Reading minds using modern statistical methods

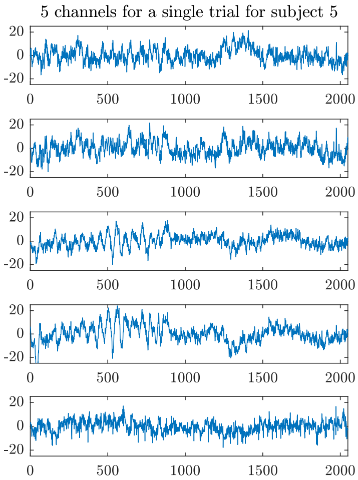
Damir Basic Knezevic and Albin Heimerson

Can we read minds? In some ways, the answer is yes—at least when encoding memories. Subjects are shown images of one of three types, and we are able to classify the type with high precision.

The human body is an immensely complex machine and one of the most complex parts is its brain. There are several theories on how neural mechanisms work within the field of *cognitive neuroscience*, and our goal is to advance the understanding of the mechanisms of memory using modern-day engineering methods.

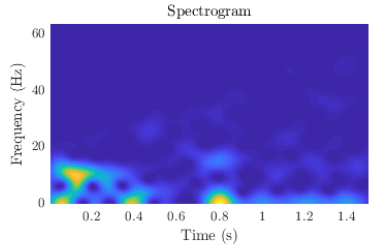
A common way to measure activity in the brain is with the electroencephalogram (EEG), where electrodes are placed over the scalp to measure voltage. It's cheap, it's fast, it's well-researched, but it's also extremely noisy, so getting useful information from EEG data is tricky. In our case, test subjects have been shown three different categories of images and we are tasked with devising an algorithm that can tell the resulting EEG data apart, both upon viewing and recollection. The idea is that some of the parts of the brain that activate during memorisation should also activate during recollection, in a process called *ecphory*.

When designing such an algorithm, there are two main parts to consider—feature extraction and classification. Feature extraction is about helping the computer see things clearer: if the aim is to distinguish circles from squares in an image, a good feature may be "amount of corners". The classification algorithm takes the features (e.g. four corners) and reduces them to a final guess (square).

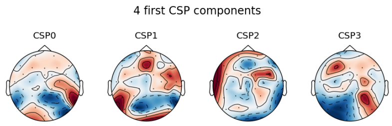


A few channels of EEG data from one of the subjects.

We mainly tested three types of features: raw data, time-frequency analysis, and common spatial patterns. Raw data is just what it sounds like. Time-frequency (TF) analysis provides information on which frequencies occur at certain points in time. This is interesting because previous research has shown that some types of neural mechanisms are encoded within certain *frequency bands*, such as for example *theta waves* oscillating at 7–10 Hz. The common spatial patterns (CSP) algorithm instead uses information about which electrodes show similar results across images of the same type. The patterns that best separate the different types are used as features.



Spectrogram of a single channel for a trial on one of the subjects.



First four CSP components from one of the subjects.

The best classification results came from using raw data (up to 82 %). We believe that the raw data retains most of the information about *when* exactly most of the activity in the brain occurred, while TF methods (60–69 %) dilute that information slightly and CSP (48–51 %) disregards it entirely. As is often the trend in machine learning research today, convolutional neural networks—while taking a long time to train—gave the highest results in terms of accuracy. However, faster algorithms like support vector machines did not have much lower accuracy.

Apart from being of interest to cognitive neuroscientists, any progress in understanding this problem can benefit for example brain-computer interface (BCI) researchers, since we are looking for general features of brain activity. Perhaps the three-class problem can be extended to even more types of images, and we can come even closer to saying that we can read minds.