Differentiable optimization of similarity scores between models and brains

Nathan Cloos (nacloos@mit.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA

Markus Siegel (markus.siegel@uni-tuebingen.de)
University of Tübingen, Tübingen, Germany

Scott L. Brincat (sbrincat@mit.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA

Earl K. Miller (ekmiller@mit.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA

Guangyu Robert Yang (yanggr@mit.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA

Christopher J. Cueva (ccueva@gmail.com)
Massachusetts Institute of Technology, Cambridge, MA, USA

What metrics should guide the development of more realistic models of the brain? One proposal is to quantify the similarity between models and brains using methods such as linear regression, Centered Kernel Alignment (CKA), and Procrustes distance. To better understand the limitations of these similarity measures we analyze neural activity recorded in five experiments on nonhuman primates, and optimize synthetic datasets to become more similar to these neural recordings. How similar can these synthetic datasets be to neural activity while failing to encode task relevant variables? We find that some measures like linear regression and CKA, differ from Procrustes distance, and yield high similarity scores even when task relevant variables cannot be linearly decoded from the synthetic datasets. Synthetic datasets optimized to maximize similarity scores initially learn the first principal component of the target dataset, but Procrustes distance captures higher variance dimensions much earlier than methods like linear regression and CKA. We show in both theory and simulations how these scores change when different principal components are perturbed. And finally, we jointly optimize multiple similarity scores to find their allowed ranges, and show that a high Procrustes similarity, for example, implies a high CKA score, but not the converse.

Keywords: Similarity analysis; Methods; Neural recordings

Introduction

In this work we study several popular methods that have been proposed to quantify the similarity between models and neural data, in particular, linear regression (Yamins et al., 2014; Schrimpf et al., 2018), Centered Kernel Alignment (CKA) (Kornblith et al., 2019), and angular Procrustes distance (Williams et al., 2021; Ding et al., 2021). We analyzed neural data from five studies on nonhuman primates, but have included only two here for space (Figure 1). In order to study what drives high similarity scores we directly optimize the synthetic datasets to maximize their similarity to the neural datasets as assessed by different methods, for example, linear regression, CKA, or angular Procrustes distance.

Comparing similarity scores across studies is challenging, primarily due to variability in naming and implementation conventions. As part of our contribution to the research community we have created, and are continuing to develop, a Python package that benchmarks and standardizes similarity measures[1].

Results

High similarity scores do not guarantee encoding of task relevant variables: We start by asking if synthetic datasets with high similarity scores relative to the neural data, encode task relevant variables, for example, the stimulus features or the response of the monkey, in the same way as the neural data. More specifically, is it possible for the synthetic datasets to have a high similarity score while failing to encode task relevant variables?

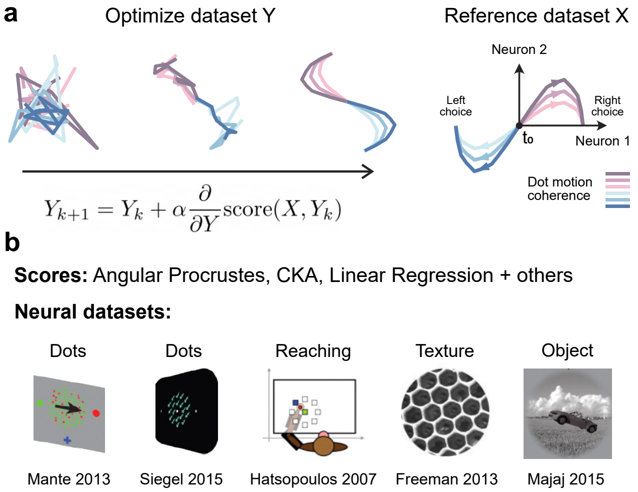

Figure 1: (a) To better understand the properties of similarity measures we optimize synthetic datasets to become more similar to a reference dataset, for example, neural recordings. (b) We analyzed similarity scores between artificial datasets and electrode recordings from five experiments on nonhuman primates. For space we present results from two neural datasets from prefrontal cortex (PFC) (Mante et al., 2013) and ventral stream V4 (Siegel et al., 2015) in monkeys performing an experimental task that required the animal to attend to either color or motion information while ignoring the non-cued feature of the stimuli. On each trial, a field of colored moving dots is shown. Monkeys are given a cue at the beginning of the trial to determine whether the dots in the stimulus are moving left vs right, or are red vs green. The monkey reported its choice with a saccade to one of two visual targets. In both datasets, we analyzed neural activity taken when the dot stimulus was presented.

Surprisingly we find that for linear regression and CKA the answer is yes, a high similarity score does not necessarily mean the synthetic datasets encode task relevant variables like the neural data. Figures 2a and 2b show the decode accuracy of a linear classifier trained to decode task relevant variables (cross-validated across different conditions) as the similarity score increases. Before optimization, the synthetic datasets initially consisted of Gaussian noise and the decode accuracy was near the baseline chance level of 0.5 as expected for the binary classifier used in this analysis.

Consider the synthetic data optimized towards the Siegel 2015 neural recordings using CKA similarity (second row and column in Figure 2). When the synthetic dataset has a high similarity score of 0.9 the decode accuracy for all the task variables is still less than that found in the neural activity (horizontal dashed lines). This is in contrast to the case where synthetic data is optimized to maximize angular Procrustes similarity (first column) and a similarity score of 0.9 yields a

[1] https://anonymous.4open.science/r/similarity-repository-03D3

dataset that encodes task variables to the same degree as the neural recordings. Note that both CKA and angular Procrustes have the same similarity scale ranging between 0 and 1 (perfect similarity).

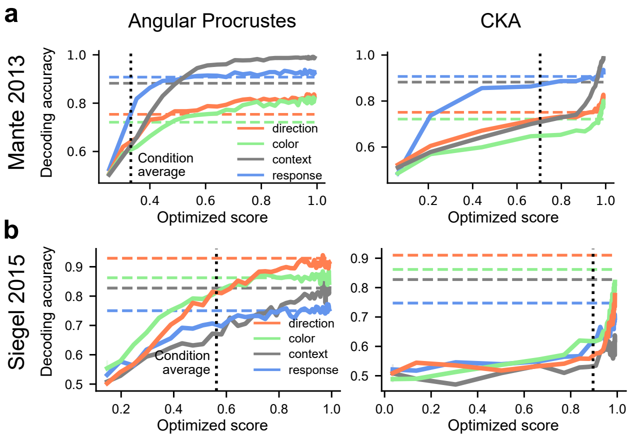

Figure 2: (a, b) Decode accuracy for experimental variables versus similarity scores. Decode is from synthetic data optimized towards greater similarity with the neural data from (a) Mante et al. (2013) and (b) Siegel et al. (2015). Horizontal dashed lines indicate the decode accuracy from the neural data.

Optimization dynamics of similarity scores: How much of the neural data must be captured by a synthetic dataset or model before the decode accuracy reaches the level seen in the neural dataset itself? One perspective on this question is to decode the task variables from neural data after projecting onto principal components 1 through N, where principal component 1 captures the most variance. In order to capture all the information about the task variables, at least several principal components must be included in the decode (analysis not shown). This motivates the following hypothesis. Perhaps the reason that CKA similarity scores can be so high while the synthetic data fails to encode task variables is because these similarity measures preferentially rely on the top few principal components.

We explore this hypothesis in the following set of analyses with a synthetic dataset based on the neural recordings from Mante et al. 2013. Figure 1a shows the reference dataset. We can think of this reference dataset as a low-dimensional neural trajectory summarizing the population activity of many neurons, or alternatively, as the firing rates of two neurons over time (shown here encoding the two task variables of choice and dot motion coherence), recorded during six different experimental conditions, with the color in Figure 1a denoting the condition. Figure 3b shows the transformation of an initially random Gaussian noise dataset as it is optimized to maximize either the angular Procrustes or CKA similarity score with respect to the reference dataset. The score increases from an initial value near 0 to a maximum near 1 as optimization progresses, with the insets at the top of the figure showing the optimized noise dataset at various points during this procedure. The yellow curve shows how well the optimized dataset captures the first principal component of the reference dataset, as quantified by R^2, throughout optimization. Notice that the second principal component, shown in purple, is only captured at a much higher optimization score for CKA versus angular Procrustes.

We show in both theory and simulations how these scores change when different principal components are perturbed (Figure 3c). And finally, we jointly optimize multiple similarity scores to find their allowed ranges, and show that a high Procrustes similarity, for example, implies a high CKA score, but not the converse (Figure 3d).

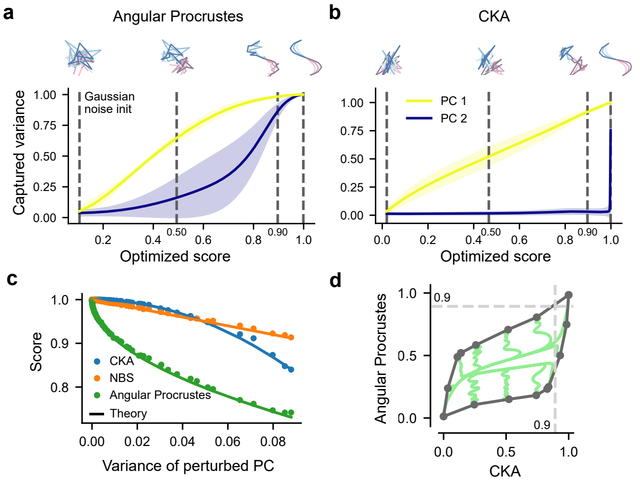

Figure 3: Different similarity measures differentially prioritize learning principal components of the data. (a, b) Gaussian random noise data is updated to maximize similarity with the reference dataset from Figure 1a, as quantified by one of the similarity measures. The transformation of the random noise dataset is shown at the top of the panels. The first principal component of the reference dataset is increasingly well captured by the optimized data as the similarity scores increase (yellow curves). The second, lower variance, component is also learned when maximizing the angular Procrustes similarity but is only captured at high similarity scores when maximizing CKA and linear regression (not shown). (c) The similarity score between a synthetic dataset and a modified version of the same dataset when a single principal component is perturbed. For all similarity measures, when small variance components are perturbed the similarity score is near 1. However, when high variance components are perturbed the Angular Procrustes score drops much more than for CKA and Normalized Bures Similarity (NBS) (Tang et al., 2020). (d) We jointly optimized the values of both angular Procrustes and CKA to illustrate the allowed ranges of both similarity scores (region enclosed by the solid lines). If angular Procrustes has a high score of 0.9 (horizontal dashed line) then CKA will have a value above this. In contrast, a high CKA score of 0.9 (vertical dashed line) does not imply a high angular Procrustes score, and a wide range of angular Procrustes scores are possible.

References

Ding, F., Denain, J.-S., & Steinhardt, J. (2021). Grounding representation similarity through statistical testing. In *Advances in neural information processing systems* (Vol. 34).

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). *Similarity of neural network representations revisited.* arXiv.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013, Nov). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*, 78–84.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, *348*(6241), 1352-1355. doi: 10.1126/science.aab0551

Tang, S., Maddox, W. J., Dickens, C., Diethe, T., & Damianou, A. (2020). *Similarity of neural networks with gradients.*

Williams, A. H., Kunz, E., Kornblith, S., & Linderman, S. W. (2021). Generalized shape metrics on neural representations. In *Advances in neural information processing systems* (Vol. 34).

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624. doi: 10.1073/pnas.1403112111