UNIVERSITY OF PADOVA

Department of General Psychology

Bachelor's Degree Course in Psychological Science

Final dissertation

Numerosity Perception in Deep Belief Networks

Supervisor
Professor Marco Zorzi

Co-supervisor
Professor Alberto Testolin

Candidate: Dunja Milovanović
Student ID number: 2081269

Academic Year 2024/2025

Table of Contents

Abstract

Numerosity perception refers to the ability to estimate the number of items in a visual scene. This thesis explores numerosity perception in ‘Deep Belief Networks’ (DBNs), hierarchical generative models that learn the underlying statistical representation of the sensory input in an unsupervised fashion. To simulate behavioral tasks, the internal representations were read-out by a supervised linear classification layer. Networks’ performance was assessed using three numerosity judgement tasks, namely pairwise numerosity comparison, fixed-reference comparison, and numerosity estimation task. Furthermore, the influences of training dataset size and type of classifier used for read-out were systematically examined. Results showed that DBNs performed accurately across tasks and conditions, capturing the key behavioral patterns observed in human empirical studies.

Key words: neural networks, deep learning, numerosity, number sense, machine learning

List of Figures

1. Introduction

1.1. Numerosity Perception

Numbers are a core part of the modern world and have been essential throughout human history. They are a crucial tool for interpreting the world around us and as such the capacities to rapidly understand, approximate, and manipulate numerical quantities seem to be an invaluable faculty of the human mind. These abilities, collectively referred to as the 'number sense' (Dehaene, 1997), are supported by the system for approximate representations of numerical magnitudes, or the approximate number system (ANS) (Feigenson et al., 2004). The ANS allows for rapid estimation of quantity without the use of language or symbols and is characterised by ratio dependence, meaning that discriminatory judgments between numerosities follow Weber's law.

The hallmarks of the system have been reported not only in human adults, but also in infants, children, and many non-human animals (Cantlon, 2012; Piantadosi & Cantlon, 2017; Starkey & Cooper, 1980; Strauss & Curtis, 1981). Infants as young as six months of age can discriminate between the numerosities with 2:1 ratio, and this competency increases substantially throughout development reaching discriminability of approximately 10:9 ratio in human adults (Halberda & Feigenson, 2008; Lipton & Spelke, 2004; Xu & Spelke, 2000). Similar capacities have also been observed in fish, chicks, and non-human primates (Agrillo et al., 2008; Merten & Nieder, 2009; Rugani et al., 2015).

The presence of these abilities both across different stages of development and among different species suggests a long evolutionary history and biological foundation of numerosity perception. Numerous studies have demonstrated the existence of specialised neural circuitry in the parietal cortex tuned for number processing. Initial indications of this came from lesion studies (Gerstmann, 1940), and have been further supported by the multitude of neuroimaging evidence for number-selective coding system in the anterior part of the intraparietal sulcus (IPS) (Cantlon et al., 2006; Piazza et al., 2004, 2007; Santens et al., 2010). Moreover, several studies have also shown the existence of number selective neurons in the IPS of non-human primates, as well as in the human medial temporal lobe (MTL), that respond preferentially to specific numerosities (Kutter et al., 2018; Nieder et al., 2002; Nieder & Miller, 2003, 2004).

The relevance of 'number sense' also stems from the notion that it has long been indicated as a foundation for symbolic mathematics (Dehaene, 1997; Wynn, 1998). This has been supported by a body of research demonstrating that number acuity, usually defined as the Weber fraction, or the smallest detectable change in numerosity, can predict later mathematical achievement (Chen & Li, 2014; Halberda et al., 2008; Jordan et al., 2010; Starr et al., 2017).

1.2. Numerosity and Continuous Magnitudes

The existence of number neurons and the evidence that numerosity perception seems to follow Weber's law would suggest that numerosity is a primary visual attribute (Burr & Ross, 2008), but the long standing debate about whether the numerosity perception is just a by- product of perception of non-numerical stimulus features is still relevant in the field.

All primary sensory attributes like size, color and distance are susceptible to adaptation and the after-effects that arise due to this process have also been observed for numerosity (Ross, 2010). Additionally, numerosity can also modulate an autonomic reflex like pupil response without an explicit task, indicating that it is a spontaneously encoded visual feature. (Castaldi et al., 2021). On the other hand, some argue that it is the perception of other sensory magnitudes, like density (Durgin, 2008), size (Henik et al., 2017), convex hull (Gebuis & Van Der Smagt, 2011), cumulative surface area (Feigenson et al., 2002) and contour length (Clearfield & Mix, 1999), that meditates perception of numerical quantities

This alternative perspective suggests that a more automatic and basic “sense of magnitude”, that enables discrimination of different continuous magnitudes, may underlie numerosity perception rather than a specific “sense of number” (Leibovich et al., 2017). Accordingly, some researchers claim that perception of numerical quantities might emerge from the integration of various continuous visual properties (Gebuis & Reynvoet, 2012).

One of the problems with both theoretical perspectives lies in the limited understanding of the computations underlying the perception of both numerosity as an independent feature, and the precise mechanism by which it might emerge from the perception of continuous visual magnitudes. Furthermore, another problem is the lack of precise definitions of these features and adequate control over them in the stimulus space. Computational modelling has shown as a promising framework for addressing these issues (Stoianov & Zorzi, 2012; Testolin et al., 2020; Zorzi & Testolin, 2018).

1.3. Computational Modelling in Cognitive Neuroscience

A crucial issue not only in the study of numerosity perception, but in the whole field of cognitive neuroscience, concerns the neural computations behind cognitive processes. A number of authors argue that in order to understand brain information processing it is necessary not to rely only on traditional hypothesis testing, but to be able to build computational models that are capable of performing cognitive tasks (Kriegeskorte & Douglas, 2018; Newell, 1973). In this view, Newell famously claimed that only through computer simulations we can truly determine how the proposed component mechanisms interact and whether these interactions can adequately model the targeted cognitive function.

In light of this critique, building of the task-performing computational models has been a crucial step in the advancements of the study of cognitive processes, as they try to satisfy a certain degree of biological plausibility through use of mathematically precise simulations (Kriegeskorte & Douglas, 2018). One of the big advancements that was brought to the field was the widespread use of neural network models that have been essential both for the study of biological neural networks, but also for introducing parallel distributed processing as a novel paradigm in cognitive science (Kriegeskorte & Douglas, 2018). Due to their capability to successfully simulate human behavioural responses across different cognitive tasks, neural networks have become a valuable tool in cognitive modelling. Their ability to successfully capture empirical findings demonstrates their value in modelling numerical cognition (Stoianov & Zorzi, 2012; Testolin et al., 2020; Zorzi et al., 2013).

1.4. Aims and Objectives

This thesis aims to explore numerosity perception in computational models based on Deep Belief Network models originally proposed by Stoianov and Zorzi (2012). Here, numerosity is suggested to emerge as a latent statistical property of the data, spontaneously captured by the network during the unsupervised learning of the generative model of the input images. A recently proposed algorithm was adopted for training the networks (Zambra et al., 2023), offering more biologically plausible accounts of processes underlying neurocognitive development. To address the problem of continuous magnitudes, stimuli were generated with an approach that allows for the disentanglement of their effects over numerosity judgements (DeWind et al., 2015). The networks were tested across three different tasks of varying difficulty: two comparison tasks and a numerosity estimation task. Moreover, to assess task performance, different supervised linear read-out methods were used and evaluated. Finally, the impact of varying the size of the dataset used for training this final layer was also examined.

2. Methods

2.1. Deep Belief Networks

Deep Belief Networks (DBNs) represent a class of deep neural networks that learn the latent causes of the data in an unsupervised manner (Hinton et al., 2006)). They are thought to have sparked the modern-day deep learning revolution and have also been widely applied in cognitive sciences as computational models inspired by cortical learning principles. DBNs are generative probabilistic graphical models composed as a stack of Restricted Boltzmann Machines (RBMs), a variant of Boltzmann machines (Ackley et al., 1985) that lack the inter-layer connections, thus allowing for the conditional independence between the units within each layer and simplification of the learning algorithm. RBMs consist of two different sets of 'neurons', namely the visible neurons (Figure 2.1, *v*), that are in direct contact with the input, and hidden neurons (Figure 2.1, *h*) that capture the latent features explaining the observed data. Learning within each RBM is performed using the contrastive divergence algorithm (Hinton, 2002), an efficient approximation method that adjusts the model's parameters to maximize the likelihood of the observed input under its generative framework.

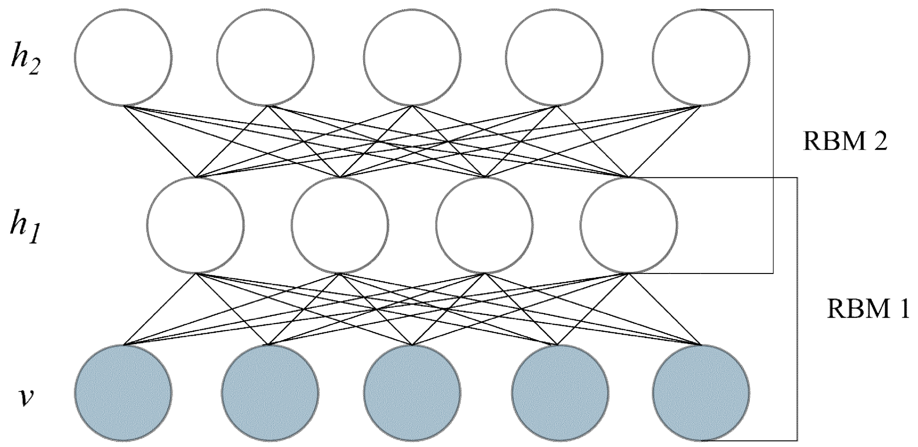

Figure 2.1. Deep Belief Network with two hidden layers

Within the DBN, each RBM is trained ‘greedily’, layer by layer, so that the hidden activations of the first RBM, after it has been fully trained, represent the input for the subsequent one. This hierarchical organization leads to increasingly complex representations of the data to emerge in the deeper layers of the network. For instance, early layers’ neurons might resemble the simple center-surround off and on detectors and edges (Lee et al., 2008; Stoianov & Zorzi, 2012), while subsequent layers might combine these into more high-order features like junctions (Lee et al., 2008), or even parts of digits and letters (Zorzi et al., 2013).

This representational hierarchy highly resembles the one encountered in the cerebral cortex, but the classical ‘greedy’ training poses a problem for the neurobiological plausibility of the model. Some limitations arise because a certain degree of neural plasticity persists throughout life and the cortical layers develop simultaneously, rather than sequentially, making this training approach inadequate for modelling human cortical development. For this reason, an alternative algorithm has been recently proposed, where the learning is instead done iteratively, so that the weights of the model are updated jointly across layers, yielding comparable results on both MNIST and numerosity datasets (Zambra et al., 2023).

In the present study, all the DBNs consisted of two RBMs. In total, 120 networks were trained and assessed, out of which half were trained with ‘greedy’ (gDBNs), and half were trained with ‘iterative’ (iDBNs) learning algorithms. The models’ architecture and hyperparameters were identical to those recommended by Testolin et al. (2020). Twelve different architectures (Figure 2.3) were examined, each with five different initializations of connection weights sampled randomly from the normal distribution with a mean of 0 and a standard deviation of 0.01. The learning rate was set to 0.15 and the weight decay factor was 0.0001. Initial momentum was 0.7, and it was updated to 0.97 after 5 epochs of training. All networks were trained for 200 epochs in total. Due to previously mentioned downsides of the ‘greedy’ learning algorithm, iDBNs were used for task simulation, while gDBNs were assessed in terms of reconstruction capabilities only.

2.2. Stimulus Space and Datasets

The stimuli construction method was adapted from DeWind et al. (2015). Here three orthogonal dimensions were used to specify all numerical and non-numerical features of each stimulus. The three dimensions were Numerosity, representing the actual quantity of dots in the array, and two mathematical constructs specifying the joint influence of all the continuous stimulus features, namely Size (combining the surface area and perimeter of each dot, total surface area and total perimeter), and Spacing (combining the field area and sparsity). These dimensions define a three-dimensional stimulus space (Figure 2.2, panel A) where any pattern of dots can be represented by a single point that reflects the Numerosity, Size, and Spacing, and through their algebraic relations, all the other continuous features of the image.

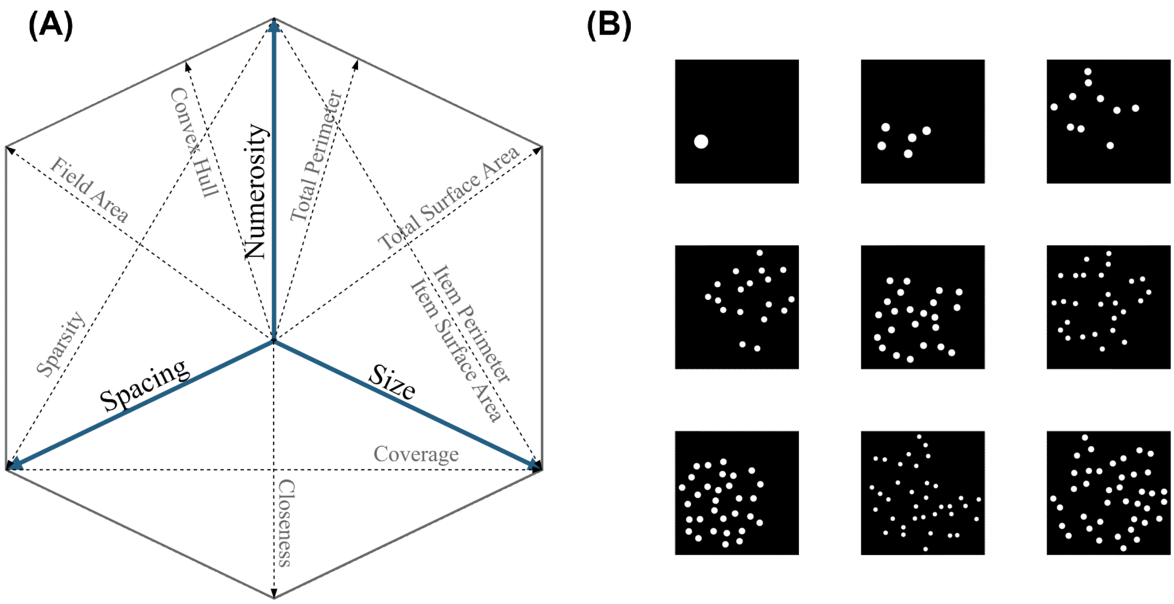

Figure 2.2. Stimulus space (A) and examples of stimuli images (B), adapted from DeWind et al. (2015)

All the images in the datasets were 100x100 pixels in size. The dataset used during unsupervised training contained all numerosities from 1 to 40, with 13 distinct levels of Size (1.6-6.4 $\times 10^5$ pixels) and Spacing (6-24 $\times 10^6$ pixels). For each combination of features, 13 images were created by randomly positioning white dots against a black background, resulting in a total of 87 880 images (Figure 2.1, panel B).

Separate training and testing datasets were created for the numerosity tasks, maintaining the 13 levels of Size and Spacing. However, for the comparison tasks explained in the next section, the ratios between the numerosities of stimuli images and the reference, or within the pairs, were constrained to fall within the range of [0.5, 2], or [-1, 1] on log_2 scale.

2.2. Tasks and Read-out Methods

Networks were tested on three different numerosity tasks often employed in human experiments. Namely:

1. **Fixed reference comparison** task, which involves judging whether the number of the objects in the image is higher or lower than a fixed reference.
2. **Pairwise comparison** task, where the goal is to determine which of the two presented images contains more objects.
3. **Numerosity estimation** task, which involves naming the exact number of objects in the image.

Internal representations of stimuli images, represented by the final hidden layer activations acquired by passing the image through the fully trained network with fixed weights, were used as the inputs for the linear classification layer performing the task specific read-out. Several different classification methods were evaluated across tasks, and the choice of the classifiers was based on task demands. For the comparison tasks, which are problems of binary nature, methods included pseudo-inverse, logistic regression, ridge regression (L2 regularised linear regression) and stochastic gradient descent (SGD). The estimation task was approached as a regression problem and here simple linear regression, lasso regression (L1 regularised linear regression), ridge regression, and SGD were used.

The pseudo-inverse method computes the Moore-Penrose inverse of the input data matrix, thus updating the weights directly. Ridge and lasso regressions expand on the ordinary least squares by adding regularisation terms that prevent overfitting and improve the generalization of the model. SGD differs from these direct methods by performing iterative, local weight updates based on gradient information, thus preserving the neural network principle of learning through repeated adjustments rather than direct solutions.

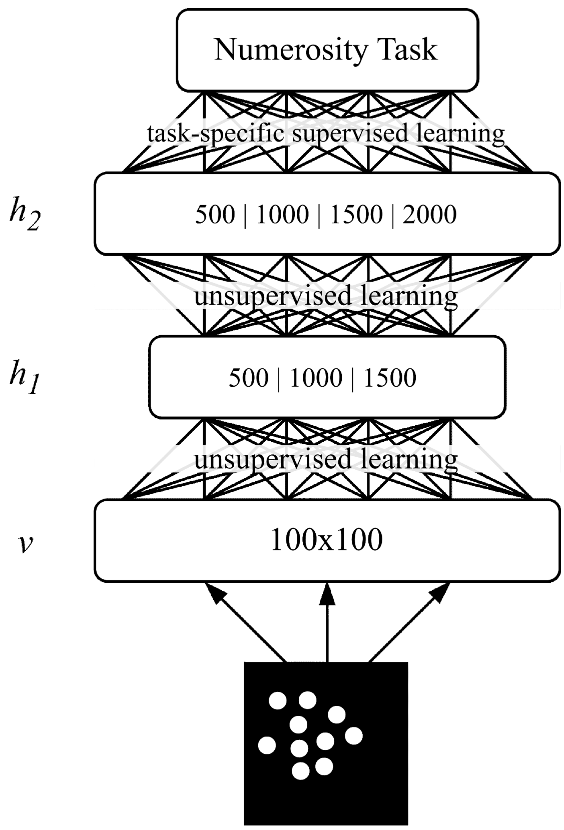

Figure 2.3. Model architecture

2.5. Generalized Linear Models

For both comparison tasks, Generalized Linear Models (GLMs) adapted from DeWind et al. (2015) were used to estimate the influence of each orthogonal feature on the model's choice. For the fixed reference comparison task, z scores of log-scaled Numerosity, Size, and Spacing were used as regressors (1), while in the pairwise comparison they were log-scaled ratios between pairs (2). The fourth regressor was added as a bias term for side/direction. The guessing term, indicated by γ, was set to 0.01, and it is specified under the assumption that a certain percentage of responses in humans would be random due to possible distractions during the experimental procedure. The model's β coefficients give the information about whether and how much each of the three orthogonal dimensions influences the choice behavior.

$$(1) \quad p(\text{ChooseLarger}) = (1-\gamma)\left[\frac{1}{2}\left(1 + erf\left(\frac{\beta_{\text{Direction}} + \beta_{\text{Num}} z_{\text{num}} + \beta_{\text{Size}} z_{\text{size}} + \beta_{\text{Spacing}} z_{\text{spacing}}}{\sqrt{2}}\right)\right) - \frac{1}{2}\right] + \frac{1}{2}$$

$$(2) \quad p(\text{ChooseRight}) = (1-\gamma)\left[\frac{1}{2}\left(1 + erf\left(\frac{\beta_{\text{Side}} + \beta_{\text{Num}} \log_2(r_{\text{num}}) + \beta_{\text{Size}} \log_2(r_{\text{size}}) + \beta_{\text{Spacing}} \log_2(r_{\text{spacing}})}{\sqrt{2}}\right)\right) - \frac{1}{2}\right] + \frac{1}{2}$$

The three β coefficients for Numerosity, Size and Spacing define the participants' discrimination vector in the three-dimensional space. In the situation of unbiased behaviour, β_{Num} would be significant ($p<0.01$), while β_{Size} and $\beta_{Spacing}$ would not ($p>0.1$). In the case of biases, the performance might be based primarily on the numerosity, or instead it might be driven by one of the non-numerical features. To inspect this, projection magnitude of the discrimination vector onto all of the dimensions in the stimulus space were calculated and evaluated. If the projection onto the numerosity axis was significantly greater than onto all the other non-numerical features ($p<0.05$), it can be considered that choices were primarily based on numerosity but with a bias for the other features. On the other hand, if the numerosity vector projection was significantly smaller than another projection ($p<0.05$), then the choices were considered to be made by primarily relying on a continuous magnitude. In the case that no significant difference was present between the numerosity vector projection and another vector projection, the response strategy was undetermined.

2.6. Dataset Size Reduction

The final aim of this thesis was to examine how reducing the training dataset size of the read-out layer affects the model's performance and how different classifiers perform under these conditions. To this end, ten different dataset sizes were examined, ranging from the full dataset used in the pairwise comparison task containing 15 200 pairs to just 1% of that size. The reduced datasets were generated using stratified random sampling, to preserve the distribution of the stimulus features and their ratios within pairs. For each of the additional 9 sizes, 10 different datasets were created. The testing was done on ten networks with varying initial weights but with the same architecture ($|H1|$ = 1500, $|H2|$ = 1000). The same four classifiers as used in both comparison tasks were evaluated: pseudo-inverse, ridge regression, logistic regression, and SGD.

3. Results

3.1. Reconstruction

In order to evaluate the learning of the DBNs it is a common practice to examine reconstructions. The task is performed with a separate test set that was never seen during the initial training. Test images are clamped to the visible layer of the network and one forward and backward pass are made while the network's weights remain fixed. Then, the Mean Squared Error (MSE) is computed between the original and reconstructed images. This was done for all 120 networks and the examples of reconstructed images (Figure 3.1) and the reconstruction loss (MSE) as a function of numerosity (Figure 3.2) can be seen below.

As expected, as the more numerous images are also more complex, the reconstruction loss increases with the increase in numerosity. Moreover, this is evident in both gDBNs and iDBNs and their performance is almost identical in terms of reconstruction capabilities.

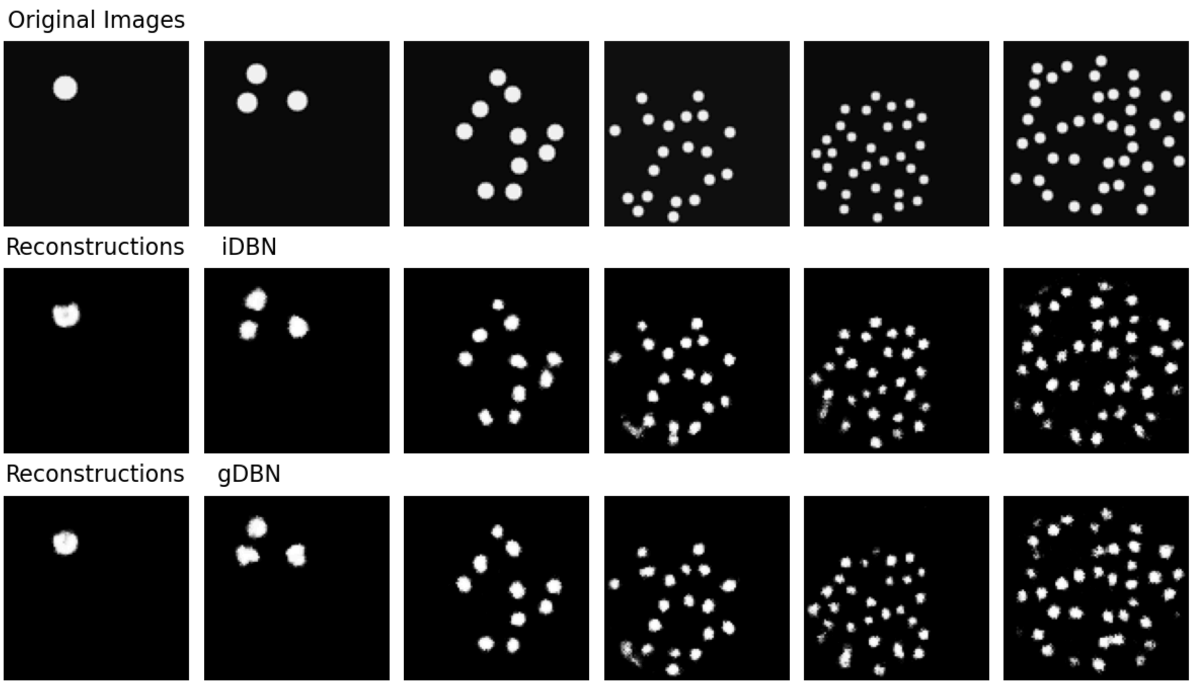

Figure 3.1. Reconstruction examples

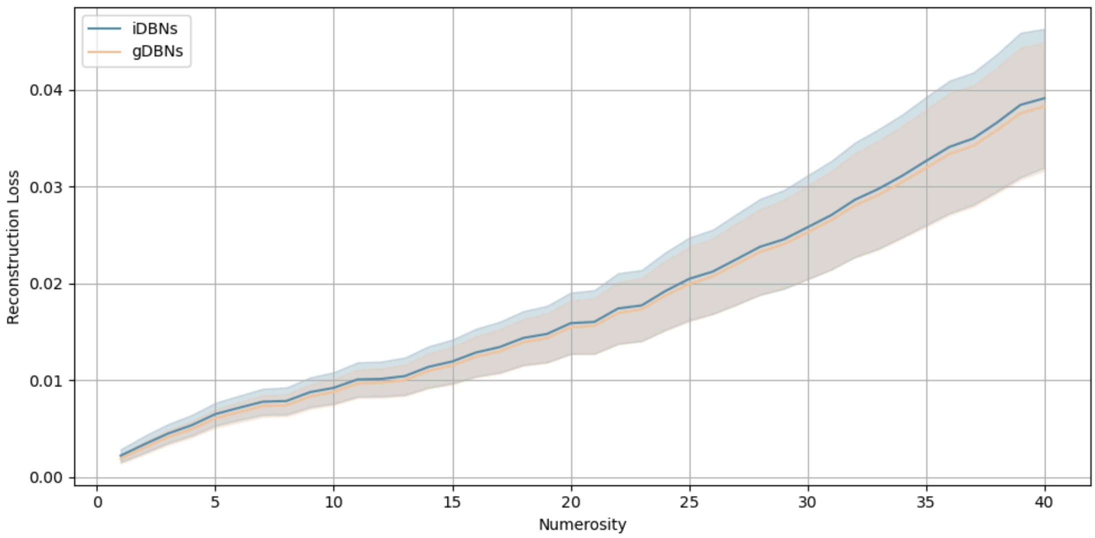

Figure 3.2. Reconstruction loss as a function of numerosity

3.2. Numerosity Tasks Performance

3.2.1. Fixed Reference Comparison

In this task, four possible read-out methods were evaluated and there were four possible references (8, 14, 16, 20). It was also the easiest task for the networks, as indicated by the mean accuracy of 95.8% (SD = 0.01). The GLM analysis of models' choice data revealed that the β_{Num} coefficients across all networks, read-out methods, and references were significantly different from zero (all ts > 39.02, all ps < 0.001), indicating that Numerosity had a strong influence on the models' choice. None of the networks showed unbiased responses with respect to Size (all ts > 4.08, all ps < 0.1), and all but two (all ts < 3.07, all ps > 0.13) for Spacing (all ts > 3.01, all ps < 0.1). The GLM fits for all classifiers and reference conditions can be seen in Figure 3.3, panel A.

Vector analysis revealed that the magnitudes of the projections onto the Numerosity axis were significantly larger than those onto any of the other dimensions (all ts > 60.38, all ps < 0.001), suggesting that Numerosity was the primary feature driving the task performance.

As a measure of number acuity, Weber fraction was calculated from the β_{Num} coefficient to reflect the discriminatory power in terms of numerosity independently from non-numerical features. The mean value obtained was 0.19 (*SD* = 0.02), consistent with those reported in previous human and modelling studies (Halberda & Feigenson, 2008; Testolin et al., 2020).

(A)

(B)

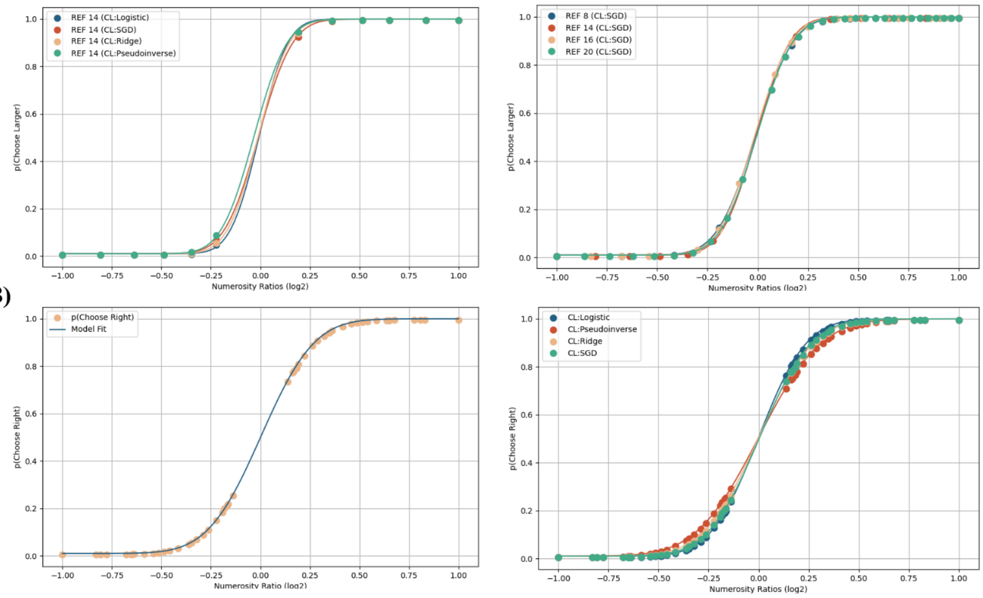

Figure 3.3. GLM fits for the fixed-reference (A) and pairwise (B) comparison tasks

3.2.2. Pairwise Comparison

The second task that the networks were tested on, also the second in terms of the task difficulty (Acc = 0.92. *SD* = 0.02), was pairwise comparison. Again, numerosity seemed to have the biggest impact on the models' choice indicated by the β_{Num} coefficients that were significantly different from zero across all networks and classifiers (all *t*s > 55.6, all *p*s < 0.001), and substantially larger than all the other coefficients. Similar patterns were observed with β_{Size} and $\beta_{Spacing}$ as in the previous task. All networks showed bias for Size (all *t*s > 2.54, all *p*s < 0.1), while in terms of Spacing only five networks showed unbiased performance (all ts < 1.55, all ps > 0.12), while the remaining networks did not (all ts > 1.65, all ps < 0.1).

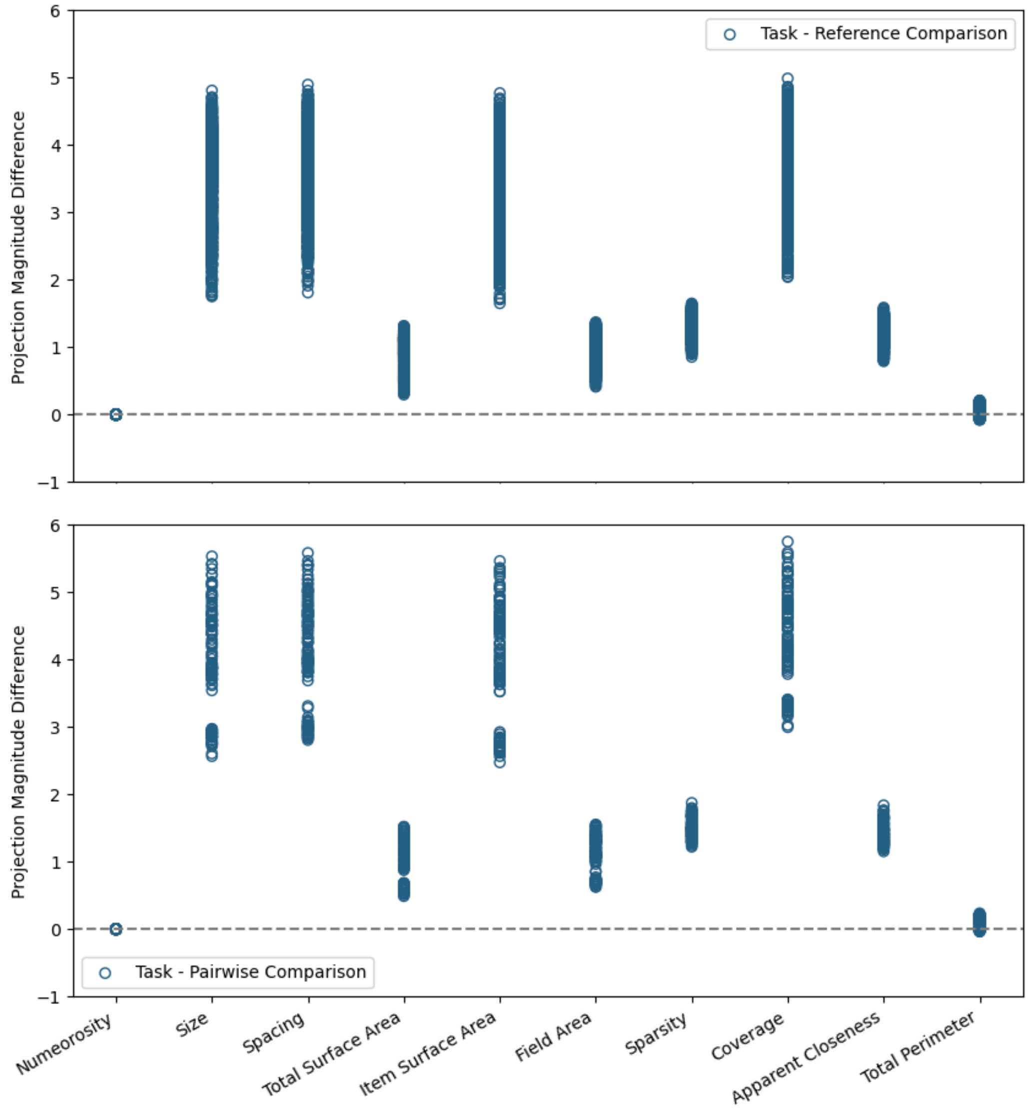

Figure 3.4. Differences between the magnitude of the discrimination vector projected onto the numerosity dimension, and the magnitude of its projection onto each non-numerical dimension.

Notably, the scatter plots of model's coefficients (Figure 3.5, panel B) closely resemble the ones from human participants and networks previously tested by Testolin and colleagues (2020) on the same task and stimuli as the ones used in this study. Once again, the discrimination vector's projection (Figure 3.4) onto the numerosity dimension was significantly larger than its projections onto any of the other dimensions (all ts > 16.86, all ps < 0.001) and the Weber fraction was 0.17 (SD = 0.03).

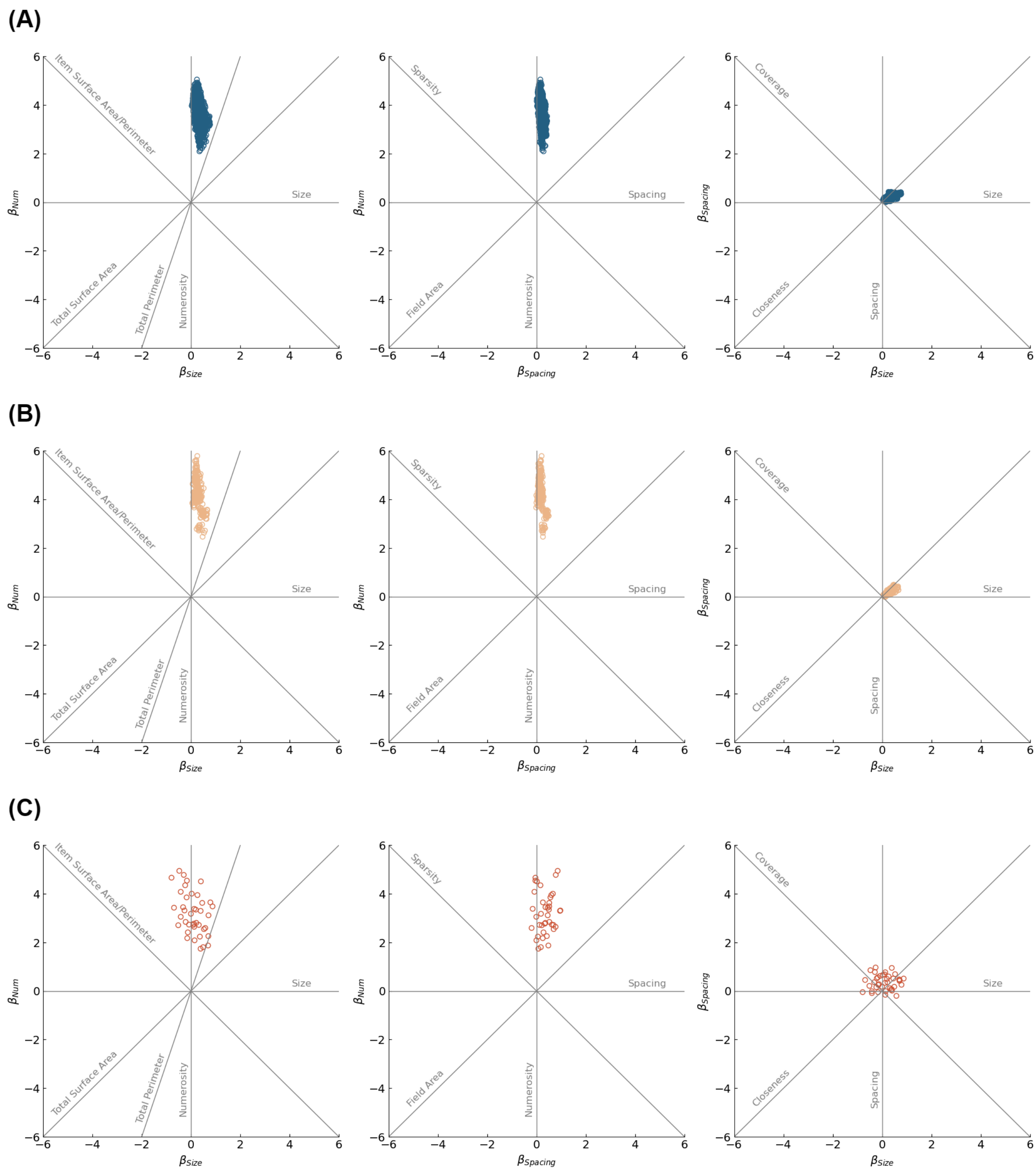

Figure 3.5. Model coefficients from the fixed reference comparison task (A), and pairwise comparison task for both DBNs (B) and human participants (C, adapted from Testolin et al. (2020))

3.2.3. Numerosity Estimation

The final task was the numerosity estimation. Performance on this task in humans is characterized by a couple of notable features. Firstly, humans are almost perfect at recognizing numerosities ranging from one to four, a range referred to as subitizing range (Kaufman et al., 1949), where there is no estimation involved. This evidence also led some researchers to believe that numerosity perception might be subserved by two core systems, the ANS described in the first chapter, and a separate system for precise representation of small quantities (Feigenson et al., 2004). Another characteristic is that as the numerosity increases, subjects' estimates become less accurate and more variable, but the errors stay proportional to the target. This is well captured by the coefficient of variation $cv = \frac{standard\ deviation}{mean}$) that should remain stable across numerosities, in accordance with Weber's law.

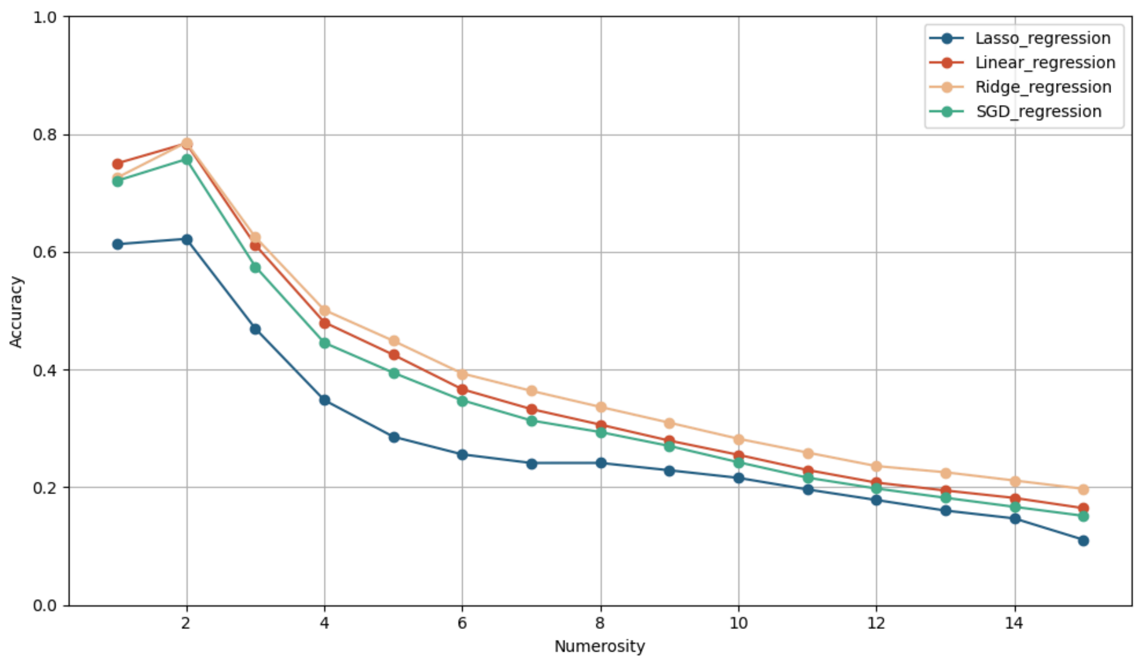

Figure 3.6. Accuracy as a function of numerosity

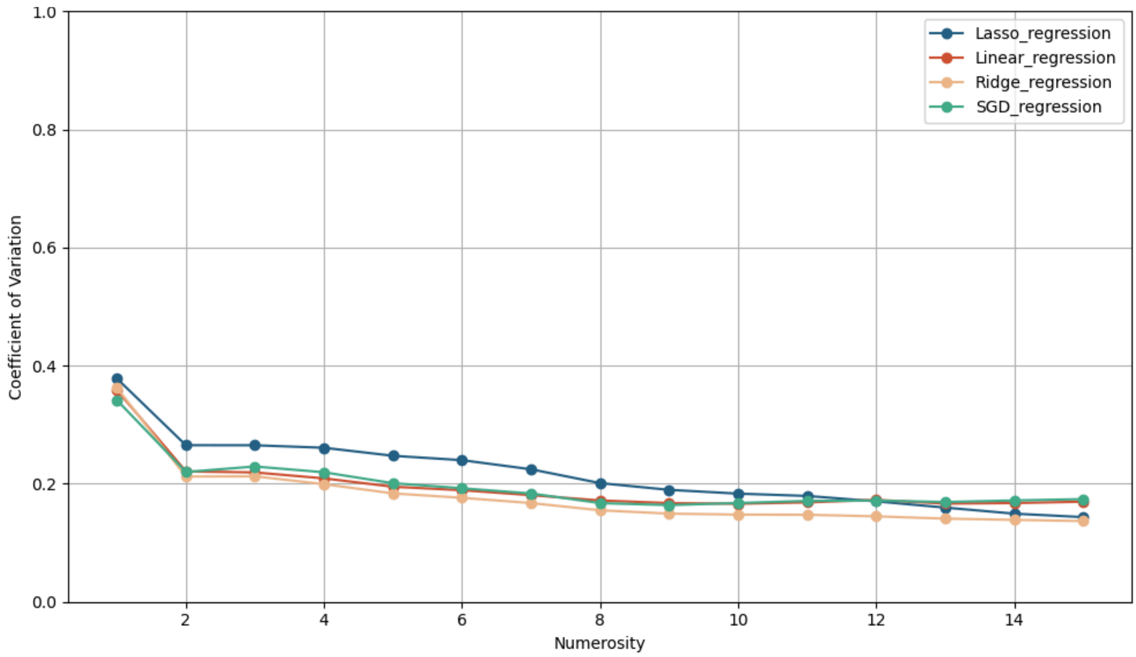

Figure 3.7. Coefficient of variation as a function of numerosity

In this task, numerosities encompassed all values from 1 to 15 and they were log-scaled. The average accuracy across all the network architectures was 35.01% (*SD* = 3.85%), which was substantially higher than the chance level (6.67%) and the one obtained by training the classifiers on the original images without passing them through the networks (max = 12%). The highest accuracy was achieved by a network with 1500 neurons in the first and 2000 neurons in the second hidden layer using ridge regression as the read-out method. This classification approach was also the best performing overall (*M* = 39.12%, *SD* = 0.06), although all the classifiers except lasso regression (*M* = 28.69%, *SD* = 0.01) yielded similar performance (mean range: 35.01% - 39.12%)

As shown in Figure 3.6, the accuracy in the subitizing range was much lower than expected. Especially surprising was that the mean accuracy for images containing a single dot was lower than that for two dots. Apart from the numerosity of one, the coefficient of variation remained relatively stable across the range.

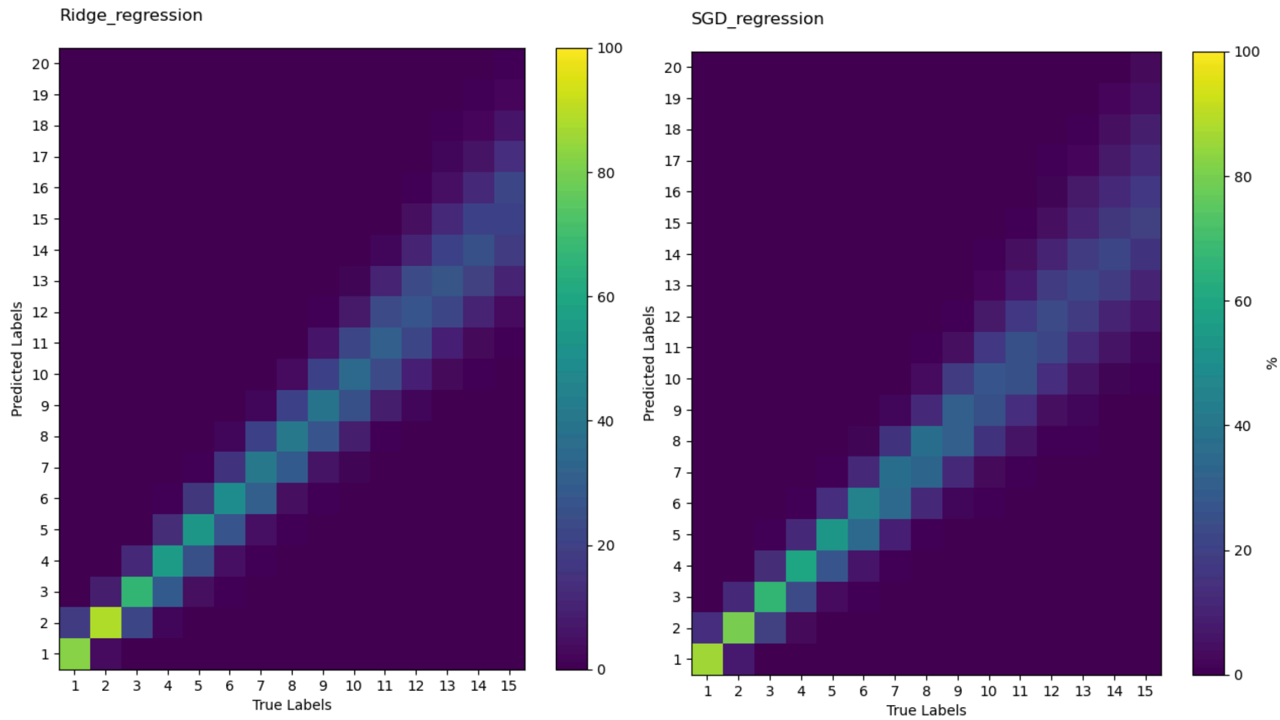

Figure 3.8. Confusion matrices

3.3. Dataset Size Reduction

There was a clear decrease in accuracy as the dataset size was reduced, although the extent varied across classifiers (Figure 3.9, panel A). At 10% of the original dataset, accuracy dropped by around 10% for logistic regression and SGD classifier. In contrast, pseudoinverse and ridge methods were performing substantially worse, possibly due to severe overfitting in limited data conditions. There was also a noticeable increase in Weber fraction, following the same pattern (Figure 3.9, panel B). Models' β_{Num} coefficients (Figure 3.9, panel C) also decreased but nevertheless were significantly different from zero for all classifiers and dataset sizes (all ts > 20.77 all ps < 0.001).

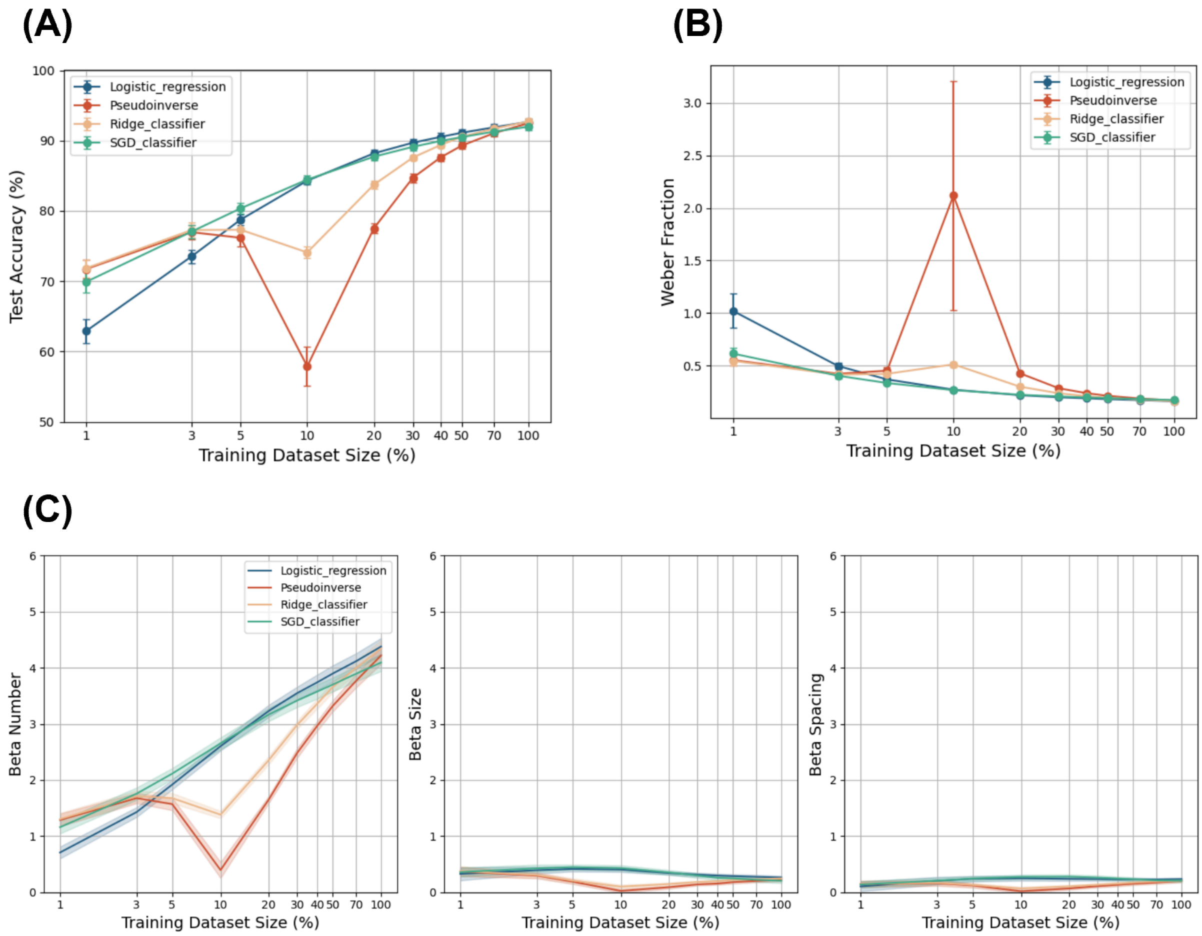

Figure 3.9. Accuracy, Weber fraction, and beta coefficients as functions of dataset size. The x-axis labels are displayed on the linear scale for readability, while the underlying scale is logarithmic.

4. Discussion and Conclusion

This thesis provided in-depth analysis of iteratively trained Deep Belief Networks as computational models of numerosity perception. Across all tasks, models showed accurate performance, mirroring the patterns previously reported in human studies. Furthermore, the performance was evaluated by using simple linear read-outs, suggesting that numerosity emerged as a spontaneously encoded feature, in line with previous findings (Stoianov & Zorzi, 2012; Testolin et al., 2020).

GLM and feature projections analysis showed that numerosity was the primary dimension driving the model response. Nonetheless, the influence of the continuous magnitudes could not be excluded as almost all networks across all tasks showed a certain degree of bias towards non-numerical features. Interestingly, this was also observed in the conditions of reduced supervised training, where even after just 152 patterns, networks performed well above the chance level. Numerosity emerging as the leading feature also in this case further supports the idea that numerical information is already present in internal representations of the networks as a result of unsupervised learning alone.

Finally, the choice of classification method seemed to have a slight impact on performance, especially in the more difficult scenarios such as those involving limited training data. While SGD was not always the best in terms of accuracy, it did seem to be the most robust across the tasks and conditions. Moreover, the use of this method is aligned with the neural network framework and together with the iterative learning algorithm contributes to the efforts of building models of numerosity perception based on biologically plausible mechanisms.

References

1. Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science*, *9*(1), 147–169. https://doi.org/10.1207/s15516709cog0901_7
2. Agrillo, C., Dadda, M., Serena, G., & Bisazza, A. (2008). Do fish count? Spontaneous discrimination of quantity in female mosquitofish. *Animal Cognition*, *11*(3), 495–503. https://doi.org/10.1007/s10071-008-0140-9
3. Burr, D., & Ross, J. (2008). A Visual Sense of Number. *Current Biology*, *18*(6), 425–428. https://doi.org/10.1016/j.cub.2008.02.052
4. Cantlon, J. F. (2012). Math, monkeys, and the developing brain. *Proceedings of the National Academy of Sciences*, *109*(1), 10725–10732. https://doi.org/10.1073/pnas.1201893109
5. Cantlon, J. F., Brannon, E. M., Carter, E. J., & Pelphrey, K. A. (2006). Functional Imaging of Numerical Processing in Adults and 4-y-Old Children. *PLoS Biology*, *4*(5), e125. https://doi.org/10.1371/journal.pbio.0040125
6. Castaldi, E., Pomè, A., Cicchini, G. M., Burr, D., & Binda, P. (2021). The pupil responds spontaneously to perceived numerosity. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-26261-4
7. Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, *148*, 163–172. https://doi.org/10.1016/j.actpsy.2014.01.016
8. Clearfield, M. W., & Mix, K. S. (1999). Number Versus Contour Length in Infants' Discrimination of Small Visual Sets. *Psychological Science*, *10*(5), 408–411. https://doi.org/10.1111/1467-9280.00177
9. Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. OUP, USA.
10. DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265. https://doi.org/10.1016/j.cognition.2015.05.016
11. Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, *18*(18), R855–R856. https://doi.org/10.1016/j.cub.2008.07.053

12. Feigenson, L., Carey, S., & Spelke, E. (2002a). Infants' Discrimination of Number vs. Continuous Extent. *Cognitive Psychology*, *44*(1), 33–66. https://doi.org/10.1006/cogp.2001.0760
13. Feigenson, L., Carey, S., & Spelke, E. (2002b). Infants' Discrimination of Number vs. Continuous Extent. *Cognitive Psychology*, *44*(1), 33–66. https://doi.org/10.1006/cogp.2001.0760
14. Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002
15. Gebuis, T., & Reynvoet, B. (2012). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, *141*(4), 642–648. https://doi.org/10.1037/a0026218
16. Gebuis, T., & Van Der Smagt, M. J. (2011). False Approximations of the Approximate Number System? *PLoS ONE*, *6*(10), e25405. https://doi.org/10.1371/journal.pone.0025405
17. Gerstmann, J. (1940). Syndrome of Finger Agnosia, Disorientation for Right and Left, Agraphia and Acalculia: Local Diagnostic Value. *Archives of Neurology & Psychiatry*, *44*(2), 398. https://doi.org/10.1001/archneurpsyc.1940.02280080158009
18. Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465. https://doi.org/10.1037/a0012682
19. Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246
20. Henik, A., Gliksman, Y., Kallai, A., & Leibovich, T. (2017). Size Perception and the Foundation of Numerical Processing. *Current Directions in Psychological Science*, *26*(1), 45–51. https://doi.org/10.1177/0963721416671323
21. Hinton, G. E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, *14*(8), 1771–1800. https://doi.org/10.1162/089976602760128018
22. Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

23. Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, *20*(2), 82–88. https://doi.org/10.1016/j.lindif.2009.07.004
24. Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The Discrimination of Visual Number. *The American Journal of Psychology*, *62*(4), 498. https://doi.org/10.2307/1418556
25. Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5
26. Kutter, E. F., Bostroem, J., Elger, C. E., Mormann, F., & Nieder, A. (2018). Single Neurons in the Human Brain Encode Numbers. *Neuron*, *100*(3), 753-761.e4. https://doi.org/10.1016/j.neuron.2018.08.036
27. Lee, H., Ekanadham, C., & Ng, A. Y. (2008). *Sparse deep belief net model for visual area V2*. *20*, 873–880.
28. Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From “sense of number” to “sense of magnitude”: The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*, e164. https://doi.org/10.1017/S0140525X16000960
29. Lipton, J. S., & Spelke, E. S. (2004). Discrimination of Large and Small Numerosities by Human Infants. *Infancy*, *5*(3), 271–290. https://doi.org/10.1207/s15327078in0503_2
30. Merten, K., & Nieder, A. (2009). Compressed Scaling of Abstract Numerosity Representations in Adult Humans and Monkeys. *Journal of Cognitive Neuroscience*, *21*(2), 333–346. https://doi.org/10.1162/jocn.2008.21032
31. Newell, A. (1973). You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium. *Machine Intelligence*, 121–146.
32. Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the Quantity of Visual Items in the Primate Prefrontal Cortex. *Science*, *297*(5587), 1708–1711. https://doi.org/10.1126/science.1072493
33. Nieder, A., & Miller, E. K. (2003). Coding of Cognitive Magnitude. *Neuron*, *37*(1), 149–157. https://doi.org/10.1016/S0896-6273(02)01144-3
34. Nieder, A., & Miller, E. K. (2004). Analog Numerical Representations in Rhesus Monkeys: Evidence for Parallel Processing. *Journal of Cognitive Neuroscience*, *16*(5), 889–901. https://doi.org/10.1162/089892904970807
35. Piantadosi, S. T., & Cantlon, J. F. (2017). True Numerical Cognition in the Wild. *Psychological Science*, *28*(4), 462–469. https://doi.org/10.1177/0956797616686862

36. Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning Curves for Approximate Numerosity in the Human Intraparietal Sulcus. *Neuron*, *44*(3), 547–555. https://doi.org/10.1016/j.neuron.2004.10.014
37. Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex. *Neuron*, *53*(2), 293–305. https://doi.org/10.1016/j.neuron.2006.11.022
38. Ross, J. (2010). Vision senses number directly. *Journal of Vision*, *10*(2), 1–8. https://doi.org/10.1167/10.2.10
39. Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, *347*(6221), 534–536. https://doi.org/10.1126/science.aaa1379
40. Santens, S., Roggeman, C., Fias, W., & Verguts, T. (2010). Number Processing Pathways in Human Parietal Cortex. *Cerebral Cortex*, *20*(1), 77–88. https://doi.org/10.1093/cercor/bhp080
41. Starkey, P., & Cooper, R. G. (1980). Perception of Numbers by Human Infants. *Science*, *210*(4473), 1033–1035. https://doi.org/10.1126/science.7434014
42. Starr, A., DeWind, N. K., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition*, *168*, 222–233. https://doi.org/10.1016/j.cognition.2017.07.004
43. Stoianov, I., & Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nature Neuroscience*, *15*(2), 194–196. https://doi.org/10.1038/nn.2996
44. Strauss, M. S., & Curtis, L. E. (1981). *Infant Perception of Numerosity*.
45. Testolin, A., Dolfi, S., Rochus, M., & Zorzi, M. (2020). Visual sense of number vs. sense of magnitude in humans and machines. *Scientific Reports*, *10*(1), 10045. https://doi.org/10.1038/s41598-020-66838-5
46. Wynn, K. (1998). Psychological foundations of number: numerical competence in human infants. *Trends in Cognitive Sciences*, *2*(8), 296–303. https://doi.org/10.1016/S1364-6613(98)01203-0
47. Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11. https://doi.org/10.1016/S0010-0277(99)00066-9

48. Zambra, M., Testolin, A., & Zorzi, M. (2023). A Developmental Approach for Training Deep Belief Networks. *Cognitive Computation*, *15*(1), 103–120. https://doi.org/10.1007/s12559-022-10085-5
49. Zorzi, M., & Testolin, A. (2018). An emergentist perspective on the origin of number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170043. https://doi.org/10.1098/rstb.2017.0043
50. Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00515