Mante et al. Synopsis Paper

Dan Birman

Prefrontal cortex remains a mysterious place for neuroscientists. Neurons in prefrontal cortex tend to be responsive to a variety of stimuli, are sometimes context dependent, and tend to show correlated activations across a wide variety of tasks. This kind of activation suggests that pFC might underlie flexible context-dependent computations and implicates them in everything from cognitive control to decision making. There is some question about how deep these computations might go: for example, in task switching sensory input needs to be selected at some level, so that a decision can be reached based only on the relevant evidence. pFC might be responsible for this but it's also plausible that this selection step occurs in an earlier sensory cortex. Regardless of selection the relevant evidence also needs to be accumulated ('integrated') to make an actual decision. Again, integration could occur in pFC or possibly in a motor cortex where evidence might be more easily translated into a motor action. These kinds of unknowns continue to make pFC a difficult brain region to understand and characterize. Mante, Sussillo, Shenoy, and Newsome used neural recordings in monkeys to try to parse out how the Frontal Eye Fields (FEF), a region in pFC, might contribute to these kinds of computations (Mante, Sussillo, Shenoy, & Newsome, 2013). They differentiate between several distinct hypotheses about how FEF might contribute to computing a choice, based on their task demands. As above they distinguish their hypothesis by whether FEF is implicated in sensory *selection* and information *integration*. But they also acknowledge that FEF may be responsible for selection (or integration) in an indirect sense if long range outputs from FEF induce selectivity in other cortical areas. They perform a two-stage analysis of their neural recordings. They first show that the responses in FEF can be characterized as a dynamic population response in a dimensionality-reduced space spanning the axes of 'choice' (towards response field or not), stimulus strength, and context (task demands). The population response follow specific trajectories through this space during the delay period which depend crucially on the current specific trial type. Their second analysis shows that a recurrent neural network, assumed to be similar in architecture to the pFC, when trained with back propagation shows qualitatively similar population dynamics. Their results point to a more complex understanding of the FEF and pFC in general, suggesting that early sensory responses are not selected by task context in advance, but are flexibly integrated by the population responses in pFC. They argue that this is only true at the population level, a parsimonious account of their findings (the line attractor and selection vector) are emergent properties of dynamical systems that cannot be observed in individual neurons. They conclude by saying that pFC may therefore use population dynamics as a way to create separable task behaviors despite constant sensory input.

Although the authors don't discuss the region that they recorded from in much detail, anatomically it corresponds well to the FEF. The FEF is responsible for voluntary saccadic eye movements, which are usually distinguished from tracking movements. Stimulation of the FEF causes saccades to specific regions in the visual field or along directed vectors (Bruce, Goldberg, Bushnell, & Stanton, G. B., 1985). Activation in the FEF occurs prior to saccades as well. The role of FEF up to this point has therefore been well characterized as a pre-motor region, where decision related inputs about eye movements converge to actually induce a movement. Note that outputs from FEF do not directly synapse with the eye muscles but are indirect inputs via the paramedian pontine reticular formation.

Although evidence of the FEF as a premotor region is strong, neural recordings from FEF during delay tasks suggest that neurons in FEF are also responsive to a remarkable number of different stimulus types and contexts. Mante et al. show that these responses are extremely general, individual neurons can rarely be categorized as responsive to a specific type of category and most units that were recorded showed dynamic responses throughout the delay period (Mante et al., 2013). Mante et al. show that there are specific sub-categories of neurons that are selective for certain stimulus types, but that the majority of FEF neurons show no particular tuning through a trial. This poses a fundamental problem for analysis of the FEF: if individual neurons show no particular tuning (but are retinotopically mapped), how should they interact to create appropriate behavior? One possibility is that the relevant information is read out at the population level as a more complex combination of individual neuron activations. This is the hypothesis that Mante et al. explore, using a number of complex analytical steps. Before discussing the results I will go through a summary of the important aspects of each analysis and potential caveats.

Behavior:

Mante et al. characterized the behavior of each monkey in the tasks as psychometric functions of the stimulus coherence values. These psychometric functions look appropriate and are well fit by sigmoidal approximations (Fig. 1). They found that the un-cued stimulus had a small effect on responses suggesting that the monkeys experienced a slight bias towards responding to strong color or motion even when attending exclusively to the other feature. In ED Fig. 2. It does appear that Monkey F is more biased towards color even in the motion context than they should be. Note also that their recurrent network model perfectly ignores the unattended feature—this is a potential concern. If a neural system is unable to avoid bias it suggests a system that has overlap between neural populations responsible for different kinds of integration and outputs, in contrast a recurrent network with perfect asymmetry across features suggests that it does not have this inherent overlap which would introduce noise. This is potentially problematic in our interpretation of whether this network represents an analogue to the FEF and pFC in general.

Population Dynamics Analysis:

To characterize the response of a population of neurons Mante et al. performed a number of analytical steps. To fully understand what their output actually represents it's crucial to understand what these steps entail. Their first step was to simply average responses for individual units (single-unit or multi-unit recordings) during correct trials, shown in ED Fig. 3. The resulting plot shows that there is a large variety of neurons in the population and only a small proportion show consistent activation, across a trial, for a particular context or stimulus type. For each unit they ran a logistic regression predicting activation using choice (to response field or not), color, motion, and context as regressors across time. The estimated beta values from this analysis therefore represent the contribution of each factor, at each time, to the neuron's overall activation or lack of activation. Note that this regression doesn't make an estimation of rate beyond the assumption of a background rate, so if information about choice/motion/color/context were rate coded in these neurons that information would now be lost. These beta weights are then projected into a 'de-noised' lower-dimensional space. This lower-dimensionality space is identified by performing principle component analysis for the original unit space (n units * t time points * c conditions). This results in an equally sized matrix but whose unit space now represents the n orthogonal principal components that capture the most to least variance in the original space. PCA can be intuitively grasped as a rotation of the axis within a coordinate frame. The 13th to nth PCs were discarded ED Fig. 4 shows the relative variance explained over time by the identified principle components for each monkey, not log scale, showing that only the first four components capture ~5% or more variance across both monkeys. The beta-value vectors were projected into this lower dimensional subspace to de-noise them by removing the variance for the 13th+ PCs. The result in ED Fig. 1 show the beta values obtained in this manner organized by the strength of the beta value for choice. Under an assumption of independent population responses for different modalities these plots would look random for motion, color, and context. For Monkey F this isn't true, indicating that monkey F's FEF conflated color and choice, possibly indicating a bias to color tasks due to prior training. This analysis is also portrayed in ED Fig. 3 showing the correlations for monkey F in some dimensions.

To portray the dynamics of the entire population as a single trajectory in state space Mante et al. chose to identify four axes primarily related to their task itself. They did this by plotting each of the beta-weight vectors in the de-noised subspace across units. This then conceptually represents the 'population' response vector for each condition. This results in four vectors that move over time. They then take the max-norm vector across time and use that for subsequent analysis. They then orthogonalize the vectors relative to the choice vector, so that each vector explains independent variance from the original subspace. Note that this introduces an interpretation difficulty in that any collinearity in the variance of the original feature space will get mapped onto the vectors in the order that they are determined. This means that the 'choice' vector may actually capture a large amount of the variance in the other conditions, iff (if and only if) they shared collinear variance originally. This is only a small problem since we know that the original subspaces were largely uncorrelated (see ED Fig. 3) but potentially problematic for Monkey F. We might therefore make the prediction that for Monkey F their population response will show no variance for color, because the choice axis will already have capture all of that variance. Note that this doesn't mean the recorded units didn't represent color—it simply means that we mis-allocated the variance to the choice axis due to our analysis. Ultimately this analysis results in a four-dimensional subspace that captures variance along each of main conditions. Time series during the delay period can then be plotted and analyzed within this subspace as trajectories as in Fig. 2.

Recurrent Neural Network Model:

The RNN model was generated by taking 100 randomly initialized nodes and giving them inputs about the color and motion coherence and the current context. A linear weighting of output nodes was used to identify the current correct choice. All of these weights (4x100 for inputs, 1x100 for outputs) were optimized through back propagation through time with a correction to deal with the vanishing gradient problem (roughly: as you back propagate errors are diminished exponentially, so propagation through time quickly causes error rates to drop to 0 requiring very high learning rate parameters). After training the network was analyzed identically to the original dataset showing both similar behavioral results and quantitatively similar 'physiological' results. More details below.

Results & Discussion:

Mante et al. show in Fig. 2 and ED Fig.7 the population dynamics that they observed from their recordings in FEF after projecting the trajectories in their four-axis subspace spanning choice, motion, color, and context. Note that for Monkey F they have to also subtract a bias factor to uncover the underlying shape. They found that sensory evidence was always represented by both contexts, suggesting that early selection was not occurring (except monkey F for color, but note the caveat noted earlier that the choice axis likely contains much of the variance in color for monkey F). This suggests that sensory evidence is directly integrated within pFC in a context-dependent manner. Take note also that the trajectories end at similar positions in all conditions, suggesting that the choice axis is not accounting for a large amount of the sensory evidence (which is stronger and weaker in different trials), although it clearly has some influence. This is important to be certain that what we are actually interpreting is related to the true input. Compare Fig. 2/7 with Fig. 5, the identical analysis performed on the recurrent neural network. Here we see a qualitatively similar outcome with graded sensory evidence and choices represented as final positions along that axis. But there are some very specific and important differences to take note of here[1]: first, the time to peak response on the coherence axes is almost instantaneous. This is likely a result of the neurons representing something more akin to the current sensory state, not accumulated sensory evidence. In other words, this recurrent network acts more like early visual cortex than like pFC, given our knowledge of stimulus responses in those regions from monkey neurophysiology. We expect an integrator to look more like the monkey data, peaking at some point in the middle of a trial and then dropping off as evidence integration is replaced by saccade preparation. The second potential issue is that the choice axis appears to be representing sensory evidence to some extent. This can be seen as the end point locations differing under different sensory strengths. This is an evidence of one of a possible failure: one failure mode is if the network itself has diverged into two sub-networks each of which selectively responds to one stimulus type. If this were the case, the same units would need to represent choice and stimulus strength, suggesting a strong correlation between their activations which would appear, in this analysis of dynamics, as co-linear variance assigned entirely to the choice axis.

Mante et al. largely skip over this considerations and focus on their results as suggesting a line attractor model of selective integration. This can be roughly understood as having parts of their population dynamics space which are 'attractors'—when trajectories go near these regions they tend to be pulled into the attractor and get stuck there. This is similar to the concept of a diffusion to bound model where once a boundary has been reached the system state is captured by that bound. They propose (and find evidence in their model results) that there is a line attractor that responds differentially in each context. In its main context (e.g. color) when presented with a color stimulus it will respond by relaxing to a new "response" state (along the line attractor), having integrated the evidence for color. But in the same context when presented with a motion stimulus the system will return to the

[1] Thanks to Dylan Cauble for pointing out the integration timing issue in our lab meeting.

initial attractor state, in essence having ignored the input entirely. The ability to relax differentially depends on a second vector in this subspace called the ‘selection vector’ which is a context dependent vector that predicts how the responses will map back onto the line attractor. See Fig. 6 for an overview of how this works, note that panel b explains the result of a motion pulse in panel a.

The results that Mante et al. compile are compelling and the concerns highlighted in this synopsis are relatively minor. The overall story is consistent: these neurons in FEF are performing selective integration at the population level and do not seem to require early selective mechanisms or integration in other networks. This result helps to consolidate earlier results from FEF and pFC in general showing neurons with a huge variability in their selectivity profiles and responses, it’s possible that earlier results should now be revisited and re-characterized in terms of population dynamics and not only individual unit activity.

References

Bruce, C. J., Goldberg, M. E., Bushnell, M. C., & Stanton, G. B. (1985). Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *Journal of Neurophysiology*, *54*(3), 714–734.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. http://doi.org/10.1038/nature12742