13].

The activations of hidden units are then fed forward to output nodes. The activation of the $k$th output node, $O_k$, is calculated by summing the weighted activations of all hidden units connected to the output node, or

$$O_k = \sum_{j=1}^{J} w_{kj} h_j \tag{3}$$

where $w_{kj}$ is the association weight between output node $k$ and reference point $j$. The probability that a particular stimulus is classified as category $C_k$, denoted as $P(C)$, is assumed equal to the activity of category $k$ relative to the summed activations of all categories, where the activations are first transformed by the exponential function [1],

$$P(C) = \frac{\exp(\phi O_c)}{\sum_k \exp(\phi O_k)} \tag{4}$$

where $\phi$ is a real-value mapping constant that controls the "decisiveness" of classification responses.

GECLE uses the gradient method to update its learnable parameters. The error function is defined as the sum of squared differences between targeted and predicted output values (i.e., $L_2$ norm), or

$$E(w,r,\Sigma^{-1}) = \frac{1}{2}\sum_{k=1}^{K} e_k^2 = \frac{1}{2}\sum_{k=1}^{K} (d_k - O_k)^2 \tag{5}$$

Then the following functions are used to update parameters.

$$\Delta w_{jk} = -\eta^w \frac{\partial E}{\partial w_{jk}} = \eta^w e_k h_j \tag{6}$$

where $\eta^w$ is the learning rate for the association weights.

$$\Delta r_j = -\eta^r \frac{\partial E}{\partial r_j} = \eta^r \sum_{k=1}^{K} e_k w_{jk} G'(D_j(x,r)) \Sigma_j^{-1} (x - r_j) \tag{7}$$

where $G'(\cdot)$ is a derivative of $G(\cdot)$. Equation (7) can be considered as a function that locates or defines prototypes of stimuli. For the exemplar-based modeling $\eta^r$ must be set to zero to maintain the static nature of reference points.

$$\begin{aligned} \Delta\Sigma_j^{-1} &= -\eta^{\Sigma} \frac{\partial E}{\partial \Sigma_j^{-1}} \\ &= -\eta^{\Sigma} \sum_{k=1}^{K} e_k w_{jk} G'(D_j(x,r))(x - r_j)(x - r_j)^T \end{aligned} \tag{8}$$

For models with global attention coverage structure, (8) should be summed over both $k$ and $j$.

### *C. Attention Mechanisms of GECLE*

As mentioned in the previous section, GECLE has very flexible attention mechanisms. The flexibility is achieved by allowing manipulations of the (a) activation transfer function (ATF) and (b) constraints on $\Sigma^{-1}$.

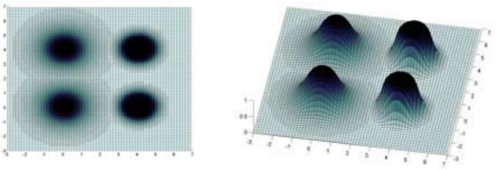


Figure 1. 2-dimensional (left panel) and 3-dimensional (right panel) figures comparing two different ATFs. ATFs for the left and right column are $(D_j^2+1)^{-1}$ and $(D_j^4+1)^{-1}$, respectively.

*1) Varieties of Activation Transfer Function*

The ATF in the GECLE can be any function as long as it is differentiable. This allows one to manipulate and compare the effects of specific characteristics of the population attention structure (e.g. fatter tail vs. thinner tail). This capability was motivated by the fact that the population attention structure can determine the effectiveness of model predictions. For example, Hanson and Gluck [14] compared RBFs with Gaussian and Cauchy activation functions, and showed that increased competition by the Cauchy's fatter tails resulted in better fit to the empirical data. Since there is not enough evidence indicating the "true" or best activation transfer function, and to enhance the flexibility of GECLE, ATF is deliberately made user-definable. Fig. 1 shows examples of activation of two different ATFs, namely $(D_j^2 + 1)^{-1}$ and $(D_j^4 + 1)^{-1}$. Note that differences in characteristics of attention coverage areas are solely created by the differences in the ATFs.

*2) Hierarchy of Constraints on Attention Parameters*

There is a hierarchy of constraints that one can impose on the attention parameters $\Sigma^{-1}$ to manipulate GECLE's attention mechanisms. There are two levels of uniqueness of $\Sigma^{-1}$ (global and local attention coverage structure), in each of which there are three levels of constraints on entries in $\Sigma^{-1}$. The following is a list of six possible levels of restriction, and Fig. 2 shows examples of the corresponding attention coverage structures. Note that regardless of the types of restriction, the entries ($s_{im}$) in $\Sigma_j$ are assumed to satisfy the following conditions: $s_{ii} \geq 0$ & $|s_{im}| \leq$ MAX($s_{ii}$, $s_{mm}$).

*a) Global Attention Coverage Structures*

**A**. Global Pure Radial (GPR): Constraints on $\Sigma_j$: $s_{ii} = s$, for all $i$: $s_{im} = 0$, for all $i \neq m$; $\Sigma_j = \Sigma$, for all reference points $j$.

**B**. Global Uncorrelated Non-radial (GUN): Constraints on $\Sigma_j$: $s_{im} = 0$, for all $i \neq m$; $\Sigma_j = \Sigma$, for all reference points $j$.

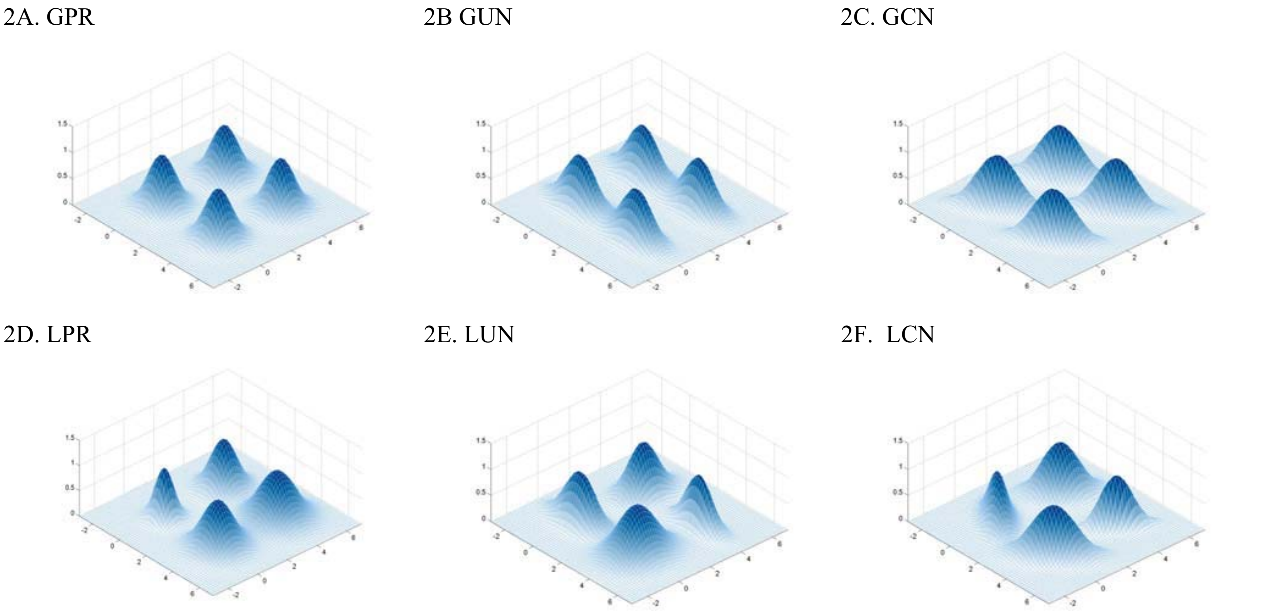
**C**. Global Correlated Non-radial (GCN): Constraints on $\Sigma_j$: $\Sigma_j = \Sigma$, for all reference points $j$.

*b) Local Attention Coverage Structures*

**D**. Local Pure Radial (LPR): Constraints on $\Sigma_j$: $s_{ii} = s$, for all $i$; $s_{im} = 0$, for all $i \neq m$.

**E**. Local Uncorrelated Non-radial (LUN): Constraints on $\Sigma_j$: $sim = 0$, for all $i \neq m$.

**F**. Local Correlated Non-radial (LCN): Constraints on $\Sigma_j$: none.



2A. GPR

2B GUN

2C. GCN

2D. LPR

2E. LUN

2F. LCN

Figure 2. Six types of attention structures in the GECEL framework. 2A: GRP - global attention structure with pure radial coverage. 2B: GUN - global, uncorrelated (orthogonal) attention structure. 2C: GCN - global correlated; 2D: LPR - local, pure radial; 2E:LUN - local, uncorrelated; 2F: LCN - local, correlated.

TABLE I: THE NUMBER OF LEARNABLE PARAMETERS FOR SEVERAL GECLE CONFIGURATIONS

| Model | Attention | RP | Weights | Total Learnable |
|---|---|---|---|---|
| GPR | 1 | J * I | J * K | 1 + J(I + K) |
| GUN | I | J * I | J * K | I + J(I + K) |
| GCN | I(I+1)/2 | J * I | J * K | I(I+1)/2+ J(I + K) |
| LPR | J | J * I | J * K | J(1 + I +K) |
| LUN | J * I | J * I | J * K | J(2I + K) |
| LCN | J * I(I+1)/2 | J * I | J * K | J{(I$^2$+3I )/2+ K} |

I: number of input feature dimensions
J: number of reference points
K: number of output categories

## III. MODEL COMPARISON METHOD

In general, as the number of parameters increases a model will have better fit for a given data set. Because exploratory modeling approaches often need to compare models with different numbers of parameters, it seems useful to have a model comparison method that is sensitive to model complexity (i.e., number of parameters). However, there are several issues that must be considered in comparing models of human category learning. In this section, the number of parameters involved in GECLE is defined first. Then, issues in model comparison are discussed, followed by a discussion on possible model comparison methods.

### *A. Number of parameters*

The total number of parameters in the GECLE framework is determined by (1) the number of user-defined parameter, which is always equal to four (i.e., $\phi$, $\eta^w$, $\eta^r$, and $\eta^\Sigma$), (2) the number of parameters for the ATF, and (3) the number of learnable parameters in the network. The number of learnable parameters can be further decomposed into: the number of parameters for (a) association weights, (b) attention strengths and orientations, and (c) locations of reference points. Table I summarizes the number of learnable parameters in various applications.

For exemplar-based modeling, the location parameters, $r_{ji}$ are static and not subject to error-minimization learning, but it is assumed that optimized locations are initially learned when the exemplars are created in memory. Thus, they are counted as learnable parameters. Similarly, although the learning rate ($\eta^r$) for the locations of exemplars in an exemplar-based model logically must be zero, it is counted as a valid parameter for the same reason.

### *B. Issues in Model Comparisons*

In the cognitive science paradigm, the computational classifier models, such as GECLE, are usually trained to categorize input stimuli used in empirical studies in human category learning. However, the models are NOT evaluated by how successfully they learn to categorize the input stimuli. Rather the models are evaluated by how similarly they behave as compared to humans in a given categorization task. Thus, in such descriptive-oriented model evaluations (i.e., how similar a model performs as compared to the empirical findings), researchers usually optimize the user-defined and ATF parameters to reproduce observed human learning curves, and let models learn to categorize stimuli. In contrast, in many standard NN applications, model comparison and or selection is based on error or risk functions defined in terms of optimal performance on the classification task itself. Therefore, many conventional model comparison or selection methods used in the standard NN applications are not directly applicable for our purpose. In addition, in the descriptive-oriented evaluations, the aspects of the simulated model training are usually highly constrained, because they must follow the procedure of the empirical studies. This further restricts the use of conventional NN model selection methods

In addition, simulations of different empirical studies may require different criteria, mainly because each study tends to have its own unique research question and thus measuring different (dependent) variables. For example, some empirical studies [15] measured only the classification response profile, while other measured learning curves of classification accuracies and attention distributions [16, 17]. The uniqueness of criterion measures across different empirical studies makes it hard to propose a standardized comparison method. In addition, if there are multiple criteria (e.g. fitting both accuracy and attention learning curves), then subjective judgment may be involved, as we must somehow weigh different criteria. On the other hand, if there is only a single fit criterion, then it has been argued that the "optimal" values for some parameters may not be easily identifiable in some cases [9].

### *C. Model Comparison*

The following function adjusting empirical risk for model complexity has been used as a fit index for model selection in the mainstream machine learning literature:

$$R \cong g(m,n) \cdot R_{emp} \tag{10}$$

where $g$ is a monotonically increasing function of model complexity (ratio of the degrees of freedom, $m$, and sample size, $n$), and $R_{emp}$ is the empirical risk [18, 19]. Final prediction error and generalized cross-validation are examples of the function $g$, and the squared error can be used as an estimate of $R_{emp}$ [19].

This model selection approach may be applied to models of human cognition with some modifications. Since in some simulation studies of human category learning there are multiple curves to be fitted (e.g. learning & attention), (10) may be modified as follows:

$$SWQR \cong \sum_{c=1}^{C} \gamma_c \cdot g(m,n_c) \cdot \Xi_c \tag{11}$$

where subscript $c$ indicates different sub fit-criteria (e.g. fit for a particular learning curve), $\gamma_c$ is a coefficient weighting each sub fit-criterion, $n_c$ is the number of data points evaluated for the *c*th criterion, and $\Xi_c$ is quasi empirical risk for the *c*th sub-fit criterion. In virtually all simulation studies in high-order human cognition, the prediction errors (i.e., risks) that are not directly minimized by the cognitive models are evaluated and compared (see section III*B*). Consequently such "indirect" risks are purposely denoted as quasi-empirical risks here to distinguish from the ordinary "direct" empirical risks. This relative fit index will be referred to as the Sum of Weighted Quasi Risks.

Here, one needs to be careful selecting the estimates of $\Xi_c$. This is because, for example, Matsuka [16] suggested that SSE could be a misleading fit index for predicted dimensional attention allocation for that particular simulation studies. In that simulation studies, there were multiple fit criteria comparing observed data against models' predicted classification accuracies and attention allocations to multiple feature dimensions. When SSE was used, a qualitatively worse fitting model with invariant dimensional attention allocation resulted in better quantitative fits as compared to a model with qualitatively better predictions. In that study, a squared correlation coefficient appeared to be a more sensible measure of fit for attention allocation.

## IV. Applications of GECLE

The flexibility and exploratory nature of the GECLE framework can make it a constructive tool that could lead to better understanding the nature of human category learning. Specifically, it enables us to conduct systematic standardized exploratory studies, comparing various types of human cognitive processes believed to be associated with category learning. Three example applications of GECLE are briefly discussed in this section.

### *A. Comparing Internal Representation assumptions.*

There has been an increasing number of studies investigating and debating how stimuli are internally represented in human cognition during the last several years (e.g., [20, 21]). Most of these debates have been based on quantitative models of categorization (i.e., models without the learning capability), and only a few have considered representational aspects of adaptive or network models of category learning. One limitation of the models of categorization is that, as Shanks [22] pointed out, its model fitting process is post hoc and thus does not generate predictions on learning processes.

Several studies [9, 16] have compared exemplar-based (EB) and prototype-based (PB) adaptive network models of category learning, but there have been no systematic comparisons of specific assumptions in EB and PB modeling. For example, the EB and PB models compared in those studies assume different attention processes and utilize reference points differently for categorizing and learning. Thus, differences in the accuracy of reproducing learning curves may not be attributed solely to the plausibility of the EB versus PB representations, but possibly to multiple factors including the models' attention mechanisms. With GECLE, one can systematically compare the plausibility of the EB and PB representations by holding the attention mechanisms constant for both models (i.e., using the same activation transfer function and the same constraints on $\Sigma^{-1}$).

### *B. Comparing Selective Attention Mechanisms.*

Selective attention processes have been suggested to play a very important role in human category learning (e.g. [23, 24]). However, only limited numbers of selective attention mechanisms have been modeled and tested. For example, virtually all recent network models of categorization assume dimensional attention processes (i.e., no attention to correlations) and global attention coverage structure (i.e., all reference points have exactly the same shape of receptive field, with independent attention allocation to dimensions).

Again, a general framework like GECLE allows systematic manipulation of models' attention mechanism, and or exploration of various types of attention mechanism that has not be tested, such as distribution of attention to correlated features dimensions. In addition, comparisons of different activation transfer functions $G(\cdot)$ can be informative for understanding how human categorize stimuli.

### *C. Investigating Interactions Between Internal representation & Attention Mechanisms.*

As a final point, it may be possible that a model with a particular internal representation system (e.g., exemplar-based) performs better with a particular attention mechanism, which does not work as well for models with other representation system (e.g., prototypes). In other words, the effectiveness of internal representation system and attention mechanism may interact with each other in the sense that the effectiveness of model's internal representation systems may depend on its attention mechanism, or the effectiveness of the models' attention mechanism may depend on its internal representation system.

I am not aware of any single study that systematically and simultaneously manipulates models' internal representation system and attention mechanism to investigate possible interaction effects between them. GECLE provides a way to systematically manipulate, by a factorial design, both internal representation system and attention mechanism to tackle this issue of interactivity of internal representation and attention mechanisms.

## V. Simulations

In this section, two simulation studies are conducted as examples showing how simulation studies with a data driven approach, such as GECLE, can provide information that may lead to better understandings of human category learning and/or some insightful alternative interpretations of previous empirical results. In particular, two simulation studies, investigating possible interactive effects of the models' internal representation and selective attention mechanisms are reported.

### A. Simulation 1: XOR problem

In Simulation 1, a simple XOR learning task is simulated with GECLE. An XOR stimulus set is one of the simplest stimuli structures with which one can expect interactions between representation system and attention mechanism. That is, a prototype-based GECLE may need to have local attention structure and the capability to pay attention to feature correlations in order to learn to categorize the XOR stimulus set with only two reference points. Whereas, an exemplar-based GECLE seems capable of categorizing the stimuli with a simpler attention mechanisms by utilizing all four unique exemplars in its memory.

#### 1) Methods

There are eight different models involved in this simulation, namely, E1: an exemplar-based (EB) model with GUN attention mechanism; E2: EB with GCN; E3: EB with LUN; E4: EB with LCN; P1: a prototype-based (PB) model with GUN; P2: PB with GCN; P3: PB with LUN; and P4: PB with LCN. All EB models had four reference points, while all PB models had two. For all eight models, the following one-parameter exponential activation transfer function was used:

$$h_j = \exp\left(-c \cdot D_j(x,r)\right) \qquad (13)$$

Note that E1 is essentially ALCOVE (Kruschke, 1992).

TABLE II: RESULTS OF PROTOTYPE-BASED GECLE

| Model | P1 | P2 | P3 | P4 |
|---|---|---|---|---|
| Attention structure | GUN | GCN | LUN | LCN |
| No. prototypes | 2 | 2 | 2 | 2 |
| No. Learnable parameters | 10 | 11 | 12 | 14 |
| SSE | 1.328 | 1.147 | 1.322 | $\varepsilon^{\dagger}$ |

$\varepsilon^{\dagger} < 10e\text{-}20$.

TABLE III: RESULTS OF EXEMPLAR-BASED GECLE

| Model | E1 | E2 | E3 | E4 |
|---|---|---|---|---|
| Attention structure | GUN | GCN | LUN | LCN |
| No. Exemplars | 4 | 4 | 4 | 4 |
| No. Learnable parameters | 18* | 19* | 24* | 28* |
| SSE | $\varepsilon^{\dagger}$ | $\varepsilon^{\dagger}$ | $\varepsilon^{\dagger}$ | $\varepsilon^{\dagger}$ |

$\varepsilon^{\dagger} < 10e\text{-}20$.

* Location parameters for exemplars were static & not subject to error-minimization learning, but it is assumed that optimized locations are learned when the exemplars are created.

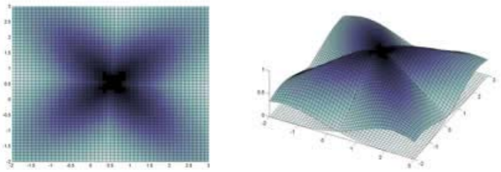


Figure 3: Two-dimensional and 3-dimensional plots for the activation areas & strengths of two reference points of P4 (correlation attentive prototype-based GECLE with the local attention structure).

#### 2) Results

Tables II and III show the results of Simulation 1. All exemplar-based models were able to learn to categorize XOR stimuli by utilizing four exemplars (i.e., all unique stimulus configurations) in their "memory", and thus complex attention mechanisms were shown to be ineffective or unnecessary for EB modeling. In fact, when the fit measure was adjusted for model complexity (i.e., number of parameters) the EB model with the simplest attention mechanism, i.e., E1, resulted in the best (relative) fit among the exemplar-based GECLE. In contrast, for prototype-based modeling, only LCN (i.e., P4) was able to learn to categorize the XOR stimulus set, suggesting that the complex attention mechanism plays an important role for the PB modeling (Fig. 3 shows activation areas and strengths produced by P4). Note that P4 also resulted in the best (relative) fit after controlling for its complexity among all eight models compared in the present simulation.

In sum, the results of the present simulation suggest that it is very likely that effectiveness of the model's attention mechanism depends on how the stimuli are internally represented by the model or vice versa; a simple GUN attention mechanism seems sufficient for EB modeling, while a complex LCN is required for PB modeling.

### B. Simulation 2: Filtration vs. Condensation

Kruschke [6] claimed that selective *dimensional* attention processes (i.e., paying attention to each dimension independently) is one of three key principles for models of category learning. His claim was based partly on the evidence that humans learn much better in "filtration" tasks, in which information from only one dimension is required for perfect categorization, than in "condensation" tasks, in which information from two dimensions is required [6, 25]. Thus, a model paying attention to correlations or having diagonal attention coverage may not be able to show the filtration advantage, because the process of allocating attention to correlation, thus rotating stimulus space, would make the model to perceive both filtration and condensation stimulus structures quite similarly, resulting in invariant classification performance. This implies that any model with a GCN or LCN attention mechanism may not be able to replicate such an advantage. If the claim is valid, then this is evidence against P4, namely the prototype-based model with diagonal localized attention coverage, as a descriptive model of human cognition, invalidating the results of Simulation 1. The main objective of the present simulation study is to test if PB model with LCN attention mechanism can replicate the filtration advantage observed in human category learning.

### 1) *Method*

In Simulation 2, I revisited Kruschke's claim regarding dimensional attention processes by simulating category learning on both filtration and condensation stimuli using the prototype-based model with LCN (and EB-LCN for a illustrative comparison). The stimulus set presented in Kruschke [6] is used in this simulation. Table IV shows the schematic representation of the stimulus set. For the filtration stimulus set, information from only Dimension 1 is required for a perfect categorization (category = A, if D1 < 2; category = B, otherwise) while information on both Dimensions 1 and 2 were required for the condensation set. The same one-parameter exponential ATF used in Simulation 1 is used in the present simulation study. The user-defined parameters were optimized using a simulated annealing method [9, 26] to reproduce observed empirical learning curves shown in Kruschke [6]. Exactly the same model configurations (i.e., user-defined parameter values) were used in simulations of the filtration and condensation tasks. It should be noted that Kruschke [6] showed that ALCOVE (i.e., an EB-GECLE with GUN) was able to reproduce the filtration advantage successfully.

### 2) *Results*

Fig. 4 shows the results of Simulation 2. The prototype-based LCN model was able to show the filtration advantage even when it paid attention to the correlation between the two input dimensions. In contrast, the exemplar-based LCN model showed no filtration advantage.

One possible reason why PB-LCN showed the filtration advantages was that PB-LCN might have been able to locate or define prototypes coverage areas more easily in the filtration task than in the condensation task. This is because while the condensation stimuli require tighter correspondence or synchronization of the "correct" movements of centroids of prototypes and the "correct" psychological scaling (i.e., attention processes) of the two feature dimensions, the filtration stimuli require "correct" movements and scaling in only one dimension. Thus, synchronization of prototype-movement and scaling was more difficult for the condensation stimuli than in the filtration task for models using prototype-like internal representation. In other words, category learning by any prototype-based network models is strongly affected by how successfully or how fast the models can find "proper" prototypes and how well psychological scaling of dimensions is synchronized with it. For correlation attentive exemplar-based models, this would not be an issue as it had exemplars in the correct locations from the beginning, resulting in no filtration advantage (Fig. 4, right panel). These results indicate that correlated (i.e., diagonally-oriented) attention coverage may be more often a required assumption for PB modeling, compared to EB modeling.

TABLE IV: SCHEMATIC REPRESENTATION OF THE STIMULUS SETS USED IN SIMULATION 2

| Stimulus Feature | | Category Membership | |
|---|---|---|---|
| Dim 1 | Dim 2 | Filtration | Condensation |
| 0 | 1 | A | A |
| 0 | 2 | A | A |
| 1 | 0 | A | A |
| 1 | 3 | A | B |
| 2 | 0 | B | A |
| 2 | 3 | B | B |
| 3 | 1 | B | B |
| 3 | 2 | B | B |

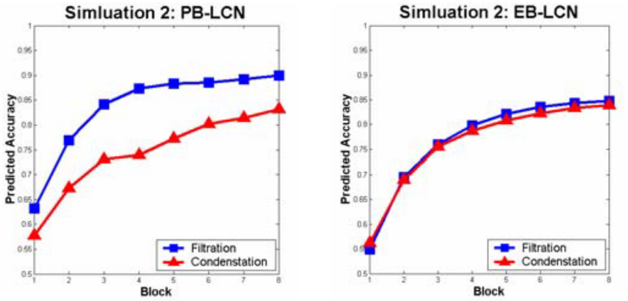


Figure 4: The results of Simulation 2. Left panel: predicted learning curves by correlation attentive prototype-based GECLE with the local attention structure. Right panel: predicted learning curves by an exemplar-based GECLE with the same attention mechanism (i.e., LCN).

As in Simulation 1, the results of the present simulation suggest that it is highly possible that stimuli's internal representation and selective attention mechanisms interact with each other. Furthermore, the present simulation studies suggest that human may allocate attention not only to individual dimensions, but also to correlations among dimensions. At the very least, the evidence of a filtration advantage, observed in human subjects, alone does not rule out the possibility that humans pay attention to correlations among feature dimensions.

## VI. DISCUSSION & CONCLUSION

### A. *Individual Differences.*

The results of some simulation studies [12, 13, 17] suggest that NN models of categorization with gradient learning methods are successful in reproducing group learning curves, but tend to underpredict variability in individual-level data in some cases. Since GECLE utilizes a gradient method for learning, it may also underpredict individual differences in some cases. However, this exploratory model is introduced to compare how well models with different representational and processing assumptions can replicate general tendencies in human category learning. Many of these general tendencies may be best described in terms of such aggregated data. Nonetheless, to account for individual differences, GECLE could be easily modified to incorporate Matsuka & Corter's [12, 13] stochastic learning algorithm for attention processes, which is shown to be more successful in reproducing individual differences in category learning.

### B. Conclusions

One of the most critical problems in evaluating recent computational models of categorization is that there is no standardized method for comparing the models' assumptions systematically. Thus, previous studies involving model comparisons have sometimes been unable to answer which element, assumption, or structure of each model was responsible for successful or unsuccessful replication of observed tendencies in human category learning. In the present study, a flexible general model is introduced, that can be used as a framework to systematically compare a limited number of assumptions at a time.

Two simulation studies are described to show how the GECLE framework can be useful in exploring issues in the field of categorization research. The results of Simulation 1 showed that a pure prototype-based category learning model (i.e., the number of prototypes is equally to that of categories) was capable of learning an XOR problem only if it incorporated a very complex attention mechanism, while the exemplar-based model was capable of learning the stimuli with a simple attention mechanism. In Simulation 2, the filtration advantage, which has been used as an argument or evidence for dimensional attention processes (i.e., paying attention to dimension independently with no attention to correlations among feature dimensions), was successfully replicated by the prototype model with a complex attention mechanism capable of paying attention to correlation. This casts some doubt on the claim that the filtration advantage shows that people pay attention to dimensions independently, without attending to correlations among dimensions.

The results of these simulations may be argued to provide new insights regarding human category learning, namely that 1) it is very likely that there are interactions between internal mental representation and attention mechanisms, and 2) people may pay attention to correlations among feature dimensions.

## Acknowledgment

T. Matsuka thanks James E. Corter, Stephen Jose Hanson, Catherine Hanson, Yasuaki Sakamoto, Areti Chouchourelou, and researchers at RUMBA for their helpful comments and suggestions.

## References

[1] Kruschke, J. E. (1992). ALCOVE: An exemplar-based connectionist model of category learning, *Psychological Review, 99.* 22-44.

[2] Kruschke, J.K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083-1119.

[3] Love, B.C. & Medin, D.L. (1998). SUSTAIN: A model of human category learning. *Proceeding of the Fifteenth National Conference on AI (AAAI-98),* 671-676.

[4] Poggio, T. & Girosi, F. (1989) A Theory of Networks for Approximation and Learning). *AI Memo 1140/CBIP Paper 31*, Massachusetts Institute of Technology, Cambridge, MA.

[5] Poggio, T. & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247,* 978-982.

[6] Kruschke, J. E. (1993). Three principles for models of category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by human and machines: The psychology of learning and motivation* (Vol. 29, pp. 57-90). San Diego, CA: Academic Press.

[7] Matsuka, T. & Corter, J. E. (2003). Neural network modeling of category learning using Generalized Radial Basis Functions. Paper presented at 36th Annual Meeting of the Society of Mathematical Psychology. Ogden, UT.

[8] Rosseel, Y. (1996). Connectionist models of categorization: A statistical interpretation. *Psychologica Belgica, 36,* 93-112

[9] Matsuka, T., Corter, J. E. & Markman, A. B. (2003). Allocation of attention in neural network models of categorization. Under review.

[10] Nosofsky, R.M., Gluck, M.A., Palmeri, T.J., McKinley, S.C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition, 22,* 352-369.

[11] Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation (2nd ed.).* Upper Saddle River, NJ: Prentice Hall.

[12] Matsuka, T & Corter, J. E. (2003). Stochastic learning in neural network models of category learning. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Boston, MA

[13] Matsuka, T. & Corter, J.E (2004). Stochastic learning algorithm for modeling human category learning. *International Journal of Computational Intelligence.* Accepted for publication.

[14] Hanson, S. J., & Gluck, M. A. (1991). Spherical units as dynamic consequential regions: Implications for attention and cue-competition in categorization. *Advances in Neural Information Processing Systems #3.* San Mateo, CA: Morgan Kaufman, 656-665.

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

[16] Matsuka, T (2002). Attention processes in computational models of categorization. Unpublished Doctoral Dissertation. Columbia University, NY.

[17] Matsuka, T. & Corter, J. E. (2003). Empirical studies on attention processes in category learning. Poster presented at 44th Annual Meeting of the Psychonomic Society. Vancouver, BC, Canada.

[18] Hardle, W., Hall, P., & Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of American Statistical Association, 83*, 86-05.

[19] Cherkassky, V. & Mulier, F. (1997). *Learning from data: Concepts, Theory, and Methods.* New York: Wiley

[20] Minda, J.P. & Smith, J.D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 275-292.

[21] Nosofsky, R.M. & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 924-940.

[22] Shanks, D.R. (1991). Categorization by a connectionist Network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433-443.

[23] Lassaline, M.E. (1990). *The basic level in hierarchical classification.* Unpublished master's thesis. University of Illinois, Champaign.

[24] Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classification. *Psychological Monograph, 75*(13).

[25] Gottwald, R. L. & Garner, W. R. (1975). Filtering and condensation tasks with integral and separable dimensions. *Perception & Psychophysics, 2*, 50-55.

[26] Ingber, L. (1989). Very fast simulated annealing. *Journal of Mathematical Modelling, 12:* 967-973.