Text mining of fMRI at full resolution

by energy coding and entropy bagging of ”resting state” series.

Onno Elzinga (s1903225)

Thesis advisors: Dr. Wouter Weeda &
Dr. Tom F. Wilderjans

MASTER THESIS

Defended on November 30, 2018

STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

Association of neurological and psychological conditions with changes in coactivation patterns of brain regions in 'resting state' is of recent interest in neuroscience. To uncover such latent functional connectivity, series of functional Magnetic Resonance Imaging (fMRI) scans are typically reduced by averaging activations in brain atlas regions. The averaged activations are further reduced to pairwise correlation in sliding fixed width time windows. Unfortunately such reduction in dimensions also reduces the scan resolution and complicates interpretation.

Changing to a text mining perspective, this thesis interprets the high dimensional scans as documents with categorical words drawn from a study bag. Consecutive scans measure the activation in V discrete voxels of brain volumes. Activation series in each voxel are segmented into stationary subsequences. Similar correlated segments within voxels and from distinct voxels are then BAGGED as WORDS. The words capture correlated activation both within- and between-voxels. Instead of being predefined in an atlas, regions emerge as neighbourhoods of voxels drawing the same word at the original scan resolution.

The word counts that document voxels draw from the bag of categorical words defines the document state. Document state transition probabilities measure the dynamics in coactivated brain locations at the original fMRI resolution, as a possible marker for a neurological condition.

This alternative fMRI activation reduction method avoids a-priori selection of regions, tuning of fixed time window widths, and selection of the number of principal components of the contrasted existing method; the alternative method allows a more direct interpretation of activations. However, the direct state switching interpretation of scan document voxels drawing categorical word counts, does not sufficiently separate subject groups for reliable classification of neurological conditions.

Index terms— fMRI, time series, energy coding, entropy, mutual information, Kullback-Leibler divergence, change point detection, latent document allocation, von Misses-Fisher, discrete Markov chain, scan document state switching.

1 Brain dynamics and dimensionality reduction

To investigate differences in functional connectivity between subject groups, studies scan subject brains using functional Magnetic Resonance Imaging (fMRI). The non-invasive fMRI method is used to locate activity in brain regions and further to infer functional connectivity between distinct coactivated brain regions. Consecutive measurements of subject brain volumes result in activation time series. The brain volumes are built up of horizontal slices. The discretized slices consist of voxels which may be seen as volumetric i.e. 3D pixels in images. Cells in voxels use oxygen supplied by blood. Akin to light in ordinary 2D pixels, fMRI measures activations in voxels as blood-oxygen-level-dependent (BOLD) values.

Due to the smooth flow of oxygenated blood through arteries and veins, consecutive BOLD activation within voxels are temporally correlated. Neuroscience attributes specific functions to distinct voxel neighbourhoods known as brain regions. Inferred from their correlation, spatially distinct coactivated regions are functionally connected (FC).

Brains of subjects who are not given a task or stimulus in a study are said to be in 'resting state'. Contrary to previous assumptions, brains in 'resting state' also show coactivated distinct regions. Studies seek to associate differences in dynamic coactivation patterns with a neurological condition using fMRI, e.g. Peraza et. al. in [14] associate fluctuations in resting state networks with dementia using fMRI.

By contrast to Electro/Magneto EncephaloGraphy (EEG/MEG) which samples on millisecond scale but with only approximate brain location, fMRI samples on second time scale with millimeter space scale location resolution. The existing approach by Leonardi et al. in [11] reduces both the spatial voxel resolution and correlates fMRI activations in sliding time windows. By contrast, the new method proposed in this thesis reduces the activations *themselves* while maintaining the observed spatial-temporal resolution in the fMRI series. Assuming that functional activation is spatially localised and temporally smooth, both the existing approach and the new method aim to uncover a latent functional connectivity in a lower dimension than the high dimensional activations.

A brief review in 1.1 of the activation data dimensions and dimension reduction approaches precedes a detailed characterisation in 1.2 of the dimension reduction steps and issues in the existing approach.

The alternative text mining method introduced in 1.3 conserves the original resolution while still capturing the between and within voxel activation correlation, it avoids tuning and interpretation issues of the existing approach.

Mining words from activation series at voxel level lets the lower dimensional functional connectivity emerge at the original voxel and time scan resolution, rather than to carve it out using parameterised templates which also coarsen activations downto a lower resolution. At original scan time resolution, scan document voxels draw mined words from the study bag.

1.1 fMRI scan dimensions and reductions

Functional Magnetic Resonance imaging (fMRI) scans of brains in “resting state” result in repeated observations. Due to e.g. cost constraints of fMRI measurements, usually the brains of few (e.g. 30) subjects S are scanned, yielding real ($\mathbb{R}$) valued activations in discretized brain voxels V. The number of voxels in the order of tens of thousands (e.g 40000) depends on resolution of the MR technology. With one scan each 2 seconds, a 10 minute session of $T = 300$ scans, yields high dimensional (HD) activation observations X of $x_{s,v,t} \in \mathbb{R}^{S \times V \times T}$.

The existing approach characterised in 1.2 reduces and approximates the voxel V and time T dimensions. The text mining method in 1.3 instead maps the real valued voxel activations $\mathbb{R}$ to categorical words in the study bag.

1.2 Voxel aggregation and time reduction in existing method

In [11], Leonardi et al., approximate functional connectivity (FC) by grouped pairwise correlation in fixed-width and synchronised windows of averaged BOLD activations in spatially distinct brain regions. They aim to associate the subjects’ activation patterns of functional connected regions with a neurological condition.

To that end, they reduce the high dimensional voxel space characterised in 1.1 into regions in 1.2.1 and further into correlated time windows in 1.2.2 before selecting p principal 'Eigenconnectivities' (EC) in 1.2.3 to approximate the group functional brain region connectivity. Enumeration 1.2.4 discusses the tuning of the reduction steps, associated assumptions, loss of resolution, and staged interpretation issues.

Given the ECs, Leonardi et al., determine the consecutive loadings over time for each subject. They found differences between patients and healthy controls in analysis of the dynamics of the loading time series. So, Leonardi et al., as they summarise in Figure 1, analyse the frequencies in the projections of BOLD activations onto a set of ECs, where the ECs serve as grouped building blocks for time windowed pairwise regional correlation.

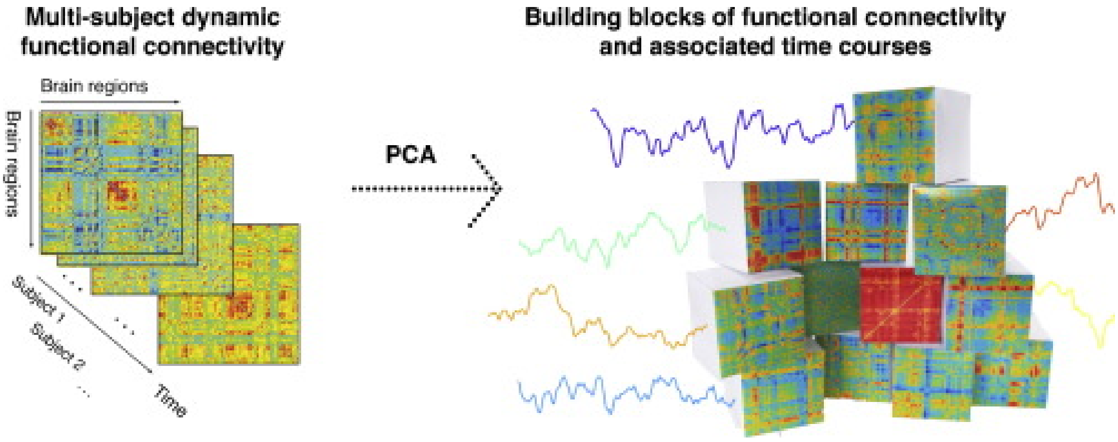

Multi-subject dynamic functional connectivity
Brain regions
Brain regions
Subject 1
Subject 2
Time
PCA
Building blocks of functional connectivity and associated time courses

Figure 1: Subject specific 'time courses' from brain region activation series. Fixed-windowed, pairwise correlation of brain atlas region averaged activation is grouped as 'functional connectivity'. The 'building blocks of functional connectivity' are found by Principal Component Analysis (PCA) approximation. The subject specific loading series on the 'building blocks' are 'the associated time courses'.

1.2.1 Atlas regions average voxel activation

Neuroscience attributes specific functions to distinct brains regions. Such regional specialisation suggests to aggregate voxels into regions. E.g. Leonardi et al., in [11], segment the brain in 88 regions R defined in the Automated Anatomical Labelling atlas [17] to reduce $\mathbb{R}^{S\times V\times T}$ to $\mathbb{R}^{S\times R\times T}$.

The assumption is that regions R are disjoint neighborhoods of voxels with similar activations so that the mean voxel activations may be taken as the region activation; the spread of the activitions is ignored.

1.2.2 Fix time windows capture pairwise regional correlation

Assuming smooth supply of oxygenated blood, consecutive BOLD activation of regions of voxels V are temporally autocorrelated. Neuroscience further assumes that coactivated regions R cooperate to perform a function of a task.

Leonardi et al., in [11], fix a window Δt of 30 consecutive images, and calculate the Pearson correlation ρ between all unique pairs $D = R\,(R-1)\,/2$ of regions in *synchronized* windows moving in steps of 2 images. This further reduces the dimensionality from $\mathbb{R}^{S\times R\times T}$ to $\rho^{S\times D\times W}$, of a sequence of $W = ((T-\Delta t)\,/2-\Delta t)$ pairwise correlations for each of the S subjects. This correlation captures time window synchronized coactivation, where identical but time lagging or partial overlapping coactivations result in smaller correlations depending on the time lag distance between windows.

1.2.3 Group eigen connectivity approximates functional connectivity

Using principal component analysis (PCA), Leonardi et al., in [11], further approximate the windowed pairwise regional correlation in 1.2.2.

They Fisher transform by inverse tangent function ($C = \tan^{-1}(\rho)$) and row center the pairwise correlations $\rho^{S\times D\times W}$ into sequentially grouped correlations $C^{S\times D\times W}$. Using the p principal

components of U from decomposition $CC^\top = U\Lambda U^\top$, they determine p loading time series $L_p^{S\times W} = U_p^\top C^{S\times D\times W}$ of length W for S subjects.
The principal components in U collect weighted pairwise regional coactivation correlations into shared 'eigenconnectivities' (EC) called 'building blocks' in Figure 1 from [11] as an approximation for functional connectivity. The ECs in U are mutually orthogonal and uncorrelated.

1.2.4 Tuning, resolution loss and indirect interpretation issues

This subsection lists decisions, some assumptions and consequences of the preceding regional reduction, temporal reduction, and approximation steps, which the text mining method in 1.3 avoids.
Leonardi et al., segment the brain in 88 **regions** R defined in the Automated Anatomical Labelling atlas [17] member of a **family** of brain atlases.
Spatial dimension reduction in 1.2.1 is reliable when a subjects' **neighboring** voxels are identically **coactivated** over all scans and **consistently** adhere to the same region.
Since individual subject regional average activation is further reduced to **group** windowed pairwise regional correlation in 1.2.3, region membership must also be stable in all subjects in the study.
The choice of the **window** size Δt in number of scans determines over how many consecutive region activation averages the pairwise correlation coefficient ρ is calculated. The correlation in 1.2.2 reduces two vectors $[t, \ldots, t+\Delta t]$ of average region activations into one real value. The correlation captures coactivation between synchronised regions, i.e. without time **lag**.
Together with the window **step size**, taken as 2 scans, Δt determines the serial reduction from T scan to a window W of length $(T - \Delta t)/2 - \Delta t$. After experimentation, Leonardi et al., fix Δt at 30 and report that smaller windows result in split 'eigenconnectivities' while larger windows collapse specific connectivity.
Their $p = 10$ largest eigen connectivities explain 34% of the grouped pairwise correlation; they find higher frequency components in loadings on the 10^{th} EC in patients than in healthy controls. The % of **variance cut-off** criterion to select the number of principal components p needs consideration, since the eigenvalues do not drop off sharply.
Finally, because group ECs are weighted sums of pairwise regional coactivation correlations, scalar subject loadings on such ECs also have a **scaled weighted sum** interpretation at **window time scale**.
The selection of atlas regions, fixed-width window size, and number of principal components needs tuning. The existing approach uses those a-priori selections to reduce the resolution of the data which complicates interpretation of the approximations of the weighted combinations of pairwise correlations at a shrunk time resolution.

1.3 Emergent connectivity at voxel and scan resolution by text mining

Where the existing approach groups pairwise averaged regional correlation in time synchronised fixed-size windows, the text mining method works on stationary individual voxel activation sequences. Assuming localised functional specialisation and smooth change, both the existing approach and the new method aim to uncover a latent functional connectivity in a lower dimension than the high dimensional real ($\mathbb{R}$) valued activations in T scans with V voxels from S subjects as $\mathbb{R}^{S\times V\times T}$.
The new method lets the lower dimensional structure emerge at the original resolution, rather than to carve it out using parameterised templates and to condense it at lower spatial and time resolution. By working on the voxel level under the stationary window assumption, the text

mining method avoids loss of spatial and temporal resolution which allows functional connectivity to emerge. Thus the text mining mapping avoids tuning and indirect interpretation issues of the existing approach listed in 1.2.4.

1.3.1 Connectivity emerges from voxel correlation in words

The existing approach condenses neighborhoods of voxels into R regions, activations in synchronised fixed windows Δt and p principal components of regional cooperation finally yielding subject loadings $L^{S \times p \times W}$.

By contrast, the new method assigns each *voxel* in each image a categorical *word* from the study 'bag of words' B, giving $B^{S \times V \times T}$. A voxel draws its next word from the bag depending on its current word and that of its neighboring voxels; words account for serial and spatially correlation in voxels. The method joins similar stationary subsequences of e.g. consecutive BOLD activations or autocorrelated noise from distinct voxels into their stationary words. The bag of words holds the stationary recurring building blocks of the complete series in a study. In consecutive scans at original resolution, the individual voxels draw different words when their activation changes. Regions emerge from voxel neighbourhoods that consistently draw the same word. Functionally connected networks emerge when distinct regions draw the same word at a given time. When regions participating in a functional network change their word, they become a node in another categorical functional network. Distinct regions drawing the same word with some time lag may be monitored at original time resolution. The number of consecutive scans during which regions form a network may be taken as a measure for its dynamic functional connectivity.

1.3.2 State switching document interpretation

Given the categorical words that its voxels draw, each MR scan resembles a *document* characterised by its categorical word counts summing to the number of observed voxels. Projected on the unit sphere, similar documents concentrate at small cosine distance around their mean *state* direction; the states are assumed to form a mixture of von Misses-Fisher distributions. Given the document projections, Expectation Maximisation determines the states' mean direction and dispersion; the Bayesian Information Criterion guides the optimal number of states in the mixture.

In a generative model, document word distributions depend on their states, where in [11] coactivations weigh in on eigenconnectivities. Instead of a loading series per eigenconnectivity, each subject scan document selects the state at smallest cosine distance. The dynamics of the BOLD series are interpreted as state occupation probabilities corresponding to Leonardi et al., who in [11] evaluate the frequency spectrum of the eigenconnectivity loading series; subjects' state occupation probabilities that deviate significantly from the group average suggests different group memberships.

1.4 Outlook

The first part of the method in 2 details resolution preserving text mining introduced in 1.3 to bag stationary words. The second part of the method interprets scans as documents drawing words from the study bag and evaluates their dynamic characteristics as a measure of dynamic functional connectivity.

Section 3 reports on method simulations, issues, and results. The method is contrasted with the current approach in 4.1 before concluding with its advantages, disadvantages, limitations, and applications. Appendix A lists mathematical notations. Appendix B defines method concepts.

2 Method: text mined emergent connectivity

Leonardi et al., consider the frequency spectrum of loading series on 'eigenconnectivities' to evaluate the possible association between neurological conditions and more frequent changes in functional connectivity. As 1.2.4 details, their reduction and approximation of functional connectivity comes with resolution loss and intricate interpretation.

To remedy, this section poses a generative model in 2.2 and develops an estimator for the subject scan state occupation parameter from the model. The innovation of this thesis is in text mining words from voxel activation series precede the categorical word document and its state switching interpretation. This alternative reduction captures the within- and between-voxel activation correlation that let functionality connected brain regions emerge at the observation time resolution of the activations. The interpretation part treats a document voxel as a variable that draws its word series from the study bag; the categorical word draws in consecutive documents are used to approximate the dynamics of functional connectivity.

2.1 Word bagging and document switching

The flow chart in Figure 2 lists the reduction steps to estimate the state switching parameter Φ in model 2.2 for a series of T consecutive scan observations on V voxels in S subjects. The first steps bag categorical words that capture correlations between real valued activations. Connected distinct brain regions emerge as networks in scan documents with voxels drawing words.
Subsequent steps interpret documents with voxels drawing their categorical word from the bag as categorical word frequency vectors to estimate document state occupation.
The summary in 2.8 concludes the method section.

2.1.1 Word bagging steps let regions emerge

Words capture both within- and between-voxel correlation. Consecutive similar subsequence activations in stationary windows capture within voxel correlation. To capture between window correlation, each window is inserted in the chain between its most strongly correlated neighbours. Too weakly correlated neighbouring windows end up in different words. Neighbouring, but separate, words break the weak bond between faintly correlated windows.
For each voxel, energy based bisection and permutation test in 2.3 segment the T consecutive real-valued $\mathbb{R}$ activations into stationary windows. The windows capture consecutive autocorrelated activation segments within voxels. The energy based coding step in 2.3.2 then codes activations in windows of width L_i into symbol probabilities $\mathbb{P}$. Using distance between codes, 2.4 chains each window between its nearest i.e. most correlated neighbour windows. At too low mutual information bonds between adjacent windows, 2.5 cuts the chain into bagged words.
In each scan document its voxels draw their current word from the study bag. Then, voxel neighbourhoods drawing the same words consistently over the series emerge as regions. Networks emerge when distinct regions draw the same word. Networks change when one or more regions switch their activation word over the scan series.

2.1.2 Categorical document state switching interpretation

Scan documents are interpreted as categorical draws from the categorical bag of words. At unit length, document word count vectors are located on the unit sphere. Documents with similar word counts cluster around their state centroid location. Dynamic documents switch state from scan to scan. Subject state switching approximates dynamics in functional connectivity.

Documents drawn up from the bag filled with categorical words in 2.5, are interpreted in 2.6 as word frequencies F to develop 'scan document state switching' with document location, document state centroid location and discrete state mixture estimation.
Finally 2.7 models state occupation with a occupation test statistic to label subjects' from subject scan document state switching rates Φ.

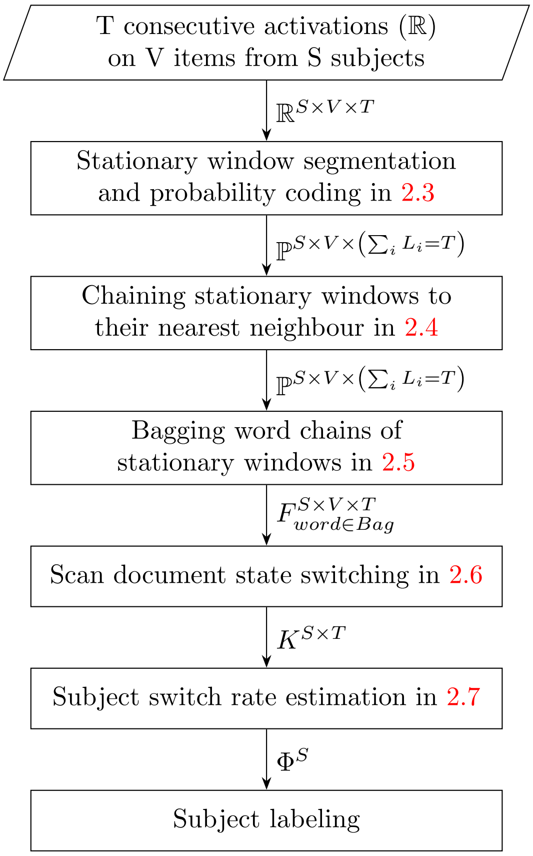

T consecutive activations ($\mathbb{R}$) on V items from S subjects
$\mathbb{R}^{S \times V \times T}$
Stationary window segmentation and probability coding in 2.3
$\mathbb{P}^{S \times V \times (\sum_i L_i = T)}$
Chaining stationary windows to their nearest neighbour in 2.4
$\mathbb{P}^{S \times V \times (\sum_i L_i = T)}$
Bagging word chains of stationary windows in 2.5
$F^{S \times V \times T}_{word \in Bag}$
Scan document state switching in 2.6
$K^{S \times T}$
Subject switch rate estimation in 2.7
Φ^S
Subject labeling

Figure 2: Word bagging and state switching interpretation steps

2.2 Model

The present alternative poses the model illustrated in factor graph Figure 3. In this model, one of the K discrete latent states generates a von Misses-Fisher word distribution W. The W vectors then generate V categorical activation observation words X for the T consecutive scan documents. A document of subject s sticks to the current state k depending on its group G specific Geometric failure probability; with success probability ϕ_s it switches to **another** categorical state with mixture probability π.
Using this scan document state switching model, the hypothesis of possible association between neurological conditions and more frequent changes in functional connectivities is evaluated using state occupation duration. The probability that the current state k is vacated equals the estimated state switching rate parameter ϕ in the model.
The model parameters are learned upwards in the directed factor graph in Figure 3 starting from

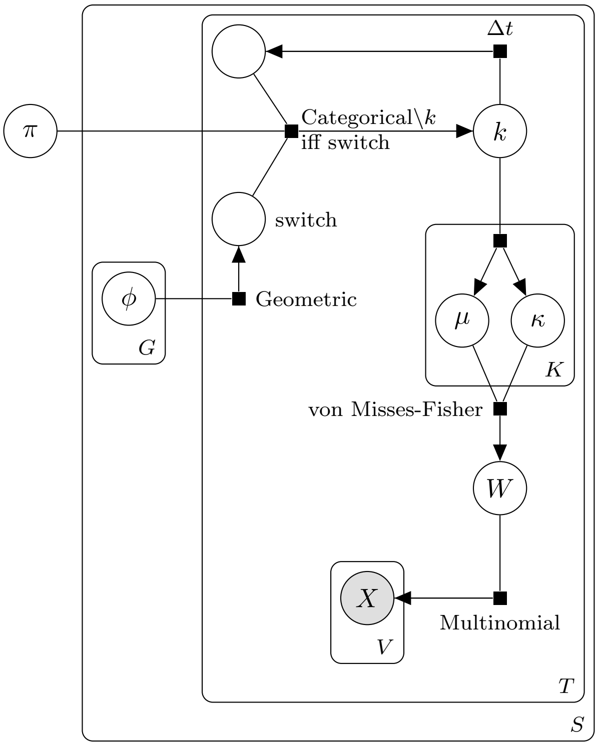


Figure 3: directed factor graph for document state switching

voxel activations X mapped via normalised word frequencies W onto one of K states and finally state change rates ϕ_g.

2.3 Consecutive similar voxel activation window coding

To bag words which capture both within- and between-voxel activation correlation, this section first captures and codes the consecutive within voxel correlation in windows.
Permutation tests in 2.3.1 find stationary windows in consecutive series; windows of stationary subsequences are mapped into symbol probability distributions using the Haar wavelet transform in 2.3.2. Using the codes, 2.4 chains windows between nearest neighbours; 2.5 cuts the shackles between stationary words for the study bag of words.

2.3.1 Consecutive window of similar voxel activations by energy permutation

In their ecp package [9] for the R language, James and Matteson (2014) implement Energy based Change Point detection. The functions in ecp are sensitive to both changes in amplitude and frequency in series. The energy divisive function `e.divisive` segments series into stationary windows by bisection and permutation tests. The computational complexity of this non-parametric method is $\mathcal{O}(cT^2)$, where c is the number of change points, and T is the number of observations.
Figure 4 shows the change points by red bars estimated using `e.divisive` with a minimum width of 16 and a quadratic energy norm. The segments are taken as windows of stationary sub-

sequences. This assumption is not critical since similar stationary subsequences are reordered before their word membership is tested and determined.

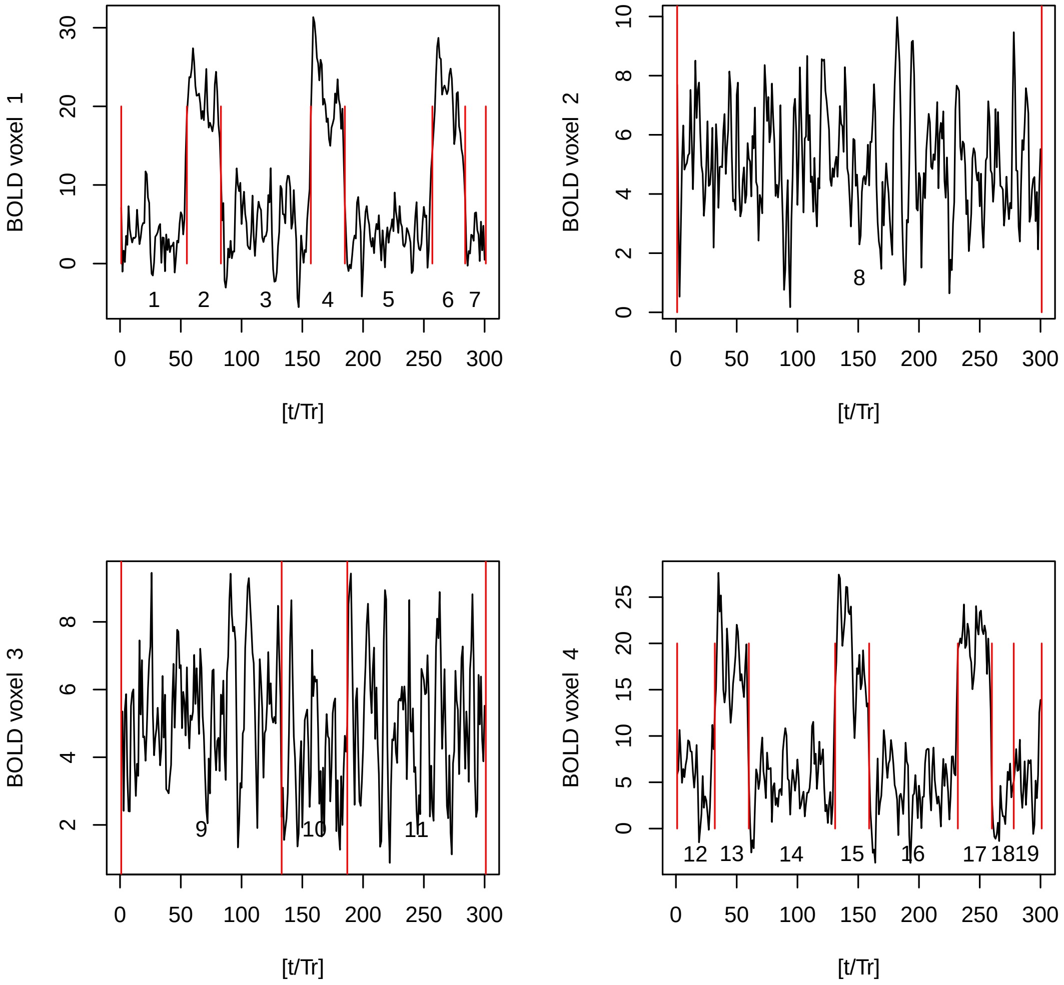


Figure 4: Simulated activation series of length 300 for 4 voxels, defined in section 3.1. Stationary subsequences separated by change points at vertical red bars are numbered sequentially. Subsequences 2, 4, 6, 13, 15 and 17 are BOLD activations. BOLD activations in voxel 4 start 25 scans before those in voxel 1.

2.3.2 Probability coding of stationary windows

To chain windows to their most similar, most correlated neighbour a distance between them is needed which is independent of window lengths found previously in 2.3.1. The distance must be

based upon a representation which is a reliable characterisation of the series. The Haar wavelet transform reviewed here, captures a series at increasing detail and, at sufficient detail, it preserves the energy of the original series in the sum of its subband energy. The energy in the subbands is normalised as a probability code for a window and later to measure distance between windows, with the aim to find the nearest neighbour window.
The stationary subsequences in windows found in 2.3.1 are mapped into symbol probability distributions using Haar wavelet transform. Haar symbols, an example and subband energy probability are reviewed in 2.3.2.1, 2.3.2.1.1 and 2.3.2.2 respectively. These probability codes are used in 2.4 to chain windows at smallest distance.

2.3.2.1 Wavelet symbols

Discrete Wavelet Transforms (DWT) capture amplitude changes in addition to differences in frequency. Discrete Wavelet Transforms map a real valued series to coefficients. The coefficients are loadings on wavelet symbols. Figure 5 illustrates Haar wavelet symbols W_l of levels $l \in \{1, \cdots, 5\}$, for the first DWT named after Alfrèd Haar, [7], in his 1910 dissertation. After removing the coarser detail at levels $k < l$ from the series, level l captures the detail in 2^l coefficients. The Haar wavelet transform yields a tree of increasingly detailed coefficients at level l.

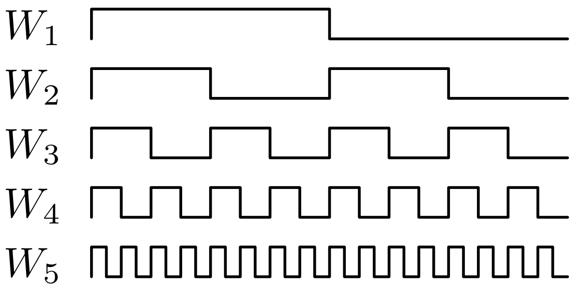


Figure 5: Haar wavelet symbols W_l. The symbol at level l allows to summarise the remaining level l signal detail using 2^l coefficients. Where the remaining signal at level l equals the remaining signal at level $l-1$ minus the coefficients at level $l-1$, for $l \geqslant 2$.

2.3.2.1.1 Wavelet transform example

Figure 6 illustrates the decomposition and recombination of the activation sequence $s = (5, 3, 7, 9)$ using components $c_{i,l|i\in\{1,\cdots,2^l\},l\in 0,1,2}$. Coefficient $c_{1,0}$ equals 6 which is the mean of sequence s. Application of W_1 to the remainder $s - c_{1,0} = (-1, -3, 1, 3)$ yields means $c_{.,1} = (-2, 2)$. Then the remainder $s - c_{1,0} - c_{.,1} = (1, -1, -1, 1)$ equals the coefficients $c_{.,l=2}$ at level 2. By adding the coefficients: $3 = 6 - 2 - 1$, $7 = 6 + 2 - 1$, etc., s is fully recovered.

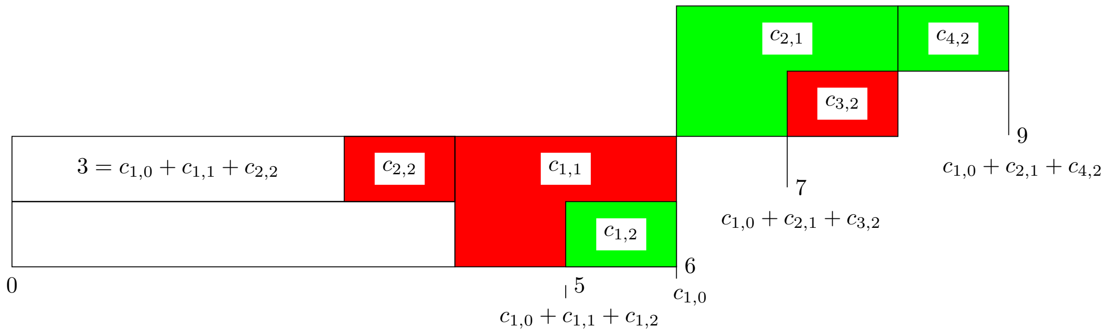

$c_{2,1}$
$c_{4,2}$
$c_{3,2}$
9
$c_{1,0} + c_{2,1} + c_{4,2}$
$3 = c_{1,0} + c_{1,1} + c_{2,2}$
$c_{2,2}$
$c_{1,1}$
7
$c_{1,0} + c_{2,1} + c_{3,2}$
$c_{1,2}$
6
0
5
$c_{1,0}$
$c_{1,0} + c_{1,1} + c_{1,2}$

Figure 6: Haar wavelet decomposition of an activation sequence $s = (5, 3, 7, 9)$ into coefficients $c_{i,l}$. Red bars indicate coefficients with a negative sign. Coefficient $c_{1,0}$ is the mean of s. Coefficients $c_{i,l|i\in\{1,\cdots,2^l\},l\in\mathbb{N}_{>0}}$ capture the remainder details on level l using symbol W_l.

2.3.2.2 Wavelet subband energy

With amplitudes of 1 and -1, Haar wavelet symbols are either of unit length for $\left(W_i^\top W_i\right)/2^i = 1$ or perpendicular for $W_i \perp W_{j\neq i}$, which means they form an orthonormal basis. Because of the orthonormal basis, the Haar wavelet transform (HWT) obeys Parseval's energy identity in that the sum of the squared function equals the sum of its squared transform coefficients. The energy in s equals $s^\top s = 164$ which in turn equals $4c_{1,0}^2 + 2c_{\cdot,1}^2 + c_{\cdot,2}^2 = 4(6^2 + 2^2 + 1^2)$.
The R package 'wmtsa: Wavelet Methods for Time Series Analysis' in [5] by Constantine and Percival, 2017, includes support for Haar wavelet transform. Its function `wavMODWT` yields the coefficients at individual subband levels l. The energy in subbands equals the sum of the squared coefficients at each level. When normalising subband energies, the energy at level l is used as the wavelet symbol W_l probability in 2.4.

2.4 Windows chain to their most correlated nearest neighbour

Although Leonardi et al., in [11] try to find correlations between 88 voxel regions and the new method between all V individual voxels, both methods aim to capture correlation between activations. Because the new method works on the smallest i.e. voxel unit, correlation cannot be naively computed in the usual matrix form, N windows are connected to their most correlated neighbour in a chain. Sufficiently correlated consecutive window sub-strings remain connected in their categorical study words separated with the chain cutter in 2.5.
The metric distance between codes in 2.4.1 is used as a similarity measure between stationary subsequences in windows. 2.4.2 defines the chain as a graph with windows as vertex and edges representing metric distance between vertices.
To find the most similar sequence among the $N - 1$ candidates by directly calculating and storing a triangular pairwise distance matrix of order $\mathcal{O}(N^2)$ is prohibitive for the N under consideration. Therefore 2.4.2.1 uses binary search on a chain of stationary subsequences with minimal between subsequence cosine distance. The nearest neighbors are identified with successive range halving using minimum edge cuts which is of order $\mathcal{O}(N \cdot \log_2 N)$ complexity. Given those nearest neighbor vertices, 2.4.2.2 inserts the candidate vertex at the maximum edge cut where the candidate adds minimal cosine distance to the chain.

2.4.1 Window neighbour is at metric cosine distance

Distance between windows is defined to find and connect a window to its nearest most correlated neighbour in the chain. Given stationary windows P and Q determined in 2.3, the complement of the inner product of their unit length probability codes $\mathbb{P}/\|\mathbb{P}\|_2$ and $\mathbb{Q}/\|\mathbb{Q}\|_2$ measures window dissimilarity by metric cosine distance $\delta_{\cos}(\mathbb{P},\mathbb{Q})$ in (1).

$$0 = \delta_{\cos}(\mathbb{P},\mathbb{P}) \leqslant 1 - \left(\frac{\mathbb{P}}{\|\mathbb{P}\|_2}\right)^\top \left(\frac{\mathbb{Q}}{\|\mathbb{Q}\|_2}\right) = \delta_{\cos}(\mathbb{P},\mathbb{Q}) \leqslant \delta_{\cos}(\mathbb{P},\mathbb{Q}\,|\mathbb{P}\perp\mathbb{Q}) = 1, \|\mathbb{P}\|_2 = \sqrt[2]{\mathbb{P}^\top\mathbb{P}} \quad (1)$$

2.4.2 Window finds and connects to its nearest neighbour

A vertex in the undirected chain graph represents a unique window with stationary sequence in a series. A chain edge represents the cosine distance $\delta_{\cos}$ between the two vertices that it connects. Except for the head of the chain, each edge connects one vertex with its only predecessor vertex. To allow binary search to half the search range, the graph is maintained as a chain of connected vertices sorted by minimum edge length.

The search space for the nearest neighbours of candidate C in the sorted chain with at least two vertices starts with the full sequence between the *Low* and *High* boundary indices; (2) positions the *Middle* index halfway the floor between the boundaries.

$$Middle = \lfloor\frac{High + Low}{2}\rfloor \quad (2)$$

2.4.2.1 Search for nearest window neighbours

Before algorithm 2 inserts a new vertex, iterations of algorithm 1 halve the search space until the boundaries have squeezed out the *Middle* index (2) which is when *Low* equals *Middle*. In algorithm 1, the smaller circle through C with its nearest boundary vertex at the origin disconnects the *Middle* vertex from the vertex at the remaining boundary index.

To clip the search range when the *Middle* vertex **is not** on the shortest edge to C, the index of the vertex on the longest boundary edge moves to *Middle*. In Figure 7 the *boundary* vertex V_L is the origin of the smaller of the tree circles through C therefore index H moves to index M.

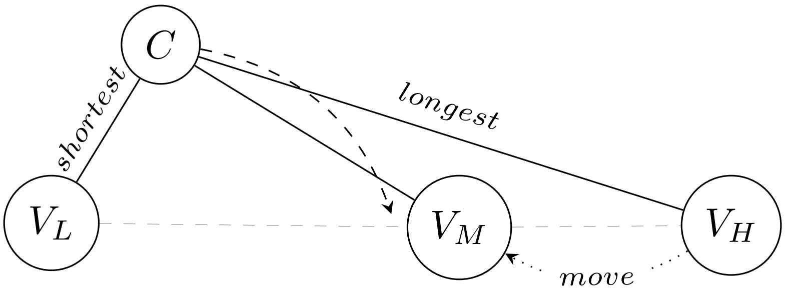


Figure 7: shortest edge at boundary stays

Otherwise *Middle* **is** the vertex on the shortest edge connecting C. Then the smallest *boundary* circle disconnects its own origin from the *Middle* vertex V_M if its origin is closer to C than to V_M, which in Figure 8 is **not** the case; in the next iteration the nearest neighbors of C are only searched between the current M and H indices in the vertex chain.

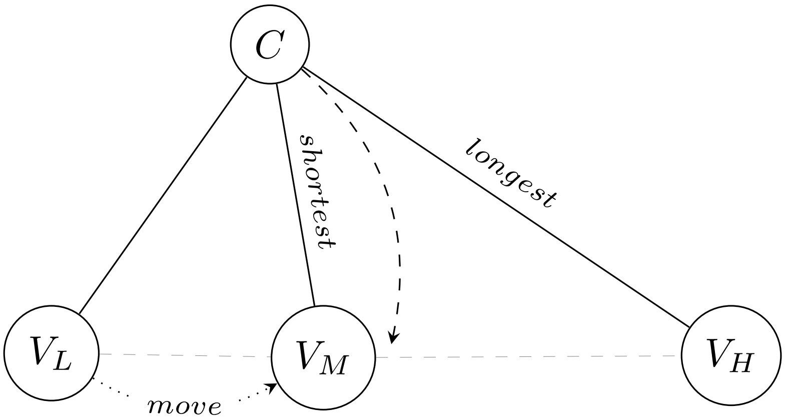


Figure 8: Middle is on the shortest edge to C, then the disconnected boundary vertex stays

2.4.2.2 Window insertion at minimum cosine distance

Algorithm 2 inserts C in the chain $(V_{L-1}, V_L, V_{L+1}, V_{H+1})$ at the cut made by the largest circle through C and the most distant of the two remaining vertices V_L and V_{L+1} at the origin.

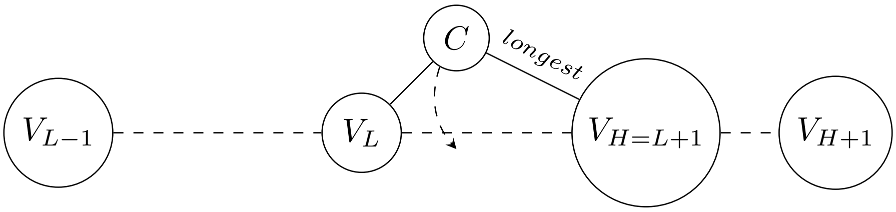


Figure 9: insert between boundaries

The larger circle disconnects both borders at circle radius length in Figure 9. C repairs the chain by connecting to both boundary vertices.

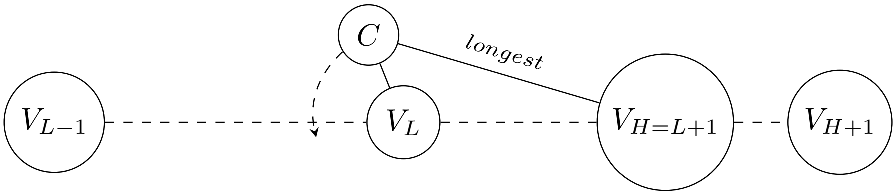


Figure 10: insert outside triangle

Figure 10 shows that C repairs the chain at its nearest boundary V_L which remains chained to its peer V_H to ensure that the average edge length does not increase.

2.5 Words separate windows with too low mutual information

Even though 2.4 chained all windows to their most correlated nearest neighbour, neighbours with too little mutual information end up in separate words in the study bag.

Algorithm 3 cuts the chain at word boundaries where the split neighbouring windows have too much unique information relative to that in their fused window. At a word boundary, the shackle between the windows is cut if the fused and split windows lack mutual information.

To develop the mutual information driven word boundary detector in 2.5.5, this section first reviews information in 2.5.1, expected information in 2.5.2, mutual information in 2.5.3, and measuring lack of mutual information in 2.5.4.

As the main result, regions emerge from document voxels V drawing words from the study bag, at the original scan resolution T in 2.5.6.

2.5.1 Probabilistic coding reveals information as optimal code length

Probabilistic coding assigns the shortest code to the most frequent symbols in series. The binary prefix code tree in Figure 11 produces a *complete* prefix code for symbols $\{S_i\}_{i=1}^{m=6}$ in a series. With a longer path from the root with more edge bits, lesser probable symbol nodes S_j are at a deeper or equal tree level than more probable symbol nodes S_i. Frequent symbols with short codes compress the series, at the cost of longer codes for rare symbols.

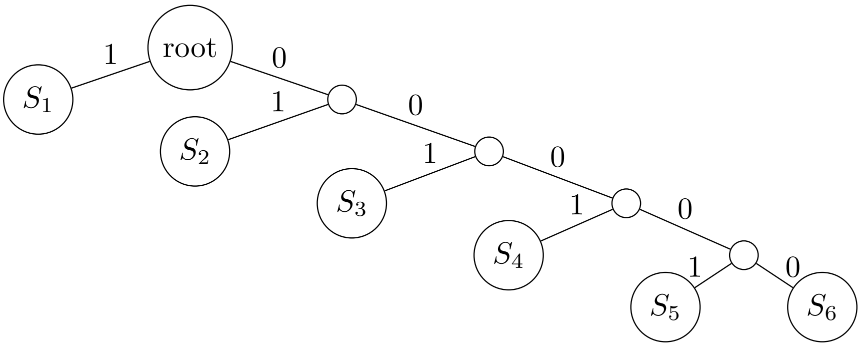


Figure 11: Complete binary prefix code tree for symbols in Table 1. The tree assigns codes to symbols at nodes S_i by joining the edge bits on its shortest path from the root node

Symbol	$\#S_i$	$\mathbb{P}_{S_i}$	$L_C(S_i)$	code	[bits]	$\mathbb{E}[bits]$
S_1	32	2^{5-6}	1	1	$32 \cdot 1$	$2^{-1} \cdot 1$
S_2	16	2^{4-6}	2	01	$16 \cdot 2$	$2^{-2} \cdot 2$
S_3	8	2^{3-6}	3	001	$8 \cdot 3$	$2^{-3} \cdot 3$
S_4	4	2^{2-6}	4	0001	$4 \cdot 4$	$2^{-4} \cdot 4$
S_5	2	2^{1-6}	5	00001	$2 \cdot 5$	$2^{-5} \cdot 5$
S_6	2	2^{1-6}	5	00000	$2 \cdot 5$	$2^{-5} \cdot 5$
sum	64	1			124	124/64

Table 1: Given the symbol counts $\#S_i$ in a series of 64 (2^6) symbols, correspondence between probability $\mathbb{P}_{S_i}$, prefix code length $L_C(S_i)$, *code*, information [*bits*] and expected information $\mathbb{E}[bits]$

Table 1 lists the given symbol counts $\#S_i$ for a series of 64 symbols. Normalising the symbol counts $\#S_i$ by their total of 64 gives probabilities $\mathbb{P}_{S_i}$. The tree in Figure 11 is constructed by keeping frequent symbols at short distance and pushing rare symbols down at longer path distance. Path distance is measured in number of edges or equivalently number of bits from the root node. All the right edges in the tree get the same bit value, the left edges are assigned the complementary binary value.

To code for symbol S_i, the tree in Figure 11 joins the bits on the shortest path from the root node to the symbol node S_i; the lengths $L_C\left(S_i\right)$ equal the number of edges. The node for symbol S_1 is just one bit edge from the root node which implies S_1 is most frequent in the given series of symbols. To distinguish the single 1 bit code for S_1, the path to S_2 *prefixes* it with a 0 bit. Code lengths for symbols equal their tree level, e.g. rare symbols S_5 and S_6 at tree level 5 have 5-bit codes 00001 and 00000 respectively. The code lengths $L_C\left(S_i\right)$ equal $-\log_2\left(\mathbb{P}_{S_i}\right)$ which are weighted by their probabilities $\mathbb{P}_{S_i}$ in the expected information in $\mathbb{E}[bits]$ or entropy in 2.5.2. A generator using probabilities $\mathbb{P}_{S_i}$ in Table 1 produces N symbols with expected information of $N \times 124/64 \approx 2N$ *bits*.

2.5.2 Expected information or entropy weighs optimal code lengths

Expected information based upon the code length is reviewed for its role in measuring lack of mutual information by the word chain cutter in 2.5.5. Developed in 2.5.1, minimum code lengths $L_{C|S\sim\mathbb{P}}(S_i)$ measure the information in symbols as $-\log_2\left(\mathbb{P}_{S_i}\right)$ bits. The *average* symbol length is the sum of symbol code lengths weighted by their probabilities of occurring in a series. This *expected* information ($\mathbb{E}$) in bits for a series of symbols S generated according to probability distribution $\{\mathbb{P}_i\}_{i=1}^{m}$, with shorthand notation $(S \sim \mathbb{P}_i)$, equals the *entropy*, developed by Shannon in 1948 [15], $H(\mathbb{P})$ defined in (3) for code C with optimal length $L_{C|S\sim\mathbb{P}}$.

$$\begin{aligned} 0 = \log\left(1\right) \leqslant H(\mathbb{P}|m) &= \sum_{i=1}^{m} \mathbb{P}_i \cdot -\log_2 \mathbb{P}_i = \sum_{i=1}^{m} \mathbb{P}_i \cdot L_{C|S\sim\mathbb{P}}\left(S_i\right) \\ &= \sum_{i=1}^{m} \mathop{\mathbb{E}}_{S_i\sim\mathbb{P}_i} L_{C|S\sim\mathbb{P}}\left(S_i\right) = \mathop{\mathbb{E}}_{S\sim\mathbb{P}} L_{C|S\sim\mathbb{P}}\left(S\right) \leqslant \log_2\left(m\right) \end{aligned} \tag{3}$$

With $\lim\limits_{x\downarrow 0} x\log(x) = 0$, a completely predictable activation, which puts **all** its probability mass on **one** of the m symbols, has unsurprisingly $H = \log(1)$. Each of the m symbols are equally likely with *uniform* probability $(1/m)$ in completely random white noise activations that maximise the entropy at $\log_2(m)$.

2.5.3 Mutual information bonds adjacent windows

Multual information as a bonding force in fused neighbouring windows is reviewed before measuring split window divergence sums in 2.5.4.

Entropies are sums of weighted symbol lengths which in turn add up as areas in e.g. relationship diagrams named after logician John Venn. The Venn diagram in Figure 12 illustrates relationships between *joint*, *marginal*, *conditional* and *mutual* information in adjacent windows. Inflating the white intersection with *mutual* information $I(\mathbb{P},\mathbb{Q})$ between windows deflates the blue and red shaded *conditional* information areas. Independent windows have a very weak bond of small mutual information. Highly correlated adjacent windows P and Q are strongly bonded by their large *mutual* information $I(P,Q)$.

Separate windows use their own *marginal* distributions $\mathbb{P}$ and $\mathbb{Q}$ to generate series with twice the mutual information. The fused *joint* entropy $H(\mathbb{P},\mathbb{Q})$ in the series equals the sum of the *marginal* entropy in a window P, which includes *partial* entropy from neighbour window Q, plus the *unique conditional* entropy $H(\mathbb{Q}|\mathbb{P})$ part of window Q.

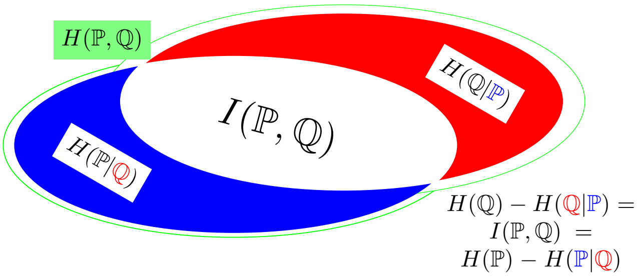

$H(\mathbb{P},\mathbb{Q})$
$H(\mathbb{Q}|\mathbb{P})$
$I(\mathbb{P},\mathbb{Q})$
$H(\mathbb{P}|\mathbb{Q})$
$H(\mathbb{Q}) - H(\mathbb{Q}|\mathbb{P}) =$
$I(\mathbb{P},\mathbb{Q}) =$
$H(\mathbb{P}) - H(\mathbb{P}|\mathbb{Q})$

Figure 12: Role of mutual information $I(\mathbb{P},\mathbb{Q})$ relative to marginal, conditional and joint information in the adjacent windows of fusion $R = (P, Q)$. Where separate windows duplicate their mutual information, their fusion uses the joint information to generate the mutual information once.

Because mutual information is symmetric in its window parameters, the marginal and conditional window order is arbitrary. By fully summing both circles, the budget *duplicates* the mutual information. To avoid duplication, the fusion generates the series from its *joint* entropy $H(\mathbb{Q},\mathbb{P})$. As the mutual information increases, the fusion decreases the budget overrun. Maximum mutual information fully covers the smaller marginal entropy area. Maximally bonded identical marginals allow the fusion to drive its cost down to the budget without overruns.

2.5.4 Relative entropy measures shortfall of mutual information

Sums of relative entropy, which is reviewed here, measure lack of mutual information in fused neighbour windows developed in 2.5.5. Where (3) measures the expected information or entropy, the *relative entropy* subtracts the entropy from another suboptimal expected code length. Kullback, [10] 1959, developed relative entropy with Leibler which is known as Kullback-Leibler divergence D_{KL}. Defined in (4) D_{KL} measures the coding inefficiency in *expected* number of *additional* bits needed when symbols generated according to distribution $\mathbb{P}$ are encoded with a code optimal for **another** symbol distribution $\mathbb{Q}$. For symbols S that are distributed as $\mathbb{P}$, the code length $L_{C|S\sim\mathbb{P}}(S)$ is optimal, any other code dissipates extra bits.

$$0 \leqslant D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \sum_{i=1}^{m} (\mathbb{P}_i \cdot -\log_2 \mathbb{Q}_i) - H(\mathbb{P}) = \mathop{\mathbb{E}}_{S\sim\mathbb{P}} \left[L_{C|S\sim\mathbb{Q}}(S) - L_{C|S\sim\mathbb{P}}(S)\right] \quad (4)$$

For example, if a receiver is not informed about the m symbol probabilities $\mathbb{P}_{Si}$ it must assume equal probabilities of a uniform distribution $\mathbb{U}_{Si}$ with code lengths $L_{C|S\sim\mathbb{U}}(S) = \log_2(m)$. For that suboptimal symbol distribution, given $\mathbb{P}_{S_i}$ from table 1, the expected information $\sum\limits_{S\sim\mathbb{P}} \mathbb{P} \cdot L_{C|S\sim\mathbb{U}}(S)$ is $2.75[bits]$. When subtracting the optimal $H(\mathbb{P}) = 124/64 \approx 2$, the diverging non-optimal uniform code costs $3/4$ extra bits per symbol on average.

2.5.5 Low mutual information splits adjacent windows at word boundaries

As illustrated with the change point detector in Figure 13, a stationary fused series $R = (P, Q)$ of length $|R|$ and distributed as $\mathbb{R}$ exceeds its budget fixed at the summed marginal entropy for the optimal symbol distributions in split window P and Q with few bits. The excess bits in the sum of the two Kullback-Leibler divergences is proportional to the negative Log Likelihood Ratio

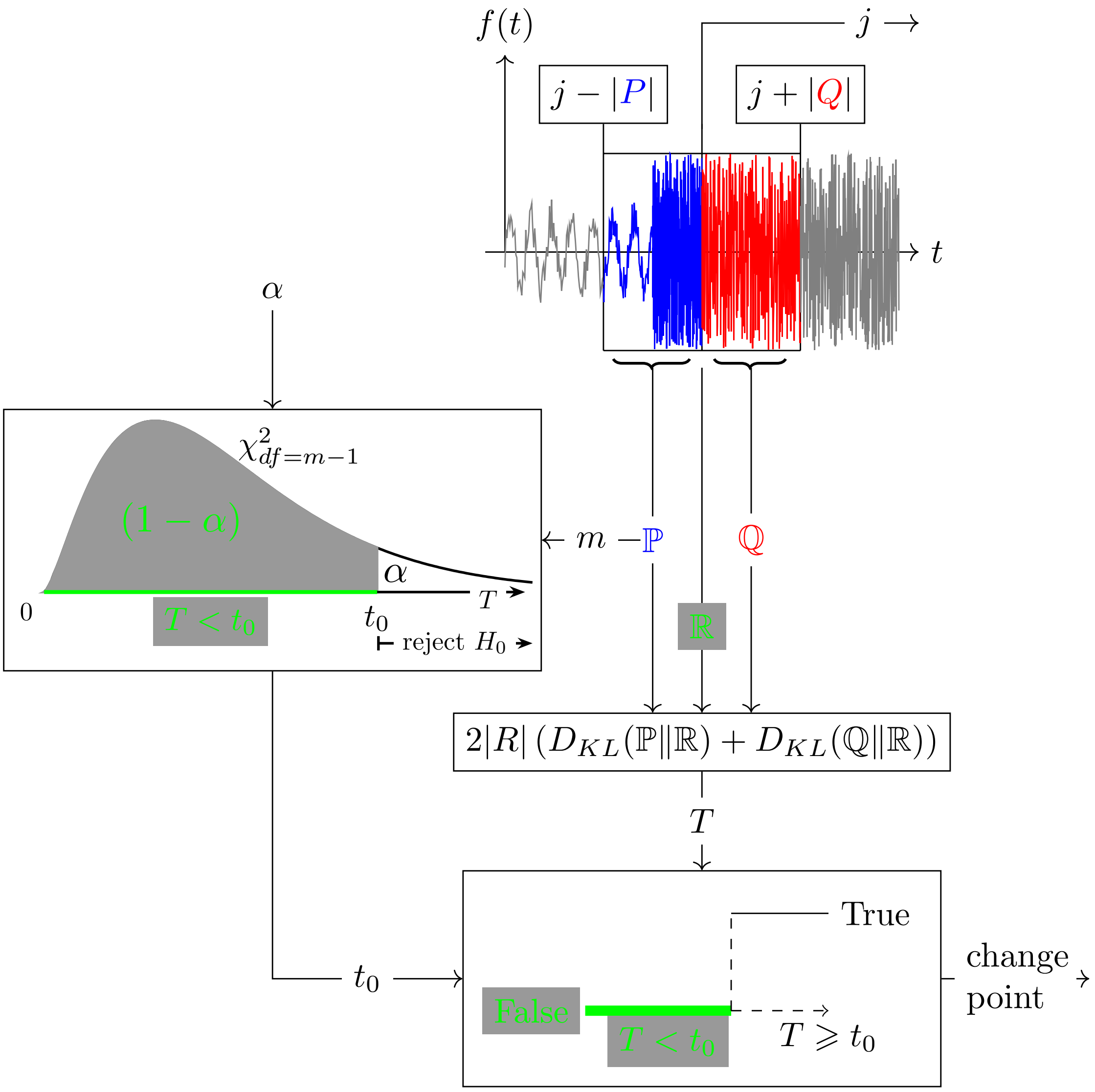


Figure 13: Change point detector of significant budget overruns T of at least t_0 bits at junction j in a series $f(t)$. Consecutive windows P and Q only share junction j. The budget is fixed at the summed entropy of the optimal symbol distributions $\mathbb{P}$ and $\mathbb{Q}$ in the individual windows P and Q respectively. The fusion $R = (P, Q)$ exceeds the budget with weighted sums of the sub optimal distribution lengths $L_{C|S\sim\mathbb{R}}$ by weights $\mathbb{P}$ and $\mathbb{Q}$. With probability $1 - \alpha$ under null hypothesis $\overset{H_0}{\sim} \chi^2_{df=m-1}$, the budget overrun is insignificant.

(LLR) (5) which asymptotically follows the χ^2 distribution in (6), appendix B of [4], Brandmaier, 2016.

$$LLR(P, Q|R) = 2|R|(D_{\text{KL}}(\mathbb{P}\|\mathbb{R}) + D_{\text{KL}}(\mathbb{Q}\|\mathbb{R})) \tag{5}$$

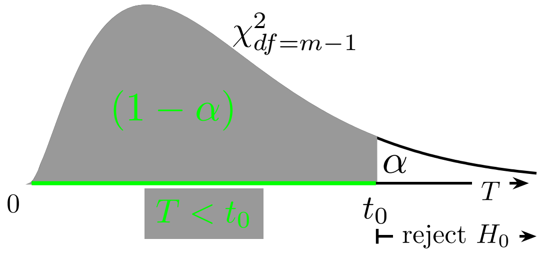
The χ^2 distributed statistic T in (6) looses one degree of freedom because the m symbol probabilities add to one.

$$T = LLR(P, Q|R) \sim \chi^2_{m-1} \tag{6}$$

When the nested null hypothesis H_0 of stationary fusion $R = (P, Q)$ is rejected for T equal of greater than t_0, the junction between P and Q is a change point.
Under the null distribution illustrated in Figure 14 for $m = 6$, the probability that the fusion entropy T matches or exceeds critical value t_0, as defined in (7), equals the desired significance level α.

$$Pr(T \geqslant t_0 | T \overset{H_0}{\sim} \chi^2_{m-1}) = \alpha \in [0, 1] \tag{7}$$

Figure 14: For $m = 6$ symbol coded windowed series, the summed divergence T of fused windows is proportional to the $\chi^2_{df=5}$ distribution with 5 degrees of freedom. Relative entropy T in excess of critical value t_0 rejects the null hypothesis H_0. With probability $1 - \alpha$ under H_0, fusion bits in excess of the summed entropy bits in the split windows are insignificant.



Algorithm 3 cuts the chain of the nearest stationary windows from 2.4 at stationary *word* boundaries, when and where H_0 is rejected by evaluating the relative entropy according to (4) for pairs of neighbours using test statistic T and critical value t_0 specified in (6) and (7) respectively. Starting from 1, Word chains of connected windows are numbered categorically and bagged. The numbers below the subsequences making up a word share a color in figure 15.

2.5.6 Detailed regions emerge from document voxels drawing study words

As assigned in 2.5.5, in the scan taken at t each of the V voxels selects the word which holds its stationary subsequence; each scan becomes a document with V words drawn from bag B. This describes unwinding $B^{S\times V\times L_i}$ with lengths L_i to $B^{S\times V\times T}$.
The original stationary sequences making up the words are colored by their word. The 4 voxels in the documents draw a word from the bag at any instance of the series. The indices of the stationary sequences for each of the voxels are colored by their word in Figure 16.
Regions emerge as neighbouring voxels drawing the same word/color. Distinct regions drawing the same word participate in functional connected networks. Dynamics in functional connectivity develops on the original scan resolution time as regions change word.

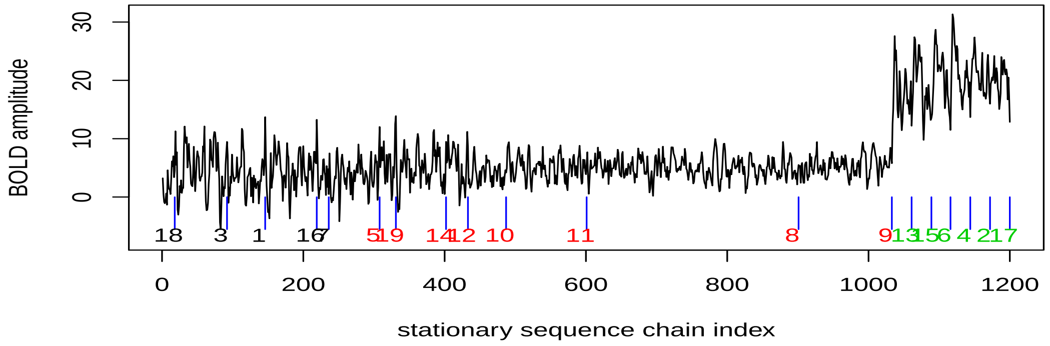


Figure 15: Subsequences reordered at smallest between subsequence cosine distance. Change points separate words with numbered sequences of like colors.

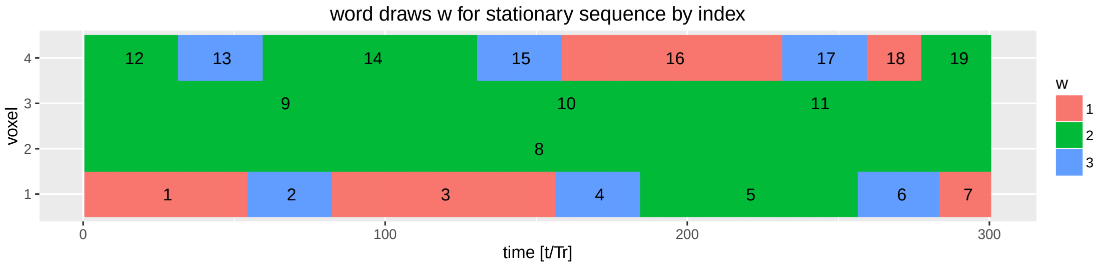


Figure 16: Stationary numbered subsequences from window segments in Figure 4 of length 300 for each of the 4 voxels colored by their word out of the Bag with three words.

2.6 Scan document state switching interpretation

Interpreting documents as categorical word count draws from the study bag filled in 2.5.5, the word count vector of unit length is a point on the unit sphere. Similar documents cluster around their state centroid. Consecutive documents switch between state at a subject specific rate.

2.6.1 Word frequencies fix document location

When represented by its word frequency vector $F^{S \times T}_{word \in B}$, a document for subject s at time t is positioned on a surface with one dimension lower than the bag size. The 4-voxel document drawing 3 words, as illustrated in Figure 16, visits the states in Figure 17. The number of words that its voxels use at a particular time defines the document state. The documents drawing words from the simulation bag, frequent a limited set of unique states located on the probability simplex in Figure 17.

States are identified as counts of words, e.g. '130' identifies the state with one word 1, three times word 2 and no word 3. E.g. in Figure 16 the 4 voxels are in state '130' during subsequences 12 and 7; after subsequence 2 but before subsequence 15; and after subsequence 4 but before

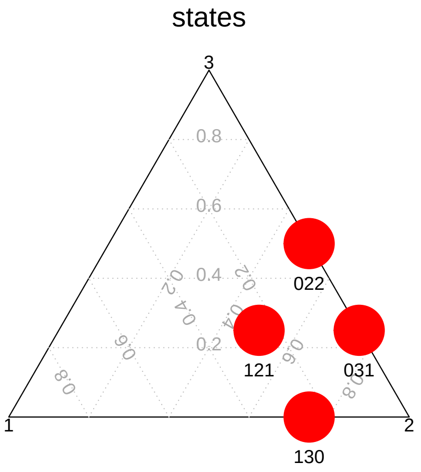


Figure 17: States for the Bag with 3 words. The number of voxels drawing word $w \in \{3, \cdots, 1\}$ make up the states, e.g. 022 is the state which results from the four-voxel document which draws zero $w = 3$, two $w = 2$, and two $w = 1$.

subsequence 17. With few simulated voxels, exact word counts fix the document state without the need to cluster them around a state centroid.
Selecting one of the words in the state vector, it activates distinct neighbourhoods of voxels, i.e. functionally connected brain regions. Thus functional connectivity emerges between distinct regions of voxel neighbourhoods sizes summing to the word count in the state vector.

2.6.2 Document state as mean location

The word frequency vector $F^{s \times t}_{word \in B}$ fixes its document location on a probability simplex as in 2.6.1. Then multiple documents may be more or less similar to each other, at small cosine distance $\delta_{\cos}$ defined in 2.4. To measure cosine similarity, vectors are scaled to their unit length. Vectors w of unit length start at the origin and end on a point on a unit sphere. When projected on the unit sphere, similar documents concentrate around their mean *state* vector at small cosine distance. A document *state* has a mean direction μ around which documents concentrate with parameter κ in the von Misses-Fisher (vMF) probability density function defined in (8), Banerjee et al., 2006 [1].

$$vMF(w|\mu, \kappa) = C_d(\kappa) e^{\kappa(\mu^T w)}, \; w \in \mathbb{R}^d, d \geqslant 2, \|w\|_2 = 1, \|\mu\|_2 = 1, \; \kappa > 0 \tag{8}$$

In the exponent of (8), the inner product $\mu^\top w$ measures cosine similarity between the d-dimensional mean and sample vectors, which is multiplied by concentration parameter κ. Near zero values of κ turns the exponential into the near constant 1 regardless of w, which allows samples to stray away uniformly around the mean direction. Whereas high values of κ increasingly reward w vectors that are similar to the mean with decreasing variance. As κ tends to infinity, it selects vectors w equal to the mean μ and pulls them towards that point concentration.

$$C_{d(\kappa)} = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} \tag{9}$$

The normalising constant $C_{d(\kappa)}$ defined in (9) ensures that (8) integrates to one as a probability density function. The modified Bessel function $I_{\frac{d}{2}-1}(\kappa)$ can only be approximated numerically.

2.6.3 Document state mixtures

Similarly to Leonardi, et al., where pairwise regional covariances in [11] change their loading on 10 'eigenconnectivities' during the windowed series, it is reasonable to allow documents to switch state over the course of the series. To that end, one of K states in a mixture of vMF ($movMF$) model in (10), by Banerjee et al., 2006 in [1], generates a document word vector.

$$\sum_{k=1}^{K} \pi_k \cdot vMF(w_i|\mu_k, \kappa_k),\ \pi_k \in [0,1]\,,\ \sum_{k=1}^{K} \pi_k = 1 \tag{10}$$

The document word vectors are generated from the mixture of states in (10), where states have their own mean direction and concentration. An independent document i is sampled by first drawing one of K states with probability $\vec{\pi}$ before the drawn mixture component k generates the document point for w_i on the sphere.

2.6.3.1 Learning the mixture states

The mixture weights and state parameters are learned using the EM algorithm described in 2.6.3.1.1 whereas the number states K are selected for mixtures with low BIC as described in 2.6.3.1.2.

2.6.3.1.1 State probabilities, means and concentrations

In an unsupervised setting the document state label vector Z is missing or hidden, making Z a random variable. In [1], Banerjee et al., find the most likely hidden Z, conditional on the most likely joint distribution of document words and state vectors.
In algorithm 4, they climb the locally concave likelihood in a non-decreasing coordinate ascent fashion with their alternating Expectation and Maximisation (EM) steps. Taking the t^{th} state mixture vector $\widehat{(\vec{\pi}^t, \vec{\mu}, \vec{\kappa})}^{t}$ from the M-step as fixed, the E-step puts all the probability mass for document w_i on the most likely state $k_i \in [1, \cdot\cdot, K]$ in line 10.
After the E-step assigns all documents w_i their hidden state $\hat{z}_{i,k_i}$, the M-step takes them as fixed to increase the likelihood with updated mixture weights $\hat{\vec{\pi}}_h^{t+1}$, more accurate means $\hat{\vec{\mu}}_k^{t+1}$, and more precise concentrations $\hat{\vec{\kappa}}^{t+1}$; the E and M steps alternate while they increment the likelihood ℓ with more than a set minimum ϵ.

2.6.3.1.2 Number of mixture states

The number of states K in the mixture is unknown. Hornik & Grün, 2014, in addition to implementing the fitting procedures described in 2.6.3.1.1, also provide a function that calculates the Bayesian Information Criterion (BIC) for fitted K-component mixtures in [8].

$$\log \ell\,(w|\mu,\kappa,\pi) = \left(\sum_{i=1}^{T} (\log(\pi_{k_i})) + \log(C_d(\kappa_{k_i})) + \kappa_{k_i}\left(\mu_{k_i}^T w_{k_i}\right)\right) \tag{11}$$

$$BIC = K \cdot \log(T) - 2\log \ell\,(w|\mu,\kappa,\pi) \tag{12}$$

Swartz developed BIC in [16] in 1978, defined in (12), BIC penalises complex models with large K while rewarding their good fit measured in log likelihood (11). A grid search selects the mixture with high likelihood and few states K at minimum BIC, to avoid too many states with few document word vector points.

2.7 Subject 'resting state' document state occupation

The last M-step in Algorithm 4 assigns the word frequencies $F^{S\times T}_{word\in B}$ of the *words* in bag B for a subject s at time t to one of K states $[1,\cdot\cdot,K]_{s,t}$. Over the series each subject traces its own state path on the unit sphere with subject specific state occupation duration.
Where Leonardi et al., in [11] analyse the dynamic nature of the loadings on 'eigenconnectivities' to estimate the group membership of subjects, here the document state occupation duration is used. Out of the K discrete states, a document enters its nearest state k where it stays during the next f scans with leave probability ϕ_s as illustrated in Figure 18. The probability for f failed escapes from the current state follows the geometric probability density function (pdf) in (13).

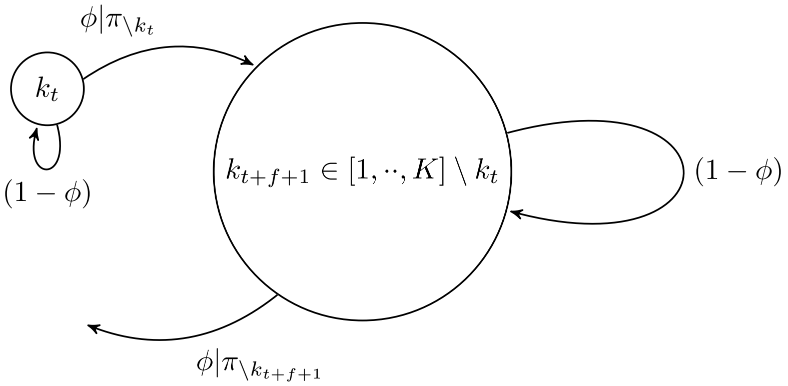

$\phi|\pi_{\setminus k_t}$
k_t
$(1-\phi)$
$k_{t+f+1}\in[1,\cdot\cdot,K]\setminus k_t$
$(1-\phi)$
$\phi|\pi_{\setminus k_{t+f+1}}$

Figure 18: Discrete time, discrete sticky state Markov chain. With probability ϕ_s, after occupying state k for f consecutive steps, a document selects another state.

$$Pr\left(F=f|f\in\mathbb{N}_{\geqslant 0}\right)=(1-\phi)^f\cdot\phi,\ 0\leqslant\phi\leqslant 1 \tag{13}$$

The concave log likelihood $\ell(\phi;f)$ for n independent state occupations equals (14).

$$\log\ell\left(\phi;f\right)=\log\left(\prod_{i=1}^{n=|f|}\left(1-\phi\right)^{f_i}\cdot\phi\right)=\sum_{i=1}^{n}\left(f_i\cdot\log\left(1-\phi\right)\right)+n\cdot\log\phi=$$
$$n\left(\bar{f}\cdot\log\left(1-\phi\right)+\log\phi\right) \tag{14}$$

The first derivative gives the slopes of the log likelihood as the score function (15).

$$u\left(\phi|f\right)=\frac{\delta}{\delta\phi}\log\ell\left(\phi;f\right)=n\left(\frac{1}{\phi}-\frac{\bar{f}}{1-\phi}\right) \tag{15}$$

Setting the score function $u(\phi)$ to zero slope yields the maximum likelihood estimator $\widehat{\phi}_{MLE}=1/(1+\bar{f})$.
The negative second derivative $-\frac{\delta u(\phi)}{\delta\phi}$ yields the Fisher information I_F in (16) as a measure of curvature of the likelihood function for the occupations f at parameter ϕ. Low curvature of $\log\ell(\phi)$ with ϕ around $\widehat{\phi}_{MLE}$ gives small confidence in the precision of that maximum

likelihood estimator.

$$-\frac{\delta u(\phi)}{\delta \phi} = I_F(\phi|f) = n \cdot \left(\frac{1}{\phi^2} + \frac{\bar{f}}{(1-\phi)^2}\right) = \frac{1}{\sigma^2(\phi)} \tag{16}$$

When near the maximum, the likelihood drops sharply for the n state visits that subject s makes, the high Fisher information and equivalently the small standard error $\sigma(\widehat{\phi}_s)$ in (16) inflates the standardised distance to the MLE in test statistic Z_s (17).

Given the confidence level, $\frac{\alpha}{2}$ equals the probability for a standard normal statistic to exceed critical value z_0. The null hypothesis H_0 is that the estimated subject state switch probability $\widehat{\phi}_s$ equals that of the study estimate $\widehat{\phi}_0 = \widehat{\phi}_{MLE}$. H_0 is rejected when the absolute value of Z_s defined in (17) matches or exceeds critical value z_0.

$$Z_s|H_0 = \frac{\left(\widehat{\phi}_s - \widehat{\phi}_0\right)}{\sigma(\widehat{\phi}_s)} \sim \mathcal{N}(0,1) \tag{17}$$

Wald developed asymptotically normally distributed Z_s in [18] in 1943.

2.8 Method summary

The dimension reduction and correlation approximation Leonardi et al., in [11], come with loss of resolution and involved interpretation. The alternative method captures all stationary within- and between-voxel activation correlation in the study bag of words without loss of resolution.

Energy based coding of similar consecutive voxel activation captures amplitude and frequency characteristics in stationary windows of variable width with a fixed length probability code. To avoid the computing complexity associated with computing and storing all pairwise correlations, windows are chained to their most correlated neighbour using a binary insertion search algorithm. The between window cosine distance metric allows efficient binary insertion sort of windows chained at their nearest neighbour. The window codes are also efficient to measure the unique information in split windows relative to the information in their fusion; excess unique information keeps windows split and thus also separates words in the study bag. Words in the study bag activate distinct voxel neighbourhoods, i.e. functionally connected regions at the scan document time resolution.

The effort associated with capturing correlation at voxel resolution yields emergent functional connectivity at observed time resolution without having to interpret weighted vectors of pairwise regional correlations at a condensed time scale.

The sliding fixed time width window that Leonardi et al., in [11] use to calculate pairwise regional correlation may miss time-lagged identical activation where the alternative method captures the lagged but highly correlated windows of variable width in the same word.

The unit length categorical word count vector interpretation clusters similar documents around their state vector, where subject documents have their own state switching dynamics which may be associated with a neurological condition.

3 Simulation and results

The brain in 'resting state' is assumed to dynamically co-activate regions over the course of fMRI series. The simulation activates 2 out of the 4-voxel series of length 300 by convolution of a sequence of onset bursts with a Haemodynamic Response Function. High frequency system noise

and low frequency noise from scanner drift, heart and respiration are added to the activations only. Due to the relatively slow changes in blood flow, the additive autoregressive temporal noise with lag 2 is relatively strong for all voxels. The goal is to find different state occupation statistics given different activation onset burst durations in two groups. The state occupation statistics in turn depend on the word count vectors that the document voxels draw from the study bag.
Temporal design parameters are defined with activation and noise parameters in 3.1.1 for individual voxel series for which 3.1.2 specifies additive noise; spatial correlation is simulated by temporally aligning voxel activations in 3.1.3.
The bold series simulated in 3.1.3 are segmented into stationary subsequence windows using the procedure in 2.3.1, which are chained at minimum between window distance using 2.4 and Bagged using the procedure in 2.5.
The 4 voxel word count is evaluated as exact state evolution in 3.3 and the occupation duration in two groups in compared in 3.4 which estimates the state mixture.

3.1 Temporal design parameters

3.1.1 Activation and noise weights

For the specified starts or **onsets**[1] each with **durations=4**, the temporal design adds **effectsize=20** amplitudes proportional to the Haemodynamic Response Function **hrf='double-gamma'** additive to voxel **base=5** amplitudes. A noise vector of **type='mixture'** with **system=.1**, scanner **drift=.01**, **temporal=.8**, **physiological=.09**, and **task=0** proportions specifies the **weights**. Because subjects in 'resting state' are not given a 'task' its weight is set to zero.
The `simprepTemporal` function in the temporal design in listing 1 takes **onsets** for convolution with the double gamma **hrf** with **durations** and **effectsize**.

3.1.2 Noise parameters

In Listing 2, the proportions of scanner noise are partitioned into system **type** and low frequency drift **freq.low=128**. The temporal noise **type='gaussian'** is auto regressive AR(p) with $p = 2$ **rho** lag parameters of .8 and $-.2$. The physiological part in the noise **mixture** applies to heart **freq.heart=1.17** and respiratory **freq.resp=.2** frequencies. Temporal **design** specifies BOLD activation signals to which `simTSfmri` adds noise proportional to the signal to noise ratio **SNR**.

3.1.3 BOLD series

Listing 3 specifies three bursts of BOLD **onsets** activations of length 25 each spaced 100 scans apart using the temporal design `td` function in listing 1. Applying `boldTS` function in listing 2 to the list of temporal designs finally results in a list of four series plotted in Figure 4; the `NULL` design yields noise only. The BOLD activations on the last voxel start 25 scans before the activations start on the first voxel.

3.2 Categorical reduction of the series

The bold series simulated in 3.1.3 are segmented into stationary subsequence windows using 2.3.1 as illustrated in Figure 4 by vertical bars. The windows between the bars are numbered sequentially. The windows are then chained to their nearest neighbour at minimum between

[1] **boldface** font indicates neuRosim parameter types

window distance using 2.4 as plotted in Figure 15. Here the window numbers are coloured by their word in the study bag. The chain is cut into words using the procedure in 2.5.
With the categorical reduction series of T document voxels draw their words from the bag as illustrated in Figure 4 where at each time point each of the voxels in a document draw their word color. Each document is thus interpreted as the state vector of document word counts summing to four for each document at each time.
When voxels within a document draw the same word they are considered co-activated, although it seems appropriate to reserve this definition for activations resulting from convolutions with a Haemodynamic Response Function (HRF). The procedure however reduces real valued stationary sequences to categorical words instead of into HRF binary classification. Figure 4 does illustrate the lagged activation between window pairs 13 and 2, 15 and 4, 17 and 6 which are due to convolutions with the HRF, which the existing approach would not capture due to its synchronised fixed width windows.

3.3 Exact state evolution

The 4-voxel documents use the states plotted on the simplex in Figure 17; as the documents progress over time they change state and follow a trajectory over the simplex. Figure 19 shows the state evolution of the documents over the course of the series.

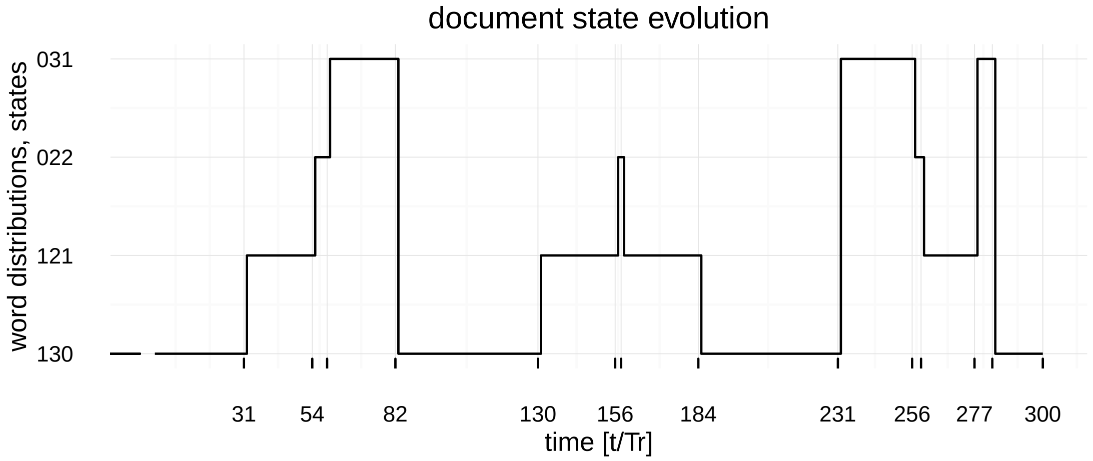


Figure 19: State evolution of shifted activations in documents in Figure 16.

Whereas in Figure 19 activations were time shifted, in Figure 20 the activations are synchronised. This can be seen as co-activated voxels from regions which are functionally connected in a ’resting state’ experiment. The state evolutions illustrate that even with four voxels in a document, some of its states are very briefly occupied, for example state ’022’ with two 2 and two 1 words is briefly visited three times. In Figure 20 which has synchronised co-activations, states ’121’ and ’031’ are briefly visited because they mark the transitions at the edges of the burst onsets.
Document states are directly defined by the word count vector that documents draw, starting with state ’130’ which means the document at $t = 1$ draws one word three/blue and three words two/green, summing to four voxel draws. When a document is made up of tens of thousands of voxels, its becomes increasingly probable that a document voxel changes its word and consequently the word count vector which fixes its state also, providing another voxel does not reverse

the counts. To remedy, the BIC puts a penalty on the number of states in the state mixture; the simulation only uses small documents which do not need state estimation from a mixture.
While ad hoc, excluding short occupations gives more stabile occupation statistics.
Despite the binary convolution of onset bursts with a Haemodynamic Response Function and categorical voxel activations, the number of documents states are likely to increase rapidly with the number of bursts and more importantly with the number of voxels, unless briefly visited states are penalised.

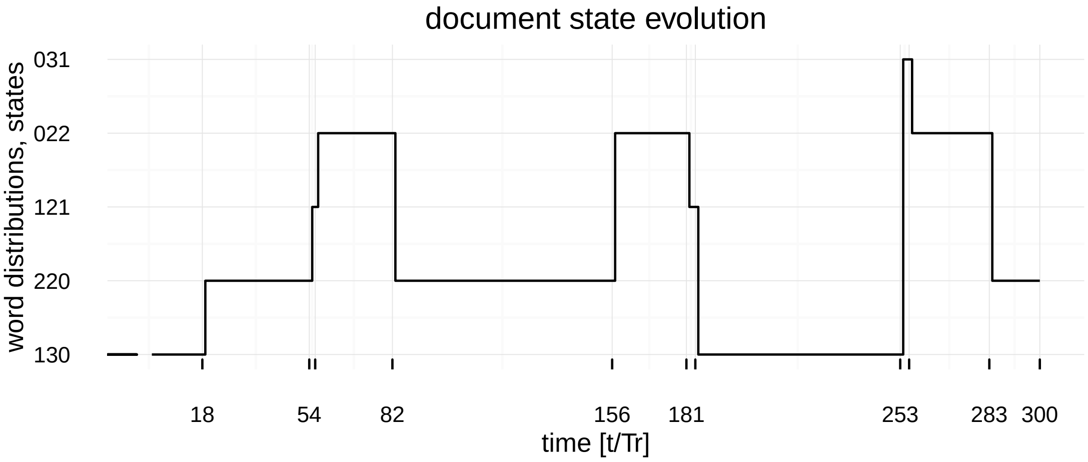


Figure 20: State evolution of documents with synchronised coactivation in two of four voxels.

3.4 Mixture state occupation

Specific activation intervals separate two groups with 30 members each. Inter-burst onsets of group specific 40 [t/Tr] and 50 [t/Tr] are offset by ± Poisson ($\lambda = 5$) draws to generate activation starts for four voxels of the members. Then following the bagging procedure in 2.4 and BIC guided state mixture estimation in 2.6.3, mean document state occupations are determined.
Each dot represents the mean document state occupation duration of a member in its respective group boxplot in Figure 21, each group with their own λ. A Welch two group sample Student-t test, at 95% confidence interval cannot reject the null hypothesis of equal mean state occupations.

3.5 Results

This section has shown that reduction of real-valued activations into categorical words facilitates interpretation at the observed time scale. Lagged activation is evident by following specific word activations in distinct voxels. However, the mean state occupation does not provide a reliable statistic for the interpretation of document word count vectors as states in a mixture. The Welch two sample t-test of mean subject state occupations for the groups specified in 3.4 does not allow rejection of the null hypothesis of equal means. With mean occupation estimates of 42.88 and 29.37 and test statistic of 1.3725, under 54.5 degrees of freedom, the difference between mean state occupations lies between -6.2 and 33.2 with 95% probability.

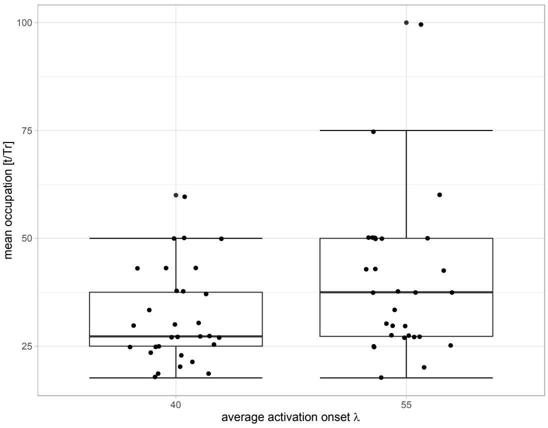


Figure 21: Mean state occupations for average activations in two groups with 30 subjects with 300 documents of 4 voxels.

4 Contrasts, characteristics and conclusions

The method's contrasting characteristics with respect the existing approach precede the relative advantages, disadvantage & limitations, application and conclusions. Bold face format in this section highlights contrasting and otherwise distinctive features.

4.1 Contrasting characteristics

The method detailed herein is presented as an alternative to time series loadings on 'eigenconnectivities'. Eigenconnectivities are principal components of synchronised fixed time window correlations between atlas brain region averaged activations in grouped subjects. The new method introduces a text mining approach to BOLD activation series in magnetic resonance scans on brains in 'resting' state at individual voxel and original time resolution.
The method captures correlated activation in and between individual voxels rather than between atlas brain regions. Instead of synchronised fixed time-width windowed averages, the method uses stationary sequences on individual voxel activations in a series. Whereas the existing approach reduces grouped time windowed regional correlations into a set of principal 'eigenconnectivities', the method bags all similar subsequences into the study 'bag of words'.
The bagged words capture both serial and spatial correlations. A scan volume is represented as a consecutive document series. A document is made up from the words that its voxels draw. The word distribution of a document determine its state on a unit probability sphere. A series of documents then switch from state to state as its word distributions change akin to a changing topic in latent topic analysis, e.g. Blei et.al, 2001, [2]. The number of document states in a discrete von Misses-Fisher mixture of states is restricted by using the Bayesian Information Criterion. Geometric state occupation statistics contrast the Fourrier analysis on the loadings in the existing approach.

4.2 Advantages

Where the existing approach relies on **fixed** time-width pairwise correlation, the new method connects distinct windows of consecutive within voxel correlation of **variable width**. The new method captures within- and between-voxel time **lagged** correlated activation, which the fix **synchronised** time-width window misses. By working on voxel level, consecutive stationary within- and between-voxel correlation is captured by words in the study bag at **individual** voxel and document **resolution**. All words are of **neurological insignificance** may be 'manually' combined into one e.g. noise word to reduce the study bag while maintaining original resolution. The existing approach **approximates** grouped pairwise regional correlation, whereas the new method chains subject idiosyncratic activation at some distance to its nearest neighbour in a separate word using a **statistical test** to cut word boundaries at excess unique information sums between split windows. Thus **subject specific activations** are fully captured whereas in the existing approach such activations may get lost in **grouped approximations**. Instead of **a-priori** averaging voxel activations in regions from a brain atlas, **regions emerge** from distinct voxel neighbourhoods that are consistently activated by the same word. Functional connectivity emerges from distinct regions drawing the same word.

The window chaining and word cutting algorithms do not depend upon specific stationary window segmentation and coding procedures, allowing **superior plug-ins** for functional and/or computation advantage.

4.3 Disadvantages & limitations

The energy divisive method in 3.1.3 may be incapable to segment consecutive stationary activations in the presence of excess **noise**, and the bisection permutation tests in 3.1.3 for stationary consecutive activations on each voxel are **computationally intensive**. A window of consecutive stationary needs a minimum width which is determined by the depth of the wavelet transform. I.e. a Haar wavelet transform in 2.3.2.1 with 4 depth levels of **detail requires a minimum window width** of $2^4 = 16$. To escape local maxima of a k -state mixture likelihood, restarts of the expectation maximisation (EM) algorithm in 2.6.3.1.1 increase the **computation cost**. Searching for an expressive state mixture with small size K in 2.6.3.1.2, requires execution of the EM algorithm over a sequence of possible K values, which bears a **computational cost**.

Interpreting consecutive documents as switching state, document word counts are used independent of which voxels contribute to the word counts. This implies that different voxel populations drawing the same word distribution **alias different documents** at the same state.

4.4 Applications

Neighbourhoods of voxels which consistently draw the same word may be used to **confirm atlas** regions.

The new method chains windows of stationary consecutive activation in distinct voxels to their nearest most correlated neighbours. Thus **time lagged** similar highly correlated activation in distinct windows is naturally captured in the same word, which allows monitoring **regions activated by one word**. Changing direction by monitoring a the words a region selects **region word sequence** switch behaviour may be analysed. **Anomaly detection** may be performed based on rare word use or extraordinary word sequence patterns in subjects. Interpreting consecutive document word vectors as switching between states, allows **analysis of dynamics** in functional connected regions.

4.5 Conclusions

By changing perspective from averaging and synchronised windowed correlation approximation to mining within- and between-voxel BOLD activation correlation, co-activated neighbourhoods

of voxels emerge. Regions emerge as neighbourhoods of voxels drawing the same categorical words, rather than from a priori definition in a brain atlas. The text mining method of reducing consecutive stationary real valued voxel activations into categorical words using statistical tests preserves the observed individual voxel and scan resolution. By chaining windows of consecutive similar activations in voxels to their nearest neighbours, computational cost of quadratic order for pairwise correlation calculation is avoided. And by noting which distinct regions subsequently use the same word, time lagged subsequent co-activation detection is simplified.
By working at individual voxel activations of each subject, separate words may capture each subject specific activation rather than by a group averaged pairwise regional correlated activation. The added precision from this preprocessing step may be exploited in subsequent analysis of subjects' dynamics in BOLD activations at the original observation resolution. However using word count vectors irrespective of origin of the voxels drawing them it too simplistic, because brain voxels are not exchangeable. Therefore the state interpretation of a word vector count per image should not be pursued. Instead voxels identified by the x, y and z coordinate location in the brain volume and their membership of neighbourhoods i.e. brain regions should be analysed used the activation words they draw as determined by the text mining reduction while fully exploiting the information in the preserved time resolution.

References

[1] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, Suvrit Sra; Clustering on the Unit Hypersphere using von Mises-Fisher Distributions Journal of Machine Learning Research, Volume 6, 2005, Pages 1345–1382, http://www.jmlr.org/papers/volume6/banerjee05a/banerjee05a.pdf.

[2] David M. Blei, David & Y. Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608.

[3] M. Blei, David & John D. Lafferty (2006), Dynamic Topic Models. ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning. 113-120.

[4] Andreas M. Brandmaier, pdc: An R Package for Complexity-Based Clustering of Time Series, 2016 Journal of Statistical Software, pp. 1-23. 10.18637/jss.v067.i05

[5] William Constantine and Donald Percival, wmtsa: Wavelet Methods for Time Series Analysis, (2017), https://CRAN.R-project.org/package=wmtsa

[6] Peter Grünwald The Minimum Description Length Principle, 2007, MIT Press, Cambridge, https://mitpress.mit.edu/books/minimum-description-length-principle

[7] Alfréd Haar, Zur Theorie der orthogonalen Funktionensysteme, Mathematische Annalen, (1910), 69 (3): 331–371, urldoi:10.1007/BF01456326

[8] Hornik, K., & Grün, B. (2014). movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions., Journal of Statistical Software, 58(10), 1 - 31. doi:http://dx.doi.org/10.18637/jss.v058.i10

[9] Nicholas A. James, David S. Matteson. ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data., Journal of Statistical Software (2014), 62(7), pp 1-25. http://www.jstatsoft.org/v62/i07/

[10] Kullback, Solomon. Information Theory and Statistics, (1959), John Wiley & Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.

[11] Leonardi, Nora, Jonas Richiardi, Markus Gschwind, Samanta Simioni, Jean-Marie Annoni, Myriam Schluep, Patrik Vuilleumier and Dimitri Van De Ville. Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest., NeuroImage 83 (2013), pp. 937-50. https://doi.org/10.1016/j.neuroimage.2013.07.019

[12] Christian A. Naesseth, Scott W. Linderman, Rajesh Ranganath, David M. Blei Variational Sequential Monte Carlo, Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. Volume 84. https://arxiv.org/pdf/1705.11140

[13] Object Management Group, 2017, Unified Modeling Language Specification, UML®, http://www.omg.org/spec/UML

[14] Peraza L. R., Kaiser M., Firbank M., Graziadio S., Bonanni L., Onofrj M., Colloby S. J., Blamire A., O'Brien J. and Taylor, J.-P., (2014). , fMRI resting state networks and their association with cognitive fluctuations in dementia with Lewy bodies., NeuroImage, Clinical, 4, 558–565., http://doi.org/10.1016/j.nicl.2014.03.013

[15] Claude E. Shannon, A Mathematical Theory of Communication, Reprinted with corrections from The Bell System Technical Journal, July, October, 1948, Vol. 27, pp. 379–423, 623–656, http://affect-reason-utility.com/1301/4/shannon1948.pdf

[16] Gideon Schwarz, Estimating the Dimension of a Model The Annals of Statistics Vol. 6, No. 2 (Mar., 1978), pp. 461-464, Institute of Mathematical Statistics. http://www.jstor.org/stable/2958889

[17] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain, NeuroImage, Volume 15, Issue 1, 2002, Pages 273-289, ISSN 1053-8119, http://www.sciencedirect.com/science/article/pii/S1053811901909784

[18] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large, Transactions of American Mathematical Society 54 (1943), 426-482 https://doi.org/10.1090/S0002-9947-1943-0012401-3

[19] Welvaert, M., Durnez, J., Moerkerke, B., Berdoolaege, G., & Rosseel, Y. (2011). neuRosim: An R Package for Generating fMRI Data.Journal of Statistical Software, 44(10), 1 - 18. doi:http://dx.doi.org/10.18637/jss.v044.i10

A Notation

$|x|$ number of elements of x

$\sum x$ sum over elements in x

$\bar{x}$ average value of x, $\frac{1}{n}\sum_{i=1}^{n=|x|} x_i$

$Pr\left(X = x\right)$ probability of random variable X taking x

$\hat{\alpha}$ estimate of α

$x^{d \times T}$ d times T dimensional vector x

$z|X = x$ value of z given variable X takes value x

$\lfloor x \rfloor$ next lower integer of x

$z \overset{H_0}{\sim} \mathbb{P}$ under H_0, z is distributed as $\mathbb{P}$

$z \approx x$ z is approximately equal to x

$z \perp x$ z is orthogonal or perpendicular to x

$x^\top y$ inner product between vectors x and y of equal length, $\sum_{i=1}^{|x|} x_i \cdot y_i$

$\|x\|_2$ Euclidean distance or L_2 norm equalling $\sqrt{x^\top x}$

$\prod x$ product of elements in x

$\frac{\delta}{\delta\theta} f(.)$ derivatives of function f with respect to variables θ

$\log(x)$ natural logarithm of x, i.e. with base e, $\log_e(x)$

$\mathbb{N}_{\geqslant 0}$ set of non-negative integers

$\mathcal{O}(N^2)$ of complexity in the order of N^2

$D_{KL}(\mathbb{P}\|\mathbb{Q})$ Kullback-Leibler divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$

$P \in Q$ P is an element of set Q

$P \cup Q$ elements in either set or both, $\{x : x \in P \vee x \in Q\}$

$P \setminus Q$ the elements of P minus those in Q, $\{x \in P : x \notin Q\}$

$\top$ Boolean logical constant TRUE

$\neg X$ complement of logical X

B Concept definitions and relations

The method proposes application of latent topic allocation to the field of magnetic resonance imaging. For application to neurological science, the method adopts and adapts the following concept definitions and their interrelationships to bridge information theoretic learning and text mining.

B.1 Concept definitions

approach existing method

study magnetic resonance measurements of brains

subject observed human in study

voxel smallest discrete 3-dimensional (3D) unit in subject brain

BOLD Blood Oxygen Level Detection

observation BOLD activation in one voxel at a specific time

series consecutive observation sequence

window boundary indices of series

junction common boundary of adjacent windows

fusion removal of junction

code transcription of series into symbols

distribution symbol probabilities

entropy expected symbol length
optimal entropy entropy in optimal symbol distribution
divergence expected sub optimal symbol length difference
fusion divergence divergence of sub optimal fusion symbols
change point junction with significant fusion divergence
stationary window fusion without change point
word stationary chain of adjacent windows
bag unordered set of words in the study
document categorical draw from the Bag
volume consecutive document series
location unit length word proportion vector coordinate
State average location

B.2 Concept relationships

Concepts and interrelationships used in the method are depicted in Figure 22, in UML®, in class diagram notation by OMG [13]. Concepts are classes in rectangles and directed lines depict relationships. In a container-part relationship such as in `study` contains S `subjects`, the $\diamond$ symbol is at the container.
The `study` contains S `subjects` each with T consecutive `documents` of V `voxels` with their `word` activation. `Documents` of V multinomal `words` draws are `located` on a probability simplex. Series of T consecutive `documents` change `location` along the `subject` specific `paths`. The lengths of consecutive `windows` in the series of activations add up to T BOLD observations on `voxels`. Adjacent `windows` are `positioned` in `word` chains at smallest `distance` between their window `codes`.

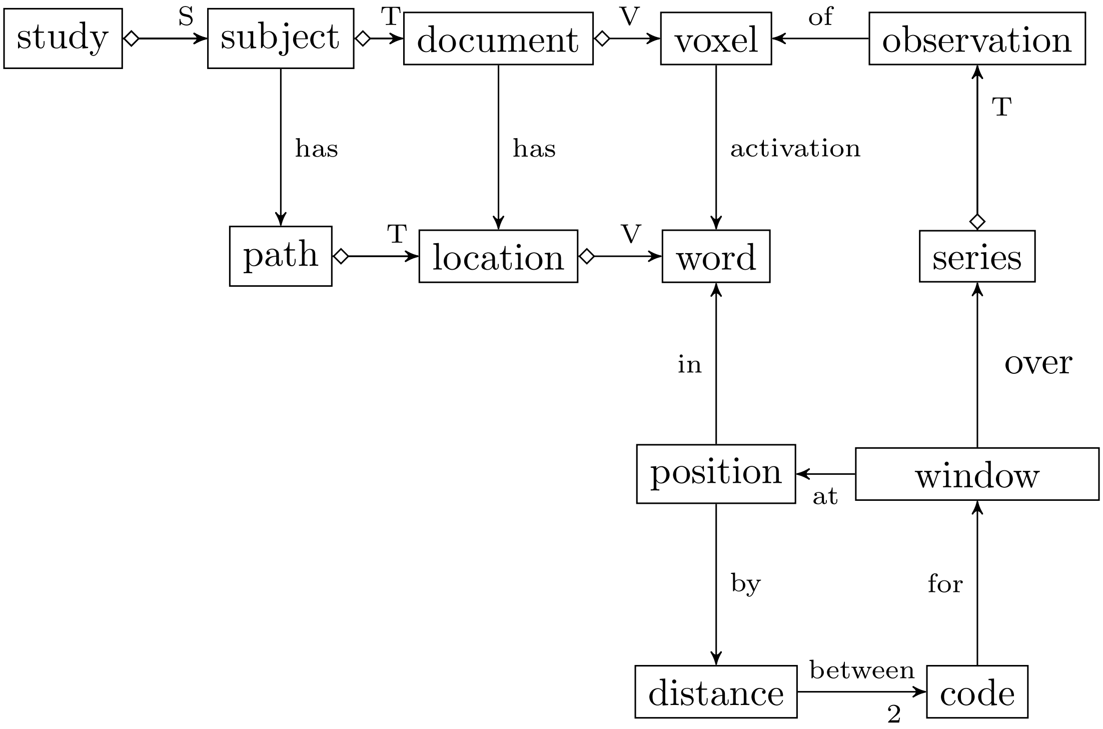


Figure 22: Method concepts and relationships

C Algorithms

Section 2.4.2.1 find nearest windows of stationary series at mimimum between window distance in a chain with algorithm 1. In 2.4.2.2 algorithm 2 inserts the window between nearest neighbours in the chain. Then in 2.5.5 algorithm 3 cuts the chain at word boundaries and bags them.
Given stationary subsequences, these algorithms implement the chaining and bagging which let voxel neighbourhoods drawing the same words emerge from the series. Given the scan document drawn up from a categorical word vector as a location on a unit sphere, expectation maximisation algorithm 4 from Banerjee, [1] determines the k states of document word vectors.

Algorithm 1 insert at smallest cosine distance by halving cuts

```
procedure minCosDistInsert($(V, C)$)
Require: $\forall i \in \mathbb{N} : i < j \leqslant |V|,\ i + 1 = \min_j \delta_{cos}(V_i, V_j)$
Ensure: $\forall i \in \mathbb{N} : i < j \leqslant |V \cup C|,\ i + 1 = \min_j \delta_{cos}(V_i, V_j)$
    $L \leftarrow 1, H \leftarrow |E|, M \leftarrow \lfloor \frac{L+H}{2} \rfloor$                ▷ (2)
    while $L < M$ do                                          ▷ L and H leave room for M
        if $\underset{i \in \{L,M,H\}}{min} \delta_{cos}(C, i) \neq M$ then
            $s \leftarrow \underset{s \in \{L,H\}}{max} \delta_{cos}(C, s)$                    ▷ move longest
        else                                                  ▷ M is on shortest edge to C
            $s \leftarrow \underset{s \in \{L,H\}}{min} \delta_{cos}(C, s)$                    ▷ shortest boundary
            if $\delta_{cos}(C, s) < \delta_{cos}(M, s)$ then
                $s \leftarrow \{L, H\} \setminus s$                            ▷ longest boundary edge
            end if
        end if
        $\{L, H\} \cup s \leftarrow M$                                ▷ pull boundary at s to Middle
        $M \leftarrow \lfloor \frac{L+H}{2} \rfloor$                                 ▷ (2) repositions Middle
    end while                                         ▷ L and H are direct neighbours in V
                                            ▷ triangle with edge lengths of cosine distance
    return InsertAtMaxBoundaryCut($(C, L, V)$)
end procedure
```

Algorithm 2 insert at cut made by longest boundary edge

```
procedure InsertAtMaxBoundaryCut($(C, L, V)$)
Require: $1 \leqslant L \leqslant |V|, \{V, C\} \in [0, 1]^m, \|V\|_2 = \|C\|_2 = 1$
Ensure: $\frac{1}{|V|} \sum_{i=1}^{|V|} \delta_{cos}(U_i, U_{i+1}) \leqslant \frac{1}{|V|-1} \sum_{i=1}^{|V|-1} \delta_{cos}(V_i, V_{i+1})$
    $H \leftarrow L + 1$
    if $|V| \leqslant 2$ then
        $U \leftarrow (V, C)$                                  ▷ need a Low and High vertex in V
    else if $\delta_{cos}(C, \{L, H\}) < \delta_{cos}(L, H)$ then
        $U \leftarrow (V_{\ldots,L}, C, V_{H,\ldots})$                                   ▷ insert between
    else if $\delta_{cos}(C, L) < \delta_{cos}(C, H)$ then
        $U \leftarrow (V_{\ldots,L-1}, C, V_{L,\ldots})$                                ▷ flip in front of L
    else
        $U \leftarrow (V_{\ldots,H}, C, V_{H+1,\ldots})$                                  ▷ flip behind H
    end if
    return $U$                                 ▷ mean distance at most equal to that of V
end procedure
```

Algorithm 3 stationary subsequence to integer word

```
procedure subSeqToInt($P, \alpha$)
Require: $\alpha \in [0,1], P \in [0,1]^{N\times m}, \|P\|_2 = \mathbf{I}_N$
Require: $\forall i \in \mathbb{N}_{>0} : i < j \leqslant N,\ i+1 = \min_j \delta_{cos}(P_i, P_j)$
Ensure: $\forall i : 1 \leqslant i < j \leqslant N, w_i \leqslant w_j, w \in \mathbb{N}^N_{\geqslant 1}$
    $i \leftarrow 1,\ j \leftarrow 2,\ w_1 \leftarrow 1, t_0 \leftarrow \chi^2_{m-1,\alpha}$    ▷ equation (7)
    while $i < N$ do
        $T \leftarrow LLR(P_i, P_j|(P_i, P_j))$    ▷ equation (6)
        $w_j \leftarrow w_i + \mathbb{1}(T \geqslant t_0)$    ▷ $\mathbb{1}\,(\neg\top, \top) \mapsto (0, 1)$
        $i \leftarrow i+1,\ j \leftarrow j+1$
    end while
    return $w$
end procedure
```

Algorithm 4 hard-movMF EM algorithm in Banerjee, [1]

```
procedure hard-movMF($w, \epsilon$)
Require: $w \in [0,1]^{T\times d},\ \|w\|_2 = \mathbf{I}_T,\ \epsilon > 0$
Ensure: $w^{N\times d} \mapsto [1, \cdot\cdot, K]^T$
    Initialise($\{\pi,\ \mu,\ \kappa\}^K$)
    repeat    ▷ find nearest K for each w in Estimating step
        for $i \in [1, \cdot\cdot, T]$ do    ▷ points on the sphere
            for $h \in [1, \cdot\cdot, K]$ do    ▷ states
                $vMF(w_i|\mu_h, \kappa_h) \leftarrow C_d(\kappa_h)e^{\kappa_h(\mu_h^T w_i)}$
                $z_{i,h} \leftarrow 0$
            end for
            $k_i \leftarrow \underset{k}{argmax}\,(\pi_k \cdot vMF(w_i|\mu_k, \kappa_k))$
            $z_{i,k_i} \leftarrow 1$    ▷ hard assignment
        end for
        for $h \in [1, \cdot\cdot, K]$ do    ▷ Increase Likelihood M step
            $\pi_h \leftarrow \frac{1}{T}\sum_{i=1}^{T} z_{i,h}$    ▷ average state membership
            $\mu_h \leftarrow \sum_{i=1}^{T} w_i \cdot z_{i,h}$    ▷ state direction sum
            $\bar{r} \leftarrow \mu_h/(T \cdot \pi_h)$
            $\mu_h \leftarrow \mu_h/\|\mu_h\|_2$    ▷ unit length state direction
            $\kappa_h \leftarrow \frac{\bar{r}d-\bar{r}^3}{1-\bar{r}^2}$    ▷ approximation (4.4) in [1]
        end for
    until $\ell_\Delta\,(w|\mu, \kappa, \pi) \leq \epsilon$    ▷ increments of (2.3) in [1]
    return $k_i|\, z_{i,\,k} = 1,\ i \in [1, \cdot\cdot, T]$
end procedure
```

D Neurosim parameterisation

Listing 1 captures the temporal design parameters in 3.1.1. Listing 2 lists the weighted temporal noise parameters defined in 3.1.2. Listing 3 generates time shifted BOLD activation on 4 voxels used in 3.1.3.

Listing 1: temporal design and series for 4 voxels

```r
require(neuRosim) # Welveart et al., 2011
samples = 300
TR = 1 # inter sample time
trT = samples*TR # total time

# temporal design
td <- function(o) { # o: onset indices
   simprepTemporal(
        totaltime = trT,
        onsets = o,
        durations = c(4),
        effectsize = c(20),  # amplitude
        TR = TR,
        hrf = "double-gamma" # convolution
        # Haemodynamic Response Function
   ) # return design for simTSfmri
}
```

Listing 2: BOLD series with noise for 4 voxels

```r
noiseProportions <- c(system=.1, temporal=.8,
drift=.01, physiological=0.09, task=0)
boldTS <- function(td) { # td: temporal design
           simTSfmri(
              base = 5, # amplitude
              nscan = samples,
              TR = TR,
              design = td,
              SNR=3.7, # signal to noise ratio
              noise = "mixture",
              weights = noiseProportions,
              type = "gaussian",
              rho = c(.8,-.2), # temporal noise
              # Auto Regressive (2) parameters
              verbose = TRUE,
              # physiological noise parameters
              freq.low = 128,
              freq.heart = 1.17,
              freq.resp = 0.2
            ) # return BOLD plus noise sequence
         }
```

Listing 3: BOLD series with noise for 4 voxels

```
fewOn <- seq(from=50, to=75, by=3) # burst
fewOn <- c(fewOn, 100+fewOn, 200+fewOn)
# voxel list with consecutive BOLD activations
# NULL design has only noise (i.e. no hrf)
lls <- lapply( # designs for 4 voxels
        list(td(fewOn), NULL, NULL,
                        td(fewOn-25)), boldTS)
```