Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data

Fabian A. Soto
Department of Psychology, Florida International University

Many research questions in visual perception involve determining whether stimulus properties are represented and processed independently. In visual neuroscience, there is great interest in determining whether important object dimensions are represented independently in the brain. Unfortunately, most previous research has only vaguely defined what is meant by "independence," which hinders its precise quantification and testing. Here we present a new quantitative framework that links general recognition theory (GRT) and encoding models from computational neuroscience, focusing on a special form of independence: perceptual separability. Without loss of generality, consider the special case in which stimuli vary along two stimulus dimensions, represented by A and B, each with only two levels indexed by $i = 1, 2$ for dimension A and $j = 1, 2$ for dimension B. A stimulus is represented by a combination of these dimension levels, A_iB_j.

In the computational neuroscience literature, an encoding model is a formal representation of the relation between stimuli and the response of a number of channels (single neurons or neural populations), represented by $\mathbf{r}$. The channel responses are assumed to be random variables, and thus the response of the model is characterized by a probability distribution $p(\mathbf{r}|A_iB_j, \theta)$, where θ represents a set of parameters describing neural noise. *Encoding separability* of dimension A from dimension B holds when encoding of the value of A does not change with the stimulus' value on B. That is, if and only if, for all values of $\mathbf{r}$ and i:

$$p(\mathbf{r}|A_iB_1, \theta) = p(\mathbf{r}|A_iB_2, \theta). \tag{1}$$

The term neural decoding refers both to a series of methods used by researchers to extract information about a stimulus from neural data and to the mechanisms used by readout neurons to extract similar information. If a dimension is encoded by N channels, then the decoded estimate of a dimensional value is $\hat{A} = g(\mathbf{r})$, where $g()$ is a function from $\mathbb{R}^N$ to $\mathbb{R}$. Because $\mathbf{r}$ is a random vector, the decoded value $\hat{A}$ is a random value that follows a probability distribution $p\left(\hat{A}|A_iB_j, \theta\right)$. *Decoding separability* of dimension A from dimension B holds when the distribution of decoded values of A does not change with the value of B in the stimulus–that is, if and only if, for all values of $\hat{A}$ and i:

$$p\left(\hat{A}|A_iB_1, \theta\right) = p\left(\hat{A}|A_iB_2, \theta\right). \tag{2}$$

It can be shown that encoding separability and decoding separability are related as summarized in Figure 1. In addition, when decoding separability is measured through the L1 distance between kernel density estimates of $p\left(\hat{A}|A_iB_j, \theta\right)$, the relations in Figure 1 hold even if decoding is performed on indirect measures of neural activity contaminated with error, as those obtained through fMRI. Figure 1 entails that when decoding separability is measured and fails, one can make the valid inference that encoding separability fails as well, but when decoding separability holds, only weak evidence of encoding separability holding has been obtained.

Importantly, it is possible to link these ideas to GRT by assuming that the perceptual representation of a stimulus dimension in GRT is the outcome of decoding a dimensional value from the activity of many channels distributed across the brain. If we also assume that the decoding scheme is the same across changes in the irrelevant dimension, perceptual separability is simply a case of decoding separability. According to Figure 1, any failure of perceptual separability documented in the literature should be reflected in a failure of encoding separability in brain regions representing the target dimension. Thus, this extended GRT framework formally specifies the relation between behavioral and neural tests of separability, providing tools for an integrative research approach in the study of independence.

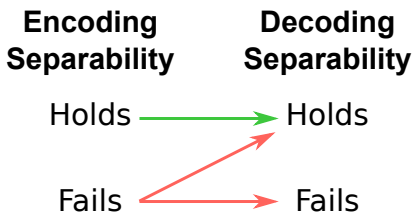

Figure 1

In addition, two commonly used operational tests of independence can be reinterpreted within this new theoretical framework, providing insights on their correct use and interpretation. It can be shown that, when some strong assumptions are met, a popular test that determines whether neural representations are orthogonal is related to the concept of perceptual independence from the traditional GRT, but it is unlikely to be related to a corresponding property of neural encoding. On the other hand, a test based on generalization of classification accuracy leads to valid inferences about encoding separability, but it provides less information than a decoding separability test. In addition, this classification accuracy test has been applied incorrectly, leading to conclusions about invariance or separability that are in general unjustified, unless one is interested in decoding separability only, and not in the separability of underlying brain representations.