USING AUTOENCODERS TO MODEL ASYMMETRIC CATEGORY LEARNING IN EARLY INFANCY: INSIGHTS FROM PRINCIPAL COMPONENTS ANALYSIS

CHRISTOPHE L. LABIOUSE, ROBERT M. FRENCH, AND MARTIAL MERMILLOD
{clabiouse, rfrench, mmermillod}@ulg.ac.be

Cognitive Science Unit, University of Liege, 4000 Liege, Belgium

[Reference to cite: Labiouse, C., French, R. M. and Mermillod, M. (2002). Using autoencoders to model asymmetric category learning in early infancy: Insights from Principal Components Analysis. *Connectionist Models of Cognition and Perception*. London: World Scientific]

Abstract

Young infants exhibit intriguing asymmetries in the exclusivity of categories formed on the basis of visually presented stimuli. For instance, infants who have previously seen a series of cats show a surge of interest when looking at dogs, this being interpreted a dogs being perceived as novel. On the other hand, infants previously exposed to dogs do not exhibit such an increased interest for cats. Recently, researchers have used simple autoencoders to account for these effects. Their hypothesis was that the asymmetry effect is caused by the smaller variances of cats' features and an inclusion of the values of the cats' features in the range of dogs' values. They predicted, and obtained, a reversal of asymmetry by reversing dog-cat variances, thereby inversing the inclusion relationship (i.e. dogs are now included in the category of cats). This reversal reinforces their hypothesis. We will examine the explanatory power of this model by investigating in greater detail the ways by which autoencoders exhibit such an asymmetry effect. We analyze the predictions made by a linear Principal Components Analysis. We examine the autoencoder's hidden-unit activation levels and, finally, we emphasize various factors that affect generalization capacities and may play key roles in the observed asymmetry effect.

1 Introduction

Given that categorization capacities are crucial to cognition, it is not surprising that such abilities develop from the earliest age. Researchers have shown that young infants only a few months old are able to segment their environment into generic categories [10, 12]. Given their adaptive learning capacities, connectionist models offer a potential means of modeling human cognitive development. Besides their ability to account for general properties of categorization and memory, autoencoder networks (encoders, for short) are able to capture certain idiosyncratic characteristics of infants' categorization behavior. Mareschal et al. [4, 6, 7, 8] used an autoencoder to provide a simple mechanistic explanation of early infant category learning. Interestingly, these autoencoders are able to model the relation between attention and representation construction. In experimental settings, categorization tasks rely on preferential looking techniques based on the fact that infants pay more attention to novel stimuli. The common interpretation is that the infants are comparing an input stimulus to an internal representation of it. As long as there is a discrepancy between the input stimuli and its internal representation, the infant continues to attend to the stimulus in order to update the internal representation. When the discrepancy disappears, attention is switched elsewhere. This process and its implementation in an autoencoder is shown in Figure 1. Studies [9] reported that 3- to 4-month-olds show an unexpected asymmetry in the exclusivity of the perceptual category representations formed for certain basic level categories, in this case, cats and dogs. Following exposure to a set of cat pictures, the infants form a perceptual representation for cats that excludes dogs. In contrast, following exposure to a series of pictures of dogs, infants form a category representation for dogs that does not exclude cats. In other words, when infants are familiarized with dogs, they perceive cats as similar to what they already know and not as a novel category, whereas when they are familiarized with cats, they perceive dogs as different from what they have previously seen, and, thus, as a novel category.

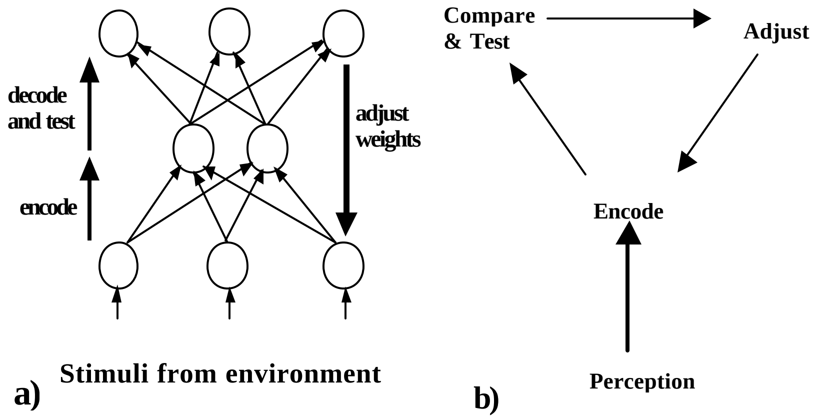

Figure 1. Learning by representation adjustment in (a) autoencoder networks and (b) infants.

Mareschal et al. [6, 8] have shown that autoencoders exhibit the same asymmetry effect. They hypothesize that statistical properties of the environment play the key role, given that infants and networks possess no semantic knowledge. The examination of the distributions of ten defining features (e.g. nose length, leg length, etc.) of the cat and dog stimuli shown to the infants reveals a potential explanation for the asymmetry effect. For most of these features, dog features have higher variances than the corresponding cat features. Crucially, the distribution of the cats' feature values is largely subsumed by the distribution of dogs' features. According to the authors, differences of within-category variance and an inclusion relationship between the two categories' distributions would cause the observed asymmetry effect. One consequence of this hypothesis is that the original asymmetry effect should be able to be reversed by reversing the inclusion relationship and the variances (i.e. variances of dog's features would be kept relatively small, whereas variances of cat's features would be made comparatively higher). This prediction was tested: both autoencoders and infants exhibited the predicted asymmetry reversal [4]. This model has predictive power given that it generates specific testable hypotheses (which, in addition, have turned out to be correct). However, predictive power does not necessarily imply explanatory power and, as a result, we propose studying in greater detail some of the fine-grained mechanisms that might rise to produce this asymmetry effect. Our hope is to be able to provide a more accurate picture of this asymmetry effect. First, we will analyze the predictions made by a linear Principal Components Analysis (PCA). Given that what a linear PCA does is similar to what autoencoders do, we believe that the insights offered by a PCA might shed light on how autoencoders produce this asymmetry effect. Then, because the hypothesis of Mareschal et al relies heavily on the nature of the internal representations developed by the autoencoder, we will examine the information provided by the hidden-unit activations. Finally, we will show that statistical tests that reveal the asymmetry effect may not be sufficient to disentangle the various alternative hypotheses for why it occurs. Specifically, we will emphasize different factors that affect generalization capacities and may play a role in the asymmetry effect.

2 Results and Predictions of a Linear PCA

Our aim is to attempt an in-depth analysis of the Mareschal et al [8] model. Unfortunately, the autoencoder contains nonlinearities that make formal analysis rather difficult. Therefore, we have chosen a somewhat easier framework that was similar enough to draw potentially informative conclusions. The closest statistical procedure to the autoassociative neural network is PCA. This technique belongs to a more general family of statistical procedures of dimensionality reduction. The standard PCA aims at reducing the dimensionality of a dataset by finding principal components which are linear combinations of the original dimensions and maximize the explained variance. It does very much what an autoencoder does by sending the input data through a smaller number of hidden units, which act as a bottleneck. Autoencoders use nonlinear activation function and online learning, whereas a PCA works with linear combinations and assumes ready-made covariance matrices. Nevertheless, PCA seems to be a good candidate to formally analyze the emergence of the asymmetry effect, given the ease of studying a linear framework with its similarity to encoders. Moreover, certain authors [1, 2, 3, 5] have mathematically demonstrated the relative equivalence between autoencoders and PCA in that the first principal components span the same space as the autoencoder's hidden layer as long as the autoencoder is trained long enough and there is no particularly complex statistical structure in the data. However, a simple mapping between a specific hidden unit and a specific principal component is not possible. In other words, the mutual orthogonality between hidden units is not guaranteed, as it is in the case of eigenvectors. A natural consequence is that there is no possible ordering of the different hidden units as in PCA, where eigenvectors and their associated eigenvalues can be ordered. On the other hand, with autoencoders, hidden units account roughly for the same amount of variance.

In a linear PCA, the total variance can be decomposed into the sum of the variance explained by the principal components and a residual variance. With zero-mean variables, the total variance can be recast in terms of distances from the origin. If V and P are, respectively, the original and the projected vectors:

$$\text{Total variance} = <||V||^2> = <||P||^2> + <||V - P||^2>$$

In the above formula[1], ||V|| and ||P|| are expressed as Euclidean distances. However, the features we used here for the cat and dog categories are not zero-mean or unit-variance variables. Thus, the total variance is not equal to $<||V||^2>$ in this case. Therefore, variance-based and distance-based analyses could lead to different, even contradictory, results. For this reason, we decided to perform distance-based analyses since the mean squared error (MSE) over output activation patterns in the network is equivalent to $<||V - P||^2>$.

2.1 Method

In Mareschal et al [8], a 10-8-10 autoencoder was used. In other words, each 10-element input vector was compressed into an 8-dimensional representation and then de-coded to another 10-dimensional output vector. The weights of the network were modified until the output approximated sufficiently the input or until 250 epochs were reached. We simulated this procedure using PCA. First, we ran a PCA on members of one category, say, Dog, using the associated co-variance matrix. The eigendecomposition gave us a matrix of eigenvectors, M_{10}^{dog}. We then created a reduced matrix composed of the 8 eigenvectors that explained the most variance for the Dog category, M_8^{dog}. For the sake of simplicity, we will designate this reduced matrix by M. We then considered the set of dogs, $\{d_i\}$, and the set of all cats, $\{c_i\}$, and calc ulated $Md_i = d_i'$ for each dog, and $Mc_i = c_i'$ for each cat (This is the equivalent of the autoencoder encoding the input at its 8-unit hidden layer.) Then we consider each vector d_i' and c_i' and "decode" these vectors with M^T. In other words for each d_i' and c_i' we calculate $M^T d_i'$ and $M^T c_i'$, respectively (This is the equivalent of the output produced by the autoencoder for each member of the Dog and Cat categories). We then compute the MSE for each category, which is equivalent to the average Euclidean distance between the original input vectors and

[1] $<||V||^2>$ indicates the average norm of all V vectors.

their respective reconstructions. We then did the same thing, starting with the Cat category and producing a matrix of eigenvectors, M_{10}^{cat} , producing a reduced 8-eigenvector matrix, etc.

2.2 Results

Using this procedure for the data of the original experiment by Mareschal et al [8] where dogs have a higher total variance compared to cats, we obtained the results summarized in Table 1.

Test with Cats		Test with Dogs		
Distance	Variance	Distance	Variance	
0.127 (4.73 %)	0.001 (0.40 %)	0.234 (7.04 %)	0.033 (9.07 %)	Residual after PCA on Cats
0.070 (2.61 %)	0.044 (19.23 %)	0.033 (0.98 %)	0.005 (1.40 %)	Residual after PCA on Dogs
2.69	0.231	3.32	0.366	Total

Table 1. Results of a linear PCA for the Quinn et al.'s data

After familiarization with dogs, the difference of mean absolute errors ("distance" in Table 1) between cats (0.070) and dogs (0.033) is not significant ($t_{(17.53)}$=1.40; p = 0.18) while, after familiarization with cats, the difference of mean absolute errors between cats (0.127) and dogs (0.234) is significant ($t_{(17.37)}$=3.06; p = 0.007). This agrees with the observation that when familiarized with dogs, newly presented dogs and cats are seen as members of the familiarized category (low amounts of error or looking time) but when familiarized with cats, newly presented cats are seen as members of the familiarized category but dogs are not and are seen as novel (significant increase of error or looking time) (See Figure 2).

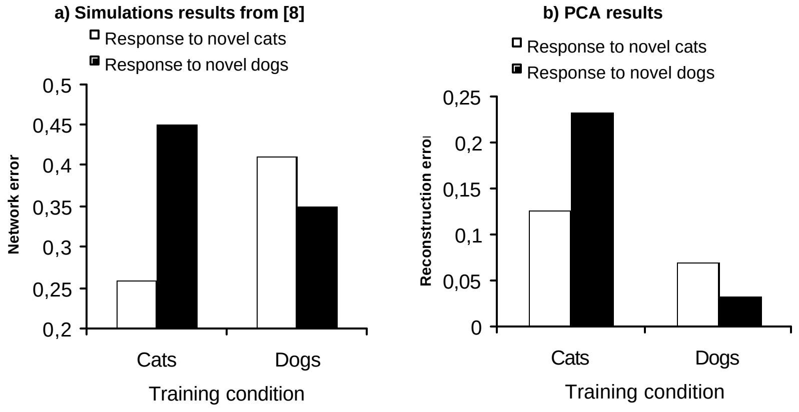

Figure 2. Results from (a) encoder simulations and (b) PCA, using data from Mareschal et al [8].

We then attempted to reproduce the asymmetry effect with the morphed data used in [4] where the authors altered the total variance of each category to obtain a reversal of the original asymmetry. The results are presented in Table 2.

Test with Cats		Test with Dogs		
Distance	Variance	Distance	Variance	
0.016 (0.35 %)	0.002 (0.75 %)	0.030 (0.82 %)	0.012 (11.95 %)	Residual after PCA on Cats
0.067 (1.48 %)	0.028 (9.56 %)	0.008 (0.21 %)	0.002 (1.18 %)	Residual after PCA on Dogs
4.53	0.294	3.62	0.095	Total

Table 2. Results of a linear PCA for the French et al.'s data

After familiarization with cats (now the "high variance" category), the difference of mean absolute errors ("distance" in Table 2) between cats (0.016) and dogs (0.030) is not significant ($t_{(21.02)}$=2.03; p = 0.055) while, after familiarization with dogs, the difference of mean absolute errors between cats (0.067) and dogs (0.008) is significant ($t_{(17.05)}$=3.55; p = 0.002). This confirms the results in [4] (See Figure 3).

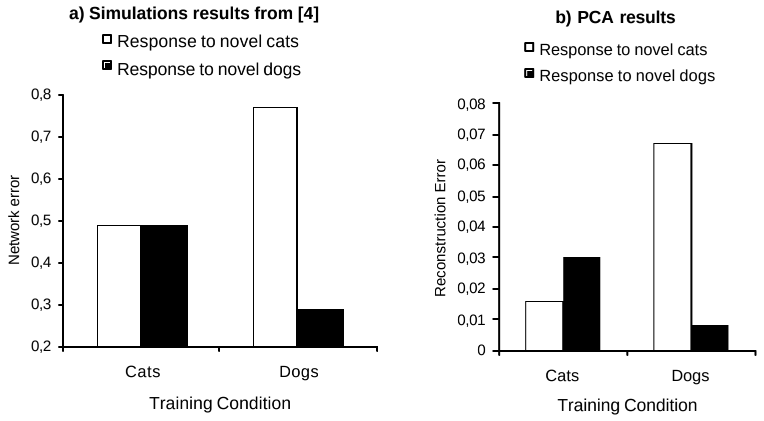

Figure 3. Results from a) encoder simulations and b) PCA, using data from French et al [4].

However, when we examine the errors strictly in terms of variance (rather than distance), the measure on which the PCA is based, we no longer reach the same conclusion. For each dataset, the asymmetry disappears: *residual variance is always higher for the novel category compared to the familiar category.* To examine the origin of the difference between the variance-based and the distance-based accounts, we need to construct virtual datasets based on the expected critical factor, which is the inclusion of the two categories' distributions of features. If the relative overlap of the two categories' distributions of features is the main critical factor, we could influence the occurrence of the asymmetry effect by building new datasets which would differ only by their relative position from each other. In other words, we would keep constant other statistical parameters, such as correlations, covariances and variances. Moreover, for zero-mean datasets distance-based and variance-based accounts will be the same.

2.3 Predictions with virtual datasets

In order to test the relative inclusion of the categories' distributions as a critical factor, we used the morphed data from [4] to build two new datasets for the Dog category and one new dataset for the Cat category. The new datasets were based on those used in [4] in which there was a higher total variance for cats compared to dogs and the feature distributions of dogs were largely subsumed by those of cats (All feature values had been normalized between 0 and 1). From these original datasets, we built the following datasets: a) a "dog" dataset with the original values increased by a constant (that may be different for each feature) to push the altered values towards the upper boundary of 1, b) a "dog" dataset with the original values shifted so that they are centered around zero (resulting in both positive and negative feature values), and c) a "cat" dataset with values centered around zero. It is important to note that in each of these cases, the correlations and covariances between features, and the variance for all features remain the same. The critical change of the means alters only the MSE. We therefore obtained 6 pairs of comparisons (i.e. a total of three "dog" datasets and two "cat" datasets) whose relative overlap is given in Table 3.

	Original cats	"Centered" cats
"Increased" dogs	Overlap (1)	No overlap (4)
Original dogs	Overlap (2)	No overlap (5)
"Centered" dogs	No overlap (3)	Overlap (6)

Table 3. Pairs of virtual datasets and their relative inclusion. Pair number is between brackets

According to the hypothesis formulated by Mareschal et al., if two datasets have an inclusion relationship such that one falls within the other, an asymmetry effect should emerge. Conversely, if two datasets are separated from each other (i.e. no overlap), the asymmetry effect should disappear. After replicating the previous PCA on the 6 pairs of comparison and considering the variance of each category, we found that only pairs 2 and 5 in Table 3 support the above hypothesis.

In other words, the linear PCA shows that inclusion does not play a role as simple as originally thought but might interact with the category position relative to the origin. It turns out that the way one codes the features (e.g. in millimeters, inches, Z-scores,…) is of crucial importance. Theoretically, the choice of initial coding should not significantly alter the results, but a linear PCA shows that these supposedly arbitrary choices can have significant consequences on the results and on the predictions of the model. Note, however, that this might not apply to the predictions from a network, which is a nonlinear model. Further testing on virtual datasets using the original encoders is needed. Nonetheless, these results indicate the need for caution when making predictions or analyzing results from this type of computer simulation given the model's sensitivity to initial coding.

3 Examination of internal representations

Mareschal et al [8] hypothesize that the key element for the emergence of an asymmetric attention effect in infants is an approximate inclusion relation of the distribution of feature values for one category within the distribution for the features of the other category. According to them, "the asymmetry inherent in the data is only translated into corresponding behavior because encoders develop internal representations that reflect the distributions of the input features. Thus, the internal representations for the narrow category [NC] are subsumed within the internal representations for the broad category [BC]… It is because the internal representations share this inclusion relationship that an asymmetry in error is observed". In other words, after familiarization with BC items, both novel BC and NC items are correctly autoassociated (low reconstruction errors). Therefore, the authors suppose that internal representations of NC items fall "inside" the space of BC items. On the other hand, after familiarization with NC items, novel NC items are, indeed, correctly autoassociated but novel BC items are not (high output errors). Therefore, they suppose that internal representations of BC items fall "outside" the space of NC items. This argument works if good generalization in autoassociation is restricted to the part of hidden-unit space covered by the familiar items and that, once you leave this part of hidden-unit space, poor generalization occurs. In this context, it would be interesting to test three points:

1. Are the internal representations of BC items (after familiarization with BC items) more scattered than those of NC items (after familiarization with NC items) ?
2. After familiarization with BC, do the internal representations of NC items fall inside the range of BC items? After familiarization with NC, do the internal representations of BC items fall outside the range of NC items?
3. Do the internal representations that fall outside the "familiarized" range produce systematically higher reconstruction errors ?

The answer to the first question is yes, in the case of autoencoders as shown in [7]. However, the last two questions are still unanswered for a nonlinear model. Given the inherent complexity of a nonlinear autoencoder, we chose to look at those issues within the linear PCA framework, which, as expected, answers the first two questions in the affirmative and the third in the negative. In fact, in a PCA, the projections of the original vectors onto the principal components are very close to the original vectors themselves (see Tables 1 & 2), which means that the principal components space is not very different from the original space. The relative distances between projected vectors are quite similar to the distances between original vectors. In other words, independent of the familiarized category, NC members are subsumed by the principal-components space (PC space) covered by BC members. On the other hand, this does not imply that internal representations that fall outside the familiarized-category range necessarily produce higher reconstruction errors on output. Thus, the examination of activation levels in the hidden-unit space might not reveal a structure similar to the structure at the output level where errors are recorded. In a PCA, reconstruction errors are independent of the relative position of the patterns in PC space. In other words, even clear-cut "categories" (i.e., vector clusters) in PC space would not necessarily predict what will happen on output.

Figure 4 shows why projected vectors of BC items far from the origin and far from the other projected vectors (in particular, far from the projected vectors of NC items) do not necessarily produce higher reconstruction errors. However counterintuitive, this interpretation is logical given that reconstruction errors arise from *distances between the original vectors and their projections onto the PC space*, regardless of where their projections lay on the principal components (see Figure 4).

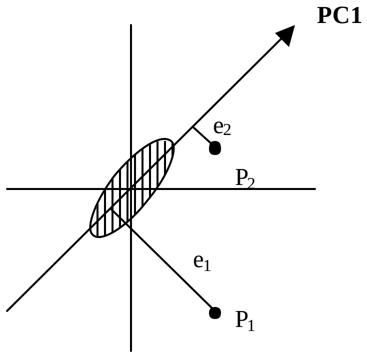

Figure 4. Even though P_1 projects into the training-exemplar zone along the first principal component and P_2 does not, there will be a smaller reconstruction error for P_2 than for P_1.

Figure 4 illustrates a simple example with principal components in two dimensions. Assume one keeps only the first PC of the covariance matrix, derived from the set of training exemplars indicated by the shaded region in the figure. Now we test two novel exemplars, P_1 and P_2. P_1 is projected to the "familiar" (shaded) zone along the first principal component, whereas P_2 is not. According to Mareschal et al [8], one would therefore expect a greater reconstruction error for P_2 than for P_1. But this would not be the case in this example. Rather, the reconstruction error is based on the residual error (e_1 for P_1 and e_2 for P_2) and, in this case, e_1 is considerably greater than e_2. As a result, the reconstruction error associated with P_1 would be greater than for P_2, in spite of the fact that P_1 projects into the training exemplar zone along the first PC, while P_2 does not.

In short, it is not necessary to posit that the asymmetry effect is caused only by the relative inclusion of sets of internal representations. Hidden-unit representations serve to encode patterns only in order to reconstruct them, not in order to classify them.

4 Factors influencing generalization

With respect to Quinn & al's data [9], Mareschal et al [8] claim that: "when presented with a cat, a network trained on dogs will recognize this item as a member of the category it has learned. In contrast, when presented with a dog, a network trained on cats will fail to generalize this item as a member of the category it has learned". Moreover, the authors add that "generalization is better in the dog-to-cat direction…". But this can be argued another way by emphasizing that cats and dogs *should*, in fact, be seen as two distinct categories. We therefore argue that familiarization with cats yields a more accurate picture of the environment, given that the new category ("dogs") is, indeed, perceived as novel, whereas familiarization with dogs does not lead to a separate categorical perception of cats.

In original simulations, researchers used a fixed-epoch criterion (i.e. learning stops after a fixed number of 250 epochs). Given it could have been argued that the asymmetry effect arises from the unequal learning of the Cat and Dog categories by the networks, the authors decided to switch from a fixed-epoch criterion to a fixed-error criterion (i.e. learning stops after the error on each output unit drops below a fixed criterion of 0.2). Under the fixed-error criterion, training with initial exemplars is as good with one category as with the other. This led the authors to conclude that unequal learning of the categories does not impact on the asymmetry effect. But high-quality learning involves not only being able to reproduce the learned exemplars, but, in addition, requires being able *to correctly generalize to new exemplars of the same category*. In fact, the differences reported in Mareschal et French [6] – namely, a larger increase in error when the network is presented with new dogs after familiarization with cats compared to a presentation of new cats after familiarization with dogs – are independent of learning (see Figure 5 for two hypothetical examples that illustrate how this might occur).

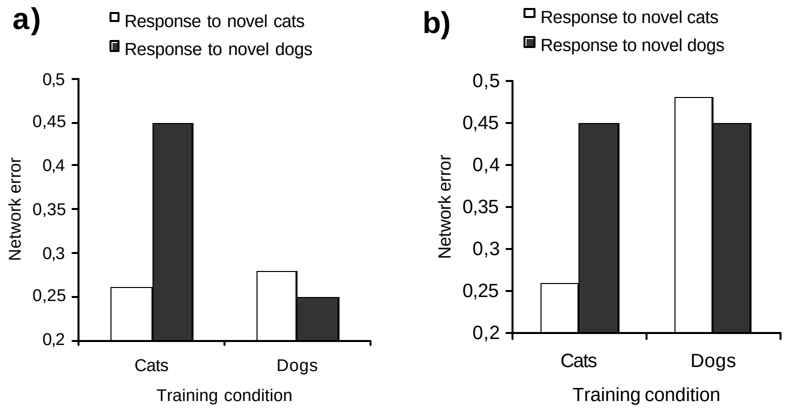

Figure 4. Two opposite asymmetry effects where generalization for dogs is (a) good or (b) poor.

Both graphs reveal a main effect of the familiarization condition; only the direction of the difference changes. Figure 5a buttresses a claim that both cats and dogs are seen as dogs after familiarization with dogs. In Figure 5b, the asymmetry effect is preserved, but familiarization yields very poor generalization, even for novel dogs.

In Figure 5a, the (unexpected) result is the small error exhibited when presented with a cat after familiarization with dogs. This would mean that networks – like infants – show over-generalization because they generalize to members from another category. This over-generalization could be due to the fact that cats are subsumed by the dog category (the explanation of Mareschal et al [8]).

On the other hand, Figure 5b shows a very large generalization error to novel dogs when trained on dogs. In this case, it could be that the cat category is simply easier to learn than the dog category and generalization errors to novel dogs are higher. Or it could be due to the difference in the distributions underlying each category and the sampling of exemplars. To achieve good generalization, the training set must be representative of the category *as a whole*. A poor set of training data may contain misleading regularities not found in the whole category This is particularly serious when the sample size is small, which is clearly the case here. Good generalization cannot be expected if a network is trained on samples from one region of the space but tested on samples from a completely different region [11]. Therefore, sampling in the more restricted (i.e., smaller variance) Cat category would necessarily yield training exemplars, as well as test exemplars, that are relatively close together in feature space. In contrast, given that variance of dogs' features is relatively high, sampling in the Dog category would result in comparatively more widely scattered training and test exemplars. This would result in significantly higher error levels.

In other words, it is crucial to emphasize the potential role of sampling in the asymmetry effect. We accept that within-category variance and relative inclusion of feature distributions of the two categories play a key role in the asymmetry effect, but are not necessarily the only relevant factors: sampling procedure and differential category complexity may also play an important role as well.

5 Conclusion

Researchers have used simple autoencoders to account for an asymmetry effect observed in infant categorization of natural categories. Their model, based on the within-category variances and the inclusion relationship of two categories' feature distributions, was able to reproduce to an categorization asymmetry observed in infants and to correctly predict a reversal of this asymmetry by reversing the variances and inclusion relationship between the two categories' feature distributions. However, questions remained about the precise mechanisms by which this asymmetry arose. Therefore, we performed a linear PCA that showed the heavy dependence of results and predictions on the choice of input coding. Further, examination of the internal representations in a linear framework demonstrates that clusters of representations at the hidden layer do not necessarily predict output errors. This potentially implies taking into consideration not only internal representations, but also the input encoding of the data itself, in order to account for the observed categorization asymmetry. Finally, the generalization ability of a network requires taking into account the fact that certain categories could be intrinsically easier to grasp than others and that the sampling procedure used to define the familiarization sets may affect the occurrence of the asymmetry effect.

6 Acknowledgements

This work was supported by a research grant HPRN-CT-1999-00065 from the European Commission. The first author is a Research Fellow of the Belgian National Fund of Scientific Research.

References

1. Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53-58.
2. Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics* , **59**, 291-294.
3. Cottrell, G., Munro, P., & Zipser, D. (1989). Image compression by backpropagation: A demonstration of extensional programming. In N. E. Sharkey (Ed.), *Models of cognition: A review of cognitive science*, (Vol. 1, pp. 208-240). Norwood, NJ:Ablex.

4. French, R. M., Mermillod, M., Quinn, P., & Mareschal, D. (2001). Reversing Category Exclusivities in Infant Perceptual Categorization. In Proceedings of the 23rd Annual Cognitive Science Society Conference. NJ: LEA, 307-312.
5. Japkowicz, N., Hanson, S. J., & Gluck, M. A. (2000). Nonlinear auto-association is not equivalent to PCA. *Neural Computation*, **12**, 531-545.
6. Mareschal, D., & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. In *Proceedings of the 19th Annual Cognitive Science Society Conference*. NJ: LEA, 484-489.
7. Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, **1**, 59-76.
8. Mareschal, D., French, R. M., & Quinn, P. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, **36**, 635-645.
9. Quinn, P., & Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, **22**, 463-475.
10. Quinn, P., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. In C. Rovee-Collier & L. P. Lippsitt (Eds.), *Advances in Infancy Research* (Vol.10, pp. 1-36). Norwood, NJ: Ablex.
11. Reed, R. D., & Marks II, R. J. (1999). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. Cambridge, MA: MIT Press.
12. Younger, B. A. (1985). The segregation of items into categories by 10-month-old infants. *Child Development*, **56**, 1574-1583.