Comparing Task-Relevant Information Across Different Methods of Extracting Functional Connectivity

Sophie Benitez Stulz[a,d], Andrea Insabato[a,b], Gustavo Deco[a,c], Matthieu Gilson[a], Mario Senden[d,e]

[a] Center for Brain and Cognition, Computational Neuroscience Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Carrer de Ramon Trias Fargas, 25-27, Barcelona, 08005, Spain

[b] The Italian Academy, Center for Theoretical Neuroscience, Columbia University, 1161 Amsterdam Ave., New York NY 10027, USA

[c] Institució Catalana de la Recerca i Estudis Avançats (ICREA), Universitat Pompeu Fabra, Passeig Lluís Companys 23, Barcelona 08010, Spain

[d] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6201BC Maastricht, The Netherlands

[e] Maastricht Brain Imaging Centre, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Abstract: 186 words
Main text: 4176 words
References: 36

Abstract

The concept of brain states, functionally relevant large-scale patterns, has become popular in neuroimaging. Not all components of such patterns are equally characteristic for each brain state, but machine learning provides a possibility of extracting the structure of brain states from functional data. However, the characterization in terms of functional connectivity measures varies widely, from cross-correlation to phase coherence, and the idea that different measures will provide the similar information is a common assumption made in neuroimaging. Here, we compare the performance of phase coherence, pairwise covariance, correlation, model-based covariance and model-based precision for a dataset of subjects performing five different cognitive tasks. We employ multinomial logistic regression for classification and consider two types of cross-validation schemes, between- and within-subjects. Furthermore, we investigate whether classification is robust for different temporal window lengths. We find that informative

links for the classification, meaning changes between tasks that are consistent across subjects, are entirely uncorrelated between correlation and covariance. These results indicate that the corresponding FC signature can strongly differ across FC methods used and that interpretation is subject to caution in terms of subnetworks related to a task.

Keywords: machine learning, functional connectivity, fMRI, task information, brain states

1. Introduction

At a macroscopic level the brain may be conceived of as a complex system of regions engaging in dynamic, interactive behaviour (Bullmore & Sporns, 2009). Neuroscience has developed various quantitative approaches to define stereotypical brain states corresponding to cognitive functions. Brain states may refer to purely spatial patterns, activity distribution across voxels or brain regions (Cabral, Kringelbach, & Deco, 2017). Alternatively, they may refer to spatio-temporal patterns and distributions functional interactions between regions (Vidaurre, Smith, & Woolrich, 2017).

Whole-brain modelling has been widely used to characterise spatio-temporal brain states and capture their multivariate distributions. This approach attempts to explain observed functional interaction in terms of models of underlying region dynamics as well as structural connections between regions. Modelling of the oscillatory behaviour in brain regions has, for instance, shown that there are differences in this local parameter across task-dependent brain states (Senden, Reuter, van den Heuvel, Goebel, & Deco, 2017). On the other hand, models estimating directed connectivity based on the functional interactions between brain regions have also revealed differences in network parameters across task-dependent brain states (Pallares et al., 2018; Senden et al., 2018).

Recently, the application of machine learning to infer brain states has also gained popularity (Naselaris, Kay, Nishimoto, & Gallant, 2011; Pallares et al., 2018; Rahim, Thirion, Bzdok, Buvat, & Varoquaux, 2017; Varoquaux et al., 2017; Xie et al., 2017). Machine learning is useful since it can extract the relevant feature patterns of brain states from multivariate data and assess the generalization capabilities of these brain states to novel data. This approach has been highly successful for inferring brain states from functional connectivity (FC). Conventionally, functional connectivity (FC) is calculated across the entire duration of a session. Recently, however, focus has shifted towards dynamic functional connectivity (dFC) which is calculated at shorter time scales in the range of tens of seconds (Gonzalez-Castillo et al., 2015; Hutchison

et al., 2013; Preti, Bolton, & Van De Ville, 2017). For example dFC can be calculated with the sliding-window approach (Cabral, Kringelbach, et al., 2017), where Pearson correlation or covariance is computed between the signals of every pair of region with a small temporal window moving along the time series. A studies using the sliding window concept of dFC could successfully distinguish between the brain states during five different cognitive tasks (Gonzalez-Castillo et al., 2015; Xie et al., 2017). At the opposite end of the spectrum of time-scales, FC can be obtained instantaneously with phase coherence (Cabral, Vidaurre, et al., 2017; Senden et al., 2017). Evidently, there is a multitude of studies using various FC metrics to investigate brain states during different tasks (Cabral, Vidaurre, et al., 2017; Gonzalez-Castillo & Bandettini, 2017; Senden et al., 2018, 2017). However, the interchangeable use of FC metrics rests on the assumption that the results are comparable across metrics. This has not been validated since varying methodologies make it impossible to compare them across studies.

Our aim is test this assumption and to systematically evaluate the task-relevant information structure of the corresponding brain states across metrics and time-scale. The tasks include rest, a n-Back task, the Flanker task, a mental rotation task, and an Odd-man-out task (Senden et al., 2018, 2017). Specifically, we want to investigate whether choice in FC metric (Pearson correlation, covariance, phase coherence) affects classification performance and whether task-dependent information is similar across metrics. Secondly, we investigate metrics across different time scales, because it is possible that certain time scales do not capture information relevant to the classification, which would not be an issue of the metric itself, but of the parameter choice for its temporal window. Also, including metrics that reach from instantaneous FC (phase coherence) until static FC (global FC) provides a broad systematic overview of the temporal spectrum.

We find that the choice of parameters and metrics for connectivity classification strongly impact the task-relevant information retrieved and call for a more careful approach towards the interpretation of such results.

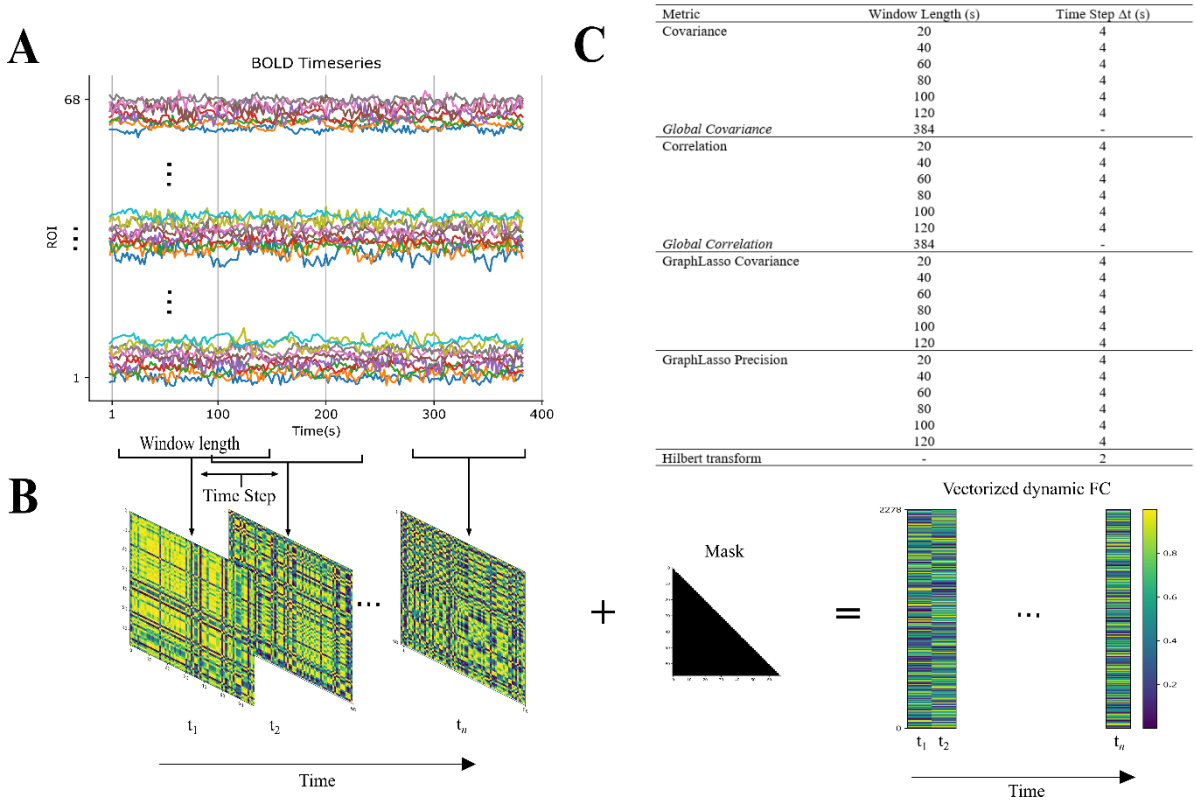
2. Material and methods

2.1 Functional MRI Data

We use an fMRI resting and task state dataset acquired in 14 subjects (8 females, M = 28.76, 22 – 43 years old) as described in a previous paper (Senden et al., 2017). The dataset comprised the blood-oxygen-level dependent (BOLD) signal of 68 Regions of Interest (ROIs) obtained in

five functional runs per subject with 192 data points each. During each run, the subjects were either resting, or engaging in one of four tasks: the Eriksen flanker task (Eriksen & Eriksen, 1974), a n-Back task (Kirchner, 1958), a mental rotation task (Shepard & Metzler, 1971), or a verbal Odd-man-out task (Flowers & Robertson, 1985). A detailed description of the stimuli used in the task paradigm can be found in Senden et al. (2017). The dataset was acquired at the Maastricht Brain Imaging Centre, (Maastricht University) on a 3 Tesla scanner (Tim Trio/upgraded to Prisma Fit, Siemens Healthcare, Germany). The data was pre-processed with BrainVoyager QX (v2.6; Brain Innovation, Maastricht, the Netherlands) using slice scan time correction, motion correction, and a high-pass filter with a frequency cut-off of .01 Hz. After subsequent wavelet de-spiking and regressing out global noise signals estimated from the ventricles, the average BOLD signal for each region was computed by taking the mean of voxels uniquely belonging to one of the 68 ROIs specified by the DK atlas (Desikan et al., 2006) with Matlab (2013a, The MathWorks, Natick, MA).

2.2 Spatiotemporal functional connectivity



Metric	Window Length (s)	Time Step Δt (s)
Covariance	20	4
	40	4
	60	4
	80	4
	100	4
	120	4
Global Covariance	384	-
Correlation	20	4
	40	4
	60	4
	80	4
	100	4
	120	4
Global Correlation	384	-
GraphLasso Covariance	20	4
	40	4
	60	4
	80	4
	100	4
	120	4
GraphLasso Precision	20	4
	40	4
	60	4
	80	4
	100	4
	120	4
Hilbert transform	-	2

Figure 1. Extracting FC from Bold signal. (A) Bold signal of 68 ROIs for 384 s of a fMRI session. Dots indicating omitted BOLD timeseries for visibility purposes. (B) FC matrices extracted from the BOLD signal in window with window length (WL). To eliminate identical values a mask is applied and (ROI*(ROI-1))/2 = 2278 features are obtained for each timepoint t. Subsequent timepoints are shifted by time step (Δt). (C) Table of FC types calculated from the Bold signal.

2.2.1 Phase Coherence. To obtain the analytical signal (Smith, 2007), a complex-valued function that has no negative frequency components, from the BOLD signal the Hilbret transformation was applied to the BOLD signal for each ROI. To calculate the instantaneous

functional connectivity (iFC) between region *i* and *j* for time *t* the cosine of the phase difference of the analytical signal of the two regions, was calculated.

$$iFC(i,j,t) = \cos(\theta(i,t) - \theta(j,t))$$

2.2.1.1 Eigenvector. To obtain the connectivity among eigenvectors we calculated the outer product of the strongest eigenvector of *iFC* as previously described in Cabral, Vidaurre, et al. (2017).

$$eigFC(i,j,t) = eig(iFC(i,t)) \otimes eig(iFC(j,t))$$

where, $iFC(t) =$ instantaneous FC at timepoint *t*.

$eig =$ largest eigenvector.

2.2.2 Covariance. The dynamic covariance (dCov) was calculated across window lengths of 20 s, 40 s, 60 s, 80 s, 100 s, 120s with a timestep of 4 s. We also computed pairwise Cov over the whole session to obtain global functional connectivity (gCov). Dynamic covariance between region *n* and *p* for time window *t* was calculated as follows:

$$Cov(i,j,w) = \left(X(i,w) - \overline{X(i)}\right) * \left(X(j,w) - \overline{X(j)}\right)$$

where, $X(k,w) =$ BOLD in region *k* in time window *w*.

$\overline{X(k)} =$ Mean BOLD in region *k*.

2.2.3 Pearson's Correlation. Dynamic pairwise Pearson correlation (dPC) was calculated with windows of 20 s, 40 s, 60 s, 80 s, 100 s, 120 s, and with a timestep of 4 s as well as within 6 s window with a timestep of 2 s to make the timescale of the PC as similar as possible to the timescale of the Hilbert transform. We also computed pairwise PC over the whole session to obtain global functional connectivity (gPC). Dynamic Pearson correlation between region *i* and *j* for time window *w* was calculated as follows:

$$Corr(i,j,t) = \frac{\left(X(i,w) - \overline{X(i)}\right) * \left(X(j,w) - \overline{X(j)}\right)}{\sqrt{\left(X(i,w) - \overline{X(i)}\right)^2 * \left(X(i,w) - \overline{X(j)}\right)^2}}$$

where, $X(k,w) =$ BOLD in region *k* in time window *w*.

$\overline{X(k)} =$ Mean BOLD in region *k*.

2.2.4 Model-based Precision and Covariance. The model-based precision and covariance (Scikit-learn, GraphLassoCV) attempts to estimate the inverse of the covariance

matrix, the precision matrix, which is proportional to the partial correlation matrix. The empirical precision matrix is not included as the covariance matrix is underdetermined, meaning it has less timepoints than regions in short time windows, and could not be calculated. The GraphLasso algorithm achieves this by enforcing sparsity on the estimation of the precision matrix by using an L1 penalty which is automatically estimated with cross-validation. More specifically, the GraphLasso algorithm (Friedman, Hastie, & Tibshirani, 2008) minimizes the following function to estimate the precision matrix *K* and the corresponding covariance matrix *S*.

$$\widehat{K} = arg_K min\,(trSK - logdetK + \alpha||K||_1)$$

where, K = precision matrix to be estimated.

S = sample covariance matrix.

$||K||_1$ = sum of absolute values of off-diagonal coefficients of *K*.

α = L1 penalty parameter.

2.3 Classification

2.3.1 Multinomial logistic regression. We use multinomial logistic regression (MLR) with a cross-entropy loss. We use an L2 penalization in combination with a limited-memory Broyden-Fletcher-Goldfarb-Shannon algorithm solver (Bishop, 2006) and an L1 penalty with a SAGA algorithm solver (Defazio, Bach, & Lacoste-Julien, 2014). The SAGA algorithm is an incremental gradient method which supports non-strongly convex problems. The penalty parameter is optimized with nested cross-validation meaning that the parameters are first optimized using cross-validation within the training set before being applied to the entire training set.

2.3.2 Cross-validation.

2.3.2.1 Within Subject. Due to temporal autocorrelation simple permutation does not give us any indication of the stability of the signal within a subject over time. Therefore, we use blocked cross-validation. For each task and subject, the samples are divided in 10 consecutive folds. The number of samples contained in each fold depends on the metric chosen. Subsequently, the decoder is trained on the first fold and tested on the second fold. Then the decoder is trained on the first and second fold and tested on the third fold. This procedure is continued until the last

fold is reached. The accuracy of the validation procedure is obtained from the mean of the testing accuracy over the 10 trained decoders.

The penalty parameter is optimized using nested cross-validation. More specifically, parameters for each training set are optimized with 2-fold temporal cross-validation on the training set (see figure 3).

2.3.2.2 Between Subject. The decoder is trained on 13 of the 14 subjects and tested on the remaining subject. This procedure is repeated with each subject being left out once. The accuracy of the validation procedure is the mean of the testing accuracy over the 14 trained decoders.

The penalty parameter is optimized using nested cross-validation. More specifically, parameters for each training set are optimized with 13-fold subject cross-validation on the training set (see figure 3).

2.3.3 Recursive feature elimination. Recursive feature elimination (RFE) iteratively removes the feature that is least important for classification. Features leading to a maximal accuracy using temporal and subject cross-validation are then deemed the best features to use for the classification. The ranking of all features obtained by the RFE is indicative of the structure of the information obtained from each FC. The number of best features was also selected within the nested cross-validation before optimizing the penalty parameter.

The classification pipeline was implemented in python using the Scikit-learn library (Pedregosa et al., 2011).

2.4 Similarity Measures

Spearman Rank. The Spearman Rank correlation r_s is a measure of non-linear correlation with a value between -1, denoting perfect anti-correlation, and 1, denoting perfect correlation (Lehman & Rourke, 2005). It quantifies how well the relationship between two variables can be expressed with a monotonic function.

$$r_s = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} * \sigma_{rg_Y}}$$

where, rg_X, rg_Y = Ranks of variables X, Y.

$cov(rg_X, rg_Y)$ = Covariance of the rank variables.

$\sigma_{rg_X}, \sigma_{rg_Y}$ = Standard deviation of the rank variables.

3. Results

3.1 Performance of the FC metrics

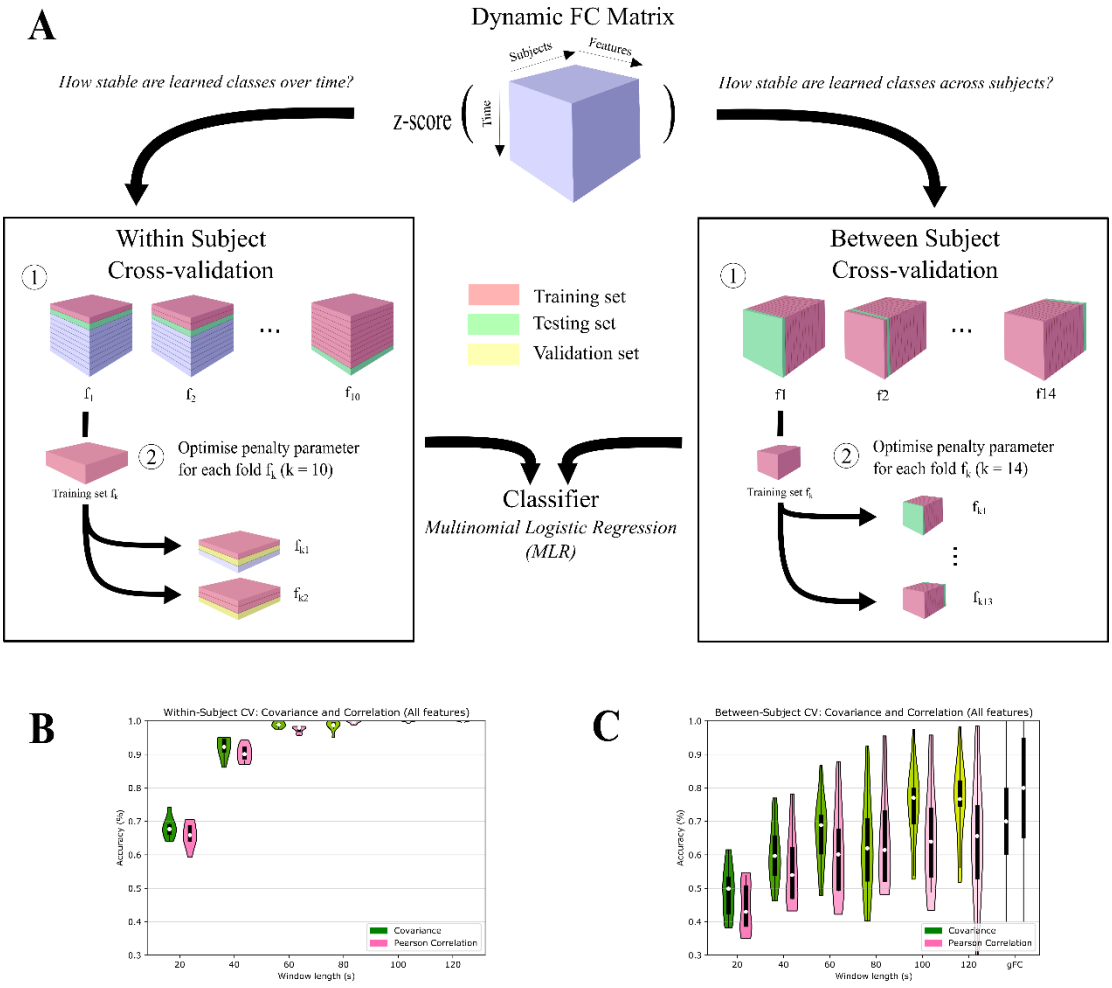


Figure 3: Within- and between-subject cross-validation procedure. Covariance outperforms correlation. (A) Within-subject and between-subject cross-validation. In within-subject cross-validation the data is split in sections along time. (B) Between-subject CV accuracy of covariance and correlation. (C) Between-subject CV accuracy of covariance and correlation.

3.1.1 Covariance

Within subject cross-validation accuracy of covariance follows a monotonically increasing trend starting from a window length of 20 s and saturates after 80 s (figure 3B). The necessity of within-subject CV to quantify the temporal stability of the classes becomes clear when compared to cross-validation with permutation sets which disregard the temporal autocorrelation (S2). While the permutation CV achieves maximal accuracy for all window lengths, within-subject CV shows a break-down of temporal stability which has also been

shown by other studies (Roberts et al., 2017). Adding variance to covariance only improves accuracy for a window length of 80 s but decreases on average by approximately 5% (S1A). Global covariance achieves a slightly higher accuracy with 0.8 (figure 3B).

Between subject cross-validation accuracy of covariance increases from a window length of 20s and saturates at 100s with a dip at 80 s (figure 3C). The performance seems to follow a growing trend (excluding 80 s) reaching a maximum at a window length of 100 s with an accuracy of 77 % and decreasing thereafter. Global covariance achieves a similar accuracy as dynamic covariance with a window length of 60 s.

3.1.2 Pearson correlation

Within-subject performance of the Pearson correlation increases from a window length of 20 s and saturates after 80 s (figure 3B). Global correlation performance is 0.8.

Between-subject cross-validation accuracy of correlation follows a monotonically increasing trend from a window length of 20 s until 120 s. Global Pearson correlation accuracy is ~15% higher than performance of dPC with a window length of 120 s. The different trends observed in empirical covariance and correlation suggests, that they are affected by the noise in the data differently. At windows until 120 s covariance generally performs better possibly because the standardization in the Pearson correlation also removes information in the variance at shorter time scales. At longer time-scales the variance likely contained more noise and the removal increases performance.

3.1.3 GraphLasso Precision

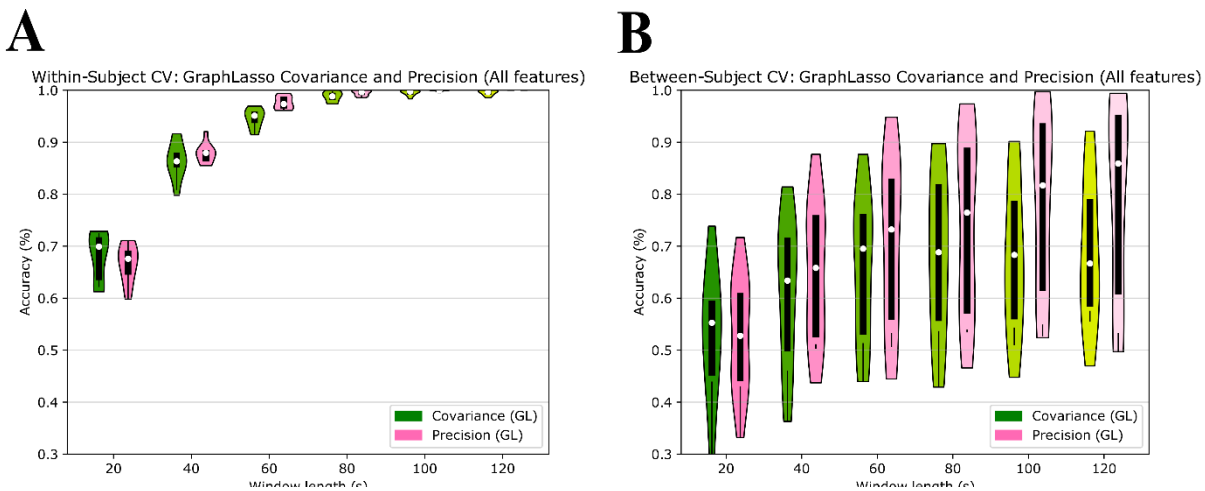


Figure 4: Performance of model-based FC measures GraphLasso covariance and precision. (A) Within-subject cross-validation accuracy using GraphLasso covariance (green) and GraphLasso precision (pink). (B) Between-subject cross-validation accuracy using GraphLasso covariance (green) and GraphLasso precision (pink). Chance level is 0.2.

Within-subject performance of the GraphLasso Precision follows an asymptotic trend towards the maximal accuracy increasing from a window length of 20 s and reaching maximal accuracy at a window length of 100 s (figure 4A).

Between-subject cross-validation accuracy of GraphLasso precision shows a growing monotonical trend continually increasing from a window length of 20 s without saturating. Similar to empirical covariance, model-based covariance does not improve at longer window lengths, suggesting, that it might be affected by noisy lower frequency fluctuations. Interestingly, removing noisy fluctuation by estimating the underlying precision performs much better than standardizing it with the variance like in the Pearson correlation.

3.1.4 GraphLasso Covariance

Within-subject performance of the GraphLasso covariance increases from a window length of 20 s and reaches approximately maximal accuracy at a window length of 100 s (figure 4A). Model-based as well as empirical metrics follow a similar asymptotic trend towards maximal accuracy, suggesting that they are affected by similar noisy temporal fluctuation at shorter time-scales.

Between-subject cross-validation accuracy of GraphLasso covariance continually increases from a window length of 20 s reaching a maximum at 60 s and decreasing again until 120 s.

3.1.5 Phase coherence.

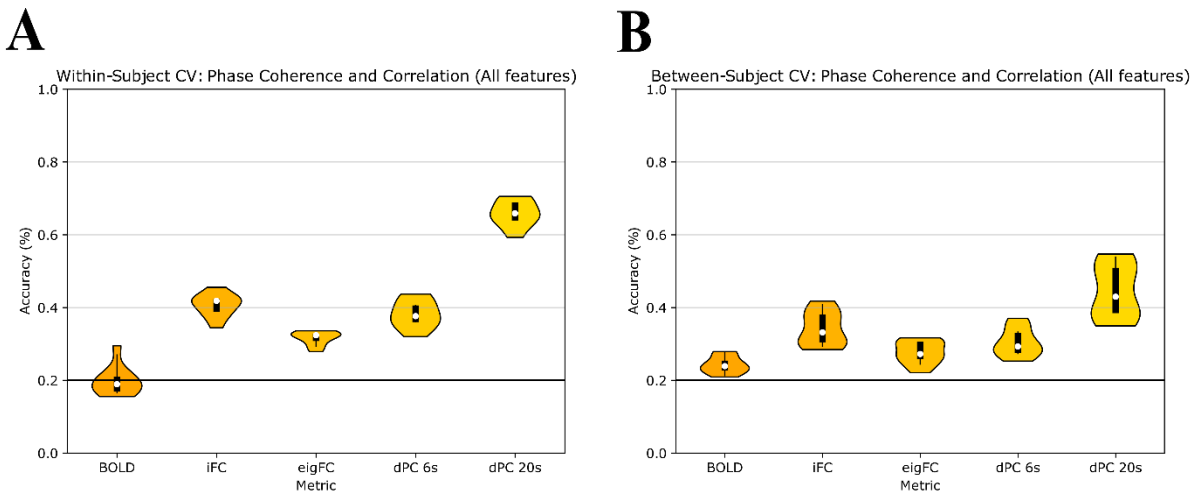


Figure 2: CV Accuracy at short time-scales. (A) Within-subject CV accuracy of the BOLD timeseries, phase coherence (iFC), the largest eigenvector of the phase coherence (iFC), Pearson Correlation with a window length of 6s and a time step of 2 s (dPC 6s) and Pearson Correlation with a window length or 20s and a time step of 4 s (dPC 20s). (B) Between-subject CV accuracy of the BOLD timeseries, phase coherence (iFC), the largest eigenvector of the phase coherence (iFC), Pearson Correlation with a window length of 6s and a time step of 2 s (dPC 6s) and Pearson Correlation with a window length or 20s and a time step of 4 s (dPC 20s). Chance level (0.2) indicated with black line.

Phase coherence performed poorly for both within- and between-subject CV. The median of the within-subject performance for phase coherence was 0.42 with chance level at 0.2 (figure 2A). The largest eigenvector of phase coherence only scored slightly above 0.32. The median of the between-subject performance of phase coherence was 0.33 and for the largest eigenvector was 0.27 (figure 2B).

The BOLD signal does not seem to carry any information to distinguish among tasks and using the eigenvector of the phase coherence leads to a decrease in accuracy and likely does not select relevant axes of the variability. Interestingly, the Pearson Correlation with a similar time-step as phase coherence and window of only 6 s did not outperform phase coherence.

3.2 Regularisation methods

Regularization is a commonly used tool to prevent a classifier from overfitting the training set leading to low testing accuracy (Bishop, 2006). However, L2 regularization did not reduce overfitting adequately as training accuracy was up to 50% higher than testing accuracy (see table S4).

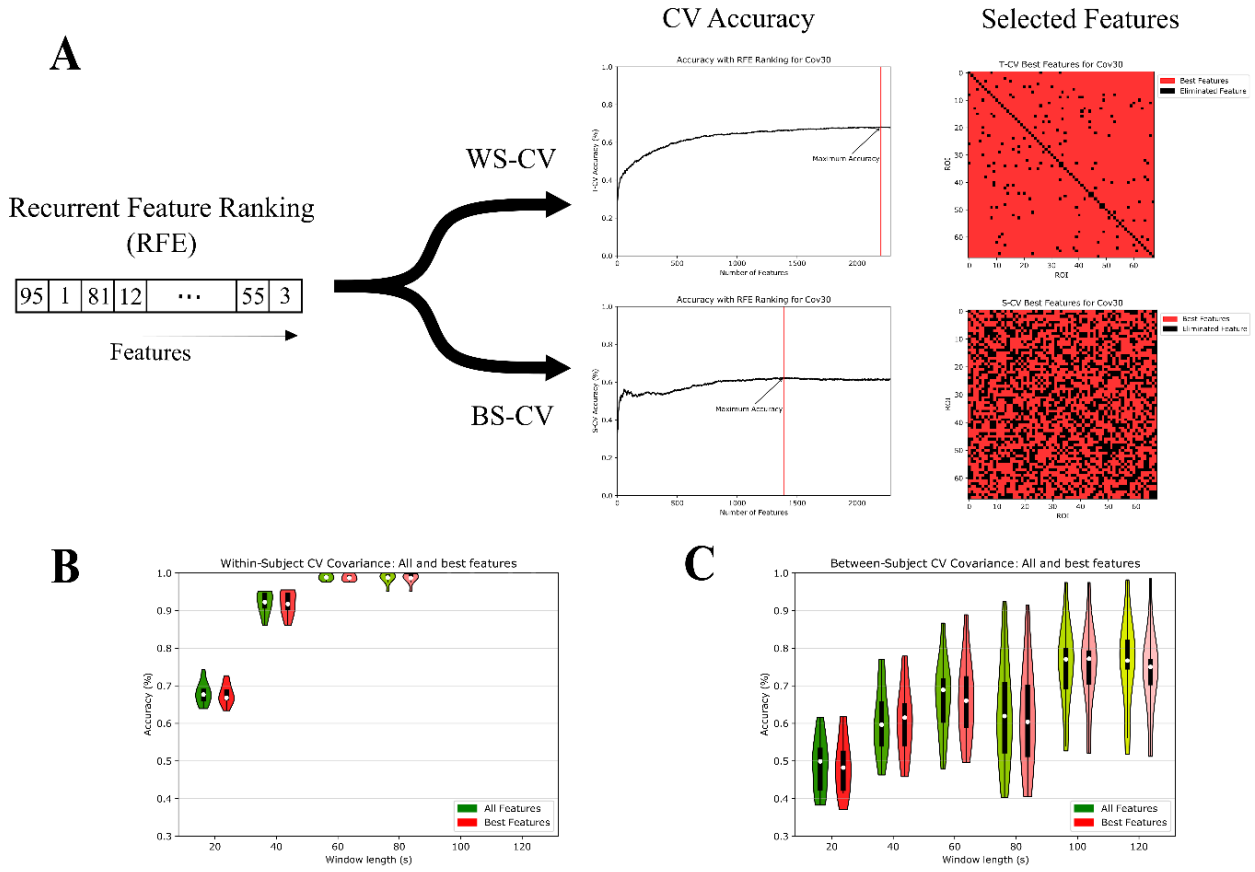


Figure 5. Feature selection performance for covariance. (A) The recurrent feature ranking (RFE) is used to test how many best features give the highest within-subject CV and between-subject CV accuracy. (B) Within-subject CV and (C) Between-subject CV accuracy of all features versus best features with covariance.

Using L1 regularization instead of L2 regularization in our classification did not improve the performance of the classifier. Rather it reduced accuracies by approximately 3% on average (see S2). Another tool that can be used to reduce dimensionality additionally is feature selection. However, this did not lead to a significant increase in within- or between-subject CV accuracy (figure 5A – C).

3.3 Task and rest are highly dissimilar

The strong decrease of accuracy towards smaller time-scales may be predominantly due to the difficulty of differentiating among tasks rather than discriminating task states from rest. Here

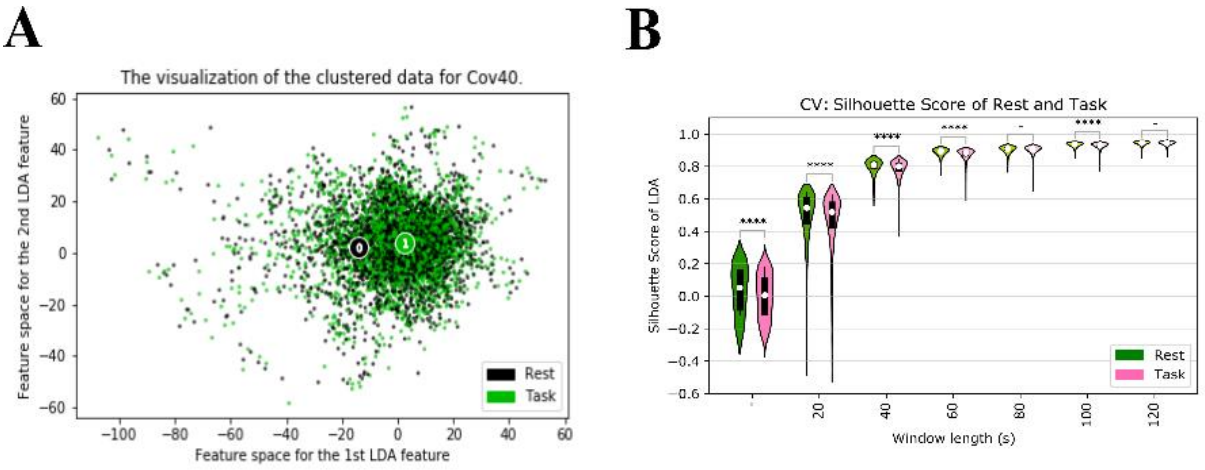


Figure 6. Silhouette scores of linear discriminant analysis (LDA) of rest and task. Features were reduced to four components with LDA and the silhouette score was calculated. (A) The first LDA component of covariance with window length 40 s is plotted on the x-axis and the second LDA component is plotted on the y-axis. Rest is plotted in black and task is plotted in green. (B) Violinplot of silhouette scores of LDA of rest and task for various FC Metrics. The metrics used were phase coherence (-), and covariance corresponding to the window lengths on the x axis. Rest is plotted in green and task is plotted in pink. Black bars indicate the inner 50 percentiles. The white dot indicated the median. A y-score of 0 indicates no clustering whereas 1 indicates strong clustering. Significance level indicated with symbols, $p < 0.0001$ (****), $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and non-significance (-).

we test this possibility by plotting the silhouette scores of phase coherence and covariance of the axes along which activity is most different between tasks, extracted with Linear Discriminant Analysis. Silhouette scores quantify if an observation (black dot) is closer to the distribution of its own class (black 2) or to the distribution of another class (green 1) as shown in figure 6A. If the observations are strongly clustered the silhouette score is high, whereas it decreases if the classes are more overlapping such as in the example given in figure 6A. Figure 6B shows that at smaller time-scales task samples have significantly lower silhouette scores, meaning that they are more similar to other classes as opposed to their own, whereas rest is more similar to itself than other classes. With increasing window length the silhouette scores increase, but the difference between rest and tasks remains except for the time window of 80 s.

3.2 The structure of task-relevant information differs strongly across time-scale and method of FC extraction.

To evaluate the distribution of information structure across various FC methods we perform recursive feature elimination for each method and compare the resulting rankings using Spearman rank correlation (figure 7). The model-based metrics (precision and covariance) as well as the empirical metrics (Pearson correlation and covariance) display a similar decrease in similarity across time scale. GraphLasso precision and covariance also retain most similarity at similar time scales. This pattern is also present for GraphLasso covariance and empirical covariance, but not for GraphLasso precision and empirical covariance. Most importantly, the feature ranking of covariance (as well as covariance-based metrics) and correlation are not correlated at any time-scale, suggesting that the task-relevant information structure retrieved by these two methods is very dissimilar. With covariance and Pearson correlation the task-relevant information structure becomes more dissimilar with increasing difference in window length. At the shortest time-scales, feature rankings obtained from iFC are slightly correlated with Pearson correlation metrics and eigFC are slightly correlated with covariance metrics. Instantaneous FC and eigFC do not seem to be correlated. The decreasing correlation with size of time window suggests task-relevant information content also differs across time-scale. Although the concurrent decrease in accuracy for shorter time-scales might also indicate that sufficiently long window lengths are necessary for a stable estimate for covariance or correlation.

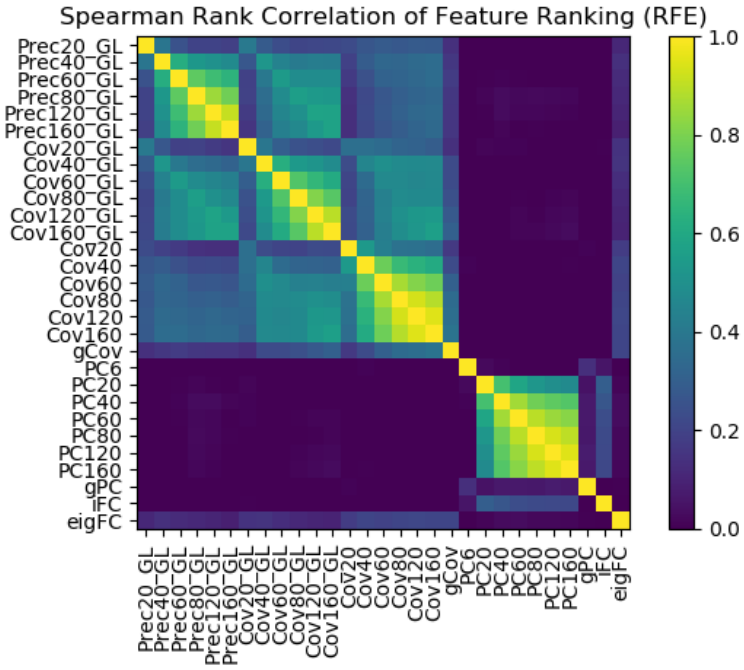


Figure 7. Task-dependent information structure can differ strongly across metric. Spearman rank correlation of all FC metrics.

4. Discussion

The aim of this paper was to evaluate if brain states can be classified with FC in a systematic manner and whether the extracted brain states are influenced by the choice of FC metric (phase coherence, Pearson correlation, covariance, GraphLasso precision, and GraphLasso covariance). Among empirical measure covariance outperformed correlation under certain conditions, in this five-task classification. Adding variance to covariance did not further increase accuracy. GraphLasso precision outperformed all empirical measures and was only outperformed by GraphLasso covariance for a window length of 20 s. Within-subject cross-validation accuracy was generally higher than between-subject cross-validation and can be conceptualized as an upper limit on accuracy. Another possibility is that more subjects are needed for between-subject cross-validation as suggested in a study by Abraham & al. (2017). They also found that accuracy increased with higher parcellation. Within- and between-subject cross-validation accuracy increased in proportion with the time-scale, which is likely due to high-frequency noise in the signal which is more likely to affect short time-scales (Cabral, Kringelbach, et al., 2017; Hutchison et al., 2013). An alternative explanation for the low accuracy at shorter time-scales is low task performance (Gonzalez-Castillo et al., 2015). Gonzalez-Castillo et al. (2015) showed that large deviations in task performance are correlated with substantial errors in classification accuracy. These deviations are more likely to bias connectivity measures at shorter time-scales. However, we did not control for this possibility. A third explanation could be that stable classification depends on specific frequency bands which would require window lengths long enough to capture these functional interactions. A study investigating the dependence of community structure on window length has already shown that different frequency bands can address distinct neuronal processes (Telesford et al., 2016). Specific neuronal processes could be better captured by models aimed at specific frequency bands such as dynamic causal modelling or the Kuramoto model (Cabral, Hugues, Sporns, & Deco, 2011; Friston, Kahan, Biswal, & Razi, 2014).

Accuracy at shorter time-scales was low for testing data, it was high for training data. This finding highlights that proper cross-validation is necessary to draw conclusions regarding classification performance since the data tends to get overfit. This is critical, since a high accuracy of the classifier on the training set is necessary, but not sufficient for high accuracy on the novel testing set. For example, in the study by Xie et al. (2017) the performance of the trained classifier was not validated with a novel dataset. Such validation would have been informative of whether the learned parameters can distinguish the brain states due to true

differences that hold at a population level or due to noise (Varoquaux et al., 2017). The problem of overfitting can generally be addressed by feature selection or regularization. Here neither feature selection nor L1 penalty regularization lead to an increase in accuracy for between- or within-subject CV. While feature selection eliminates features, the L1 penalty forces their weights to 0 indicating that the task-relevant signatures may be more distributed, because the classification improves of no features are discarded. Note that we did not perform an exhaustive search of the parameter space for the optimal combinations of feature number and L2 penalty parameter. Instead, we searched the parameter space serially: We optimized the feature number and then optimized the penalty parameter.

The strong decrease of accuracy at shorter time-scales was primarily driven by the difficulty of distinguishing tasks from each other rather than distinguish task states from rest. This suggests that the brain at rest is very dissimilar to the brain engaging in a task. This is in line with previous studies using whole-brain modelling (Ponce-Alvarez, He, Hagmann, & Deco, 2015; Senden et al., 2018, 2017). However, it could be argued that this stems from the fact that the stimuli used here were all visual, making the classification entirely reliant on non-sensory processes. It is, therefore, quite possible for other classification problems to reach better accuracies at smaller time-scales and with different FC methods. Another limiting factor could be the context-dependence of the features used in the multinomial classification. A feature can be crucial for distinguishing between task A and B, but not between task A and C. If the classification problem only includes tasks A and C the task-relevant information structure that is extracted by the classifier changes depending on the tasks included.

Task-relevant features that are crucially depends on which tasks are included in the classification. le specific functional interactions might be relevant in a pairwise discrimination between two tasks, they could become irrelevant in a multinomial discrimination depending on the tasks among which the classifier is discriminating.

The most important finding, however, is that the task-relevant information structure differs strongly not only across time-scale, but also across connectivity measures. The absence of any similarity in information structure retrieved from correlation and covariance is a counterintuitive and problematic result. Correlation is merely normalized covariance and evidence that such closely related methods can provide very different information contradicts the implicit assumption that similar methods should lead to similar conclusions. That this is not the case is problematic for the interpretation of any results obtained for different measures and

time-scales since there is no ground-truth on task-relevant functional interactions. For example, how would one interpret evidence from studies using network theory to detect communities based on different FC methods (Fuertinger & Simonyan, 2016; Najafi, Mcmenamin, Simon, & Pessoa, 2016; Sporns, 2013)? This underlines the need for alternative, better defined metrics such as model-based FC, where the relationships between the various metrics are better defined (Cabral et al., 2011; Friston et al., 2014; Pallares et al., 2018; Senden et al., 2018, 2017). However, the optimal metric may still strongly depend on the classification problem itself. Consequentially, this will impact the research design, for example when attempting to classify switching trials. Here, the task intervals have to be long enough for windows to only contain a single task.

In conclusion, the following suggestions can be made for classification in neuroscience. (1) When one is interested in groups and wished to obtain results which generalize to new subjects, accuracy model-based FC metrics should be used and precision should be preferred except for window lengths around 20 s. (2) When one is interested in individual subjects, empirical covariance should be preferred for classification. (3) Generally, larger window lengths should be preferred. (4) For MLR classifiers, L2 regularization should be preferred.

The pipeline developed here can be applied to other neuroimaging tools as well such as electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS). Quantifying the performance of a classifier is furthermore especially important in clinical settings when aiming to identify pathological brain states in new patients. Predictive decoders, for example in the case of brain-computer interfaces, can be implemented with FC metrics, but should be tuned within-subject as the performance is better and more stable. The main result of this study, namely, the dissimilarity of information-structure across FC methods, calls for greater care in the selection of FC method with respect to the aim of a study as well as a more careful interpretation of results in neuroscience using different FC methods in the future.

References

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, *147*, 736–745. https://doi.org/10.1016/J.NEUROIMAGE.2016.10.045

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning*. *Springer*. https://doi.org/10.1117/1.2819119

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(4), 312–312. https://doi.org/10.1038/nrn2618

Cabral, J., Hugues, E., Sporns, O., & Deco, G. (2011). Role of local network oscillations in resting-state functional connectivity. *NeuroImage*, *57*(1), 130–139. https://doi.org/10.1016/j.neuroimage.2011.04.010

Cabral, J., Kringelbach, M. L., & Deco, G. (2017). Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: Models and mechanisms. *NeuroImage*, (March). https://doi.org/10.1016/j.neuroimage.2017.03.045

Cabral, J., Vidaurre, D., Marques, P., Magalhães, R., Silva Moreira, P., Miguel Soares, J., … Kringelbach, M. L. (2017). Cognitive performance in healthy older adults relates to spontaneous switching between states of functional connectivity during rest. *Scientific Reports*, *7*(1), 5135. https://doi.org/10.1038/s41598-017-05425-7

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. Retrieved from http://arxiv.org/abs/1407.0202

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. https://doi.org/10.1016/j.neuroimage.2006.01.021

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Flowers, K., & Robertson, C. (1985). The effect of Parkinson's disease on the ability to maintain a mental set. *Journal of Neurology Neurosurgery, and Psychiatry*, *48*, 517–529. https://doi.org/10.1136/jnnp.48.6.517

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Friston, K. J., Kahan, J., Biswal, B., & Razi, A. (2014). A DCM for resting state fMRI. *NeuroImage*, *94*(100), 396–407. https://doi.org/10.1016/j.neuroimage.2013.12.009

Fuertinger, S., & Simonyan, K. (2016). Stability of Network Communities as a Function of Task Complexity. *Journal of Cognitive Neuroscience*, *28*(12), 2030–2043. https://doi.org/10.1162/jocn_a_01026

Gonzalez-Castillo, J., & Bandettini, P. A. (2017). Task-based dynamic functional

connectivity: Recent findings and open questions. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.08.006

Gonzalez-Castillo, J., Hoy, C. W., Handwerker, D. A., Robinson, M. E., Buchanan, L. C., Saad, Z. S., & Bandettini, P. A. (2015). Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proceedings of the National Academy of Sciences*, *112*(28), 8762–8767. https://doi.org/10.1073/pnas.1501242112

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., … Chang, C. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *Neuroimage*, *80*, 5–79. https://doi.org/10.1016/j.neuroimage.2013.05.079

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Lehman, A., & Rourke, N. O. (2005). *JMP for Basic Univariate and Multivariate Statistics A Step-by-Step Guide*. *Analysis*. Retrieved from http://books.google.com/books?id=1nlApuloc0AC&pgis=1

Najafi, M., Mcmenamin, B. W., Simon, J. Z., & Pessoa, L. (2016). Overlapping communities reveal rich structure in large-scale brain networks during rest and task conditions HHS Public Access. *Neuroimage*, *135*, 92–106. https://doi.org/10.1016/j.neuroimage.2016.04.054

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Pallares, V., Insabato, A., Sanjuan, A., Kuehn, S., Mantini, D., Deco, G., & Gilson, M. (2018). Subject- and behavior-specific signatures extracted from fMRI data using whole-brain effective connectivity. *Doi.Org*, 201624. https://doi.org/10.1101/201624

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://scikit-learn.org/stable/about.html#citing-scikit-learn

Ponce-Alvarez, A., He, B. J., Hagmann, P., & Deco, G. (2015). Task-Driven Activity Reduces the Cortical Activity Space of the Brain: Experiment and Whole-Brain Modeling. *PLoS Computational Biology*, *11*(8), 1–26. https://doi.org/10.1371/journal.pcbi.1004445

Preti, M. G., Bolton, T. A., & Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, *160*, 41–54. https://doi.org/10.1016/j.neuroimage.2016.12.061

Rahim, M., Thirion, B., Bzdok, D., Buvat, I., & Varoquaux, G. (2017). Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage*, *158*, 145–154. https://doi.org/10.1016/j.neuroimage.2017.06.072

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Senden, M., Reuter, N., van den Heuvel, M. P., Goebel, R., & Deco, G. (2017). Cortical rich club regions can organize state-dependent functional network formation by engaging in

oscillatory behavior. *NeuroImage*, *146*(October 2016), 561–574. https://doi.org/10.1016/j.neuroimage.2016.10.044

Senden, M., Reuter, N., van den Heuvel, M. P., Goebel, R., Deco, G., & Gilson, M. (2018). Task-related effective connectivity reveals that the cortical rich club gates cortex-wide communication. *Human Brain Mapping*, *39*(3), 1246–1262. https://doi.org/10.1002/hbm.23913

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Smith, J. O. (Julius O. (2007). *Mathematics of the discrete Fourier transform (DFT) : with audio applications* (2nd editio). W3K Publishing.

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*, *15*(3), 247–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24174898

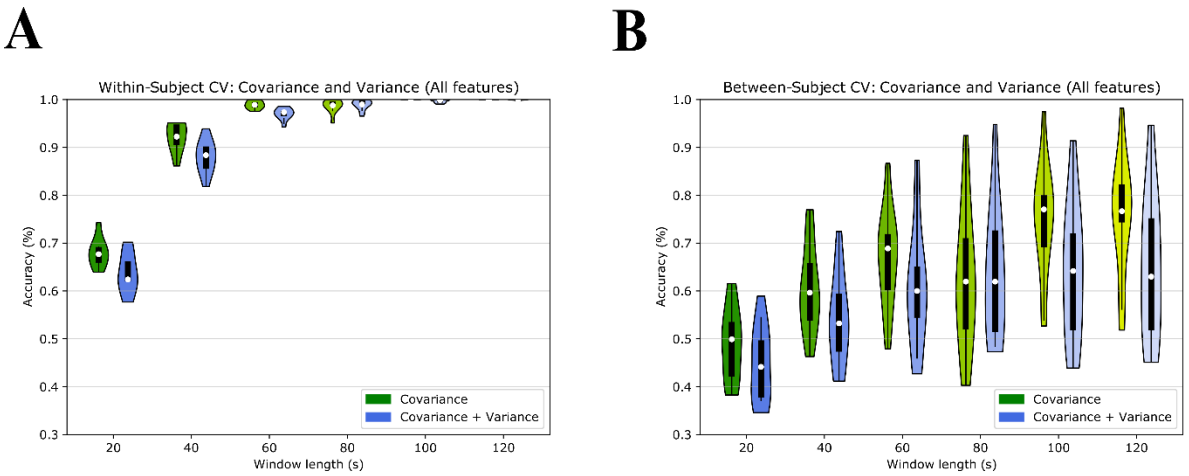
Telesford, Q. K., Lynall, M.-E., Vettel, J., Miller, M. B., Grafton, S. T., & Bassett, D. S. (2016). Detection of functional brain network reconfiguration during task-driven cognitive states. *NeuroImage*, *142*, 198–210. https://doi.org/10.1016/j.neuroimage.2016.05.078

Varoquaux, G., Reddy Raamana, P., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. https://doi.org/10.1016/j.neuroimage.2016.10.038

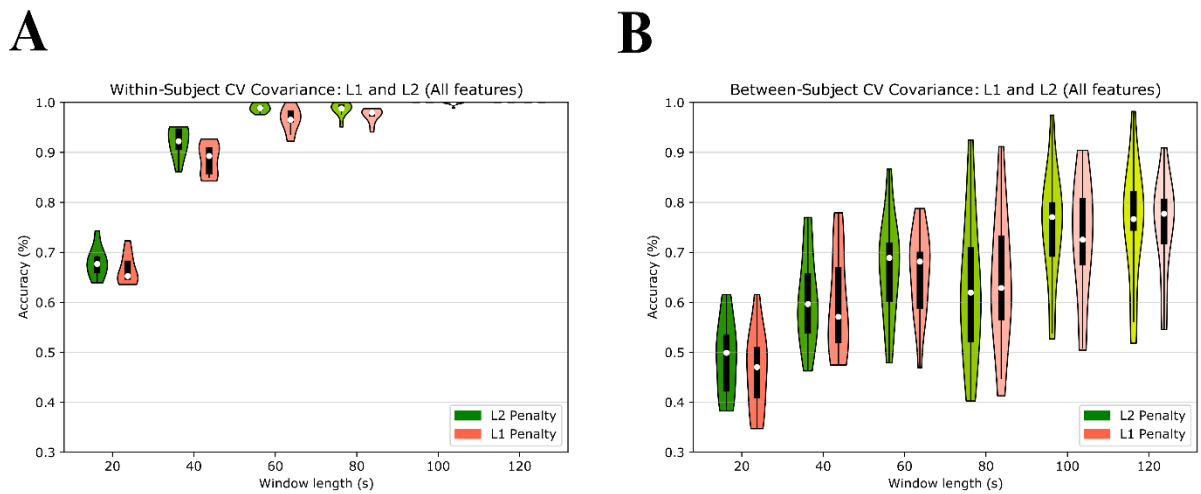
Vidaurre, D., Smith, S. M., & Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *PNAS*, *114*(48), 12827–12832. https://doi.org/10.1073/pnas.1705120114

Xie, H., Calhoun, V. D., Gonzalez-Castillo, J., Damaraju, E., Miller, R., Bandettini, P. A., & Mitra, S. (2017). Whole-brain connectivity dynamics reflect both task-specific and individual-specific modulation: A multitask study. https://doi.org/10.1016/j.neuroimage.2017.05.050

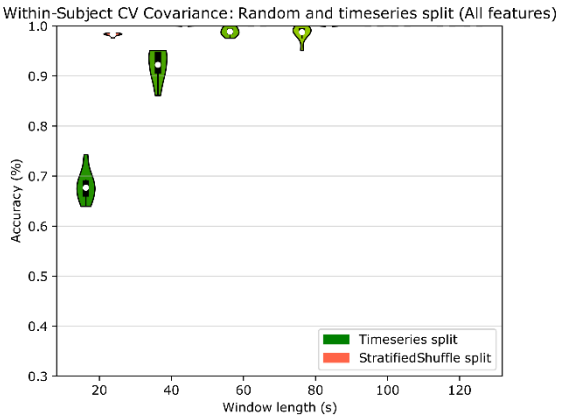
Supplementary Material



Supplementary Figure 1: Adding variance to covariance does not outperform covariance alone. (A) Within-subject cross-validation accuracy using only covariance (green) and covariance + variance (blue). (B) Between-subject cross-validation accuracy using only covariance (green) and covariance + variance (blue). Chance level is 0.2.



Supplementary Figure 2: Using L1 regularization instead of L2 regularization does not improve accuracy for covariance. (A) Within-subject cross-validation accuracy using L2 penalty (green) or L1 penalty (orange). (B) Between-subject cross-validation accuracy using L2 penalty (green) or L1 penalty (orange). Chance level is 0.2.



Supplementary Figure 3: Random split overestimates the cross-validation accuracy within a timeseries. Within-subject cross-validation within a run using covariance with a time series split (green) and a stratified shuffle split (orange). Chance level is 0.2.

Table S4

Results of parameters for metric and cross-validation using all features

Cross-Validation	Features (All/ Best)	Metric	Penalty parameter C (Mean)	Penalty parameter C (Std)	Testing Accuracy (Median)	Testing Accuracy (Std)	Training Accuracy (Median)	Training Accuracy (Std)
T	AF	Cov20	1012.72	1907.481	0.6768	0.0286	1	0
S	AF	Cov20	0.0008	0.0008	0.4989	0.0724	0.8052	0.0956
T	AF	Cov40	106.1225	35.3507	0.9224	0.0291	1	0
S	AF	Cov40	592400.6	2079076	0.5965	0.087	1	0
T	AF	Cov60	59.8244	58.09	0.9888	0.0079	1	0
S	AF	Cov60	591719.8	2079269	0.6889	0.0973	1	0
T	AF	Cov80	94.332	47.148	0.9878	0.0137	1	0
S	AF	Cov80	1155225	2829641	0.6195	0.1389	1	0
T	AF	Cov100	70.7443	57.7611	1	0	1	0
S	AF	Cov100	577996	2082457	0.7704	0.1128	1	0
T	AF	Cov120	24.1652	46.8836	1	0	1	0
S	AF	Cov120	757.4603	1661.841	0.7667	0.1172	1	0
S	AF	gCov	14186.89	50856.45	0.7	0.229	1	0.0713
T	AF	Covvar20	530.6649	1432.906	0.6241	0.0393	1	0
S	AF	Covvar20	2324511	3644483	0.4413	0.0751	0.9879	0.0129
T	AF	Covvar40	577.2371	1416.75	0.8837	0.033	1	0
S	AF	Covvar40	591795.2	2079248	0.5322	0.0897	1	0.0008
T	AF	Covvar60	117.906	0	0.9735	0.0118	1	0
S	AF	Covvar60	2325932	3643578	0.6	0.1218	1	0
T	AF	Covvar80	94.332	47.148	0.9898	0.0106	1	0

S	AF	Covvar80	1155621	2829479	0.6195	0.1347	1	0
T	AF	Covvar100	59.8244	58.09	0.9988	0.0035	1	0
S	AF	Covvar100	578701.7	2082262	0.6417	0.1353	1	0
T	AF	Covvar120	71.0459	57.3969	1	0.0015	1	0
S	AF	Covvar120	578365.5	2082355	0.6299	0.1443	1	0
T	AF	PC6	828418.9	2420044	0.3765	0.0365	0.9205	0.0647
S	AF	PC6	0.0016	0.0004	0.2931	0.0365	0.8337	0.0692
T	AF	PC20	59.5502	58.3639	0.6589	0.034	1	0
S	AF	PC20	14455.62	50797.02	0.4297	0.0716	0.9877	0.0044
T	AF	PC40	117.906	0	0.901	0.0242	1	0
S	AF	PC40	14119.31	50875.18	0.5395	0.109	0.9991	0.0004
T	AF	PC60	83.1174	53.1455	0.9796	0.0094	1	0
S	AF	PC60	42.1202	56.4874	0.6012	0.1339	0.9999	0.0003
T	AF	PC80	106.4035	34.5077	0.9976	0.0042	1	0
S	AF	PC80	16.8893	41.2399	0.6145	0.1448	1	0
T	AF	PC100	82.8226	53.596	1	0.0024	1	0
S	AF	PC100	8.6683	30.3058	0.6394	0.1538	1	0
T	AF	PC120	48.0339	57.0611	1	0.0019	1	0
S	AF	PC120	14144.6	50868.18	0.6561	0.1779	1	0.0052
S	AF	gPC	362.2131	1238.769	0.8	0.1767	1	0
T	AF	iFC	0.8653	1.3192	0.4185	0.0301	0.8386	0.0943
S	AF	iFC	0.0015	0.0006	0.3318	0.0448	0.7501	0.0773
T	AF	eigFC	808640.7	2425920	0.3239	0.0181	0.7442	0.1186
S	AF	eigFC	577600.3	2082567	0.2724	0.0299	0.633	0.0026
T	AF	Bold	0.2888	0.8639	0.1895	0.0384	0.252	0.0267
S	AF	Bold	8.6388	30.3142	0.2391	0.0224	0.2575	0.004
T	BF	Cov20	59.5502	58.3639	0.6679	0.0259	1	0
S	BF	Cov20	14110.8827	50877.5085	0.4824	0.0732	0.8012	0.1035
T	BF	Cov40	83.1174	53.1455	0.9173	0.0315	1	0
S	BF	Cov40	1170345.445	2823915.009	0.6151	0.0889	1	0.0002
T	BF	Cov60	82.8295	53.5854	0.9867	0.0083	1	0
S	BF	Cov60	578315.0282	2082369.074	0.6605	0.104	1	0
T	BF	Cov80	94.332	47.148	0.9867	0.0133	1	0
S	BF	Cov80	577970.3163	2082464.378	0.6039	0.1353	1	0
T	BF	Cov100	70.7443	57.7611	1	0	1	0
S	BF	Cov100	412.5183	1225.3701	0.7718	0.1126	1	0
T	BF	Cov120	24.1652	46.8836	1	0	1	0
S	BF	Cov120	592442.9543	2079064.265	0.75	0.1155	1	0
T	AF	Prec10_GL	39510.52	79020.91	0.876	0.0127	0.9963	0.0035
S	AF	Prec10_GL	0.0017	0	0.5275	0.1068	0.9867	0.0014
T	AF	Prec20_GL	0.6255	1.1277	0.9934	0.002	0.9992	0.0003
S	AF	Prec20_GL	0.0213	0.031	0.6581	0.1298	0.9986	0.0005
T	AF	Prec30_GL	0.0223	0.0315	1	0.0007	1	0
S	AF	Prec30_GL	101.4738	40.2506	0.7321	0.159	1	0
T	AF	Prec40_GL	0.0017	0	1	0	1	0

S	AF	Prec40_GL	51.5701	57.458	0.7645	0.17	1	0
T	AF	Prec50_GL	0.0086	0.0206	1	0.0008	1	0
S	AF	Prec50_GL	59.9869	57.9263	0.8169	0.1697	1	0
T	AF	Prec60_GL	0.0017	0	1	0	1	0
S	AF	Prec60_GL	14498.76	50784.77	0.8591	0.1859	1	0
T	AF	Cov10_GL	0.9134	1.2877	0.9349	0.0081	0.9963	0.0002
S	AF	Cov10_GL	345.5998	1242.707	0.5527	0.1188	0.9965	0.0005
T	AF	Cov20_GL	519.7243	1436.427	0.9967	0.0022	0.9992	0.0001
S	AF	Cov20_GL	1413.447	2158.975	0.6337	0.1353	0.9991	0.0001
T	AF	Cov30_GL	507.0977	1440.45	1	0.0012	1	0
S	AF	Cov30_GL	117.906	0	0.6951	0.1365	1	0
T	AF	Cov40_GL	0.0566	0.0275	1	0.0012	1	0.0006
S	AF	Cov40_GL	413.1352	1225.163	0.6882	0.1519	1	0
T	AF	Cov50_GL	0.3102	0.8573	1	0	0.9998	0.0001
S	AF	Cov50_GL	577651.8	2082552	0.6831	0.1412	1	0
T	AF	Cov60_GL	0.2896	0.8636	1	0.0009	0.9995	0.0002
S	AF	Cov60_GL	43.961	55.1154	0.6667	0.1351	1	0