Quantifying Temporal Correlations: A Test-Retest Evaluation of Functional Connectivity in Resting-State fMRI

Mark Fiecas, University of California - San Diego
Hernando Ombao, University of California - Irvine
Dan van Lunen, Brown University
Richard Baumgartner
Alexandre Coimbra, et al.

Available at: https://works.bepress.com/mfiecas/5/
Quantifying Temporal Correlations: A Test-Retest Evaluation of Functional Connectivity in Resting-State fMRI

Mark Fiecas\textsuperscript{a,}\textsuperscript{*}, Hernando Ombao\textsuperscript{b}, Dan van Lunen\textsuperscript{c}, Richard Baumgartner\textsuperscript{d}, Alexandre Coimbra\textsuperscript{e}, Dai Feng\textsuperscript{d}

\textsuperscript{a}Department of Psychiatry, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92037
\textsuperscript{b}Department of Statistics, University of California at Irvine, Bren Hall 2206, Irvine, CA 92697
\textsuperscript{c}Division of Applied Mathematics, Brown University, Providence, RI 02912
\textsuperscript{d}Biometrics Research Department, Merck Research Laboratories, Rahway, NJ 07065
\textsuperscript{e}Imaging Research, Merck Research Laboratories, West Point, PA 19486

Abstract

There have been many interpretations of functional connectivity and proposed measures of temporal correlations between BOLD signals across different brain areas. These interpretations yield from many studies on functional connectivity using resting-state fMRI data that have emerged in recent years. However, not all of these studies used the same metrics for quantifying the temporal correlations between brain regions. In this paper, we use a public-domain test-retest resting-state fMRI data set to perform a systematic investigation of the stability of the metrics that are often used in resting-state functional connectivity (FC) studies. The fMRI data set was collected across three different sessions. The second session took place approximately eleven months after the first session, and the third session was an hour after the second session. The FC metrics comprised of cross-correlation, partial cross-correlation, cross-coherence, and parameters based on an autoregressive model. We discussed the strengths and weaknesses of each metric. We performed ROI-level and full-brain seed-based voxelwise test-retest analyses using each FC metric to assess its stability. For both ROI-level and voxel-level analyses, we found that cross-correlation yielded more stable measurements than the other metrics. We discussed the consequences of this result on the utility of the FC metrics. We observed that for negatively correlated ROIs, their partial cross-correlation is shrunk towards zero, thus affecting the stability of their FC. For the present data set, we found greater stability in FC between the second and third sessions (one hour between sessions) compared to the first and second sessions (approximately 11 months between sessions). Finally, we report that some of the metrics showed a positive association between strength and stability. In summary, the results presented in this paper suggest important implications when choosing metrics for quantifying and assessing various types of functional connectivity for resting-state fMRI studies.

\textit{Keywords:} Functional connectivity; Frequency domain; Bivariate autoregressive model; Resting-state fMRI; Test-retest analysis

\textsuperscript{*}Correspondence to: Mark Fiecas, University of California, San Diego, Department of Psychiatry, 9500 Gilman Drive, La Jolla, CA 92037

Email address: mfiecas@ucsd.edu (Mark Fiecas)
1. Introduction

Functional connectivity (FC) analyses using resting-state fMRI data have shown the potential of FC as a biomarker for various diseases (Fox & Raichle, 2008; Fox & Greicius, 2010). Despite the fact that a number of studies already suggest the potential clinical utility of resting-state FC, the overall scope of clinical applications for FC has yet to be defined. One difficulty with FC analyses is that, although there is agreement on FC being “conceptually” defined as the level of temporal correlation in the fluctuations of the blood oxygenation level-dependent (BOLD) signal across different regions in the brain (Friston et al., 1993; Salvador et al., 2005b), there is no established standard on how to quantify FC, and thus numerous metrics have been proposed to express FC. Our contribution is a systematic test-retest study to evaluate the stability of the most commonly used metrics for FC.

The popular metrics for quantifying FC include cross-correlation (Friston et al., 1993; Worsley et al., 2005), partial cross-correlation (Marrelec et al., 2006; Salvador et al., 2005a), cross-coherence (Chang & Glover, 2010; Cordes et al., 2001; Sun et al., 2004) and partial cross-coherence (Salvador et al., 2010; Sun et al., 2004), and metrics based on the vector autoregressive model (Liao et al., 2011; Roebroeck et al., 2005; Uddin et al., 2009). Each of the proposed FC metrics measures some form of dependency between the time courses of the BOLD signal, but the precise nature of dependency differs across metrics. Some metrics capture oscillatory synchronization at a frequency band of interest (e.g., cross-coherence, partial directed coherence), others capture temporally lagged relationships (e.g., directed influence), and yet another set of metrics quantify conditional (“direct”) dependencies (e.g., partial cross-correlation, partial coherence) as opposed to marginal (“indirect”) dependencies. Thus, when one claims that two regions are “functionally connected”, one should precisely define how FC was quantified.

Recent resting-state fMRI studies have taken a graph-theoretic approach to FC analyses. Graph theory defines networks of vertices which may be interconnected by edges. In this theoretical framework vertices correspond to different regions of the brain and an edge connecting two vertices is drawn when the two corresponding regions are “functionally connected”, where FC is quantified by some FC metric and then edges are assigned between two vertices of the graph if the metric exceeds some a priori defined threshold. Lynall et al. (2010) used both cross-correlation and cross-coherence, Salvador et al. (2005b) used partial mutual information, and Liao et al. (2011) used vector autoregressive models. Using simulated data, Smith et al. (2011) used many of the current methods for quantifying FC to evaluate the sensitivity and specificity of each metric for reconstructing the simulated graphs.

Various metrics for quantifying FC have been investigated in a number of recent articles. An overview of many of the popular metrics was given by Zhou et al. (2009). David et al. (2004) used a neural mass model to show that different FC metrics are sensitive to different types of signal dependencies and recommended that
when the nature of signal dependencies is not well-understood, one should apply a variety metrics. Astolfi et al. (2007) compared different metrics of FC based on the vector autoregressive model and investigated their performance at various levels of signal-to-noise ratios and amount of data present. Quiroga et al. (2002) analyzed EEG signals and compared and contrasted results using different measures of linear and nonlinear dependency. Smith et al. (2011) performed an extensive simulation study to evaluate the performance of FC metrics on reconstructing the underlying connectivity network under numerous simulation settings, but because the results were based on simulated data where the true underlying dependency structure was known, then the conclusions drawn are limited to situations that closely follow the simulation model.

It is important to understand not only the nature of dependency each metric quantifies but also to assess the stability of each FC measure. In this context, stability refers to the variability of each metric obtained from the same subjects measured at different sessions and without any changes to the scanning protocols and FC analysis per session. Ideally, the dependencies that are being measured by each metric do not change with respect to the session under the assumption that the brain dynamics remain unchanged across sessions. Here, we investigated the effects of the elapsed time between sessions.

Bennett & Miller (2010) gave an overview of test-retest analyses of various types of fMRI analyses. Specifically to resting-state FC analyses, Meindl et al. (2010) showed that the “default mode network” (DMN), as obtained using independent component analysis, is stable in healthy young subjects, and Braun et al. (2012) showed the stability of graph-theoretical properties of the resting brain’s underlying network. With the same data used in this paper, Wang et al. (2011) showed the stability of various graph-theoretic measures in a network analysis of resting-state FC. Zuo et al. (2010) showed the stability of the components derived from an independent components analysis. Finally, Shehzad et al. (2009) showed the stability of FC as quantified by cross-correlation.

In the present work, we focus on investigating the stability of inter-regional time-course dependencies as quantified by various FC metrics: cross-correlation, partial cross-correlation, cross-coherence, and autoregressive model-based metrics. We give an overview of each of these metrics and elaborate on the type of time course dependency they measure. To quantify stability across sessions, we used the random-effects ANOVA model under which we can rigorously define the amount of between and within-subject variability in the measurements. Finally, we make some recommendations based on the stability profile encountered for each metric.

2. Materials and Methods

2.1. Participants

We used a resting-state fMRI data set of 25 participants (mean age 29.44 ± 8.64, 10 males) that is publicly available at NITRC (http://www.nitrc.org/projects/trt). This data set has been used to examine the
stability of cross-correlations (Shehzad et al., 2009), ICA and dual regression (Zuo et al., 2010), low-frequency oscillations (Zuo et al., 2010b), and graph-theoretic network properties (Wang et al., 2011).

2.2. Data Acquisition

A Siemens Allegra 3.0-Tesla scanner was used to obtain three resting-state scans for each participant. Each scan consisted of $T = 197$ contiguous EPI functional volumes with a time repetition (TR) = 2000 ms; time echo (TE) = 25 ms; flip angle (FA) = 90°; 39 number of slices, matrix = 64 × 64; field of view (FOV) = 192 mm; voxel size $3 \times 3 \times 3$ mm$^3$. Scans 2 and 3 were conducted in a single session 45 minutes apart and were 5-16 months (mean 11 ± 4 months) after scan 1. During each scan, each participant was asked to relax and remain still with eyes open during the scan. For spatial normalization and localization, a high-resolution T1-weighted magnetization prepared gradient echo sequence was obtained (MPRAGE, TR = 2500 ms; TE = 4.35 ms; inversion time = 900 ms; FA = 8°, number of slices = 176; FOV = 256 mm).

2.3. Preprocessing

The data were preprocessed using both FSL (version 4.1, http://www.fmrib.ox.ac.uk) and AFNI (version AFNI_2010_10_19_1028, http://afni.nimh.nih.gov/afni). The images were 1) motion corrected using FSL’s mcflirt (rigid body transform; cost function normalized correlation; reference volume the middle volume) and then 2) normalized into the Montreal Neurological Institute space using FSL’s flirt (affine transform; cost function mutual information). FSL’s fast was then used to 3) obtain a probabilistic segmentation of the brain to obtain white matter and cerebrospinal fluid (CSF) probabilistic maps, thresholded at 0.99. AFNI’s 3dDetrend was then used to 4) remove the nuisance signals, namely the six motion parameters, white matter and CSF signals, and the global signal. The effects of removing the global signal has been discussed in the literature (see, e.g., Fox et al. (2009) and Murphy et al. (2009)), and so we repeat our analyses without removing the global signal to investigate the impact this preprocessing step has on our results. Finally, using FSL’s fslmaths, the volumes were 5) spatially smoothed using a Gaussian kernel with FWHM = 6mm.

2.4. Quantifying Functional Connectivity

We now present an overview of the metrics for quantifying FC that we studied in the present work. These metrics are summarized in Table 1. We also give a short overview of some of the popular metrics that were omitted.

**Cross-correlation and partial cross-correlation.** Cross-correlation (CCor) is the simplest and most popular metric and has been well-studied and used in statistics and neuroimaging. CCor measures the contemporaneous linear relationship between two signals. It is possible to compute *lagged* cross-correlations, but
one must specify the lag. Throughout this work, unless otherwise stated, the Fisher-z transform (hyperbolic arctangent function) of the cross-correlation values was utilized. This transform is known to stabilize the asymptotic variance of the estimators (Brockwell & Davis, 1998) so that the asymptotic variance of the transformed sample cross-correlation, under Gaussianity, no longer depends on the population cross-correlation which is an unknown quantity. Cross-correlation is a measure of marginal (indirect) relationship between two signals. Thus, it is possible that the two signals are only indirectly related, i.e., the linear dependence is being driven by other signals. By removing the linear effects of the other signals in the data and then cross-correlating the residuals, then one obtains another metric called partial cross-correlation (PCCor).

An attractive feature of PCCor is that it quantifies the conditional (direct) linear relationship between two signals, i.e., conditional on the other signals in the analysis. To illustrate this, consider the following simple example. Suppose we have three regions-of-interest (ROIs), call them X, Y, and Z, such that the information flows from X to Y and then to Z; as a directed graph we can illustrate this as \( X \rightarrow Y \rightarrow Z \). In this setup, the signals from ROIs X and Z will be temporally correlated because the signal in Z is carrying information from the signal X. However, this is an indirect relationship because the information flowed through ROI Y, i.e., the signals from X and Z are related conditional on the signal from Y. If we were to remove the effects of the signal in Y on each of the signals in X and Z, then X and Z’s adjusted signals will no longer be correlated. In this setup, the signals in X and Z will have zero partial cross-correlation because they are conditionally uncorrelated.

**Regularized partial cross-correlation.** All pair-wise sample partial cross-correlations can be efficiently estimated using the precision matrix, which is the inverse of the sample variance-covariance matrix of the data (Marrelec et al., 2006). However, this can be problematic when the number of ROIs is large and the time course is relatively short because the sample variance-covariance matrix will be ill-conditioned and hence difficult to invert. Consequently, small perturbations in the data can result in large changes in the entries of the precision matrix (Medkour et al., 2009; Fiecas et al., 2010), i.e., the estimates of the partial cross-correlations will be numerically unstable. One approach to alleviate this problem is to regularize the sample variance-covariance matrix. We regularized the sample variance-covariance matrix using an \( \ell_1 \) penalty (Friedman et al., 2007). In particular, given a sample variance-covariance matrix \( S \), we seek an estimator of the precision matrix, denoted \( \hat{\Omega} \), that maximizes the penalized log-likelihood

\[
\log(\det(\hat{\Omega})) - \text{tr}(S\hat{\Omega}) - \lambda\|\hat{\Omega}\|_1,
\]

where the matrix norm \( \| \cdot \|_1 \) is the sum of the absolute values of the elements of the matrix, including the diagonal elements. With this approach, the variance-covariance matrix is shrunk towards the identity matrix (i.e., the off-diagonals are shrunk to zero), which in turn improves its condition number.
In Equation (1), one would need to select a tuning parameter $\lambda$ which determines how much weight to put on the penalty term: larger values of $\lambda$ give greater amount of regularization and setting $\lambda = 0$ is equivalent to no regularization. In this work, we picked $\lambda$ by maximizing the penalized log-likelihood in Equation (1) using a 10-fold cross-validation procedure (Friedman et al., 2007). For implementation, we used the glasso library in R for obtaining an $\ell_1$-regularized precision matrix (Witten et al., 2011), and from here we extract the partial cross-correlations (PCCorL1).

**Cross-coherence.** Whereas cross-correlation measures the linear relationship between the two time courses in the time domain, cross-coherence (CCoh) measures the linear relationship in the frequency domain and quantifies the synchronization of the time courses at frequencies (or frequency bands) of interest. As demonstrated in Ombao & van Bellegem (2008), cross-coherence is asymptotically equivalent to the cross-correlation between a pair of filtered signals. Thus, CCoh can identify the frequencies (or frequency bands) that drive the linear association between the two time courses. An attractive feature of cross-coherence is that it captures the linear relationship at all possible lags. For computing CCoh, we used the MATLAB toolbox (http://groups.google.com/group/fc-toolbox) provided by Zhou et al. (2009), whose implementation used Welch periodograms for estimation of the spectral density matrix, from which CCoh can be extracted. One can also estimate the spectral properties of the time course data using various other methods, e.g., for spectral density estimation one can use periodogram smoothing (Brockwell & Davis, 1998) or the multitaper estimator (Walden, 2000).

Throughout this work, we took the square root of the modulus of cross-coherence (also known as zero-phase coherency) so that we were looking only at the strength of the relationship. Our estimates of cross-coherence were averaged within the frequency band [0.01,0.10] Hertz because the low frequencies are known to carry relevant information to resting-state FC (Salvador et al., 2005b). We point out that other frequency bands may, in fact, be of interest and can affect the analysis. For instance, Lynall et al. (2010) showed that there exists differences in resting-state FC between schizophrenia and healthy controls by quantifying FC via CCoh in the frequency band [0.060, 0.125] Hertz. Finally, we applied the Fisher-$z$ transform because it is a variance-stabilizing transform of the estimates of CCoh (Ombao & van Bellegem, 2008).

**Partial coherence.** Partial coherence (PCoh) is the frequency domain analog of partial cross-correlation. PCoh quantifies the conditional linear relationship between two time courses at frequencies (or frequency bands) of interest. While PCCor was estimated by first inverting the sample variance-covariance matrix of the data, PCoh can be estimated by inverting the spectral density matrix of the data, which is the frequency domain analog of the variance-covariance matrix (Fiecas et al., 2010). However, estimates of the spectral density matrix are often not invertible or suffer from numerical instability, and thus, shrinkage methods
were developed to address these problems (Medkour et al., 2009; Fiecas et al., 2010). To obtain estimates of PCCoh, we first estimated the spectral density matrix by smoothing the periodogram matrix using a 31-point Hamming window, and this smoothed periodogram matrix was regularized using the shrinkage method developed by Fiecas et al. (2010). From here, estimates of PCCoh can then be extracted because the regularized smoothed periodogram matrix is guaranteed to be invertible.

**Autoregressive model-based metrics.** Bivariate autoregressive models capture influence by relating the present value of one signal \( X \) to the past values of another signal \( Y \). This extra layer of complexity in bivariate autoregressive models admit a good statistical model of the temporal dynamics of each signal in the system as well as allows one to measure both contemporaneous relationships and temporally lagged relationships between signals. Moreover, bivariate autoregressive models also allow one to deduce the direction of the dependency. This type of dependency is known as Granger-causality (Geweke, 1982). We emphasize that this type of causality is only in the “statistical” sense (not necessarily in the physiological sense of the word) and we say that \( Y \) “Granger-causes” \( X \) if knowledge of past values of \( Y \) improves prediction of the current value of \( X \).

To quantify FC in the time domain, one can look at each direction of influence, the joint influence, and the difference in directional influences (Goebel et al., 2003; Roebroeck et al., 2005; Roebroeck et al., 2005) advocated to look at the difference in directed influences to investigate dependency, claiming that it is a more robust metric. For each pair of ROIs, we fit a bivariate autoregressive model of the first order (the order was selected by the Bayesian information criterion for a large majority of the ROI pairs) and then calculated the difference in directional influences (DI). In the frequency domain, we looked at partial directed coherence, which captures a frequency-specific measure of directed dependency based on the autoregressive model (Baccalà & Sameshima, 2001) (PDC).

A vector autoregressive model where each ROI represents one component of the model can be used to capture the dependencies between two ROIs conditional on the other ROIs in the model. In particular, partial directed coherence is meant to model the frequency-specific conditional dependencies via a vector autoregressive model. However, there were not enough data to estimate all of the parameters of such a model, and thus, following Smith et al. (2011), we only used pairwise bivariate autoregressive models.

To compute any of these metrics, one would need to set the order of the autoregressive model and estimate the parameters of the model. Estimation and inference procedures are well-known and well-studied for these models. However, one would need to decide if the same order should be used between all pairs of ROIs or voxels, as many studies have done (Goebel et al., 2003; Liao et al., 2011; Roebroeck et al., 2005; Uddin et al., 2009). Using the same order simplifies the comparisons of the FC estimates between ROI pairs or voxel pairs because the estimates come from comparable models. However, this approach runs the risk of potentially
underfitting or overfitting the time courses from some pairs of ROIs or voxels. Alternatively, one can vary the model orders across each pair but this approach is time-consuming especially for voxelwise analyses and introduces complications when comparing FC metrics.

**An overview of omitted metrics.** Nonlinear dependencies can be captured using mutual information (MI; Shannon & Weaver (1948)). This metric quantifies the amount of information shared by the two signals; if the two signals are independent, then the MI between the two signals is zero, and if one signal is completely determined by the other, then the MI between the two signals is infinity. Investigating nonlinear interactions can be useful in resting-state fMRI studies, as shown by Salvador et al. (2005b) and Salvador et al. (2005a). MI is a function of the joint probability density function between the two signals and the marginal probability density functions of each of the two signals and thus estimating MI is non-trivial since one would need to estimate these underlying probability density functions. If we assume Gaussian signals, then MI is easier to compute because it reduces to a monotonic transform of CCoh (Salvador et al., 2005b; Zhou et al., 2009). Because of this relationship between MI and CCoh when assuming Gaussian signals, a test-retest FC analysis using MI assuming Gaussian signals will yield identical results to CCoh; we have, in fact, verified this, but we omit the results in the present work.

Dynamic causal modeling (DCM; Friston et al. (2003)) and structural equation modeling (SEM; Wright (1921); McIntosh & Gonzales-Lima (1994)) were not investigated in this work because of their need for a hypothesized network prior to model fitting. Moreover, the established DCM methods require “input” timings to be specified, which are not known for resting-state fMRI data.

**2.5. Test-retest Analysis**

We used the intraclass correlation coefficient (ICC) to assess stability of subject-specific measures across the three sessions. To calculate the ICC, we used the following two-way random-effects ANOVA model,

\[ \rho_{jk} = \rho + \alpha_j + \beta_k + \epsilon_{jk}, \]  

(2)

where \( \rho_{jk} \) is the estimated FC quantified by one of the metrics for subject \( j \) in session \( k \), \( \rho \) is the population mean connectivity, \( \alpha_j \sim N(0, \sigma^2_\alpha) \) is the subject-effect, \( \beta_k \sim N(0, \sigma^2_\beta) \) is the session-effect, and \( \epsilon_{jk} \sim N(0, \sigma^2_\epsilon) \) is noise (Shrout & Fleiss, 1979). The ICC is then defined as

\[ \text{ICC} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\epsilon}, \]

(3)

which is the proportion of variability in the FC measurements that is attributed to the subjects. The value of ICC lies between 0 and 1; ICC is close to 0 if the variability in the FC is driven by the effect of the session.
<table>
<thead>
<tr>
<th>FC Metric</th>
<th>Marginal vs Conditional</th>
<th>Frequency-specific</th>
<th>Directed</th>
<th>Necessary Parameters</th>
<th>Type of Dependency Quantified</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCOr</td>
<td>Marginal</td>
<td>No</td>
<td>No</td>
<td>Lag</td>
<td>Linear association</td>
</tr>
<tr>
<td>CCoh</td>
<td>Marginal</td>
<td>Yes</td>
<td>No</td>
<td>Smoothing kernel and bandwidth</td>
<td>Frequency-specific linear association</td>
</tr>
<tr>
<td>Regularized PCoh</td>
<td>Conditional</td>
<td>Yes</td>
<td>No</td>
<td>Smoothing kernel, bandwidth and regularization parameter</td>
<td>Frequency-specific conditional linear association</td>
</tr>
<tr>
<td>PCCor</td>
<td>Conditional</td>
<td>No</td>
<td>No</td>
<td>Lag</td>
<td>Conditional linear association</td>
</tr>
<tr>
<td>PCCorL1</td>
<td>Conditional</td>
<td>No</td>
<td>No</td>
<td>Lag and regularization parameter</td>
<td>Conditional linear association</td>
</tr>
<tr>
<td>DI</td>
<td>Marginal</td>
<td>No</td>
<td>Yes</td>
<td>Model order and model parameters</td>
<td>Instantaneous and lagged linear associations</td>
</tr>
<tr>
<td>PDC</td>
<td>Conditional</td>
<td>Yes</td>
<td>Yes</td>
<td>Model order and model parameters</td>
<td>Directed frequency-specific conditional linear association</td>
</tr>
</tbody>
</table>

Table 1: A summary of the metrics used in the present work. Each of these metrics are described in the text.
or the noise or both; and ICC is close to 1, if the variability in the FC is driven by the variation across subjects. Note that ICC = 1 indicates perfect stability, i.e., there is no variation across sessions and the within-subject-session estimates do not contain noise.

To estimate ICC for the ROI-level analysis, we used the `lmer` procedure in the `lme4` package in R [http://www.r-project.org/], and for the voxelwise analysis we used the `3dICC_REML.R` program in AFNI which also uses the `lmer` procedure. In both analyses, the estimates of the parameters of the two-way random-effects ANOVA model were obtained using restricted maximum likelihood (REML) which guarantees that estimates of ICC fall in the interval [0,1]. Some test-retest articles use the method-of-moments estimator which could give negative values of ICC. In fact, negative estimates were either reported as they were (Shehzad et al., 2009; Zuo et al., 2010b; Wang et al., 2011) or deliberately set to zero (Braun et al., 2012). However, the `lmer` procedure can yield estimates of 0 for ICC. This can happen when the ANOVA model in Equation (3) is not valid, e.g., subject and session effects may not be additive.

We investigated long-term ICC for Sessions 1 and 2 by setting \( k = 1, 2 \) in Equation 2 above and short-term ICC for Sessions 2 and 3 (again by setting \( k = 2, 3 \) above) to investigate the effect of elapsed time on the ICC. We also investigated the overall (Sessions 1, 2, and 3) ICC.

### 2.6. ROI Analysis

To obtain anatomically defined ROIs, we used the Anatomical Automatic Labeling atlas (AAL) which gives an anatomical parcellation of the whole brain into 45 regions in each hemisphere (Tzourio-Mazoyer et al., 2002). These anatomical regions are listed in Table 2. To investigate the sensitivity of our analysis with respect to the choice of the ROIs, we also used the functional ROI parcellation, which also yields a total of 90 ROIs (Shirer et al., 2012). Each region’s mean time course was obtained for each individual by averaging the fMRI time series over all of the voxels within the region. Each regional time course was then detrended and standardized to unit variance. Furthermore, for the FC metrics in the time-domain (CCor, PCCor, PCCorL1, and DI) we applied a 4-th order Butterworth filter with passband [0.01 0.10] Hertz; the frequency-based FC metrics (CCoh, PCoh, and PDC) were calculated on unfiltered signals because these metrics are already frequency-specific.

To perform the ROI analysis, we calculated each of the FC metrics given in Table 1. The subject and session-specific estimates were then exported to R where we used the random effects ANOVA to investigate the stability of FC. In particular, we calculated the ICCs as described above to investigate how much each source of variability contributed to the overall variability in each of the calculated FC metrics. Only data from Sessions 1 and 2 were used to compute the long-term ICCs and only data from Sessions 2 and 3 were used to compute the short-term ICCs. To investigate the difference between the mean long-term and short-term stability, we used the permutation test which was conducted as follows. Under the null hypothesis of no difference between the mean long-term and mean short-term ICCs, then, iterating over the subjects, we
randomly assign the session pairs (Sessions 1 and 2 for long-term ICC and Sessions 2 and 3 for short-term ICC) to a group used to calculate either the long-term or the short-term ICCs. After the permutation, the difference between the mean long-term and short-term ICCs were computed. This procedure was repeated 1000 times to obtain a distribution of differences under the null hypothesis. The difference between the mean long-term and short-term ICCs computed using the unpermuted data was then compared to the distribution of differences in order to obtain an approximate p-value. We proceeded with the ROI analysis as follows. We show the relationship between cross-correlation (CCor) and partial cross-correlation (PCCor and PCCorL1) and show how this relationship changes as the penalty parameter in Equation 1 increases. We estimated FC at the population level by averaging each of the metrics over all twenty-five subjects per session. We then compared these population-level estimates between sessions to investigate between-session stability in the estimates. Finally, we investigated the relationship between the strength of FC with stability as measured by the ICC.

2.7. Voxelwise Seed Analysis

We used the seeds in the posterior cingulate cortex (PCC), supplemental motor area (SMA), and intraparietal sulcus (IPS) defined as a cube having side length 6 mm centered at MNI coordinates (-6, -58, 28), (2, 10, 48) and (26, 58, 48), respectively. These seeds are at the same location as that used by Salvador et al. (2005b). The seed time course was defined as the average of the voxel time courses within the cube, and was then detrended, standardized to unit variance, and then filtered with a 4-th order Butterworth filter with stop band at 0.08 Hertz. To perform the seed analysis, we take each voxel in the data and similarly

Table 2: A list of the cortical and subcortical regions as anatomically defined by the the AAL atlas, along with their abbreviations as used by Salvador et al. (2005b). Left-hand and right-hand sided homologies are suffixed by L and R, respectively.
detrended, standardized, and filtered its time course, and then calculated the CCor, CCoH, and DI between the preprocessed time course with the preprocessed seed time course.

After computing each FC metric per voxel, we used the the 3dICC_REML.R program in AFNI to compute the ICCs. Due to the large number of voxels in the analysis, we only investigated the ICC values and did not investigate how much each source of variability contributed to the overall variability. With these results, we used the cluster program in FSL to identify the spatial locations in the brain, and consequently the ROI, that had the highest ICC values. group-level Z statistics per voxel for each metric per session. The Z statistics were computed using OLS were adjusted for the underlying smoothness in the calculated FC metrics using random field theory (RFT). Next, we compared each of these metrics across sessions to investigate between-session stability. Finally, we investigated the relationship between the strength of FC with stability as measured by the ICC.

3. Results

3.1. ROI-Level FC Stability

**ROI-level ICCs.** We show the distribution of the ICCs over the 4005 possible pairs of ROIs in Figure 1 and this distribution is summarized in the last column of Table 3. In Table 3 we report the mean and standard deviation of the empirical distribution of the 4005 estimated ICCs and the estimated variance from each source of variability (subject, session, and noise). CCor had the highest ICCs. The stability of CCor is similar to that obtained by Shehzad et al. (2009), with the differences coming from our different procedures for data preprocessing and calculation of the ICC. For all metrics, a permutation test showed that the mean ICC was higher in the short-term than in the long-term. As we will discuss later, while CCor and CCoH were essentially estimating the same type of dependency, the mean of the distribution of overall ICCs for CCor was higher than the mean of the distribution of the overall ICCs for CCoH ($p < 0.001$ from a permutation test). Similarly, PCCorL1 and PCoh were essentially estimating the same type of dependency, though there was no evidence of a difference in the mean of their respective distribution of overall ICCs ($p = 0.214$ from a permutation test).

Using the functional ROI parcellation, the ICCs, which are shown in Supplementary Table 1, were different, but CCor again yielded the highest ICCs. The effects of removing the global signal can be seen in Supplementary Table 2, where we see a substantial increase in the mean ICCs for CCoH. Looking at the distribution of the ICCs in Supplementary Figure 1, we see that the increase in the mean ICCs for CCoH occurred because there were not as many null ICCs.

**The relationship between cross-correlation and partial cross-correlation.** In Figure 2 we see the relationship between CCor with each of PCCor and PCCorL1. The partial cross-correlation values were
Table 3: ROI-level FC stability, decomposed into the three sources of variation (subject, session, and noise). The estimated variances and the standard deviation of their empirical distribution reported in this table are ×10^5 the estimated value in the analysis. A permutation test was carried out to test for differences in the mean ICC for short-term and long-term stabilities. The p-values for the permutation test to test for differences in the mean short-term and long-term ICCs are as follows: *p < .05 **p < .01.
smaller than their corresponding cross-correlation values, as expected. When we regularized the variance-covariance matrix, the partial cross-correlation values decreased towards zero in magnitude in a nonlinear fashion. We also note that all of the negative cross-correlation values were shrunk towards zero relative to their corresponding partial cross-correlation counterpart, suggesting that in this data set if two ROIs are marginally negatively correlated, then they are conditionally uncorrelated so that the linear dependencies between two negatively correlated ROIs are being driven by the signals in the other ROIs in the analysis.

**The effects of regularization.** A large-scale variance-covariance matrix is ill-conditioned when the sample size (number of volumes) is small and the dimension (number of ROIs) is large. This leads to poor estimates of partial cross-correlation. Looking again at Table 3 we see the effects of the $\ell_1$ regularization on the estimated parameters of the ANOVA model and consequently on the ICC. The numerical instability of the variance-covariance matrix is clear because all sources of variation are much larger compared to the regularized variance-covariance matrices.

**Group-level stability between sessions.** For each of the three sessions we averaged the metrics over all subjects. In Figure 3 and Supplementary Figure 2, we can see that the metrics based on the autoregressive model (DI and PDC) and PCCor yielded the lowest ICCs.

**The relationship between strength of FC and ICC.** In Figure 4 and Supplementary Figure 3 we show the relationship between ICC and the strength of FC. To investigate the relationship between the strength
of FC and ICC, we used a lowess curve to give a smooth estimate of the potentially nonlinear relationship (Cleveland, 1979). Looking at the overlayed lowess curves, the trend we see is that ICC is greater for stronger absolute values of FC. However this trend is not observed for negative values of partial cross-correlation. Tests of significance of this trend for positive and negative values of FC are shown in Supplementary Table 3.

3.2. Voxel-level FC Stability

Voxelwise ICCs. In Figure 5, we show on axial slices the overall ICC per voxel on a seed-based FC analysis, where the PCC was picked as the seed. Long-term and short-term ICCs are shown in Supplementary Figures 4 and 5, respectively. Overall ICCs for the seed-based FC analysis where the seed is in the IPS and SMA are shown in Supplementary Figures 6 and 7, respectively. CCor seemed to be the most stable of the three FC metrics, as there were many null voxels for CCoh, and DI.

We summarized Figure 5 in Table 4, where we identified the peaks in the overall ICC. We drew a sphere with radius 6 mm centered at the coordinates of these peaks and computed the mean overall, long-term, and short-term ICCs. We report only those results whose peak overall ICC exceeded 0.80 (for CCor and CCoh) or 0.65 (for DI); if an ROI had multiple peaks in the overall ICC, we reported only the largest peak. From Figure 5, we can see that the ICC values are not spatially smooth over the volume, and this is more apparent in Table 4 where we see drastic differences between the overall peak ICC and the mean overall ICC. Comparing between the FC metrics, we see that even restricting to only these regions that yielded the highest ICC values per FC metric, CCor had the highest ICC values and DI generally had the lowest ICC values.
Figure 4: A smoothed density scatter plot of the strength of FC against long-term stability for each of the FC metrics. Darker shades of blue denote a greater density of points in that area. A lowess curve is overlayed in dashed red lines, and the zero vertical line is in dashed black lines. For each FC metric, the lowess curve suggests a positive association between stability and absolute strength of FC. The results of a correlation analysis to quantify and test this association is shown in Supplementary Table 3. A smoothed density scatter plot of the strength of FC against short-term stability for each of the FC metrics is shown in Supplementary Figure 3.

Figure 5: Voxelwise overall ICCs for seed-based FC analysis, where PCC was the seed. The ICC values were thresholded below at .001. The long-term and short-term ICCs for this seed-based FC analysis is shown in Supplementary Figures 4 and 5, respectively.
Table 4: Listed are the voxels whose ICC for FC with the PCC seed exceeds 0.80 for CCor and CCoh or exceeds 0.65 for DI. The mean ICCs were computed from the voxels within a sphere of radius 6 mm centered at the given MNI coordinates.

<table>
<thead>
<tr>
<th>FC Metric</th>
<th>ROI</th>
<th>Peak Overall ICC</th>
<th>Mean Overall ICC</th>
<th>Mean Long-term ICC</th>
<th>Mean Short-term ICC</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCor</td>
<td>MTG_L</td>
<td>0.848</td>
<td>0.469</td>
<td>0.454</td>
<td>0.532</td>
<td>20</td>
<td>37</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>MFG_R</td>
<td>0.826</td>
<td>0.488</td>
<td>0.494</td>
<td>0.513</td>
<td>62</td>
<td>87</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>PCUN_R</td>
<td>0.824</td>
<td>0.428</td>
<td>0.433</td>
<td>0.474</td>
<td>50</td>
<td>38</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>MOG_L</td>
<td>0.814</td>
<td>0.377</td>
<td>0.390</td>
<td>0.445</td>
<td>23</td>
<td>25</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>MTG_R</td>
<td>0.808</td>
<td>0.427</td>
<td>0.373</td>
<td>0.541</td>
<td>66</td>
<td>34</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>CUN_L</td>
<td>0.805</td>
<td>0.392</td>
<td>0.314</td>
<td>0.454</td>
<td>45</td>
<td>26</td>
<td>46</td>
</tr>
<tr>
<td>CCoh</td>
<td>SFGdor_R</td>
<td>0.793</td>
<td>0.287</td>
<td>0.286</td>
<td>0.294</td>
<td>56</td>
<td>84</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>MTG_R</td>
<td>0.764</td>
<td>0.291</td>
<td>0.308</td>
<td>0.317</td>
<td>73</td>
<td>53</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>CUN_R</td>
<td>0.752</td>
<td>0.325</td>
<td>0.350</td>
<td>0.302</td>
<td>46</td>
<td>27</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>CUN_L</td>
<td>0.744</td>
<td>0.457</td>
<td>0.477</td>
<td>0.426</td>
<td>45</td>
<td>27</td>
<td>53</td>
</tr>
<tr>
<td>DI</td>
<td>SFGmed_L</td>
<td>0.809</td>
<td>0.074</td>
<td>0.030</td>
<td>0.071</td>
<td>42</td>
<td>95</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>ITG_R</td>
<td>0.786</td>
<td>0.165</td>
<td>0.046</td>
<td>0.199</td>
<td>70</td>
<td>52</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>ORBsupmed_L</td>
<td>0.718</td>
<td>0.285</td>
<td>0.107</td>
<td>0.380</td>
<td>43</td>
<td>90</td>
<td>31</td>
</tr>
</tbody>
</table>

Figure 6: A smoothed density scatter plot showing the group-level long-term stability for each metric, where PCC was the seed. Darker shades of blue denote a greater density of points in that area. The identity line is overlayed in dashed black lines. Group-level short-term stability for each metric is shown in Supplementary Figure 8.

**Group-level stability per voxel.** For each session we averaged each metric over all subjects. In Figure 6 and Supplementary Figure 8 we plotted each of the metrics between sessions to evaluate long-term and short-term stability. CCor and CCoh yielded high ICCs, and DI had the lowest ICC.

**The relationship between strength of FC and ICC in the voxelwise analysis.** In Figure 7 we show the relationship between ICC and the strength of FC. From Figure 5 there were many voxels that had very small or null ICCs, and thus, we omit the voxels whose ICC was less than 10^{-10}. We still see the increasing trend in ICC as the strength in FC increases. We point out that, in CCoh, the behavior in the lowess curves is being driven by the outlying voxels and edge effects from the lowess smoother.

4. Discussion

We compared the stability of the different FC metrics using the public domain resting-state fMRI data [Shehzad et al., 2009]. The FC metrics studied were cross-correlation, partial cross-correlation, cross-coherence, and autoregressive model based metrics. A random-effects ANOVA model was used to quantify stability and the amount of between and within-subject variability in the measurements.
Figure 7: A smoothed density scatter plot showing the strength of FC against long-term stability for each metric, where PCC was the seed. Darker shades of blue denote a greater density of points in that area. From Figure 5, many voxels yielded very small or null ICCs. Thus, voxels whose ICC was less than $10^{-10}$ was not considered in this analysis. A lowess curve is overlayed in dashed red lines, and the zero vertical line is in dashed black lines. The strength of FC against short-term stability for each metric is shown in Supplementary Figure 9.

Regularization of the variance-covariance and spectral matrix. To estimate conditional dependencies using, e.g., partial cross-correlation, partial cross-coherence, or partial mutual information, then it would be efficient to first estimate the inverse of the variance-covariance matrix or the inverse of the spectral density matrix. However, these matrices are numerically unstable, and increasing the number of ROIs in the analysis only exacerbates their numerical instability. The numerical instability is due to multicollinearity present in the data, and with the spectral density matrices, spectral leakage is also another culprit (Medkour et al., 2009; Fiecas et al., 2010). Performing regularization procedures on the matrix shrinks the matrix towards the identity matrix, which consequently improves its numerical stability. As we have shown in the present work, regularization substantially decreased the amount of variability in the estimates, yielded larger ICCs than the unregularized estimate, and even improved the stability in the estimates between sessions.

In practice one would need to properly balance the amount of regularization with the bias of the resulting estimates of the FC metrics. If one regularizes too much (or equivalently via Equation (1), setting the penalty too large), then one will be biasing the estimates of partial cross-correlation towards the null. In other words, the more one regularizes the variance-covariance matrix, the poorer the sensitivity of the FC metric to detecting conditional dependencies. Medkour et al. (2009) and Fiecas et al. (2010) recommended some form of regularization of the covariance and spectral matrices and the use of a data-driven approach for determining how much regularization should be performed. This approach was carried out in the present work.

On the stability and utility of the measured dependencies. There are many factors that can affect the stability of the results, and an overview of many of these factors was given by Bennett & Miller (2010). Our ROI-level analysis suggests that the variability in the measured dependencies is being driven by both the variability between subjects and variability due to noise. Various studies have shown evidence that resting-state FC is a dynamic function of consciousness, and physiological variations in the state of consciousness
could be driving the noise in our results (Greicius et al., 2008; Martuzzi et al., 2010; Waites et al., 2005). We point out that, even though the total amount of variability in the measured dependencies varied across the different FC metrics, we cannot conclude from our test-retest analysis that some FC metrics are more robust than the other metrics to the noise in the data. As pointed out by one of the reviewers of this manuscript, the ICC as a measure of stability only indicates how the FC metrics behave with respect to the noise, and not with respect to the signal; this can be seen in Equation (3). We speculate that one can investigate the behavior of the FC metrics with respect to changes in the signal by, e.g., using the bootstrap to resample the mean time courses per subject within a session to obtain a distribution of each of the estimated FC metrics; the variance of the bootstrapped distribution can describe how the estimated FC metrics behave with respect to perturbations in the data. This, however, is beyond the scope of the present work.

In the present work we saw that CCor was generally more stable than the other types of dependencies. Ombao & van Bellegem (2008) showed that coherency is asymptotically equivalent to the cross-correlation of two band-passed signals. In our analysis, CCor was estimated by cross-correlating two band-passed signals and CCoh was computed by extracting the modulus of the coherency between two unfiltered signals, and so CCor and CCoh were estimating the same type of dependency between the two signals. Thus, we conclude that the present data set favors the time domain estimation approach (CCor) over the frequency domain estimation approach (CCoh) for quantifying the frequency-specific linear association between the signals because CCor was shown to be the more stable metric. Similarly, PCCor, PCCorL1, and PCCoh were estimating the same type of dependency between the two signals. As discussed earlier, when obtaining estimates of conditional dependencies we recommend regularization of estimates of the variance-covariance matrix or the spectral density matrix. Thus, for ROI-level analyses we recommend against the use of PCCor. On the other hand, PCCorL1 and PCCoh seemed to yield identical results in the test-retest analysis. Finally, we cannot compare DI with PDC because they quantify different types of dependencies. These two FC metrics, however, rank low in terms of ICC relative to the other FC metrics.

Even though in terms of stability CCor outperforms the other metrics we investigated in this work, in the context of clinical utility, we do not discourage the use of other FC metrics. Various metrics have been successfully used to discriminate between patients and controls. For example, Lynall et al. (2010) used both cross-coherence and mutual information to discriminate between patients with schizophrenia and healthy controls. Similarly, Salvador et al. (2010) used partial cross-coherence to discriminate between patients with schizophrenia and healthy controls, and Salvador et al. (2005a) used partial cross-correlation to show the differences between FC in healthy subjects with a subject with brainstem ischaemia. Thus, it is of clinical importance to investigate the sensitivity, specificity, and stability of each type of dependency for identifying patient groups.
**Per metric group-level stability.** If we were to perform a two-stage approach for testing for FC, we would first estimate each subject’s FC using some metric and then average the estimates over all of the subjects. We saw that by averaging over the subjects, the stability between sessions of many of the estimated FC metrics was high. However, when we estimated the ICCs, we also saw that the variability in each of the FC metrics per session and per subject was highly driven by noise. By averaging over subjects, we were able to investigate the stability in the estimated dependencies at the population level between sessions. The group-level discordance in PCCor at the ROI level is likely due to the ill-conditioning problem because the group-level partial cross-correlations obtained from the regularized variance-covariance matrices showed more stability between sessions. It is unclear the reason behind the group-level discordance in both DI and PDC. It is possible, though, that accounting for directionality in an FC analysis introduces more uncertainty (more variability) in the estimates, which drive down the ICCs.

In a voxel activation study, Caceres et al. (2009) showed that there exists a weak association between the strength of activation and stability. In particular, they showed that even null voxels can be stable. We observed the same phenomenon in our analysis. The increasing trend in the relationship between the strength of FC with stability is circumstantial evidence that the phenomena related to signal dependency, when they exist between regions of the brain, is more stable than other unrelated phenomena affecting the BOLD signal.

**Limitations of the present study.** Various parameters must be set before proceeding with the FC analysis, and there is no consensus on how these parameters must be set. The preprocessing steps we have taken can impact the results and consequently the stability of the results. In particular, the effects of removing the global signal on resting-state networks have been widely discussed in the literature; see, e.g., Fox et al. (2009) and Murphy et al. (2009) for two contrasting views on the topic. Fox et al. (2009) argued that the global signal has a biological basis and is a nuisance signal that should be removed, whereas Murphy et al. (2009) argued that removing the effects of the global signal will introduce the observed negative correlations in the data. In this work, we removed the global signal prior to the analysis. We then showed that if partial cross-correlations were used to quantify FC, the large negative (marginal) cross-correlations were shrunk towards the null, but the large positive (marginal) cross-correlations remained. Further study is needed to investigate this phenomenon. In our analysis where we did not remove the effects of the global signal, we saw that in some of the metrics, in particular, CCor and CCoh, the stability increased substantially, though the other metrics their ICCs decreased. The distribution of the ICCs for CCoh were substantially affected, with both having more non-zero ICCs. Because the effect of the removal of the global signal is not consistent across the different FC metrics, or even a subset of the metrics (marginal or conditional, directional or non-directional), we cannot make a broad statement on the effects of the global signal on FC analyses, and thus,
our analysis only adds fuel to the debate of the importance of removing the global signal; unfortunately, our results do not favor one position over the other.

The other parameters that must be set in an FC analysis are the parameters associated with each metric themselves, which we discussed earlier. How these parameters can impact the results is still unknown. Moreover, the choice of the seed or the ROIs in the analysis can impact the results. We performed the analysis on the functional ROI parcellation described by Shirer et al. (2012). Even though the relative stability of FC between metrics was qualitatively similar using this parcellation compared to when we used the AAL atlas, the absolute stability of FC within a metric still changed between parcellations. Thus, the choice of ROIs can impact the results of FC analyses. The simulation study by Smith et al. (2011) showed that the estimated dependency structure can be incorrect if the ROIs were not properly constructed. For voxelwise analyses, Shehzad et al. (2009) showed the stability of cross-correlation in a full-brain voxelwise analysis using three different seed voxels. We used the same seed voxel coordinate for the PCC given by Shehzad et al. (2009), which may favor the results of CCor.

To assess the stability of the dependencies between time courses, we used the ICC as estimated by a random-effects ANOVA model. The method-of-moments estimator for estimating ICC that various studies have used can yield negative estimates, which are sometimes set to zero (Shehzad et al., 2009; Zuo et al., 2010; Braun et al., 2012; Wang et al., 2011). With our approach of using a random-effects ANOVA model, one would not obtain negative estimates of ICC. Implementation is easy using readily available software (3dICC_REML.R in AFNI and the lme4 package in R). However, zero estimates of ICC are still possible. This can happen if the likelihood surface is relatively flat, which can be caused by extremely noisy estimates in the estimated dependencies or model misspecification. We can see this problem in some of the results in the voxelwise analysis. A potential solution in that setting is to take into account the underlying spatial smoothness of the data into the likelihood. However, this will not completely solve the problem because this problem was also present in the ROI analysis. One could then consider other types of models for test-retest analysis, but this is beyond the scope of this work.

5. Conclusions

We presented a test-retest analysis of various FC metrics in resting-state fMRI. We investigated the stability of each FC metric and showed how the sources of variability may affect the stability of each metric. Since each metric quantifies a different type of dependency between time courses, we agree with David et al. (2004) and Quiroga et al. (2002) that it would be prudent to use a variety of metrics in order to obtain a more complete picture of the underlying dependency structure in the observed signals. However, for quantifying linear association within a frequency band, our analysis suggests that cross-correlating the signals after applying a band-pass filter yields will yield more stable results compared to computing the cross-
coherence between the unfiltered signals. For quantifying conditional linear associations, we advocate the use of regularization techniques either on the variance-covariance matrix to obtain partial cross-correlation or on the spectral density matrix to obtain partial coherence.

6. Acknowledgements

We thank the Handling Editor and three anonymous referees for their suggestions that have led to an improved paper.

References


