Dynamic Changes in Neural Manifolds During Working Memory

Santo-Angles A[1], Gyurkovics M[2,3], Palva M[4], Thut G[2], Palva S[1]
1. University of Helsinki (Finland), 2. University of Glasgow (UK), 3. University of East Anglia, Norwich (UK), 4. Aalto University (Finland)

Abstract

How cognitive operations supporting working memory, the ability to actively hold and manipulate information, are implemented through neural computations remains a topic of ongoing debate. Neuronal recordings in non-human primates suggest that working memory contents are encoded and maintained in low-dimensional representations, known as neural manifolds, with cognitive operations occurring through dynamic changes within these manifolds. In the present study, we tested this hypothesis using magneto- and encephalography data from human participants collected during a working memory task. In the task, a retro-cue required maintaining, inhibiting or updating memory representations of two stimulus features: orientation and shape. We found that the size of neural subspaces encoding stimuli during the delay period dynamically changed with task demands, expanding or shrinking based on shifts in the relevance of stimulus features. These findings support the role of manifold dynamics in working memory.

Keywords: working memory; MEG; EEG; neural geometry; manifolds

INTRODUCTION

The neural systems involved in working memory and the cognitive operations they support are well-established (D'Esposito & Postle, 2015; Sreenivasan & D'Esposito, 2019). However, the neural mechanisms by which these operations are implemented remain a topic of debate (Barbosa et al., 2020; Miller et al., 2018). Recent evidence suggests that cognitive operations are implemented through the dynamics of low-dimensional neural manifolds. For example, multiple items of the same stimulus feature can be maintained simultaneously without interference because its neural representations are encoded in orthogonal subspaces (Panichello & Buschman, 2021; Xie et al., 2022). In this study, we explore the geometry of working memory representations when two distinct stimulus features, orientations and shapes, are maintained simultaneously. More precisely, we asked how inhibition or updating of these representations reshapes neural manifolds.

METHODS

We invited 50 healthy participants (28 females, 25±5 years old, all right-handed) to perform a working memory task (Figure 1A). Subjects had to hold in mind all the presented stimuli, gratings and polygons, until the retro-cue indicated which features would be probed after a delay. In the control condition, all features remained relevant, which could be gratings (orientation), polygons (shape) or both simultaneously. In the inhibition condition, only one feature, either orientation or shape, remained relevant, while the other became irrelevant. In the updating condition, one feature remained relevant, and one of its items had to be replaced with a new stimulus presented alongside the cue. All conditions involved presenting either one or two stimuli per feature, with analyses averaged across stimulus loads. We recorded concurrent magnetoencephalography (MEG) and electroencephalography (EEG) as participants performed the task. MEG-EEG data was preprocessed (Ferrante et al., 2022), source modelled with minimum-norm estimate (Gramfort et al., 2014), and source timeseries were parcellated into 400 brain parcels (Schaefer et al., 2018) with fidelity weighting (Korhonen et al., 2014).

We constructed subject- and condition-specific neural activity matrices X (M × N), where M=7 stimulus (4 orientations, 3 shapes) and N=400 parcels. Each matrix captured stimulus-specific time-domain activity during the delay periods, estimated using Lasso regression ($\lambda = 0.01$) applied independently for each channel and stimulus feature, using only trials with correct responses. Feature-specific matrices were then merged into a joint X matrix (7 × 400). To account for individual variability, acknowledging that shared manifolds may be implemented or sampled differently across subjects, we aligned the joint X matrices into a common space to minimize representational misalignments (Barbosa et al., 2025; Haxby et al., 2020).

We reduced the channel dimensionality of the joint X matrix using PCA and projected the data onto the top k=3 components, yielding a Z matrix (7 conditions × 3 components). This was split by feature into $Z_{orientation}$ and Z_{shape}. For each feature, we applied a second PCA to define the best-fitting plane by selecting the top two eigenvectors. Subspace alignment was quantified by the principal angle (PA) between planes, and stimulus separability by Euclidean distances between stimuli within each subspace. We used these metrics to assess the effect of feature load across conditions, and the effect of cue within each condition, by means of dependent-samples t-test. P-values were FDR corrected.

RESULTS

Orientation and shape representations laid in oblique neural subspaces during encoding (PA: 56° ± 21°) and maintainance (PA: 55° ± 23°) (Figure 1B). We found no effects of load or cue in the PA. Separability between stimuli of the same feature were significantly reduced when two stimulus features (orientation and shapes) were presented simultaneously (Figure 1C). We found no effect of cue in the control condition. In the inhibition condition, separability increased for the relevant feature after the cue, while it decreased for the non-relevant feature (Figure 1D). In the updating condition, separability also increased for the relevant feature after the cue, but no effect was observed for the non-relevant feature (Figure 1E).

CONCLUSION

Neural activity supporting working memory was confined to a lower-dimensional neural manifold, consistent with neuronal recordings in non-human primates (Jahn et al., 2024; Panichello & Buschman, 2021; Tian et al., 2024; Xie et al., 2022). We observed that neural subspaces for orientation and shape occupied oblique subspaces but did not dynamically shift with task contingencies, suggesting that the alignment between subspaces remained stable, likely due to the processing of these features in distinct cortical circuits (Haque et al., 2024). In contrast, we found that the size of subspaces changed with task demands. When multiple stimulus features were presented simultaneously, each subspace occupied a relatively smaller portion of the representational space, likely due to size limitations within the space. Additionally, we observed that feature subspaces did not remain static over time but expanded or shrank depending on their relevance, supporting the neural manifold hypothesis, according to which neural computations are carried out throughout dynamical changes in the low-dimensional manifolds (Langdon et al., 2023; Thibeault et al., 2024).

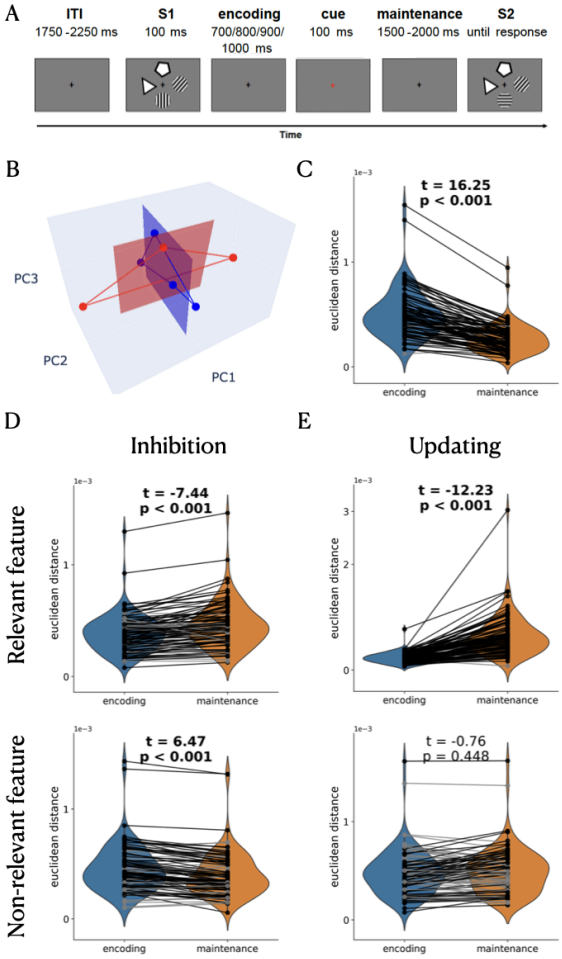

Figure 1. A) Task design. B) Neural subspaces representing orientation (blue) and shape (red) for one subject. C) Effect of feature load during encoding period in control condition. D-E) Effect of cue in inhibition (D) and updating (E) for relevant (top) and non-relevant features (bottom). Black lines connect observations of the same participant aligned with the group; grey lines indicate opposing effects.

References

Barbosa, J., Nejatbakhsh, A., Duong, L., Harvey, S. E., Brincat, S. L., Siegel, M., Miller, E. K., & Williams, A. H. (2025). Quantifying differences in neural population activity with shape metrics. In *bioRxiv*. https://doi.org/10.1101/2025.01.10.632411

Barbosa, J., Stein, H., Martinez, R. L., Galan-Gadea, A., Li, S., Dalmau, J., Adam, K. C. S., Valls-Solé, J., Constantinidis, C., & Compte, A. (2020). Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature Neuroscience*, *23*(8), 1016–1024.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115–142.

Ferrante, O., Liu, L., Minarik, T., Gorska, U., Ghafari, T., Luo, H., & Jensen, O. (2022). FLUX: A pipeline for MEG analysis. *NeuroImage*, *253*(119047), 119047.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460.

Haque, H., Wang, S. H., Siebenhühner, F., Robertson, E., Palva, J. M., & Palva, S. (2024). Synchronization networks reflect the contents of visual working memory. In *Research Square*. https://doi.org/10.21203/rs.3.rs-3853906/v1

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*. https://doi.org/10.7554/eLife.56601

Jahn, C. I., Markov, N. T., Morea, B., Daw, N. D., Ebitz, R. B., & Buschman, T. J. (2024). Learning attentional templates for value-based decision-making. *Cell*, *187*(6), 1476–1489.e21.

Korhonen, O., Palva, S., & Palva, J. M. (2014). Sparse weightings for collapsing inverse solutions to cortical parcellations optimize M/EEG source reconstruction accuracy. *Journal of Neuroscience Methods*, *226*, 147–160.

Langdon, C., Genkin, M., & Engel, T. A. (2023). A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews. Neuroscience*, *24*(6), 363–377.

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475.

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *28*(9), 3095–3114.

Sreenivasan, K. K., & D'Esposito, M. (2019). The what, where and how of delay activity. *Nature Reviews. Neuroscience*, *20*(8), 466–481.

Thibeault, V., Allard, A., & Desrosiers, P. (2024). The low-rank hypothesis of complex systems. *Nature Physics*, *20*(2), 294–302.

Tian, Z., Chen, J., Zhang, C., Min, B., Xu, B., & Wang, L. (2024). Mental programming of spatial sequences in working memory in the macaque frontal cortex. *Science (New York, N.Y.)*, *385*(6716), eadp6091.

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., & Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science (New York, N.Y.)*, *375*(6581), 632–639.