**Using sensitivity analysis to interpret machine learning applied to fMRI data**

Recently, machine learning has been employed as an analysis tool in fMRI. Algorithms are trained to classify or predict the stimulus from voxel time series. These multivariate and non-linear algorithms can potentially uncover complex and distributed patterns of activation. However, relating their results to brain function is difficult.

Classifier performance is a lower bound on mutual information between the voxels and stimulus. If the voxels can be used to predict the stimulus then some, but not necessarily all, of those voxels must be driven by that stimulus. However, classification performance of a single voxel is not sufficient to show that that voxel is not involved due to potential conditional dependencies; two voxels may be able to better predict the stimulus when taken jointly rather than independently. This is consistent with the concept of population encoding.

We want a method that can isolate the minimum distributed network of voxels that best encodes the stimulus without evaluating every permutation of voxels. We use a form of analysis that measures the sensitivity of the outputs of a classifier with respect to its inputs. In feed-forward neural networks, this can be solved analytically. We train the network, calculate the sensitivity of the network to each voxel, and remove the voxels with low sensitivity. We repeat this process until the performance of the network degrades.

We tested our approach on fMRI data from a complex visual stimulus. Data was collected with a whole brain prescription while subjects viewed a virtual world that contained a variable number (1—6) of animated characters. We trained a feed-forward neural network to predict the number of characters present in each frame with a classification performance over 50%. We performed Sensitivity analysis and voxel culling until the performance of the network degraded. We smoothed and averaged the sensitivity measurements across subjects and projected the results onto the cortical surface. The distributed network of voxels sufficient for classification by the network is presented in figure 1.

