Dynamic Identification of Focal Cluster ROIs Using Time-lagged Mutual Information

Tim R. Mullen

CS182: The Neural Basis of Thought and Language
Final Project
Spring, 2006
Professor: Srini Narayanan

Abstract

An ongoing problem in systems neuroscience is identification of temporal and causal components of cortical networks. Ongoing research in the application of probabilistic and statistical methods to neuroscience has led to breakthroughs in this domain. Information Theory may also have much to contribute to this problem. At Helen Wills Neuroscience Institute, Galit Fuhrmann Alpert and I have been exploring the use of time-delated mutual information to address this issue using EEG data collected intracortically from seizure patients prior to surgery. The method we are using is based on that developed by Fuhrmann for extracting spatio-temporal information from fMRI data. An important component of this process is the identification of regions of interest (ROIs) that share processing of task-dependent information. In this paper I will show that a hierarchical clustering algorithm, utilizing a distance function that clusters by relative optimal delay and MI, may effectively identify focal clusters dedicated to a task-dependent process.

Theory

A well-known measure in information theory, mutual information (MI) measures the mutual dependence between two random variables. The formula is given by

\[ I(X;Y) = H(X) - H(X|Y) \]  

where \( X, Y \) are random variables (e.g. time series) and \( H(X) \) is the Shannon entropy, which quantifies the amount of uncertainty one has about the value of a variable. This is given by \( H(X) = -\sum_x p(x) \log_2 p(x) \), where the sum extends over all values of \( X \). The conditional entropy \( H(X|Y) \) is the amount of uncertainty one has about \( X \), given that \( Y \) is known. Thus, if \( Y \) is statistically correlated with \( X \), then knowing \( Y \) should reduce the uncertainty about \( X \) and thus \( H(X|Y) < H(X) \). The conditional entropy is given by \[ H(X|Y) = \sum_y p(y) H(X|Y=y) = -\sum_y p(y) \sum_x p(x|Y=y) \log_2 p(x|Y=y) \]. The mutual information, \( I(X;Y) \) thus measures the reduction in entropy of a process or variable \( X \), given that we have information about \( Y \); if \( X \) and \( Y \) are dependent, we will see a reduction in entropy and thus a high mutual information value.

Note that this does not capture any measure of directional influence or temporal dynamics. We only know that \( X \) and \( Y \) are correlated, nothing more. However, by introducing a time lag in one of the variables, we can capture a sense of directionality:

\[ I(X;Y) = -\sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \]  

where \( x, y \) are random variables and \( \log_2 \) is the logarithm to base 2.

(2)
where \(x, y\) extend over possible values of \(X, Y\) and \(p(x_k, y_{k-t})\) is the joint probability that an observation in \(x\) at time \(k\) also appears in \(y\) at time \(k-t\). In our method, \(X\) and \(Y\) are EEG time-series for two intracranial electrodes and \(t\) extends over \([-\tau, \tau]\) for some \(\tau \in \mathbb{R}\) (generally, \([-50, 50]\) ms).

A central goal is to find the temporal delay \(t\) that maximizes the mutual information between \(X\) and \(Y\). This optimal delay can be thought of as the propagation time of information between \(X\) and \(Y\). For example, a high MI value of \(I(X;Y)_{t=0} = 0.4\) for \(t=5.5\) suggests that \(Y\) propagated information to \(X\) over a period of 5.5 time units (here, milliseconds). Note that if \(t\) were reversed in sign, this would indicate that information propagated in the opposite direction (\(X \rightarrow Y\)). Our method performs bi-directional pairwise analysis of MI between channel (electrode) time-series data, identifying the direction and delay of propagation with the aim of elucidating the properties of task-dependent cortical networks.

In this paper, I will not go into further detail on the application and results of using this specific method on EEG data, but instead will now focus on the subproblem of identifying regions of interest (ROIs) using time-delayed MI.

One implication that arises from equation 1, combined with established knowledge of neural conduction properties, is that an optimal delay of \(t=0\) between channels \(X\) and \(Y\) may represent simultaneous shared processing of information between these two regions. Consider a case where we observe a high value of \(I(X;Y)_{t=0}\). Clearly it is not neurally possible for information to be conducted between two cortical regions in a span of 0 ms (or even < 1ms, due to action potential discharge constraints). Furthermore, a high MI value indicates that \(X\) and \(Y\) possess similar information. Thus, a reasonable conclusion is that we are not observing the transmission of information between \(X\) and \(Y\) but rather that \(X\) and \(Y\) are processing the same information in parallel. A likely explanation for this effect is that \(X\) and \(Y\) are both simultaneously receiving information from a third source (e.g., another focal cluster). Another possibility is that \(X\) and \(Y\) share background information and it is this shared information that is reflected by \(I_t\). Methods involving \(n\)-way co-information\(^4\) analysis can be used to distinguish between these (and other) cases, but in either case, it seems reasonable to assume that \(X\) and \(Y\) jointly form a focal cluster that collectively processes some aspect of the information resonating through a cortical network.

**Method**

To identify these and other kinds of clusters, I use a single-linkage hierarchical clustering algorithm (other algorithms such as K-Means\(^5\) are also applicable with varying degrees of success). Hierarchical clustering\(^6\) is a fairly simple agglomerative clustering method that builds up clusters by iteratively merging together similar clusters. In single-linkage clustering, each iteration merges the two ‘nearest’ clusters, where distance is computed as the shortest distance between any member of one cluster and any member of the other cluster. Through this process a hierarchical tree is constructed where progressively deeper levels correspond to progressively more distinct clusterings. For example, level 0 would correspond to a single cluster consisting of all datapoints, level 1 would have a few relatively large clusters, and level \(N\) (the deepest level) would have a separate cluster for each datapoint. Specifying an upper bound on the desired number of clusters constrains the choice of possible clustering combinations; the level that yields the number of clusters closest to the desired number of clusters is chosen.
Arguably the most critical part of the clustering process is defining an appropriate distance function. In our case, the aim is to cluster channels that share a low optimal delay (dt) and high MI. A simple solution is to use the between-channel dt weighted by the pair’s MI. However, because many pairs only differ slightly in dt and MI, we wish to encourage small changes in MI and dt to result in large overall changes in distance. Furthermore, MI should inversely weight the distance such that a low MI between pairs should increase the overall distance between the pairs. To satisfy these constraints the following distance function is proposed:

$$D[X;Y] = e^{\beta \cdot dt(X;Y)} \cdot e^{-\alpha \cdot I(X;Y)} = e^{(\beta \cdot dt(X;Y) - \alpha \cdot I(X;Y))}$$

(3)

where, $dt, I$ is the optimal delay and corresponding MI value between X and Y; $\beta, \alpha$ are free parameters that allow the researcher to control the relative contributions of dt and MI to the overall distance metric. Note that MI is always real-valued over [0, 1]. In the interest of consistency (and to make it easier to select appropriate values for $\beta$ and $\alpha$), we set $dt$ to also be real-valued over [0, 1] using the transformation $dt_{\text{new}} = dt / \max(DT)$, where DT is the set of all possible delays.

In theory, MI is symmetric, so $I(X;Y) = I(Y;X)$. This means that the delays that maximize MI should likewise be symmetric and sign reversed: $dt(X;Y) = -dt(Y;X)$. However, due to the constraints of estimating probability distributions with limited data sets, this is not always the case in practice. To resolve this issue, we define a consistent delay pairing as one in which the delays in each direction differ only within some predefined interval: $|dt(X;Y)| - |dt(Y;X)| < \Phi$, where $\Phi$ is some small value (here, 2 ms). We then define $dt(X;Y)$ to be the mean of the optimal delays in both directions: $dt(X;Y) = \text{mean}(|\text{delay}(X;Y)|, |\text{delay}(Y;X)|)$. A similar metric is applied to MI. Inconsistent pairings are marked such that they do not affect the overall clustering process.

### Application and Results

We applied this hierarchical clustering function with the above distance function to a data set yielded by an exhaustive pairwise MI and dt comparison of 64 intracranial electrode time-series (4096-64 = 4032 pairs) from a single seizure patient. The physical spacing between each electrode pair on the surface of the cortex is 1 cm. The subject was engaged in a passive-listening + attention task where hand and mouth action words were presented along with proper names and the subject was asked to press a button with the right index finger if the word presented was a proper name. Here we only look at processing for the action words, and ignore the proper names (thus eliminating artifacts of motor processing). We also group together hand and mouth action words so as to retain enough trials to generate reasonable probability distributions. The epoch window is [0, 800] ms. Each electrode is common-averaged with respect to reference electrode 64.

The choice of $\beta$ and $\alpha$ in $D[X;Y]$, as well as the expected number of clusters (n), has a significant impact on overall clustering. To encourage smaller clusters to form, we set $n=35$. A high value of $\alpha$ imposes a higher “penalty” for low MI, while higher values of $\beta$ impose higher penalties for high delay values. Because we are primarily interested in optimal delay relationships (while retaining MI as a factor), we choose a $\beta$ that is approximately twice that of $\alpha$.

Figure 1 shows the results from this clustering analysis overlaid on a nearest-neighbor directed graph showing pairwise dt and MI values. Although we ran the clustering analysis on all
4032 pairs, here we only explicitly label nearest-neighbor connections for the 64 channels. Each edge connection is labeled with the delay in each direction; so the format is “dt(X \rightarrow Y) : dt(Y \rightarrow X)”, where X is the leftist channel along the edge between X and Y; if X and Y are along the same vertical line, then X is the downmost channel. To provide a concrete example, between channels 2 and 11 we have the edge label “2.5 : -3.” This means that dt(2 \rightarrow 11) = 2.5 ms and dt(11 \rightarrow 2)= -3 ms. Thus information was propagated from channel 2 to channel 3 in approx. 2.75ms. In addition, edge line width is weighted by the MI between the two connected channels, using the function

$$width(X; Y) = \varphi \left[ 1 - \left( \frac{0.01}{\varphi} \right)^{I(X; Y)} \right]$$

where \(\varphi\) = the maximum width in pixels (here, \(\varphi=5.0\)) such that \(I(X; Y) = 1.0\) would correspond to a line width of 5 pixels, while \(I(X; Y) = 0.0\) corresponds to a line width of 0 (no line). For rapid identification, channels belonging to the same cluster are color-coded with the same colors. Channels that did not belong to any cluster with \(n>1\) are not colored. For additional reference, each channel’s cluster number appears in parentheses to the right of the channel number. The entire graph is furthermore overlaid on a sulci map which shows the physical relation of each channel to major sulci on the cortex of this subject.

Turning our attention to Figure 1, we see that certain prominent clusters have emerged. It is immediately apparent that these clusters mostly appear to correspond to very low-dt, high-MI connections between contributing channels, which is precisely what we might have expected. Furthermore, it is apparent that there are three major clusters that roughly comprise Brodmann’s areas 44, 45; 22; and 6 (and part of Area 4). Areas 44, 45 are also known as “Broca’s Area” and are known to be implicated in word processing and production. Area 22, also known as “Wernicke’s Area” is another region highly involved in word comprehension. Area 6 and 4 form the premotor/motor processing area. It has been suggested that motor/premotor cortex is also involved in semantic processing of words. Other notable clusters include the temporal pole (chs 4, 3), Area 20 (chs 17, 35) and Area 46 (chs 7, 8; shown on Figure 2). Some of the auditory belt/parabelt region (chs 27, 35, 34) share a low-dt cluster with “Wernicke’s” channels suggesting that they are closely related functionally. Channel 35 is identified as primary auditory cortex.

As a secondary measure, we ran an analysis with \(\alpha >> \beta\) so that MI would be more of a determining factor than dt (although dt still plays a significant role). The results (Figure 2), show an increased separation of clusters with clusters more tightly focused around a few central channels. Note that here channels 7 and 8 now appear clustered.

As a side note, some might argue that inter-channel delays of 1-2ms may seem too low for intercortical communication. However, it has been shown that neurons in neighboring regions can communicate within intervals as low as 1 ms.

**Conclusions**

Mutual information remains a simple, yet powerful, method for elucidating the directional and temporal properties of task-dependent cortical networks. While we were able to effectively utilize a clustering algorithm to identify task-dependent ROIs within this network, some problems still remain. Hidden variables remain a constant concern. We cannot be certain that a high delay between two channels reflects their membership in separate clusters: it may be the
case that a third source is propagating information to these channels at varying time-delays to each channel. This may cause regions that are, in fact, part of the same functional cluster to appear disparate due to (artifactually) long delays between them. Weighting the delay by MI partially resolves this, by reducing the long-delay penalty if the regions appear to be processing the same information; but it is not fool-proof. 3-way co-information analysis can be used to resolve a 3-channel configuration, but the possibility always exists that some $n^{th}$ variable may be the “real” source. A second issue is that, while the clustering method we have used is an exclusive clustering algorithm, it is likely that localized networks may share membership in more than one focal cluster. Thus, fuzzy or probabilistic clustering algorithms may yield more accurate results.

One of the most promising applications of clustering as we have proposed is that it allows us to identify members of ROIs while imposing highly specific sets of constraints on what should determine ROI membership. A further step is to then take a normalized average of the time-series data for ROI members (perhaps weighted by MI) and collapse these into a single “channel” which can then be compared to other ROIs or individual channels across the cortex. Due to the sensitivity of MI, even a single false member of an ROI could drastically skew the overall distribution and thus provide false MI estimates when comparing this to other regions. The ability of this clustering method to eliminate such false members provides a natural way to reduce the possibility of such mishaps and provide cleaner, more reliable results.

Much thanks to Srini Narayanan and Galit Fuhrmann Alpert for advice and assistance.

---


8 E. Izhikevich, *Polychronization: Computation with spikes* (Neural Computation (in press)).
Figure 1

Graphical representation of edge colors and legend:
- Black: no edge
- Blue: consistent w/o OPDR > 0
- Green: consistent w/ OPDR > 1
- Red: not consistent

Cluster members are defined by the graph's layout and color-coding.

Center of Salience
Lateral of Salience

Legend: (Black) - no edge, (Blue) - consistent w/o OPDR > 0, (Green) - consistent w/ OPDR > 1, (Red) - not consistent.
Figure 2
of Channel - 7, β = 1; α = 7
emphasizing MI over delay
line width denotes MI.
Cluster members are defined by shared color.