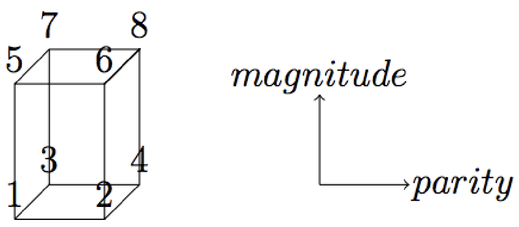
Investigating the structure of abstraction in neural networks with population-level hypothesis testing

Anastasia Dmitrienko[1], Sean R. Bittner[2], John P. Cunningham[1]
[1]Department of Statistics and [2]Department of Neuroscience, Columbia University

Motivation

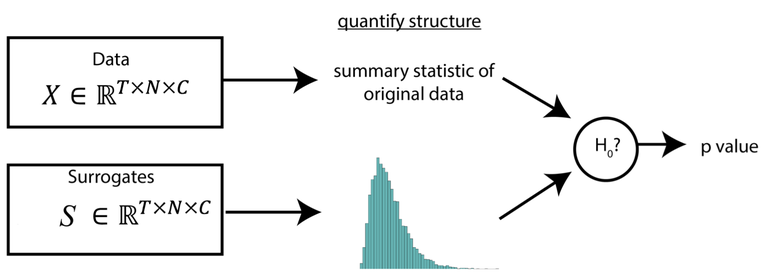
- Research in statistical neuroscience focuses on finding structure in large-scale neural recordings.
- Theoretical models of brain computation are designed to reflect this neural population structure.
- A methodological bottleneck is generating of surrogate neural data for population-level hypothesis testing.
- We use methods from Elsayed et al. (2017) [1] to investigate if primary features give rise to geometrical abstraction in neural networks classifying both magnitude and parity of MNIST digits (Bernardi et al. (2018) [2]).



- Geometrical abstraction is the brain's representation of abstract variables to enable conceptual generalization.

Population-level hypothesis testing

- Mechanics of a hypothesis test:



- We use **tensor maximum entropy** (TME) to control for primary features of neural data.
- TME samples surrogate datasets $S \in \mathbb{R}^{T\times N\times C}$ from a probability distribution $p(S)$ that maximizes Shannon entropy with the average primary features of the data.
- Maximum entropy objective:

$$\hat{p}(S) = \underset{p(S)}{\operatorname{argmax}} - \int p(S) log(p(S)) dS$$

- Constraints:

$$\int p(S)dS = 1 \qquad E_p[S] = M$$

$$\bar{S} = S - M$$

$$\Sigma_T = E_p\Big[\sum_{n=1}^{N}\sum_{c=1}^{C} \bar{S}(:,n,c)\bar{S}(:,n,c)^T\Big]$$

*This expectation is analogous for N and C dimensions.

- Additionally, we used the corrected Fisher randomization method (CFR) as a control, which also preserves the tensor-marginal means and covariances.

Cognitive abstraction

- To perform complex tasks, neuronal ensembles must represent multiple variables simultaneously.
- A serial reversal learning task requires inferring how to switch back and forth between two contexts defined by sets of variable mappings (figure below).
- Bernardi et al. (2018) [2] trained monkeys to perform this task, and analyzed the geometrical representation of
  - task context
  - operant responses
  - reinforcement outcomes

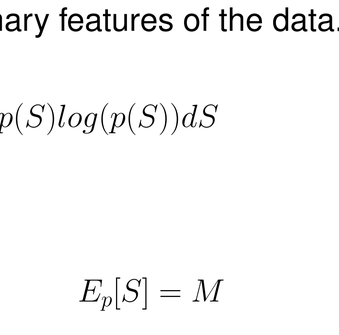


Figure credit: Fig. 1b from Bernardi et al. (2018) [2]

- An analysis of the geometrical structure of the points representing the experimental conditions revealed an abstract representation of each variable, separate from randomly distributed representations in firing rate space.

Abstraction metrics

CCGP

(Cross Conditional Generalization Performance)

- Ability of linear readout to generalize multiple variables simultaneously

$$\text{CCGP}(X) = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{a_{ij}}{16}$$

a_{ij} is test set accuracy for train set i from dichotomy side 1 and train set j from side 2

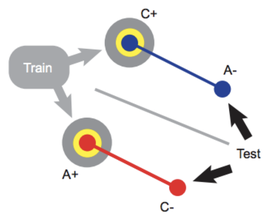


Figure credit: Fig. 4a from Bernardi et al. (2018) [2]

Parallelism Score

- The degree to which the coding directions determined when training a decoder are parallel for different sets of training conditions

$$\text{PS}(X) = \max_{y} \sum_{i=1}^{4}\sum_{j>i}^{4} \frac{\cos(\theta_{i,j})}{6}$$

Maximum sum of cosines between all planes i and j for y, where y is one of 24 dichotomy pairings

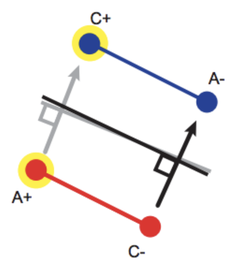
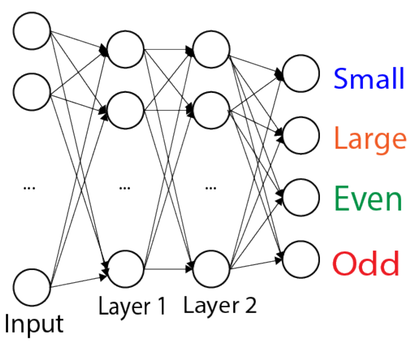


Figure credit: Fig. 4b from Bernardi et al. (2018) [2]

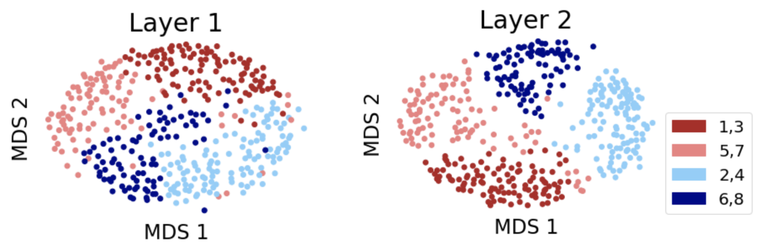
Multi-tasking networks

- To mimic the computational process required in the serial-reversal learning task, Bernardi et al. (2018) [2] trained a 2-layer neural network to classify MNIST digits by both magnitude and parity.
  - 100 neurons per hidden layer (tanh activation, Adam optimizer)
  - Cross entropy was minimized for two classification tasks, parity and magnitude, simultaneously.

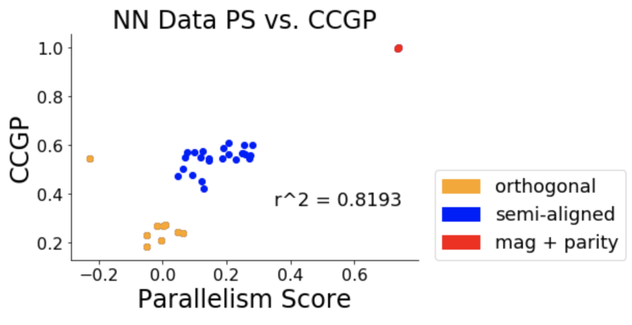


Structure of multi-tasking network activity

- We aim to measure the level of abstraction across all possible abstract variables, or dichotomies of digits.
  - 35 possible balanced dichotomies, from the ways to split 8 digits into 2 groups of 4: $\binom{8}{4}/2$.
  - **Orthogonal** dichotomies are balanced in magnitude and parity (2 small/large, 2 even/odd digits)
  - **Semi-aligned** dichotomies are imbalanced



- Digits of different parities and magnitudes are more separated in Layer 2 than Layer 1. This projection reveals orthogonal coding directions for the two variables.
- Orthogonal dichotomies have lowest parallelism score and CCGP for neural network output (Layer 2).

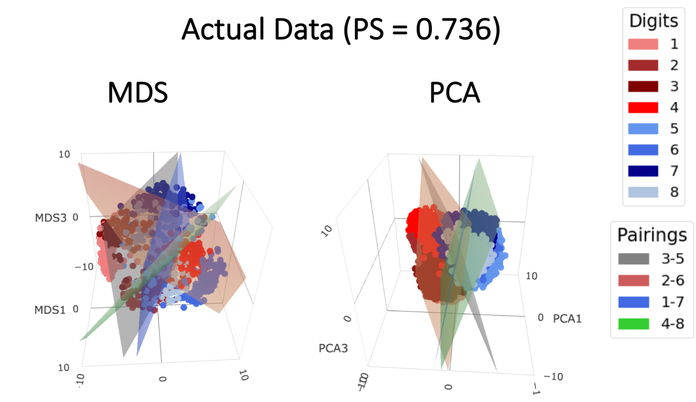


- Orthogonal dichotomies are least similar to the magnitude and parity baselines, representing half of the two trained dichotomies, so their linear separability is lower than others.
- Thus, dichotomies the network wasn't trained to classify are not represented abstractly.

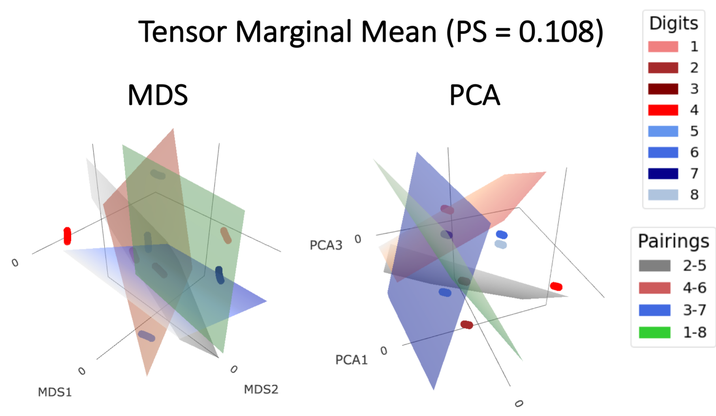
Results

Parallelism in magnitude dichotomy

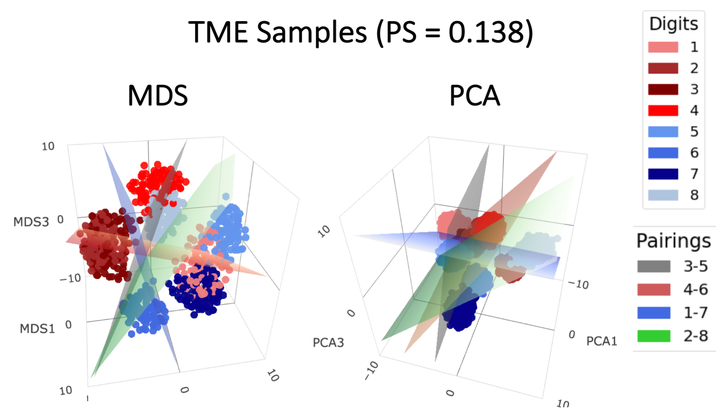
- For the best pairing of conditions, planes are relatively parallel for the dichotomy, resulting in a high PS.



- The tensor marginal mean lacks parallel structure in this (and any) dichotomy.

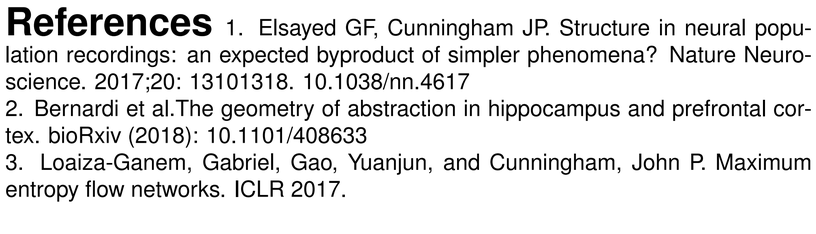


- Like TMM, the TME sample data has similar clustering and skewed hyperplanes.



Controlling for primary features

- When examining the 4 pairings of conditions that minimizes angles between all 4 hyperplanes, the PS for neural network data is higher. This is because the NN data is clustered by the trained abstract variables and this improves linear separability.
- However, the surrogate data does not contain this population structure, so the digit clusters are distributed randomly in 100-D space.



*Each data set is ordered according to its own dichotomy rank.

- Neural correlation structure and tuning across conditions are not enough to account for degree of abstraction measured in neural networks by CCGP and PS.
- The trained dichotomies in NN data are significantly higher for CCGP and PS than in TME and CFR data.

Discussion

- Our testing across primary features of the neural readout using TME and CFR suggests that the surrogate datasets **do not** have the same higher-order structure as the neural network responses
- This gives greater contextualization to the statistical significance of the conclusions by Bernardi, et al. (2018) [2]
- Future research can investigate which non-linear metrics, such as pairwise distances, give rise to abstracted neural representations.

References

1. Elsayed GF, Cunningham JP. Structure in neural population recordings: an expected byproduct of simpler phenomena? Nature Neuroscience. 2017;20: 13101318. 10.1038/nn.4617
2. Bernardi et al.The geometry of abstraction in hippocampus and prefrontal cortex. bioRxiv (2018): 10.1101/408633
3. Loaiza-Ganem, Gabriel, Gao, Yuanjun, and Cunningham, John P. Maximum entropy flow networks. ICLR 2017.

Code can be found online at github.com/admitrienko/MNIST-Abstraction-Testing