Fingerprinting using KL divergence

Charles Zheng and Yuval Benjamini

May 21, 2018

0.1 Motivation

Functional connectivity (FC) matrices are a useful statistic for summarizing fMRI resting-state data. Let $\mathbf{X}^t$ be a vector of ROI-averaged activations for time t. From observations $\mathbf{X}^1, \cdots, \mathbf{X}^T$, we construct the functional connectivity matrix $\mathbf{R}$ via the empirical correlation matrix, with entry R_{ij} indicating the correlation between ROI X_i and X_j,

$$R_{ij} = \frac{\hat{\text{Cov}}(X_i, X_j)}{\sqrt{\hat{\text{Var}}(X_i)\hat{\text{Var}}(X_j)}}.$$

Often it is useful to find the *nearest neighbors* of a given FC matrix $\mathbf{R}^*$ from a collection of other FC matrices $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}$. Given a similarity metric $S(\cdot, \cdot)$, the nearest neighbor is the FC matrix $\mathbf{R}^{(i)}$ with maximal similarity to the query,

$$\text{argmax}_{i=1}^m S(\mathbf{R}^*, \mathbf{R}^{(i)}).$$

Some applications include:

- *Individual identification (fingerprinting).* Each of n individuals is scanned for two sessions, yielding for individual i the FC matrices $\mathbf{R}^{(i,1)}, \mathbf{R}^{(i,2)}$, one from each session. An individual can be *identified* from session 1 to session 2 if $\mathbf{R}^{(i,1)}$ is the nearest neighbor to $\mathbf{R}^{(i,2)}$ among all of the FC matrices for n subjects with the same session number, i.e., if

  $$i = \text{argmax}_{j=1}^m S(\mathbf{R}^{(i,1)}, \mathbf{R}^{(j,2)}).$$

  The identification rate (from session k to session l) is therefore

  $$\text{IDRate}(k \to l) = \frac{1}{m} \sum_{i=1}^m I\{i = \text{argmax}_{j=1}^m S(\mathbf{R}^{(i,k)}, \mathbf{R}^{(j,l)})\}.$$

Interesting variations include when session k and l are not the same type of scan, e.g. session k may be resting-state and l may be task.

- *Classification.* Each of n individuals has a phenotype y_i as well as an FC matrix $\mathbf{R}^{(i)}$. We wish to predict the phenotype of a new individual based on their FC matrix $\mathbf{R}^*$. A simple classification rule is k-nearest neighbor, defining $i_1, \ldots, i_k$ as top k the indices maximizing $S(\mathbf{R}^*, \mathbf{R}^{(i)})$ and then predicting y^* as the majority label among $\{y_{i_1}, \ldots, y_{i_k}\}$.

0.2 Symmetric KL works better than elementwise correlation

A variety of similarity scores can be considered for the application.

1. *Elementwise Correlation.*

$$S_{Cor}(\mathbf{R}, \mathbf{R}') = Corr(vec(\mathbf{R}), vec(\mathbf{R}')).$$

2. *Elementwise MSE.*

$$S_{MSE}(\mathbf{R}, \mathbf{R}') = -\|vec(\mathbf{R}_a) - vec(\mathbf{R}_a))\|.$$

3. *Gaussian KL divergence.*

$$S_{KL}(\mathbf{R}, \mathbf{R}') = \mathrm{tr}(\mathbf{R}'\mathbf{R}^{-1}) - \log(|\mathbf{R}'\mathbf{R}^{-1}|).$$

4. *Symmetrized KL divergence.*

$$S_{SKL}(\mathbf{R}, \mathbf{R}') = \min S_{KL}(\mathbf{R}, \mathbf{R}'), S_{KL}(\mathbf{R}', \mathbf{R}).$$

The most commonly used similarity measure is elementwise correlation. S_{Cor} works much better than S_{MSE} in practice, due to benefits of normalization.

However, a more principled approach from the theoretical point of view is Gaussian KL divergence. This is because assuming that the fMRI signal $\mathbf{X}_t$ has a multivariate Gaussian distribution with covariance Σ, the KL divergence provides a theoretically optimal (under large-sample conditions) test statistic for testing whether a given fMRI time series $\mathbf{X}_1, \ldots, \mathbf{X}_T$ has a given covariance Σ_0. Since fMRI signals do not have meaningful units, we can assume that the signals are normalized, and so the population covariance matrix is the same as the population correlation matrix. And while the population correlation is unknown for a given individual, one can take the empirical correlation (or FC matrix) $\mathbf{R}$ as an estimate. Therefore, to test if another FC matrix $\mathbf{R}'$ was obtained from the same individual, it makes sense to test for whether that matrix was generated from a Gaussian distribution with covariance $\mathbf{R}$. This implies the use of the KL divergence as a similarity measure, $S_{KL}(\mathbf{R}, \mathbf{R}')$.

Note however that KL divergence is not symmetric. Empirically, we find that symmetrized KL S_{SKL} divergence yields much better identification rates than S_{KL}. We do not yet have a theoretical explanation of this.

0.3 Properties of KL divergence

One can understand KL divergence in terms of the eigendecomposition of the FC matrices. First note that

$$\operatorname{tr}(\mathbf{R}'\mathbf{R}^{-1}) - \log(|\mathbf{R}'\mathbf{R}^{-1}|) = \operatorname{tr}(\mathbf{R}^{-1/2}\mathbf{R}'\mathbf{R}^{-1/2}) - \log(|\mathbf{R}^{-1/2}\mathbf{R}'\mathbf{R}^{-1/2}|)$$

Hence, defining

$$\Delta = \mathbf{R}^{-1/2}\mathbf{R}'\mathbf{R}^{-1/2}$$

we have

$$S_{KL}(\mathbf{R}, \mathbf{R}') = S_{KL}(I, \Delta).$$

Now, let $\delta_1, \ldots, \delta_R$ be the eigenvalues of Δ. (R is the number of ROIs). Then, we have

$$S_{KL}(I, \Delta) = \sum_{i=1}^{R} (\delta_i - \log \delta_i).$$

The function $f(x) = x - \log(x)$ is minimized by $x = 1$, hence we see that S_{KL} is minimized if $\Delta = I$, implying $\mathbf{R} = \mathbf{R}'$.

However, what this implies in general is that KL divergence can be decomposed as a sum of how much the eigenvalues of $\mathbf{R}^{-1/2}\mathbf{R}'\mathbf{R}^{-1/2}$ differ from 1. An interesting possibility is to change the weighting of the terms in the sum to see if one can find even more effective similarity metrics.