Inferring latent task strategy from prefrontal activity

Yichen Qian[a], Roger Herikstad[b], Camilo Libedinsky[a, b]

[a] *Department of Psychology, National University of Singapore, Singapore*

[b] *N.1 Institute of Health, National University of Singapore, Singapore*

1. Introduction

Working memory is the ability to maintain and manipulate information derived from past sensory experiences, while motor preparation is the ability to retain and manipulate information derived from past decision-making processes. The lateral prefrontal cortex (LPFC) and the prearcuate cortex (PAC) are key regions in the brain network associated with working memory and motor preparation [1, 2]. The LPFC contains both working memory and motor preparation signals [3, 4], while the PAC is primarily involved in motor preparation [5].

While much is known about the neural mechanisms of the maintenance of working memory and motor preparation information, less is known about how this information is manipulated. Human behavioral studies have shown that tasks that require working memory updating can be performed employing different strategies. One strategy is to retrieve from passive storage at the time of recall (R@R) [6]. Neurologically, Panichello and Buschman [7] showed that the selection of one out of multiple memory items, similar to the R@R strategy, led to a rotation process that transferred the selected information from a "memory" subspace that encoded multiple pre-selected memory items, to a "readout" subspace that was common for the selected memory items.

Another strategy is to rehearse and update the memory online as the items are shown (R&U), i.e. the replacement of one memory for another [8]. Rehearsing and updating online would presumably involve directly encoding and updating memories in a common activity space (e.g., motor preparation), but few studies has investigated this mechanism.

To quantitatively understand the representational properties of neural networks employing different strategies, we trained artificial recurrent neural network models (RNNs) to solve a spatial working memory task and compared the hidden unit activities to the neural activities recorded in the LPFC and the PAC of monkeys. In both brain regions we observed patterns that were consistent with the RNNs trained with the R&U strategy. This study shows that latent behavioral strategies can be inferred using RNNs.

2. Substantial section

We trained two monkeys with a 2-item delayed response task (Fig. 1). Two memory items were sequentially presented on 1 out of 4 spatial locations on the screen, each followed by a 1s delay. Item 1 (I1) in each trial was always a target (red), while Item 2 (I2) could be either a new target (red, T/T) or a distractor (green, T/D). A saccadic response was required at the end of each trial to the location of the most recent target (i.e., I2 in T/T, I1 in T/D). We then trained RNNs to solve the same problem with two different strategies, namely the R@R and the R&U.

We first investigated the cross-temporal decodability characteristics of memory items by the full space neural activities from RNN hidden units and prefrontal populations of monkeys. Dynamic coding of target was only found in RNNs trained with the R@R strategy during the delay 1 (D1), whereas the R&U RNNs, LPFC and PAC exhibited stable code throughout the trial. In T/D trials, target code was found morphed significantly between D1 and D2 (i.e. lack of generalisation across time) only in the R@R RNNs, but not in the R&Us and the prefrontal populations.

Next we examined the geometry of the encoding subspaces of memory items at different stages of task. The encoding subspace geometries were predicted to be different with different strategies: specifically in the T/T trials where retargeting was required, the R&U strategy predicted the encoding space of I1 during D1 ($\mathrm{I1D1}_{\mathrm{T/T}}$) to be equivalent to the space of I2 during D2 ($\mathrm{I2D2}_{\mathrm{T/T}}$), while the R@R predicted these two spaces to be varied. (For conceptual illustration, see Fig 2.) Consistent with the predictions, we found $\mathrm{I1D1}_{\mathrm{T/T}}$ equivalent (coplanar, aligned, and with transferable code) to the subspace of $\mathrm{I2D2}_{\mathrm{T/T}}$ in R&U RNNs but not in the R@Rs (Fig. 2). LPFC and PAC also obtained equivalent geometry between $\mathrm{I1D1}_{\mathrm{T/T}}$ and $\mathrm{I2D2}_{\mathrm{T/T}}$, suggesting that target information was directly encoded and updated on a shared subspace in these regions and were similar to the R&U models.

Finally we directly estimated the shared subspace and projected the state evolvement across time. Based on the previous findings, we hypothesised that memory update in the R&U model would involve drifted projections from the representational locus of location of I1 to I2; in contrast, projection drifts were expected from the R@R models regardless of the need of updating memory. As expected, we found that R&U RNNs obtained projections unchanged between D1 and D2 in the T/D trials and drifted only in T/T trials (Fig. 3). LPFC and PAC exhibited similar drifting patterns to the R&U model. In contrast, the R@R RNNs had drifted projections during D2 in both T/T and T/D trials. This result suggests that memory update under the R&U strategy, and also in the monkey prefrontal populations, could be manifested as projection drifts on the readout space.

Overall, our results revealed the neural mechanism of working memory rehearsal and update. In the present study, prefrontal populations ofs monkeys showed activity patterns that were analogous to the models based on the rehearse and update strategy rather than the retrieve at recall strategy. Our results also show the potential use of RNN simulations in inferring latent cognitive strategies.

Inferring latent task strategy from prefrontal activity

Yichen Qian[a], Roger Herikstad[b], Camilo Libedinsky[a,b]

[a] *Department of Psychology, National University of Singapore, Singapore*

[b] *N.1 Institute of Health, National University of Singapore, Singapore*

2.2 Figures and tables

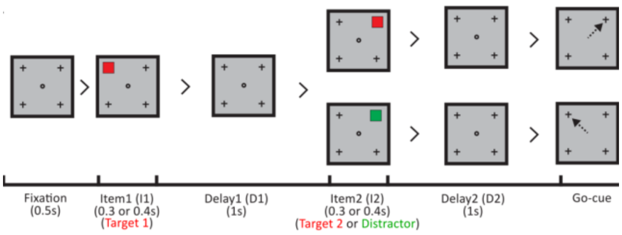
Fixation (0.5s)
Item1 (I1) (0.3 or 0.4s) (Target 1)
Delay1 (D1) (1s)
Item2 (I2) (0.3 or 0.4s) (Target 2 or Distractor)
Delay2 (D2) (1s)
Go-cue

Fig. 1: Task Design.

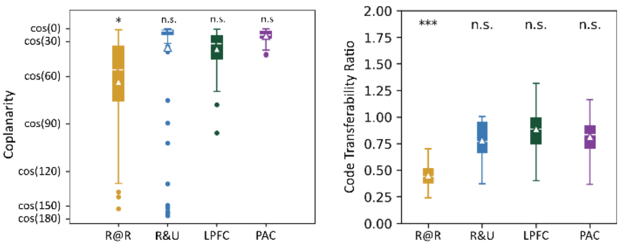

Figure 2. Quantifications of Subspace Geometry between I1D1 $_{T/T}$ and I2D2 $_{T/T}$. Left: Coplanarity; Right: Code Transferability.

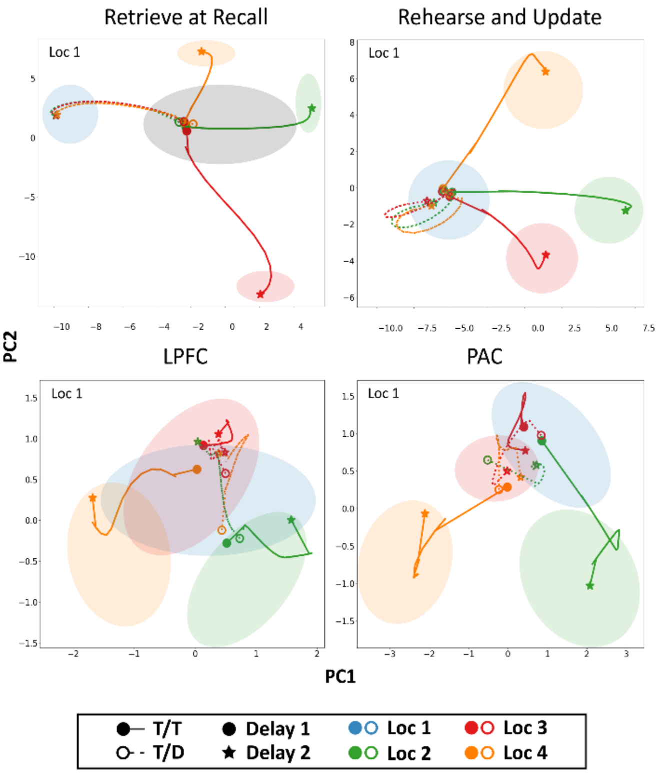

Figure 3. Projection drifts in example RNNs/cortical populations on the “readout” subspace for trials with I1 presented at location 1. Round dots and stars represent the projections at the end of D1 and D2, respectively. Solid and dash (hollow) lines/marks represent average of T/T and T/D trials. Representational locus of each spatial location is color-outlined: location 1 (blue), location 2 (green), location 3 (red), location 4 (orange).

References

[1] Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey’s dorsolateral prefrontal cortex. Journal of Neurophysiology, 61(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

[2] Fuster, J. M., & Alexander, G. E. (1971). Neuron Activity Related to Short-Term Memory. Science, 173(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

[3] Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nature Neuroscience, 20(12), 1770–1779. https://doi.org/10.1038/s41593-017-0003-2

[4] Tang, C., Herikstad, R., Parthasarathy, A., Libedinsky, C., & Yen, S.-C. (2020). Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. eLife, 9, e58154. https://doi.org/10.7554/eLife.58154

[5] Jonikaitis, D., Noudoost, B., & Moore, T. (2023). Dissociating the Contributions of Frontal Eye Field Activity to Spatial Working Memory and Motor Preparation. https://doi.org/doi: 10.1523/JNEUROSCI.1071-23.2023

[6] Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: a reconciliation. Journal of experimental psychology. Learning, memory, and cognition, 31(6), 1235–1249. https://doi.org/10.1037/0278-7393.31.6.1235

[7] Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. Nature, 592(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w

[8] Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. Neuroscience, 139(1), 23-38.