Modeling Category Identification Using Sparse Instance Representation

Shunan Zhang
Michael D. Lee
{szhang, mdlee} @uci.edu
Department of Cognitive Sciences
University of California, Irvine

Meng Yu
Jack Xin
{myu3, jack.xin} @uci.edu
Department of Mathematics
University of California, Irvine

Abstract

We study the problem of category identification, which involves making inferences about category membership (e.g., a 'cat') given a set of features (e.g., has a tail, has four legs). We note that this problem is closely related to classification problems in machine learning, for which standard methods exist, and new methods continue to be developed. Using a large database of associations of features to animal stimuli, made by different people, we test several standard benchmark methods, including nearest neighbor, decision tree, and logistic regression methods. We also apply a new classifier, developed for image processing, which we call Sparse Instance Representation. We show that it is the best-performed, especially when constrained in a novel psychologically interpretable way. We conclude that our results argue for sparse exemplar-based representations of category structures.

Keywords: category identification, sparse representation, machine learning, category learning, exemplar representation

Category Identification

Suppose your friend tells you they are thinking of a particular animal, asks you what type it is, and starts listing its features: it has a tail, has four legs, lives inside, and so on. You are now facing a *category identification* problem, which requires you to infer the category of a given instance of that category (Kemp, Chang, & Lombardi, 2010). In other words, when your friend describes the features, you are being asked to identify a cat.

Category identification is clearly closely related to two other cognitive capabilities. One of these is *identification*, which is the problem of inferring which of a set of instances is being presented, such as recognizing Jack the cat among a group of individual cats. Identification has been widely studied in various subfields of the cognitive sciences, including psychology (Nosofsky, 1986), machine learning and statistics (Bunge & Fitzpatrick, 1993), and linguistics and philosophy (Michie, Spiegelhalter, & Taylor, 1994). The other related cognitive ability is *categorization*, which is problem of inferring the category membership of presented stimuli, such as deciding whether something is a cat or a dog. This has also been widely studied, especially in cognitive psychology (e.g., Nosofsky, 1986; Kruschke, 1992; Love, Medin, & Gureckis, 2004)

The difference between identification and category identification is that the former is about instances, and not their categorical structure. The difference between category identification and categorization is much more subtle. The key issue, as described by Kemp et al. (2010, p. 230) relates to the sorts of features used to present the stimulus. In categorization, features are presented at the level of instances (e.g., Jack the cat has a tail). In category identification, features are presented at the level of categories (e.g., has a tail). Intuitively, categorization is about deciding whether a stimulus belongs to a family, whereas category identification is about which family of stimuli as a whole is being described.

It is clear that category identification is an important capability, because it allows us to think about stimuli described by features in terms of their category membership. Nestled between identification and categorization, category identification blends psychologically interesting aspects of both. It maintains the focus on differences and individual instances inherent in identification, while incorporating the focus on sameness and coherence of conceptual structure inherent in categorization. In particular, category identification offers an interesting window onto the structure of mental representations, since it involves the relationship between categories and features, and so requires the representation of both what makes stimuli different, and what makes them the same.

In this paper, we use existing data relating stimuli to features that can be interpreted in terms of category identification. We explore a number of models of these data based on classification approaches from machine learning. We consider a benchmark set of standard classifiers, as well as a new method developed in the image processing literature, which we call Sparse Instance Representation, that makes interesting, and psychologically interpretable, representational assumptions. We show that one variant of Sparse Instance Representation is the best-performed model of the data, and, based on the results, we draw some conclusions about the way people may represent categories.

Data

Our data come from the Leuven Natural Concept Database (DeDeyne et al., 2008). As summarized by Storms, Navarro, and Lee (2010), this database involves more than 400 stimulus words, distributed over 16 se-

	P1	P2	P3	P4
is mammal	1	1	1	1
can fly	0	0	0	0
is small	1	1	1	1
has small ears	0	0	1	0
has pointy ears	0	1	1	1
has large eyes	0	0	0	1
has round eyes	1	1	0	0
is hairy	1	1	0	1
is friendly	0	0	1	1
live alone	0	1	1	0

Table 1: Examples of feature applicability judgments.

Figure 1: Feature-judgment discrepancy for each participants, on each animal, grouped by animal family.

mantic categories: two food categories (fruits and vegetables), two activity categories (professions and sports), six animal categories (amphibians, birds, fish, insects, mammals, and reptiles), and six artifact categories (musical instruments, tools, vehicles, clothing, kitchen utensils, and weapons). For every stimulus word, the database contains data for a large number of variables, including typicality ratings, goodness ratings, goodness rank order, exemplar generation frequencies, exemplar associative strength, category associative strength, estimated age of acquisition, word frequency, familiarity ratings, imageability ratings, and pairwise similarity ratings.

Exemplar by Feature Data

In addition, the database incorporates a large feature generation study, in which 1003 student participants (about half participating for course credit, and half paid the equivalent of $10 per hour) wrote down around 10 features for 6–10 stimuli. Features were generated for each of the stimulus words by at least 20 participants. After tallying generation frequencies, all features that were generated at least four times were selected. These features were rated for their importance in defining the different categories to which the corresponding stimulus words belonged.

Most importantly for our modeling, the stimuli and features were combined in a feature verification task, in which four participants (two students, two adults with university degrees, paid the equivalent of $10 per hour, and not including any of the authors of this paper) judged whether or not each of the features belonged to each of the stimuli. This resulted in two large exemplar by feature applicability matrices for each participant, one for the animal domain, with 129 animal stimulus words and 765 animal features, and the other for the artifact domain, with 166 artifact stimulus words and 1295 artifact features. The subset of the database we use involve the four exemplar by feature matrices for the animal domain. Originally, DeDeyne et al. (2008) categorized the animal into six families (mammals, birds, fish, insects, reptiles, amphibians), but also pointed out that people found it
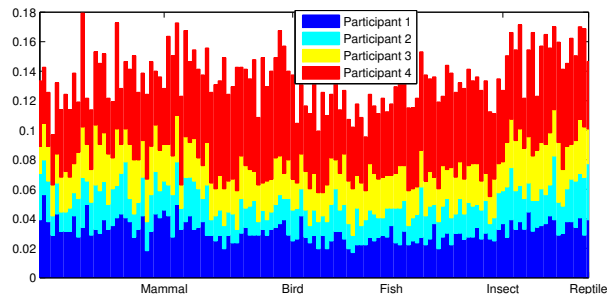
difficult to distinguish reptiles and amphibians. We combined the amphibian family and the reptile family, so that we had 30 animals from the mammal family, 30 for birds, 23 for fish, 26 for insects and 20 for amphibians and reptiles.

Interpretation as Category Identification Data

In this paper, we treat the exemplar by feature judgments of each of the four participants as providing data relevant to category identification. This is certainly not the only way these data could be interpreted, but we think it is a reasonable interpretation. The participant is being asked to decide whether or not each of a large list of features applies to a word describing a category. Thus, we can use the set of features a participant assigned to a word—"can fly", "is small", and so on—as the input to a category identification problem, where the task is to identify the category associated with that list of features.

A challenge for the four participants in doing this task is that many of the category-level features in the Leuven set do not have clear answers. This intuition is made more concrete by the example in Table 1, which shows the feature judgments for the stimulus word "cat" made by all four participants for some selected features. The first two features—"can fly" and "is small"—are good examples of the features where participants give unanimous assessments. Cats as a category cannot fly, and are small.

But the remaining features in Table 1 focus on the features where there are reasonable differences in the assessments of the participants. Whether a cat "is short haired", "is friendly", "lives alone", and so on, is less clear. The differences in individual assessments highlighted in Table 1 are evident throughout the data. Figure 1 shows, for each participant and each animal, the proportion of features for which that participant was different from all others. It is clear that these differences occur for all animals and all participants.

Machine Learning Classifiers

The psychological problem of category identification bears a close formal relationship with supervised classification methods developed in machine learning. In gen-

eral, supervised classifiers are algorithmic procedures for assigning a new case into one of a set of pre-defined classes, on the basis of observed attributes (Duda, Hart, & Stork, 2000). Typically, the input to a classifier is a vector of features, and the output is the indicator of the assigned class, or the probability of assignment to each class.

Thus, it is straightforward to map classifiers to the psychological task of category identification. The problem is one of taking a set of features—has a tail, has four legs, lives inside, and so on—and mapping them to a category like 'cat'. This is exactly the psychological problem of category identification. It is also straightforward to use the Leuven dataset to evaluate different classifiers with a standard machine learning methodology. Specifically, we split the exemplar by feature matrices for the four participants into training and test sets in a four-fold, leave-one-out, cross-validation. This means, in each validation, we train the classifier using data from three participants, so that it can learn which features are associated with each animal, and then test on the data from the participant left out, so that it has to classify a presented set of features as one of the animals.

We believe this link of theory and methods could constitute one useful starting point for understanding how people do category induction. It provides set of sophisticated methods for doing the task, and a natural way of evaluating them as benchmarks. It is also possible to identify psychological assumptions implicit in many of these methods—such as the nature of the representational assumptions they make—to help guide psychological theorizing and model building.

We used versions of three standard machine learning methods—nearest neighbor, decision trees, and multinomial regression—which we describe only briefly, since they are well documented in the literature (e.g., Bishop, 2006).

Benchmark Methods

1-nearest-neighbor (1NN)   1-nearest-neighbor (1NN) assigns the test sample to the same class as its closest training sample in the feature space. We implemented two versions of 1NN. In the first version, we combined the three training matrices to be one with dimension of 764 features by 387 animal instances, with 3 instances for each animal that come from different training matrices. In the second version, we found the nearest neighbor of the test sample in each of the training matrices separately, and took the majority vote for classification. In both versions, distance between two feature vectors are calculated using the l_1 norm.

Decision tree (DT)   Decision tree methods classify the test sample based on a learned tree-structure model. Each interior node corresponds to one of the features, branching to different paths based on the value of the current feature. Each leaf represents a decision class given the values of the input features represented by the corresponding path. The algorithm works top-down by choos-

ing a feature at each step, trying to split the data into subsets that belong to the same class. Different criteria for splitting are available, and we used information gain in our implementation (Bishop, 2006).

Sparse multinomial logistic regression (SMLR) Multinomial logistic regression is a multi-class generalization of standard binary logistic regression. It generates the logistic distribution of multiple classes using a linear combination of per-class weights on each of the dimensions of the input. We employed a multinomial logistic regression method that enforces sparsity using a l_1 regularization (Krishnapuram, Figueiredo, Carin, & Hartemink, 2005), and the implementation was done using the Princeton Multi-Voxel Pattern Analysis Toolbox (MVPA)[1]. The output is a set of probabilities of membership of 129 classes. The test sample is then assigned to the class with the highest probability.

Sparse Instance Representation

In machine learning, sparse representation has proven to be a powerful tool for representing high-dimensional input with high fidelity (Bruckstein, Donoho, & Elad, 2009). Some methods, such as SMLR discussed above, implement sparsity by selecting only a small subset of features for classification. Other methods select a small number of observations, rather than features. In Support Vector Machine (Vapnik, 1995), for example, only a small subset of relevant training samples are selected to characterize the decision boundary between classes.

A new and interesting machine learning method, developed in the image classification literature, uses the second approach (Wright, Yang, Ganesh, Shankar Sastry, & Ma, 2009). We call this new method Sparse Instance Representation (SIR), because it represents test samples in terms of a small number of the training samples themselves. Specifically, the test samples are represented as a linear combination of just a few training samples from the same class. This representation is naturally sparse, involving only a small fraction of the overall input. Instead of using sparsity to identify a relevant model or relevant features that can later be used for classifying all test samples, it uses the sparse representation of each individual test sample directly for classification. In this sense, it can be considered a generalization of nearest neighbor approaches. While SIR has been successful in the image applications for which it was developed, we believe ours is the first attempt to apply it to a cognitive problem.

Mathematical framework   Mathematically, in a typical SIR formulation, a *dictionary D* is constructed as $D = [d_1, d_2, ..., d_n]$, where each $d_i \in R^m$ is a feature vector of the ith instance. D is an over-complete dictionary if the number of instances n is much larger than the dimension of the feature vector m. To reconstruct an instance in terms of its feature vector y, SIR uses the equation $y = D\theta$, where a regularization is enforced on θ, such that only a small number of instances from the dictionary D

---

[1] Available from http://www.pni.princeton.edu/mvpa

are selected to describe y. A test sample is assigned to the class with the smallest residual in presenting y as a linear combination using all instances from the corresponding class.

For our data, the dictionary D is a $m \times n$ matrix with 0-1 entries, where $m = 764$ is the total number of features and $n = n_o \times n_p$ is the total number of instances from all training matrices, where $n_o = 129$ is the number of instances for an individual training matrix, and $n_p = 3$ is the number of training matrices. Thus, each class (category) has n_p instances, in the form of three feature vectors (columns) in D. A test sample is in the form of a feature vector y of the size $m \times 1$.

By re-aligning the dictionary matrix D, columns are grouped by animal. Thus D can be rewritten as $D = [D_1, D_2, ..., D_{n_o}]$, where each D_k, $k = 1, 2, ..., n_o$, has dimension $m \times n_p$. For a given test sample with feature vector y, SIR assumes that y can be expressed by a linear combination of columns in subset D_{k_0} that is of the same class as y.

$$y = \sum_{i=1}^{n_p} \theta_{k_0}^i D_{k_0}^i, \quad (1)$$

where $D_{k_0}^i$ is the ith column of D_{k_0}. Since the class membership of y is yet unknown, it is necessary to consider global linear combination of all the columns in D, thus

$$y = D\theta = \sum_{k=1}^{n_o} \sum_{i=1}^{n_p} \theta_k^i D_k^i. \quad (2)$$

However, only those instances from the same class (e.g. k_0) of y are highly relevant, whereas the features of other instances are much less relevant. In theory, only n_p feature vectors of training samples that belong to k_0 contribute to the expression of y, which means globally the linear expression weights in θ are sparse.

The convex objective function is

$$\theta^* = \arg\min_{\theta} \|\theta\|_1, \quad y = D\theta.$$

where $\|\theta\|_1$ denotes the l_1 norm of θ. However, based on the data, $m > n$ for the dimension of matrix D. This means the dictionary D is non-overcomplete, because the equation $y = D\theta$ is overdetermined, where the number of equations is larger than the number of unknown variables. In this case, the equation usually does not hold. Instead, we place it in the objective function by incorporating a trade-off parameter μ, where μ can be tuned for speed of convergence, we fixed μ value in this study. Thus the objective becomes,

$$\theta^* = \arg\min_{\theta} \frac{1}{2} \|y - D\theta\|_2^2 + \mu \|\theta\|_1. \quad (3)$$

Decision criterion   After obtaining the sparse solution θ^*, we calculate the residuals

$$r_k(y) = \|y - D_k\theta_k\|_2^2, \quad k = 1, 2, ..., n_o, \quad (4)$$

Table 2: Cross-validation results.

Method	Accuracy				
	Test 1	Test 2	Test 3	Test 4	Ave
1NN V1	.605	.674	.605	.612	.624
1NN V2	.628	.682	.643	.674	.657
DT	.388	.558	.543	.426	.480
SMLR	.612	–	.775	.411	.599
SIR	.659	.729	.744	.605	.684
NonNeg SIR	.760	.760	.783	.659	.740

where θ_k has size $n_p \times 1$. Its elements are n_p linear weights of the n_p instances of the kth animal, features of one instance thus receive the same weight. Finally, the test sample is identified by

$$\text{Index}(y) = \arg\min_{k} r_k(y) \quad (5)$$

Non-Negative Variant   A novel and psychologically interpretable variant of SIR places a non-negative constraint on the linear expression weights in θ. Therefore, the l_1 regularized unconstrained convex optimization in Equation 3 becomes a non-negative penalized l_1 regularized unconstrained convex optimization:

$$\theta^* = \arg\min_{\theta} \frac{1}{2} \|y - D\theta\|_2^2 + \mu \|\theta\|_1, \quad \theta \geq 0. \quad (6)$$

The natural psychological interpretation is that this constraint forces representations that include only instances that provide evidence for a category identification decision.

Results

We used the Split Bregman method (Goldstein & Osher, 2009) to solve the optimization task, both without and with non-negative restriction.[2] We describe overall performance of the classifiers, then focus on the details of the SIR classifier.

Accuracy

We measured the performance of each method using *accuracy*, which is simply the proportion of correctly classified animals. Table 2 details the accuracies for all of the classifiers on each of the four cross-validations, as well as average accuracy.[3] It is clear that SIR outperformed the other benchmark methods, especially with the inclusion of the non-negativity constraint. Nearest neighbor classifiers were the next best performed, followed by sparse multinomial logistic regression, and decision trees.

We think the variation in accuracy across the cross-validations may be interpretable in terms of individual

---

[2] A technical note regarding details of implementation is available at http://www.socsci.uci.edu/∼szhang/research.htm
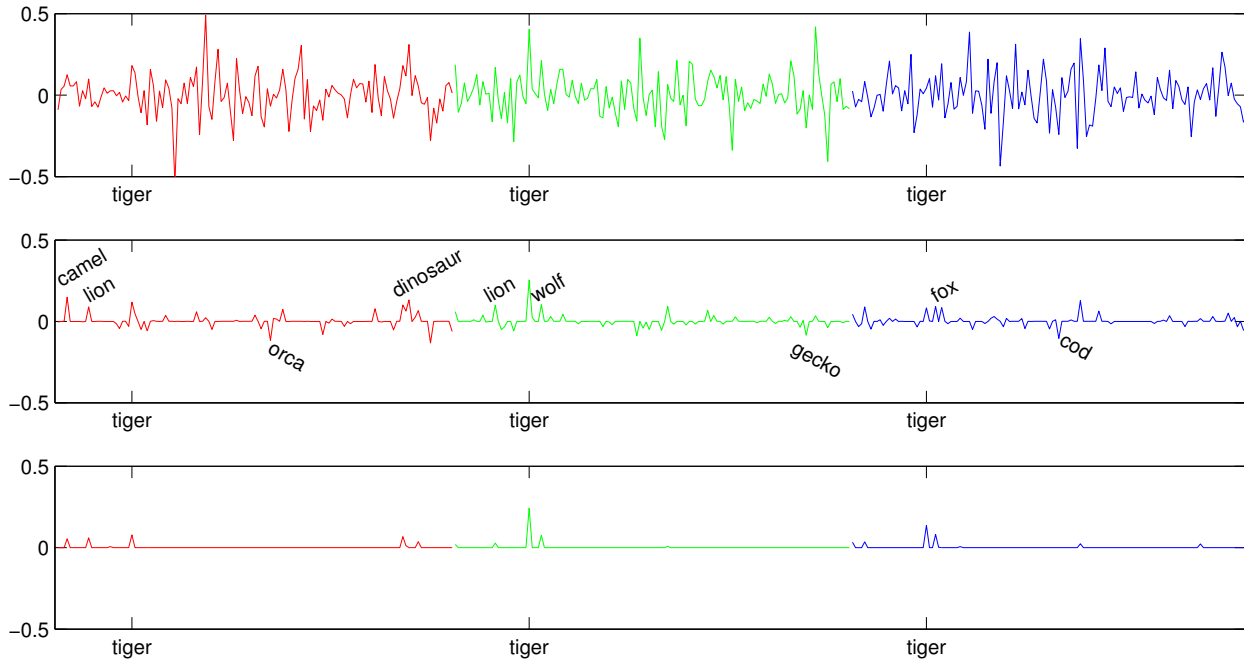
[3] Note that SMLR did not converge on Test 2.

Figure 2: Weights on all instances in the training data learned for 'tiger', using (upper panel) no regularization, (middle panel) sparse regularization, and (lower panel) non-negative sparse regularization, colored by participant.

differences in representation. Classification accuracy was almost always lowest when the feature assignments of participant 4 were used for testing. Referring back to the individual differences in agreement in Figure 1, we note that participant 4 shows more discrepancy in their representations of animals.

Selected Instances

Figure 2 gives an example of the instance-based representations used by SIR, for the example 'tiger'. Each panel shows the weights for all of the animals for the three training participants concatenated together. The top panel shows the case when no regularization is used. The middle panel shows the case when sparse regularization is used. The bottom panel shows the case when the non-negativity constraint is placed on sparse regularization.

To give some intuitions about the instances selected by the regularization processes, the middle panel labels a number of animals besides tiger that receive significant positive or negative weights. For example, the third participant, to the right, uses both 'fox' and 'cod' as well as 'tiger' in their representation, with fox features contributing positive evidence and cod features negative evidence. In the non-negative regularization, only fox continues to contribute. The other participants use other animals to represent tiger, again showing individual differences, but use animals that are easily interpreted in terms of the evidence their features provide for identification.

Another analysis is presented in Figure 3, which shows the weight distribution across all pairs of a test category (vertical axis) and any potential category (hori-

zontal axis). For each test category and a potential category, we summed the estimated weights of all 3 instances of the potential category, resulting in a weight-sum associated with the specific pair. For each pair, values across tests are further summed to yield the value shown.

Clearly, the sparse weights are generally assigned to instances from the correct category, illustrated by overall larger values along the diagonal, whereas instances from wrong categories received much lower weights, illustrated by the shaded areas off the diagonal. Another interesting pattern is shown by the five squares along the diagonal, each containing all pairs within an animal family (mammal, bird, fish, insect and reptile). This illustrates the within-category similarity in weighting that reflects the natural conceptual structure.

Discussion

The critical representational assumption of SIR is that sparsity in enforced in terms of instances. All features from the same instance receive the same weight, but different instances receive different weights. The insight is that, although features are naturally very high dimensional, instances belonging to the same class lie approximately in low-dimensional feature subspaces. If a collection of representative samples can be found, it is possible to represent a typical sample with respect to the basis they form.

The decision-making assumptions of SIR are simple, and assume a linear combination of the basis instances in doing category identification. One potentially important
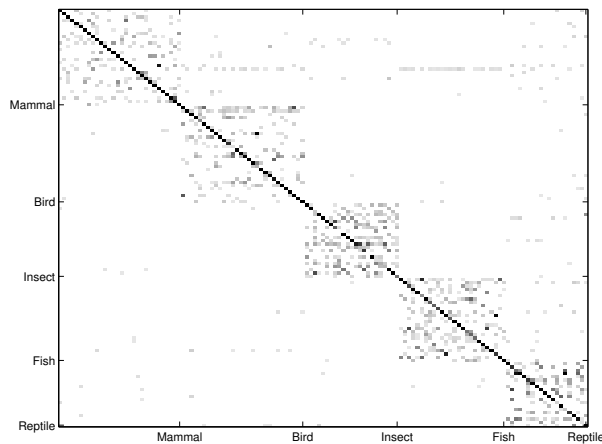
Figure 3: Overall weight distribution for SIR.

contribution we have made is to identify a non-negativity constraint on the weights as leading to better SIR performance on category identification data. This focus on instances that are *evidence for* a category mirrors findings in the similarity modeling literature that emphasize the role of positively weighted common features in stimulus representation (e.g., Navarro & Lee, 2004).

Thus, the superior performance of SIR in our evaluations support the idea that people represent stimuli in terms of sparse sets of the relevant instances. This is a natural extension of prominent and successful exemplar theories of concept representation (Nosofsky, 1986). It assumes that specific stimuli are the basis of concept representation, but implies that relatively few key stimuli are used. This is what is done—by various specific mechanisms–by a number of existing models of category learning, including the original ALCOVE model (Kruschke, 1992), SUSTAIN (Love et al., 2004), and the Varying Abstraction Model (Vanpaemel & Storms, 2008). Useful next steps are to apply these sorts of psychological models to account for the category identification behavior, and to explore their formal relationship to machine learning methods like SIR, and related case-based reasoning systems (e.g., Aamodt & Plaza, 1994).

Acknowledgments

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications 7*, *1*, 39–52.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bruckstein, A., Donoho, D. L., & Elad, M. (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, *51*, 34–81.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association*, *88*, 364–373.

DeDeyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., & Voorspoels, W. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 213–231.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Wiley.

Goldstein, T., & Osher, S. (2009). The split Bregman algorithm for l_1 regularized problems. *SIAM Journal of Imaging Science*, *2*, 323–343.

Kemp, C., Chang, K. M., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*, 216–233.

Krishnapuram, B., Figueiredo, M., Carin, L., & Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 957-968.

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

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

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. Prentice Hall.

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified versionof the contrast model. *Psychonomic Bulletin & Review*, *11*(6), 961–974.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental psychology: General*, *115*, 39-57.

Storms, G., Navarro, D. J., & Lee, M. D. (2010). Introduction to the special issue on formal modeling of semantic concepts. *Acta Psychologica*, *133*, 213–315.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732–749.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Wright, J., Yang, A. Y., Ganesh, A., Shankar Sastry, S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 210–227.