Identifying good practices for detecting inter-regional linear functional connectivity from EEG

Franziska Pellegrini[a,b,*], Arnaud Delorme[c], Vadim Nikulin[d], Stefan Haufe[e,f,a,b,**]

[a] *Charité-Universitätsmedizin Berlin, Charitéplatz 1, Berlin, 10117, Germany*
[b] *Bernstein Center for Computational Neuroscience, Philippstraße 13, Berlin, 10117, Germany*
[c] *Swartz Center for Computational Neuroscience, 9500 Gilman Dr., La Jolla, California, 92903-0559, United States*
[d] *Max Planck Institute for Human Cognitive and Brain Sciences Leipzig, Stephanstraße 1a, Leipzig, 04103, Germany*
[e] *Technische Universität Berlin, Straße des 17. Juni 135, Berlin, 10623, Germany*
[f] *Physikalisch-Technische Bundesanstalt Braunschweig und Berlin, Abbestraße 2–12, Berlin, 10587, Germany*

Abstract

Aggregating voxel-level statistical dependencies between multivariate time series is an important intermediate step when characterising functional connectivity (FC) between larger brain regions. However, there are numerous ways in which voxel-level data can be aggregated into inter-regional FC, and the advantages of each of these approaches are currently unclear.

In this study we generate ground-truth data and compare the performances of various pipelines that estimate directed and undirected linear phase-to-phase FC between regions. We test the ability of several existing and novel FC analysis pipelines to identify the true regions within which connectivity was simulated. We test various inverse modelling algorithms,

*Corresponding author: franziska.pellegrini@charite.de
**Principal corresponding author: haufe@tu-berlin.de

strategies to aggregate time series within regions, and connectivity metrics. Furthermore, we investigate the influence of the number of interactions, the signal-to-noise ratio, the noise mix, the interaction time delay, and the number of active sources per region on the ability of detecting phase-to-phase FC.

Throughout all simulated scenarios, lowest performance is obtained with pipelines involving the absolute value of coherency. Further, the combination of dynamic imaging of coherent sources (DICS) beamforming with directed FC metrics that aggregate information across multiple frequencies leads to unsatisfactory results. Pipeline that show promising results with our simulated pseudo-EEG data involve the following steps: (1) Source projection using the linearly-constrained minimum variance (LCMV) beamformer. (2) Principal component analysis (PCA) using the same fixed number of components within every region. (3) Calculation of the multivariate interaction measure (MIM) for every region pair to assess undirected phase-to-phase FC, or calculation of time-reversed Granger Causality (TRGC) to assess directed phase-to-phase FC. We formulate recommendations based on these results that may increase the validity of future experimental connectivity studies.

We further introduce the free ROIconnect plugin for the EEGLAB toolbox that includes the recommended methods and pipelines that are presented here. We show an exemplary application of the best performing pipeline to the analysis EEG data recorded during motor imagery.
Keywords: Electroencephalography, Inter-regional Functional Connectivity, Simulation, Source Reconstruction, Linearly-constrained Minimum Variance Beamforming, Multivariate Interaction Measure,

1. Introduction

In recent years, the field of functional neuroimaging has seen a shift from the mere localization of brain activity towards assessing interaction patterns between functionally segregated and specialized brain regions (Friston, 2011; Schoffelen and Gross, 2019). Functional connectivity (FC), in contrast to structural connectivity, expresses a statistical dependency between two or more neuronal time series. It has been proposed that FC reflects inter-areal brain communication (Fries, 2015). Moreover, empirical FC estimates have been linked to various cognitive functions (Schoffelen and Gross, 2019) and show pathological alterations in many neurological diseases like Parkinson's Disease, Alzheimer's Disease, and epilepsy (Van Diessen et al., 2015).

Electroencephalography (EEG) and Magnetoencephalography (MEG) are suitable tools for recording neural activity non-invasively with high temporal resolution. Pipelines for analysing inter-regional FC from M/EEG recordings typically consist of a series of processing steps: artifact cleaning, source projection, aggregation of signals within regions of interests (ROIs), and, finally, FC estimation. At each step, researchers can choose between a huge selection of processing methods, where every decision has the potential to crucially affect the final result of an analysis and its interpretation (Wang et al., 2014; Colclough et al., 2016; Mahjoory et al., 2017). This not only complicates the comparison of results from different FC studies, it also raises the question: which pipelines are suitable for reliable source-level FC detection from M/EEG?

In the absence of a robust ground truth on information flow patterns in the human brain, computer simulations are a straightforward way to address such questions (Ewald et al., 2012). Indeed, numerous works have aimed to validate parts or aspects of M/EEG FC methodologies by employing simulated activity. Several studies have focused on assessing the accuracy of different inverse solutions (Grova et al., 2006; Haufe et al., 2008, 2011; Castaño-Candamil et al., 2015; Bradley et al., 2016; Hincapié et al., 2017; Anzolin et al., 2019; Halder et al., 2019; Jaiswal et al., 2020; Hashemi et al., 2021; Allouch et al., 2022). Others have tested the performance of different FC metrics (Astolfi et al., 2007; Silfverhuth et al., 2012; Haufe et al., 2013; Anzolin et al., 2019; Sommariva et al., 2019; Allouch et al., 2022); however, not always on source-reconstructed data exhibiting realistic levels of source leakage.

Many studies aim at aggregating FC within physiologically defined ROIs (Supp et al., 2007; Palva et al., 2010, 2011; Schoffelen et al., 2017; Basti et al., 2020; Idaji et al., 2021). This approach has various advantages. First, it is computationally more tractable (both memory- and time-wise) than the computation of FC between many pairs of individual sources, and it can avoid numerical instabilities for FC metrics that require full-rank signals. Second, interpreting or even visualizing FC between thousands of separate sources is almost impossible. Third, statistical testing is far easier due to a much reduced number of multiple comparisons. And, forth, across-subject statistical analyses are eased by working on a standardized set of regions rather than in individual anatomical spaces lacking a common set of source locations.

There have been various suggestions on how to reduce the signal dimensionality within ROIs. While some approaches focus on selecting one source for each ROI that best represents the activity of all sources in it (Hillebrand et al., 2012; Ghumare et al., 2018; Perinelli et al., 2022), others involve some kind of averaging or weighted averaging over all source time series of a ROI (Palva et al., 2010, 2011; Korhonen et al., 2014). This approach can be made more general by using the strongest principal component (PC) of all sources of a ROI as a representative time series of that ROI (Supp et al., 2007; Hillebrand et al., 2012; Ghumare et al., 2018; Rubega et al., 2019; Basti et al., 2020). The assumption behind this is that the projection of the data that captures the highest amount of variance within a ROI (its strongest PCs) also reflects the connectivity structure of that ROI best. While most works use only the first PC per region, the use of multiple components has also been suggested (e.g. Schoffelen et al., 2017). For this approach, the subsequent FC estimation is usually calculated between pairs of multivariate time series. Another approach, used for example in Schoffelen et al. (2017), is to apply a multivariate FC metric (here, a multivariate extension of Granger causality, Barrett et al., 2010) to the first C PCs of each pair of ROIs. Comparable undirected metrics are the multivariate interaction measure (MIM) and the maximized imaginary coherency (MIC) (Ewald et al., 2012; Basti et al., 2020), which are currently already in use for source-to-source FC estimation (e.g. D'Andrea et al., 2019). These are promising approaches towards more reliable FC estimation. But their virtue in the context of inter-regional FC estimation is still unclear. Moreover, a comprehensive approach evaluating entire data analysis pipelines rather than individual steps is still lacking (see

Mahjoory et al., 2017; Haufe and Ewald, 2019).

Consequently, this work addresses the following questions: First, which pipelines are promising candidates for inferring phase-to-phase FC? Second, which pipelines are promising candidates for inferring the directionality of an interaction? And, most importantly, which pipelines are not suitable to detect FC from data that is corrupted by signal mixing? In addition, we investigate how the number of PCs per ROI affects FC estimation. Finally, we evaluate how the performance of detecting ground-truth interactions varies depending on crucial data parameters like the signal-to-noise ratio (SNR), the number of ground-truth interactions, the noise composition, and the length of the interaction delay. All pipelines are tested within an EEG signal simulation framework that builds on our prior work (Haufe and Ewald, 2019). Note that we focus here on 1:1-phase-to-phase coupling with non-zero time delay, which is the most commonly studied type of FC. Other coupling types including phase–amplitude, amplitude–amplitude, phase–frequency, frequency–frequency, and amplitude–frequency coupling (e.g., Jirsa and Müller, 2013) are not studied here.Further note that we do not intend to propose a realistic model of EEG data or the whole brain. Rather, we aim to identify metrics and pipelines that can accurately reconstruct ROI-level functional connectivity (FC) in the presence of signal mixing, which heavily affects popular metrics used to infer directed and undirected linear FC. That is, we don't address the question of whether networks estimated using FC metrics provide an accurate depiction of actual brain networks.

The best-performing methods and pipelines identified in this study are

implemented in the free ROIconnect plugin for the EEGLAB toolbox. We describe the functionality of ROIconnect and apply it to investigate EEG phase-to-phase FC during left and right hand motor imagery.

2. Methods

2.1. Data generation

We generate time series at a sampling rate of 100 Hz with a recording length of three minutes ($N_t = 100 \cdot 60 \cdot 3 = 18000$ samples). For spectral analyses, we epoch the data into $N_e = 90$ segments of $T = 200$ samples (2 seconds) length.

Ground-truth activity of interacting sources (c.f. Figure 1a) is generated as random white noise filtered in the alpha band (8 to 12 Hz). Throughout, we use zero-phase forward and reverse second-order digital band-pass Butterworth filters. The interaction between two regions is modeled as unidirectional from the sending region to the receiving region. This is ensured by defining the activity at the receiving region to be an exact copy of the activity at the sending region with a certain time delay (see Section 3). Additionally, pink (1/f scaled) background noise is added to the sending and receiving regions independently. More specifically, both the ground-truth signal and the pink background noise are first normalized to have unit-norm in the interacting frequency band. To this end, every interacting ground truth signal time series $\mathbf{g}_x \in \mathbb{R}^{N_t}$ at region x is divided by its ℓ_2-norm: $\mathbf{g}_{x\,\mathrm{n}} = \frac{\mathbf{g}_x}{||\mathbf{g}_x||_{\ell_2}}$. Every pink background noise time series $\mathbf{p}_x \in \mathbb{R}^{N_t}$ is filtered in the interacting frequency band to obtain $\mathbf{p}_{x\,\text{8-12Hz}} \in \mathbb{R}^{N_t}$. The unfiltered noise time series is then divided by the ℓ_2-norm of its filtered version: $\mathbf{p}_{x\,\mathrm{n}} = \frac{\mathbf{p}_x}{||\mathbf{p}_{x\,\text{8-12Hz}}||_{\ell_2}}$.

Subsequently, a weighted sum of the normalized signal time series and the normalized noise time series is calculated:

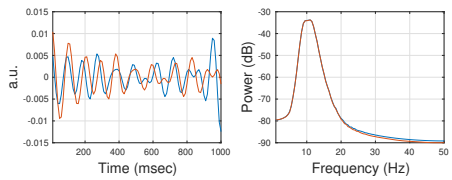
$$\mathbf{s}_x = \theta \mathbf{g}_{x\,\mathrm{n}} + (1-\theta)\mathbf{p}_{x\,\mathrm{n}} \ \in \mathbb{R}^{N_t} \tag{1}$$

The result is called the (interacting) *signal* (Figure 1b). The parameter θ takes values between 0 and 1 and defines the source-level SNR in decibel (dB): $SNR_\theta = 20 * log_{10}(\frac{\theta}{1-\theta})$. The source-level SNR is set to 3.5 dB (θ=0.6). The transposed column vectors of all $2N_I$ signal time series $\mathbf{s}$ form the signal sources $\tilde{\mathbf{J}}_I \in \mathbb{R}^{2N_I \times N_t}$, with N_I region pairs containing the $2N_I$ interacting signals.

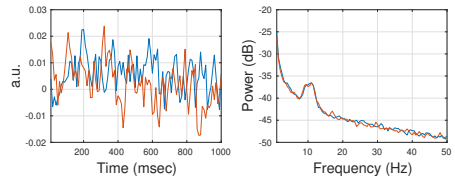
In contrast, activity of a non-interacting source at region y $\mathbf{b}_y \in \mathbb{R}^{N_t}$ – referred to as *brain noise* (Figure 1c) – is generated using random pink noise only without additional activity in the alpha band. The transposed column vectors of all $R - 2N_I$ brain noise time series $\mathbf{b}$ form the brain noise sources $\tilde{\mathbf{J}}_b \in \mathbb{R}^{R-2N_I \times N_t}$, with R denoting the number of regions.

We use a surface-based source model with $N_v = 1895$ dipolar sources placed in the cortical gray matter. Regions are defined according to the Desikan-Killiany atlas (Desikan et al., 2006), which is a surface-based atlas with $R = 68$ cortical regions. Depending on the number of interacting voxels (see Experiment 6, Section 3), one or two time series per region are generated. Every ground-truth time series is placed in a randomly selected source location within a region, so that every region contains the same number of ground-truth time series. The N_I region pairs containing the $2N_I$ interacting signals are chosen randomly, and all other regions contain time series with brain noise.

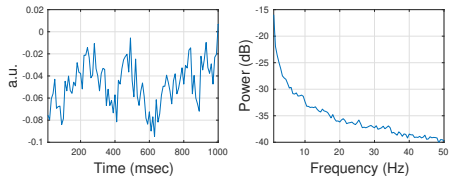
In the next step, source activity is projected to sensor space by using a



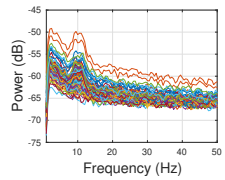
(a) Ground-truth interacting sources.



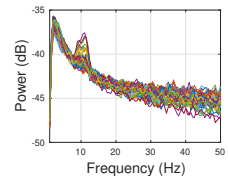
(b) Noisy interacting sources.



(c) Brain noise.



(d) Sensor activity.



(e) Reconstructed source activity.

Figure 1: Example of simulated data in time and frequency domain. (a) Ground-truth activity at two interacting sources was generated as random white noise filtered in the alpha band (8 to 12 Hz). Left: the one-second window of data in the time domain. Right: Power spectral density (PSD). (b) Two interacting signals, generated as a mixture of the ground truth activity and pink background noise ($\text{SNR}_\theta = 3.5$ dB). Left: one-second window of data in the time domain. Right: PSD. (c) Brain noise, generated as random pink noise without additional activity in the alpha band (shown is the activity of an exemplary non-interacting source). Left: one-second window of data in time domain. Right: PSD. (d) PSD of activity at the sensor level is generated by mixing white sensor noise, and the interacting signal, and the brain noise at the sensor level (SNR = 3.5 dB). (e) PSD of reconstructed source-level activity. Shown are PSDs of the first principal component of all 68 regions.

physical forward model of the electrical current flow in the head, summarized by a leadfield matrix. The leadfield describes the signal measured at the sensors for a given source current density. It is a function of the head geometry and the electrical conductivities of different tissues in the head. The template leadfield is obtained from a BEM head model of the ICBM152 anatomical head template, which is a non-linear average of the magnetic resonance (MR) images of 152 healthy subjects (Mazziotta et al., 1995). We use Brainstorm (Tadel et al., 2011) and openMEEG (Gramfort et al., 2010) software to generate the headmodel and leadfield. $N_s = 97$ sensors are placed on the scalp following the standard BrainProducts ActiCap97 channel setup. Note that the spatial orientation of all simulated dipolar sources is chosen to be perpendicular to the cortex surface, so the three spatial orientations that define the dipole orientation of the source activity orientations are summarized into one. This assumption implies a scalar leadfield $\mathbf{L}_\mathrm{s} \in \mathbb{R}^{N_s \times N_v}$. We denote the columns of $\mathbf{L}_\mathrm{s}$ that correspond to the interacting sources by $\mathbf{L}_I \in \mathbb{R}^{N_s \times 2N_I}$ and those corresponding to the brain noise sources by $\mathbf{L}_b \in \mathbb{R}^{N_s \times R-2N_I}$. Signal sources $\tilde{\mathbf{J}}_I$ and brain noise sources $\tilde{\mathbf{J}}_b$ are then separately projected to sensor space:

$$\tilde{\mathbf{Q}}_I = \mathbf{L}_I * \tilde{\mathbf{J}}_I \tag{2}$$

$$\tilde{\mathbf{Q}}_b = \mathbf{L}_b * \tilde{\mathbf{J}}_b, \tag{3}$$

with $\tilde{\mathbf{Q}}_I$ and $\tilde{\mathbf{Q}}_b \in \mathbb{R}^{N_s \times N_t}$.

At sensor level, we mix the different signal and noise components. We generate white sensor noise $\tilde{\mathbf{Q}}_s \in \mathbb{R}^{N_s \times N_t}$ with equal variance at all sensors. The multivariate sensor-space time series corresponding to all three signal components—brain noise, interacting signals, and sensor noise—are divided

by their Frobenius norms with respect to the interacting frequency band (8-12 Hz):

$$\tilde{\mathbf{Q}}_{I\,\mathrm{n}} = \frac{\tilde{\mathbf{Q}}_I}{||\tilde{\mathbf{Q}}_{I\,\text{8-12Hz}}||_{\ell 2}}, \tag{4}$$

$$\tilde{\mathbf{Q}}_{b\,\mathrm{n}} = \frac{\tilde{\mathbf{Q}}_b}{||\tilde{\mathbf{Q}}_{b\,\text{8-12Hz}}||_{\ell 2}}, \tag{5}$$

$$\tilde{\mathbf{Q}}_{s\,\mathrm{n}} = \frac{\tilde{\mathbf{Q}}_s}{||\tilde{\mathbf{Q}}_{s\,\text{8-12Hz}}||_{\ell 2}}, \tag{6}$$

with $\tilde{\mathbf{Q}}_{I\,\text{8-12Hz}}, \tilde{\mathbf{Q}}_{b\,\text{8-12Hz}}$ and $\tilde{\mathbf{Q}}_{s\,\text{8-12Hz}} \in \mathbb{R}^{N_s \times N_t}$. Then the three components are combined as follows: first, we add brain noise and sensor noise with a specific brain noise-to-sensor noise-ratio (BSR) to obtain the total noise $\tilde{\mathbf{Q}}_n$ and normalize it with respect to the interacting frequency band:

$$\tilde{\mathbf{Q}}_n = \theta_{\mathrm{bsr}} \tilde{\mathbf{Q}}_{b\,\mathrm{n}} + (1 - \theta_{\mathrm{bsr}}) \tilde{\mathbf{Q}}_{s\,\mathrm{n}} \tag{7}$$

$$\tilde{\mathbf{Q}}_{n\,\mathrm{n}} = \frac{\tilde{\mathbf{Q}}_n}{||\tilde{\mathbf{Q}}_{n\,\text{8-12Hz}}||_{\ell 2}}. \tag{8}$$

The default BSR value is set to 0 dB, i.e., $\theta_{\mathrm{bsr}} = 0.5$. Second, we sum up signal and total noise with a specific global (sensor-level) SNR:

$$\tilde{\mathbf{Q}} = \theta_{\mathrm{snr}} \tilde{\mathbf{Q}}_{I\,\mathrm{n}} + (1 - \theta_{\mathrm{snr}}) \tilde{\mathbf{Q}}_{n\,\mathrm{n}} \tag{9}$$

$$\tilde{\mathbf{Q}}_{\mathrm{n}} = \frac{\tilde{\mathbf{Q}}}{||\tilde{\mathbf{Q}}_{\text{8-12Hz}}||_{\ell 2}}. \tag{10}$$

The default SNR value is set to 3.5 dB, i.e., $\theta_{\mathrm{snr}} = 0.6$. An example of the power-spectral density of the resulting activity on sensor level is illustrated in Figure 1d. As a last step, we high-pass filter the generated sensor data with a cutoff of 1 Hz.

2.2. Source reconstruction

We test four different inverse solutions for source reconstruction: 'Exact' low-resolution electromagnetic tomography (eLORETA), linearly-constrained minimum variance beamforming (LCMV), dynamic imaging of coherent sources (DICS), and Champagne. Inverse source reconstructions are based on the same leadfield used to simulate the signals. Full 3D currents are estimated for each source dipole. That is, prior information about the dipoles' orientation is not used. A normal direction could in principle be estimated from the reconstructed cortical surface mesh (which we used here for signal generation); however, such estimation is considered to be rather unstable, since we do not have a good estimate of the cortical surface orientation in practice. The aggregation of the three spatial dimensions is discussed in Section 2.3.

'Exact' low-resolution electromagnetic tomography

The starting point to solve the source localization problem is the linear forward model $\tilde{\mathbf{Q}} = \mathbf{L}_{\mathrm{v}}\tilde{\mathbf{J}}$, where $\tilde{\mathbf{Q}} \in \mathbb{R}^{N_s \times N_t}$ stands for the sensor measurements, $\tilde{\mathbf{J}} \in \mathbb{R}^{3N_v \times N_t}$ is the vector-valued activity of the dipolar brain sources to be recovered, and $\mathbf{L}_{\mathrm{v}} \in \mathbb{R}^{N_s \times 3N_v}$ is the vector-valued linear leadfield matrix that maps the electrical activity from sources to sensor level. Here, $3N_v$ stand for the three spatial dimensions that together define the dipole orientation of the source activity. The solution of this equation is ill-posed since the number of brain sources N_v is much smaller than the number of measurement sensors N_s. Therefore eLORETA imposes the constraint of spatially smooth current density distributions (Pascual-Marqui, 2007; Pascual-Marqui et al., 2011). Briefly, eLORETA uses a weighted minimum norm criterion to

estimate the source distribution:

$$\hat{\mathbf{J}} = \arg\min_{\tilde{\mathbf{J}}} \left[||\tilde{\mathbf{Q}} - \mathbf{L}_{\mathrm{v}}\tilde{\mathbf{J}}||^2 + a\tilde{\mathbf{J}}^\top \mathbf{W}\tilde{\mathbf{J}} \right], \tag{11}$$

where $a \geq 0$ denotes a regularization parameter, and $\mathbf{W}$ is a block-diagonal symmetric weight matrix:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{W}_{N_v} \end{bmatrix} \in \mathbb{R}^{3N_v \times 3N_v}, \tag{12}$$

where $\mathbf{0}$ is the 3×3 zero matrix and $\mathbf{W}_v$ the 3×3 weight matrix at the v-th voxel defined in Equation (15). The solution of Equation (11) is given by:

$$\hat{\mathbf{J}} = \mathbf{W}^{-1}\mathbf{L}_{\mathrm{v}}^\top(\mathbf{L}\mathbf{W}^{-1}\mathbf{L}_{\mathrm{v}}^\top + a\mathbf{K})^\dagger\tilde{\mathbf{Q}} = \mathbf{P}^{\mathrm{E}^\top}\tilde{\mathbf{Q}}, \tag{13}$$

where $\mathbf{K} \in \mathbb{R}^{N_s \times N_s}$ is a centering matrix re-referencing the leadfield and sensor measurements to the common-average reference, $A^\dagger$ is the Moore-Penrose pseudo-inverse of a matrix A, and $\mathbf{P}^{\mathrm{E}} \in \mathbb{R}^{N_s \times 3N_v}$ is the eLORETA inverse filter. eLORETA then first computes

$$\mathbf{M} = (\mathbf{L}_{\mathrm{v}}\mathbf{W}^{-1}\mathbf{L}_{\mathrm{v}}^\top + a\mathbf{K})^\dagger \tag{14}$$

and then for $v = 1, ..., N_v$, calculates weights

$$\mathbf{W}_v = [\mathbf{L}_{\mathrm{v}v}^\top \mathbf{M}\mathbf{L}_{\mathrm{v}v}]^{1/2}, \tag{15}$$

with $\mathbf{L}_{\mathrm{v}v} \in \mathbb{R}^{N_s \times 3}$ denoting the leadfield for a single source location. It then iterates Equation (14) and (15) until convergence and use the final weights to calculate $\hat{\mathbf{J}}$. eLORETA has been shown to outperform other linear solutions

in localization precision (Pascual-Marqui, 2007; Halder et al., 2019; Allouch et al., 2022).

In this study, we choose the regularization parameter based on the best result in a five-fold spatial cross-validation (Hashemi et al., 2021) with fifteen candidate parameters taken from a logarithmically spaced range between $0.01 * \mathrm{Tr}(\mathbf{Cov}_{\tilde{\mathbf{Q}}})$ and $\mathrm{Tr}(\mathbf{Cov}_{\tilde{\mathbf{Q}}})$, where $\mathrm{Tr}(A)$ denotes the trace of a matrix A and $\mathbf{Cov}_{\tilde{\mathbf{Q}}} \in \mathbb{C}^{N_s \times N_s}$ denotes the sample covariance matrix of the sensor-space data.

Linearly-constrained minimum variance beamforming

The LCMV (Van Veen et al., 1997) filter $\mathbf{P}^{\mathrm{L}} \in \mathbb{R}^{N_s \times 3N_v}$ belongs to the class of beamformers. It estimates source activity separately for every source location. While LCMV maximizes source activity originating from the target location, it suppresses noise and other source contributions. Let $\mathbf{L}_{\mathrm{v}v} \in \mathbb{R}^{N_s \times 3}$ and $\mathbf{P}_v^{\mathrm{L}} \in \mathbb{R}^{N_s \times 3}$ denote the leadfield and projection matrix for a single source location, respectively. The LCMV projection filter minimizes the total variance of the source-projected signal across the three dipole dimensions:

$$\mathbf{P}_v^{\mathrm{L}} = \arg\min_{\mathbf{P}_v} \mathrm{Tr}({\mathbf{P}_v}^\top \mathbf{Cov}_{\tilde{\mathbf{Q}}} \mathbf{P}_v) \tag{16}$$

under the unit-gain constraint

$${\mathbf{P}_v}^\top \mathbf{L}_{\mathrm{v}v} \;=\; \mathbf{I}_{3\times 3}\ . \tag{17}$$

The source estimate $\hat{\mathbf{J}}_v \in \mathbb{R}^{3 \times N_t}$ at the v-th voxel is given by

$$\hat{\mathbf{J}}_v = \left[({\mathbf{L}_{\mathrm{v}v}}^\top {\mathbf{Cov}_{\tilde{\mathbf{Q}}}}^{-1} \mathbf{L}_{\mathrm{v}v})^{-1} {\mathbf{L}_{\mathrm{v}v}}^\top {\mathbf{Cov}_{\tilde{\mathbf{Q}}}}^{-1} \right] \tilde{\mathbf{Q}} = {\mathbf{P}_v^{\mathrm{L}}}^\top \tilde{\mathbf{Q}}. \tag{18}$$

Previous simulations indicated that LCMV overall shows a higher connectivity reconstruction accuracy than eLORETA but is more strongly affected

by low SNR (Anzolin et al., 2019). We show a power spectrum of exemplary LCMV-reconstructed source activity in Figure 1e.

Dynamic imaging of coherent sources

DICS (Gross et al., 2001) is the frequency-domain equivalent of LCMV. In contrast to LCMV, DICS estimates spatial filters separately for each spectral frequency. The DICS filter $\mathbf{P}^{\mathrm{D}}$ is evaluated for a given frequency f using the real part of the sensor-level cross-spectral density matrix $\mathbf{S}_{\mathbf{Q}}$:

$$\mathbf{P}_v^{\mathrm{D}}(f) = \left(\mathbf{L}_{\mathrm{v}v}^{\top}\mathbf{S}_{\mathbf{Q}}(f)^{-1}\mathbf{L}_{\mathrm{v}v}\right)^{-1}\mathbf{L}_{\mathrm{v}v}^{\top}\mathbf{S}_{\mathbf{Q}}(f)^{-1}. \tag{19}$$

with

$$\mathbf{S}_{\mathbf{Q}}(f) = \left\langle \mathbf{q}(f,e)\mathbf{q}^{*}(f,e)\right\rangle_e \in \mathbb{C}^{N_s \times N_s}, \tag{20}$$

where $(\cdot)^*$ denotes complex conjugation and $\mathbf{q}(f,e)$ denotes the Fourier transform of the sensor measurements $\tilde{\mathbf{q}}(t,e)$. That is, the time-domain sensor signal $\tilde{\mathbf{Q}}$ is cut into N_c epochs of T time samples to derive $\tilde{\mathbf{q}}(t,e)$, then multiplied with a Hanning window of length T, and Fourier-transformed epoch by epoch to derive $\mathbf{q}(f,e)$.

The beamformer filter $\mathbf{P}^{\mathrm{D}}(f) = [\mathbf{P}_1^{\mathrm{D}}(f), \ldots, \mathbf{P}_{N_v}^{\mathrm{D}}(f)]$ can then be used to project the sensor cross-spectrum to source space:

$$\mathbf{S}_{\mathbf{J}}(f) = \mathbf{P}^{\mathrm{D}^{\top}}(f)\mathbf{S}_{\mathbf{Q}}(f)\mathbf{P}^{\mathrm{D}}(f) \in \mathbb{C}^{3N_v \times 3N_v} \tag{21}$$

Based on previous literature described above, we hypothesize that the beamformer solutions (LCMV and DICS) perform better than eLORETA when used in combination with undirected FC measures. However, since directed FC measures need to aggregate information across frequencies, we

hypothesize that the estimation of such measures might be negatively affected by DICS source reconstruction. Concretely, we expect that DICS' ability to optimize SNR per frequency and, thereby, to reconstruct different sources for each frequency can be counterproductive in cases where in fact the same pairs of sources are interacting at multiple frequencies. In contrast, we expect that LCMV, which reconstructs a single set of sources by optimizing the SNR across the whole frequency spectrum, would yield more consistent source cross-spectra and, therefore, better directed FC estimates than DICS.

Champagne

Champagne (Wipf et al., 2010) uses hierarchical sparse Bayesian inference for inverse modelling. Specifically, it imposes a zero-mean Gaussian prior independently for each source voxel. The prior source covariance is given by

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Gamma}_{N_v} \end{bmatrix} \in \mathbb{R}^{3N_v \times 3N_v} , \tag{22}$$

where $\mathbf{\Gamma}_v$ is the 3×3 covariance of the v-th voxel. Here we use a Champagne variant that models each $\mathbf{\Gamma}_v$ as a full positive-definite matrix

$$\mathbf{\Gamma}_v = \begin{bmatrix} \gamma_{v,1} & \gamma_{v,4} & \gamma_{v,5} \\ \gamma_{v,4} & \gamma_{v,2} & \gamma_{v,6} \\ \gamma_{v,5} & \gamma_{v,6} & \gamma_{v,3} \end{bmatrix} \tag{23}$$

with six parameters. The prior source variances and covariances in $\mathbf{\Gamma}$ are treated as model hyperparameters and are optimized in an iterative way.

For any given choice of $\mathbf{\Gamma}$, the posterior distribution of the source activity is given by (Wipf et al., 2010):

$$p(\tilde{\mathbf{J}}|\tilde{\mathbf{Q}}, \gamma) = \prod_{t=1}^{N_t} \mathcal{N}(\hat{\mathbf{j}}(t), \mathbf{\Sigma_j}) \ , \quad \text{where} \tag{24}$$

$$\hat{\mathbf{j}}(t) = \mathbf{\Gamma}\mathbf{L}_{\mathrm{v}}^{\top}(\mathbf{\Sigma_q})^{-1}\tilde{\mathbf{q}}(t) = \mathbf{P}^{\mathrm{C}}\tilde{\mathbf{q}}(t) \tag{25}$$

$$\mathbf{\Sigma_j} = \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{L}_{\mathrm{v}}^{\top}(\mathbf{\Sigma_q})^{-1}\mathbf{L}_{\mathrm{v}}\mathbf{\Gamma} \tag{26}$$

$$\mathbf{\Sigma_q} = \sigma^2\mathbf{I} + \mathbf{L}_{\mathrm{v}}\mathbf{\Gamma}\mathbf{L}_{\mathrm{v}}^{\top} \ , \tag{27}$$

and where σ^2 denotes a homoscedastic sensor noise variance parameter. The posterior parameters $\hat{\mathbf{j}}(t)$ and $\mathbf{\Sigma_j}$ are then used to obtain the next estimate of γ by minimizing the negative log model evidence (Bayesian Type-II likelihood):

$$\mathcal{L}^{II}(\gamma) = -\log\, p(\tilde{\mathbf{Q}}|\gamma) = \frac{1}{N_t}\sum_{t=1}^{N_t} \tilde{\mathbf{q}}(t)^{\top}\mathbf{\Sigma_q}^{-1}\tilde{\mathbf{q}}(t) + \log||\mathbf{\Sigma_q}|| \ . \tag{28}$$

This process is repeated until convergence. Importantly, the majority of source variance parameters converges to zero in the course of the optimization, so that the reconstructed source distribution becomes sparse.

In the original Champagne version, a baseline or control measurement is used to estimate noise covariance in sensor data. Since baseline data are not available in our study, we use a homoscedastic noise model in which all sensors are assumed to be perturbed by uncorrelated Gaussian white noise with equal variance, and estimate the shared variance parameter using five-fold spatial cross-validation (Hashemi et al., 2021). Again, fifteen candidate parameters

are taken from a logarithmically spaced range between $0.01 * \mathrm{Tr}(\mathbf{Cov}_{\tilde{\mathbf{Q}}})$ and $\mathrm{Tr}(\mathbf{Cov}_{\tilde{\mathbf{Q}}})$.

2.3. Dimensionality reduction

To aggregate time series of multiple sources within a region, an intuitive approach would be to take the mean across sources within each spatial dimension. However, this approach has two disadvantages: First, it assumes a high homogeneity within all voxels of a pre-defined region, which is not always given. Second, it does not offer a solution for aggregating the three spatial dimensions, since averaging across these might lead to cancellations due to different polarities.

Principal component analysis

An alternative approach is to reduce the dimensionality of multiple time series by employing a singular value decomposition (SVD) or, equivalently, principal component analysis (PCA), and to subsequently only select the C strongest PCs accounting for most of the variance within a region for further processing. Let $\tilde{\mathbf{J}}_{\mathrm{r}} \in \mathbb{R}^{N_t \times 3R}$ denote the reconstructed broad-band source time courses of R dipolar sources within a single region r after mean subtraction. The covariance matrix $\mathbf{Cov}_{\mathrm{r}} = {\tilde{\mathbf{J}}_{\mathrm{r}}^{\top} \tilde{\mathbf{J}}_{\mathrm{r}}}/{N-1} \in \mathbb{R}^{3R \times 3R}$ is a symmetric matrix that can be diagonalized as

$$\mathbf{Cov}_{\mathrm{r}} = \mathbf{V}\mathbf{B}\mathbf{V}^{\top}, \tag{29}$$

where $\mathbf{B} \in \mathbb{R}^{3R \times 3R}$ is a diagonal matrix containing the eigenvalues λ_v (variances) of the PCs, which are, without loss of generality, assumed to be given

in descending order, and $\mathbf{V} \in \mathbb{R}^{3R \times 3R}$ is a matrix of corresponding eigenvectors in which each column contains one eigenvector. The jth PC can then be found in the jth column of $\tilde{\mathbf{J}}_{\mathrm{r}}\mathbf{V}$.

In practice, the PCs are calculated using an SVD of the zero-mean data matrix $\tilde{\mathbf{J}}_{\mathrm{r}}$ as

$$\tilde{\mathbf{J}}_{\mathrm{r}} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top} \ . \tag{30}$$

Using the 'economy version' of the SVD, $\mathbf{U} \in \mathbb{R}^{N_t \times 3R}$ is a matrix of orthonormal PC time courses, $\mathbf{D} \in \mathbb{R}^{3R \times 3R}$ is a matrix of corresponding singular values, and $\mathbf{V} \in \mathbb{R}^{3R \times 3R}$ is the matrix of eigenvectors (or, equivalently, singular vectors) defined above. Note that the square of the elements of $\mathbf{D}$, divided by $N_t - 1$ are identical to the variances of the corresponding PCs (eigenvalues of $\mathbf{Cov}_{\mathrm{r}}$). Each squared singular vector, normalized by the sum of all singular vectors, thus corresponds to the variance explained by the corresponding singular vector. We will use this property for the two VARPC pipelines (Section 2.5).

Comparing PCA and SVD, one can easily see that

$$\mathbf{Cov}_{\mathrm{r}} = \mathbf{V}\mathbf{D}\mathbf{U}^{\top}\mathbf{U}\mathbf{D}\mathbf{V}^{\top} = \mathbf{V}\frac{\mathbf{D}^2}{N-1}\mathbf{V}^{\top}, \tag{31}$$

and $\lambda_v = \frac{d_v^2}{N-1}$. Thus, the PCs can also be calculated with SVD:

$$\tilde{\mathbf{J}}_{\mathrm{r}}\mathbf{V} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top}\mathbf{V} = \mathbf{U}\mathbf{D} \ . \tag{32}$$

To reduce the dimensionality of the voxel data within one region, we keep only the strongest C PCs, i.e., the columns of $\mathbf{UD}$ that correspond to the largest eigenvalues. For a more extensive overview of the relationship between SVD and PCA, we refer to Wall et al. (2003). Note that in this study, we

applied SVD on the time-domain source signals $\tilde{\mathbf{J}}_r$ for most of the pipelines. However, we applied PCA on the real part of the source-level cross-spectrum, summed across frequencies, in case of DICS. For the ease of reading, we will stick to PCA terminology for all pipelines in the following.

It has been popular in the literature (Friston et al., 2006; Basti et al., 2020) to select only the first PC for every region and subsequently employ a univariate FC measure for further processing. We describe this approach further in Section 2.5, pipeline FIXPC1.

2.4. Connectivity metrics

There are numerous approaches to estimate FC (Schoffelen and Gross, 2019). One key distinction can be made between FC metrics that measure undirected (symmetric) interactions between signals and those that also measure the direction of FC.

It has been shown that the estimation of both undirected and directed FC from M/EEG recordings is complicated by the presence of mixed noise and signal sources (Nolte et al., 2004; Haufe et al., 2013; Bastos and Schoffelen, 2016; Wang et al., 2018; Schaworonkow and Nikulin, 2021). Due to volume conduction in the brain, signal sources from all parts of the brain superimpose at each M/EEG sensor. Projecting the sensor signals to source space can help disentangling separate signal sources. However, a signal reconstructed at a specific source voxel may still contain contributions from other sources in its vicinity. This phenomenon is called source leakage (Schoffelen and Gross, 2009).

Volume conduction and source leakage can lead to spurious FC despite the absence of genuine interactions (Nolte et al., 2004; Haufe et al., 2013).

To overcome this problem, *robust* FC metrics have been developed (Nolte et al., 2004, 2008; Haufe et al., 2013; Winkler et al., 2016). Robustness is here referred to as the property of an FC measure to converge to zero in the limit of infinite data when the observed data are just instantaneous mixtures of independent sources (Nolte et al., 2004). Robust FC metrics use that spurious interactions due to signal mixing are instantaneous, while physiological interactions impose a small time delay. Robust FC metrics are therefore only sensitive to statistical dependencies with a non-zero time delay while eliminating zero-delay contributions.

We here test six different FC measures, four to detect undirected FC (coherence, iCOH, MIC, and MIM), and two measures that estimate the direction of interaction between two sources (multivariate GC and TRGC). This selection includes four robust FC metrics (c.f. Section 1) and two non-robust ones (coherence and GC). Based on the literature described above, we hypothesize that robust metrics will perform better than non-robust metrics. Please note that all tested FC metrics are frequency-resolved. That is, all metrics output an $N_{roi} \times N_{roi} \times N_{freq}$ tensor that contains the estimated FC for all region pairs at all frequencies. However, since we expect the interaction to be located in the interacting frequency band between 8 and 12 Hz (see Section 2.1), we select only those frequency bins within this band and average the FC scores across them. As a result, we obtain an $N_{roi} \times N_{roi}$ matrix.

All tested FC metrics are derived from the cross-spectrum. Let $\tilde{\mathbf{x}}(t, e) \in \mathbb{R}^K$ and $\tilde{\mathbf{y}}(t, e) \in \mathbb{R}^L$ be two multivariate time series where $t \in \{1, \ldots, T\}$ indexes samples within epochs of 2 seconds length and e indexes epochs. Often, $K = L = 3$ represents the three dipole orientations of two reconstructed

current sources. In other cases, K and L denotes the number of retained data dimensions of two brain regions after (e.g., PCA) dimensionality reduction. These time-domain data are then multiplied with a Hanning window and Fourier transformed into $\mathbf{x}(f, e)$ and $\mathbf{y}(f, e)$, where $f \in \{0, 0.5, \ldots, 50\}$ indexes frequencies. The joint cross-spectrum is then computed from the Fourier transformed data as

$$\mathbf{S}_{[\mathbf{xy}]}(f) = \begin{bmatrix} \mathbf{S}_{\mathbf{xx}}(f) & \mathbf{S}_{\mathbf{xy}}(f) \\ \mathbf{S}_{\mathbf{yx}}(f) & \mathbf{S}_{\mathbf{yy}}(f) \end{bmatrix} \in \mathbb{C}^{(K+L)\times(K+L)} , \tag{33}$$

where $\mathbf{S}_{\mathbf{xy}} = \langle \mathbf{x}(f, e)\mathbf{y}^*(f, e)\rangle_e \in \mathbb{C}^{K\times L}$.

Coherence and imaginary part of coherency

(Absolute) coherence (COH) and iCOH are measures of the synchronicity of two time series. Both coherence and iCOH are derived from the complex-valued coherency, which is a generalization of correlation in the frequency domain. As such, coherency quantifies the linear relationship between two time series at a specific frequency. Its phase expresses the average phase difference between the two time series, whereas its absolute value expresses the stability of the phase difference.

Complex-valued coherency $\mathbf{C}_{\mathbf{xy}} \in \mathbb{C}^{K\times L}$ is the normalized cross spectrum (Nunez et al., 1997):

$$\mathbf{C}_{\mathbf{xy}}(f) = \frac{\mathbf{S}_{\mathbf{xy}}(f)}{\left(\mathbf{S}_{\mathbf{xx}}(f)\mathbf{S}_{\mathbf{yy}}(f)\right)^{1/2}} . \tag{34}$$

Based on the terminology of Nolte et al. (2004), we define *coherence* as the absolute part of coherency: $\mathbf{COH}_{\mathbf{xy}}(f) = \left|\mathbf{C}_{\mathbf{xy}}(f)\right| \in \mathbb{R}^{K\times L}$, where $|\cdot|$ denotes the absolute value. Coherence captures both zero-delay and non-zero-delay synchronization between two time series. This can be problematic

in the context of M/EEG measurements, where substantial zero-delay synchronization can be introduced by signal spread due to volume conduction or source leakage in absence of genuine interactions between distinct brain areas (Nolte et al., 2004). In contrast, the imaginary part of coherency is a robust FC measure since it is only non-zero for interactions with a phase delay different from multiples of π (Nolte et al., 2004). Here, we use the absolute value of the imaginary part of coherency, $\mathbf{iCOH_{xy}}(f) = \left|\mathbf{C}^{\mathfrak{I}}_{\mathbf{xy}}(f)\right| \in \mathbb{C}^{K\times L}$, as a measure of synchronization strength, where $\mathbf{C}^{\mathfrak{I}}$ denotes the imaginary part of $\mathbf{C}$.

Note that both coherence and iCOH are not designed to aggregate FC between two multivariate time series into one FC score. A single FC score can be obtained by taking the average across all elements of $\mathbf{COH_{xy}}$ or $\mathbf{iCOH_{xy}}$, respectively.

Multivariate interaction measure and maximized imaginary coherency

The multivariate interaction measure (MIM) and maximized imaginary coherency (MIC, Ewald et al., 2012) are multivariate generalizations of iCOH and are therefore also robust against source leakage.

MIM is defined as follows:

$$\mathbf{MIM_{xy}}(f) = \mathrm{Tr}\left[\left(\mathbf{C}^{\mathfrak{R}}_{\mathbf{xx}}(f)\right)^{-1}\mathbf{C}^{\mathfrak{I}}_{\mathbf{xy}}(f)\left(\mathbf{C}^{\mathfrak{R}}_{\mathbf{yy}}(f)\right)^{-1}\left(\mathbf{C}^{\mathfrak{I}}_{\mathbf{xy}}(f)\right)^{\top}\right], \qquad (35)$$

where $\mathbf{C}^{\mathfrak{R}}$ denotes the real part of $\mathbf{C}$. In contrast, MIC aims at maximizing iCOH between the two multivariate time series. That is, MIC finds projections from two multi-dimensional spaces to two one-dimensional spaces such

that iCOH between the projected signals becomes maximal:

$$\mathbf{MIC}_{\mathbf{xy}}(f) \;=\; \max_{\mathbf{a},\mathbf{b}} \left(\frac{\mathbf{a}^{\top} \tilde{\mathbf{S}}^{\mathfrak{I}}_{\mathbf{xy}}(f) \mathbf{b}}{|\mathbf{a}|\,|\mathbf{b}|} \right) , \tag{36}$$

where $\tilde{\mathbf{S}}$ is a whitened version of the cross-spectrum $\mathbf{S}$ (Ewald et al., 2012), and where $\mathbf{a} \in \mathbb{R}^{K\times 1}$ and $\mathbf{b} \in \mathbb{R}^{L\times 1}$ are projection weight vectors corresponding to the subspaces, or regions, of $\mathbf{x}$ and $\mathbf{y}$, respectively. Note that, while the imaginary part itself can be positive or negative, flipping the sign of either $\mathbf{a}$ or $\mathbf{b}$ will also flip the sign of the imaginary part. Thus, without loss of generality, maximization of Eq. (36) will find the imaginary part with strongest magnitude.

All undirected FC metrics (COH, iCOH, MIC, and MIM) are bounded between 0 and 1.

Multivariate Granger causality and time-reversed Granger causality

Granger Causality (GC) defines directed interactions between time series using a predictability argument (Granger, 1969; Bressler and Seth, 2011). Considering two univariate time series $\tilde{x}(t)$ and $\tilde{y}(t)$, we say that $\tilde{y}$ Granger-causes $\tilde{x}$ if the past information of $\tilde{y}$ improves the prediction of the presence of $\tilde{x}$ above and beyond what we could predict by the past of $\tilde{x}$ alone. That is, GC does not only assess the existence of a connection but also estimates the direction of that connection. We here use a spectrally resolved multivariate extension of GC (Geweke, 1982; Barrett et al., 2010; Barnett and Seth, 2014), which allows us to estimate Granger-causal influences between groups of variables at individual frequencies. There are multiple strategies to arrive at spectral Granger causality estimates. Here, we follow recommendations made in Barnett and Seth (2014, 2015); Faes et al. (2017); Barnett et al. (2018)

that ensure stable and unbiased estimates, and use Matlab code provided by the respective authors.

We first transform the joint cross-spectrum into an autocovariance sequence $\mathbf{G}_{[\mathbf{xy}]}(p) \in \mathbb{R}^{(K+L)\times(K+L)}$ with lags $p \in \{0, 1, \ldots, N_P\}$, $N_P = 20$, using the inverse Fourier transform. The autocovariance spectrum is further used to estimate the parameters $\mathbf{A}(p) \in \mathbb{R}^{(K+L)\times(K+L)}, p \in \{1, \ldots, N_P\}$ and $\mathbf{\Sigma} = \mathrm{Cov}_t\left[\boldsymbol{\epsilon}(t)\right] \in \mathbb{R}^{(K+L)\times(K+L)}$ of a linear autoregressive model

$$\begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix} = \sum_{p=1}^{N_P} \mathbf{A}(p) \begin{bmatrix} \mathbf{x}(t-p) \\ \mathbf{y}(t-p) \end{bmatrix} + \boldsymbol{\epsilon}(t) \tag{37}$$

of order N_P using Whittle's algorithm (Whittle, 1963; Barnett and Seth, 2014). Autoregressive model parameters are next converted into a state-space representation $(\bar{\mathbf{A}}, \bar{\mathbf{C}}, \bar{\mathbf{K}}, \bar{\Sigma})$ corresponding to the model

$$\mathbf{z}(t) = \bar{\mathbf{A}}\mathbf{z}(t) + \bar{\mathbf{K}}\boldsymbol{\varepsilon}(t) \tag{38}$$

$$\begin{bmatrix} \bar{\mathbf{x}}(t) \\ \bar{\mathbf{y}}(t) \end{bmatrix} = \bar{\mathbf{C}}\mathbf{z}(t) + \boldsymbol{\varepsilon}(t)\ , \tag{39}$$

using the method of Aoki and Havenner (1991), where $\bar{\mathbf{x}}(t) = [\bar{\mathbf{x}}^\top(t), \bar{\mathbf{x}}^\top(t-1), \ldots, \bar{\mathbf{x}}^\top(t-N_P)]^\top$ and $\bar{\mathbf{y}}(t) = [\bar{\mathbf{y}}^\top(t), \bar{\mathbf{y}}^\top(t-1), \ldots, \bar{\mathbf{y}}^\top(t-N_P)]^\top$ are temporal embeddings of order N_P, $\mathbf{z}(t) \in \mathbb{R}^{(K+L)N_P}$ and $\boldsymbol{\varepsilon}(t) \in \mathbb{R}^{(K+L)N_P}$ are unobserved variables, and all parameters are $(K+L)N_P \times (K+L)N_P$ matrices. Subsequently, the transfer function $\mathbf{H}(z) \equiv \mathbf{I} - \bar{\mathbf{C}}(\mathbf{I} - \bar{\mathbf{A}}z)^{-1}\bar{\mathbf{K}}z \in \mathbb{C}^{(K+L)N_P\times(K+L)N_P}$ of a moving-average representation

$$\begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix} = \mathbf{H}(z) \cdot \boldsymbol{\varepsilon}(t) \tag{40}$$

of the observations is derived, where $\mathbf{I} \in \mathbb{R}^{(K+L)N_P \times (K+L)N_P}$ denotes the identity matrix and where $z = e^{-i4\pi f/T}$ for a vector of frequencies $f \in \{0 \text{ Hz}, 0.5 \text{ Hz}, \ldots, 50 \text{ Hz}\}, T = 200$, and a factorization of the joint cross-spectrum is obtained as $\mathbf{S}_{[\mathbf{xy}]}(f) = \mathbf{H}(f)\bar{\Sigma}\mathbf{H}^*(f)$ (Barnett and Seth, 2015). Frequency-dependent *Granger scores*

$$\mathcal{F}_{\mathbf{x}\to\mathbf{y}}(f) = \log \frac{||\mathbf{S}_{\mathbf{yy}}(f)||}{||\mathbf{S}_{\mathbf{yy}}(f) - \mathbf{H}_{\mathbf{yx}}(f)\bar{\Sigma}_{\mathbf{xx}|\mathbf{y}}\mathbf{H}^*_{\mathbf{yx}}(f)||} \tag{41}$$

and (analogously) $\mathcal{F}_{\mathbf{y}\to\mathbf{x}}(f)$ are then calculated, where $\mathbf{H}(f)$ and $\bar{\Sigma}$ are partitioned in the same way as $\mathbf{S}(f)$, where $\bar{\Sigma}_{\mathbf{xx}|\mathbf{y}} \equiv \bar{\Sigma}_{\mathbf{xx}} - \bar{\Sigma}_{\mathbf{xy}}\bar{\Sigma}^{-1}_{\mathbf{yy}}\bar{\Sigma}_{\mathbf{yx}}$ denotes a partial covariance matrix, and where $|| \cdot ||$ denotes matrix determinant (Barnett and Seth, 2015). Finally, differences

$$\mathcal{F}^{\text{net}}_{\mathbf{x}\to\mathbf{y}}(f) \equiv \mathcal{F}_{\mathbf{x}\to\mathbf{y}}(f) - \mathcal{F}_{\mathbf{y}\to\mathbf{x}}(f) \tag{42}$$

and $\mathcal{F}^{\text{net}}_{\mathbf{y}\to\mathbf{x}}(f) = -\mathcal{F}^{\text{net}}_{\mathbf{x}\to\mathbf{y}}(f)$ summarizing the net information flow between the multivariate time series $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}(t)$ are calculated (Winkler et al., 2016).

Just like coherence, GC is not robust, i.e. can deliver spurious results for mixtures of independent sources as a result of volume conduction or source leakage (e.g., Haufe et al., 2012, 2013). This can be easily acknowledged by considering a single source that spreads into two measurement channels, which are superimposed by distinct noise terms. In that case, both channels will mutually improve each other's prediction in the sense of GC (Haufe and Ewald, 2019). This problem is overcome by a robust version of GC, time-reversed GC (TRGC), which introduces a test on the temporal order of the time series. That is, TRGC estimates the directed information flow once on the original time series and once on a time-reversed version of the time series.

If GC is reduced or even reversed when the temporal order of the time series is reversed, it is likely that the effect is not an artifact coming from volume conduction (Haufe et al., 2012, 2013; Vinck et al., 2015; Winkler et al., 2016). Formally, multivariate spectral GC as introduced above can be evaluated on the time-reversed data by fitting the autoregressive model in Eq. (37) on the transposed autocovariance sequence $\mathbf{G}^{\mathrm{TR}}_{[\mathbf{xy}]}(p) = \mathbf{G}^{\top}_{[\mathbf{xy}]}(p), p \in \{0, 1, \ldots, N_P\}$. This yields net GC scores $\mathcal{F}^{\mathrm{TR\ net}}_{\mathbf{x}\to\mathbf{y}}(f)$ for the time-reversed data, which are subtracted from the net scores obtained for the original (forward) data to yield the final time-reversed GC scores:

$$\mathcal{F}^{\mathrm{TRGC}}_{\mathbf{x}\to\mathbf{y}}(f) \equiv \mathcal{F}^{\mathrm{net}}_{\mathbf{x}\to\mathbf{y}}(f) - \mathcal{F}^{\mathrm{TR\ net}}_{\mathbf{x}\to\mathbf{y}}(f) \tag{43}$$

and (analogously) $\mathcal{F}^{\mathrm{TRGC}}_{\mathbf{y}\to\mathbf{x}}(f) \equiv \mathcal{F}^{\mathrm{net}}_{\mathbf{y}\to\mathbf{x}}(f) - \mathcal{F}^{\mathrm{TR\ net}}_{\mathbf{y}\to\mathbf{x}}(f) = -\mathcal{F}^{\mathrm{TRGC}}_{\mathbf{x}\to\mathbf{y}}(f)$.

2.5. Pipelines

In the following section, we describe the processing pipelines that were tested. All pipelines take the sensor measurements $\tilde{\mathbf{Q}}$ as input. Then all pipelines calculate and apply an inverse model $\mathbf{P}$ to project sensor data to source level. From there, we aggregate voxel activity within regions by employing PCA and estimate inter-regional FC with various FC metrics described above. We describe several strategies of combining PCA with the calculation of FC in the following subsections. This step results in a $N_{roi} \times N_{roi} \times N_{freq}$ FC matrix which is then averaged across the frequency bins within the interaction frequency band (8-12 Hz). The output of all pipelines is one connectivity score for every region combination. We describe the processing exemplarily for the calculation of FC between two regions X and Y.

Pipelines FIXPC1 to FIXPC6: Fixed number of principal components

The first six pipelines use PCA dimensionality reduction. Afterwards, depending on the pipeline, a fixed number C of either one, two, three, four, five, or six strongest PCs are selected for further processing. Then, FC is calculated: in case of univariate measures (i.e., coherence and iCOH), we first calculate FC scores between all PC combinations of the two regions X and Y and then average across all pairwise FC scores. In case of multivariate FC measures, we directly calculate a single FC score between the PCs of region X and those of region Y. This approach has been used previously (e.g. Schoffelen et al., 2017).

Pipelines VARPC90 and VARPC99: Variable numbers of principal components

Pipelines VARPC90 and VARPC99 are equivalent to the FIXPC pipelines, with the difference that we do not select the same fixed number of PCs for every region. Instead, we select the number of PCs such that at least 90% (VARPC90) or 99% (VARPC99) of the variance in each ROI is preserved (c.f. Section 2.3). Thus, an individual number of PCs is chosen for each region. FC is then calculated analogously to pipelines FIXPC1 to FIXPC6. The idea of selecting the number of PCs such that a pre-defined fraction of the variance is retained has been used in previous literature (e.g. Gómez-Herrero et al., 2008).

Pipeline MEANFC: Mean first FC second

In this pipeline, the time series of all voxels within one region are averaged separately for the three orthogonal dipole orientations. Then, for univariate

FC measures, FC is calculated between all 3*3 dimension combinations of the 3D-time series of region X and region Y. Afterwards, the average of these nine FC scores is taken. Multivariate FC measures are directly calculated between the 3D time series.

Pipeline CENTRAL: Central voxel pick

In this pipeline, we select only the central voxel of each region for further processing. The central voxel of a region is defined as the voxel whose average Euclidean distance to all other voxels in the region is minimal. To calculate the FC score between the 3D time series of the central voxel of region X and the 3D time series of the central voxel of region Y, we proceed analogous to pipeline MEANFC: in case of univariate FC measures, the FC score for all combinations of dipole orientations is calculated and then averaged. In case of multivariate FC measures, only one FC score is calculated between the two 3D time series. Selecting the time series of the central voxel as the representative time series for the region is an idea that has been used in previous studies already (Perinelli et al., 2022).

Pipeline FCMEAN: FC first mean second

In pipeline FCMEAN, the multivariate FC between each 3D voxel time series of region X with each voxel time series of region Y is calculated first. That is, if R_X is the number of voxels of region X and R_Y is the number of voxels in region Y, $R_X * R_Y$ FC scores for all voxel combinations are calculated. To obtain a single FC score between region X and region Y, we then average all $R_X * R_Y$ FC scores. Due to computational and time constraints, we test this pipeline only for MIM and MIC. This approach has

also been used in the literature before (Babiloni et al., 2018).

Pipeline TRUEVOX: True voxel pick

This pipeline is used as a baseline. Here we select the voxel for further processing that indeed contains the activity of the given ROI—i.e. the ground-truth voxel (see Section 2.1). All further processing is analogous to pipeline CENTRAL. In configurations with two active voxels per region (see Section 3, Experiment 6), FC scores are calculated for $2 * 3 * 3$ voxel- and dipole orientation combinations.

2.6. Performance evaluation

We use a rank-based evaluation metric to assess the performance of the pipelines. All processing pipelines result in one FC score for every region–region combination. To evaluate the performance of a pipeline, we first sort all FC scores in a descending order and retrieve the rank $r \in \mathbb{R}^{N_I}$, with $N_I \in \{1, 2, 3, 4, 5\}$ denoting the number of ground-truth interactions. Based on this rank vector, we calculate the percentile rank (PR):

$$PR' = \frac{\sum_i^{N_I} \left(1 - \frac{r_i}{F}\right)}{N_I}, \tag{44}$$

with F denoting the total number of FC scores. The PR' is then normalized to the perfect-skill PR_{ps} and no-skill PR_{ns} cases, and is therefore defined between 0 and 1:

$$PR_{ps} = \frac{\sum_i^{N_I} \left(1 - \frac{i}{F}\right)}{N_I} \tag{45}$$

$$PR_{ns} = \frac{\sum_i^{N_I} \left(1 - \frac{F-i+1}{F}\right)}{N_I} \tag{46}$$

$$PR = \frac{PR' - PR_{ns}}{PR_{ps} - PR_{ns}}. \tag{47}$$

We report all PR values rounded to the second decimal. In case of the phase-based FC metrics, the PR is calculated on the original FC scores. In case of GC and TRGC, we separately evaluate each pipeline's interaction detection ability, and its ability to determine the direction of the interaction. For evaluating the detection, we calculate the PR on the absolute values of the FC scores, whereas for evaluating the directionality determination performance, we calculate the PR only on the positive FC scores. Note that this is sufficient for the anti-symmetric directed FC measures used here.

2.7. Statistical assessment

In Experiment 1C, we provide a suggestion on how to statistically assess the presence of FC. Here, we obtain p-values by testing against a surrogate distribution consistent with the null hypothesis of zero interaction between all region pairs. The 10,000 samples of the surrogate distribution are drawn by shuffling epochs relative to each other when computing the cross-spectrum. More specifically, we calculate the cross-spectrum between the time series of one region and the shuffled time series of another region with the Welch method, where the diagonal entries of the cross-spectrum (spectral powers) are obtained without shuffling. From the shuffled cross-spectrum, MIM is calculated. We obtain p-values by counting the number of shuffled MIM-samples that are higher than the true MIM score and dividing this number by the total number of samples in the null distribution. FDR-correction (α-level = 0.05) is used on the upper triangle of the region–region p-value matrix to set a significance threshold.

2.8. ROIconnect toolbox

Based on our experimental results (see Section 3), we identified a set of recommended methods and pipelines. These have been implemented in a Matlab toolbox and are made available as a plugin to the free EEGlab package[1]. This toolbox also contains code for analyzing spectral power in EEG source space, and for visualizing power and FC results in source-space. A comprehensive description of the functionality and usage of the toolbox is provided in Appendix A. Moreover, an exemplary application of the toolbox to the analysis of a real EEG dataset is provided in Section 4.

3. Experiments and Results

We conducted a set of experiments to assess the influence of the different pipeline parameters on the reconstruction of ground-truth region-to-region FC. We describe the general experimental setting in Figure 2. Each experiment consisted of the following steps: (1) Signal generation. (2) Source projection. (3) Dimensionality reduction within regions. (4) Functional connectivity estimation. (5) Performance evaluation. Each experiment was carried out 100 times (= iterations). If not indicated otherwise, all experiments had the following default setting:

- LCMV inverse solution
- SNR = 3.5 dB
- BSR = 0 dB

[1]https://github.com/arnodelorme/roiconnect

- number of interactions = 2
- time delay of the interaction = 50 to 200 ms
- number of generated sources per region = 1

If not stated otherwise, the following parameters were drawn randomly in each iteration: ground-truth interacting (seed and target) regions (two distinct regions uniformly drawn between 1 and N_{roi}), ground-truth active voxel(s) within regions (uniformly drawn between 1 and R_{roi}), time delay (uniformly drawn between 50 and 200 ms). Furthermore, brain noise and sensor noise, as well as the signal were generated based on (filtered) random white noise processes as described above.

Figure 3 to Figure 11 show the results of experiments 1–6. In addition, all main results are summarized in Table 1. All figures (plotting code adapted from Allen et al., 2019) follow the same scheme: in every subplot, the 100 dots on the right side mark the performance, i.e. the PR, measured in each of the 100 iterations. On the left, a smooth kernel estimate of the data density is shown. The red and black lines represent the mean and median PR of the experiment, respectively, and the boxcar marks the 2.5th and 97.5th percentiles. Please note that the Y-axis is scaled logarithmically in all plots. We tested differences between pipeline performances with a one-sided Wilcoxon signed-rank test. Please note that a p-value $p_{\mathrm{A,B}}$ corresponds to a one-sided test for $B > A$.

Matlab code to reproduce all experiments is provided under[2].

[2]https://github.com/fpellegrini/FCsim

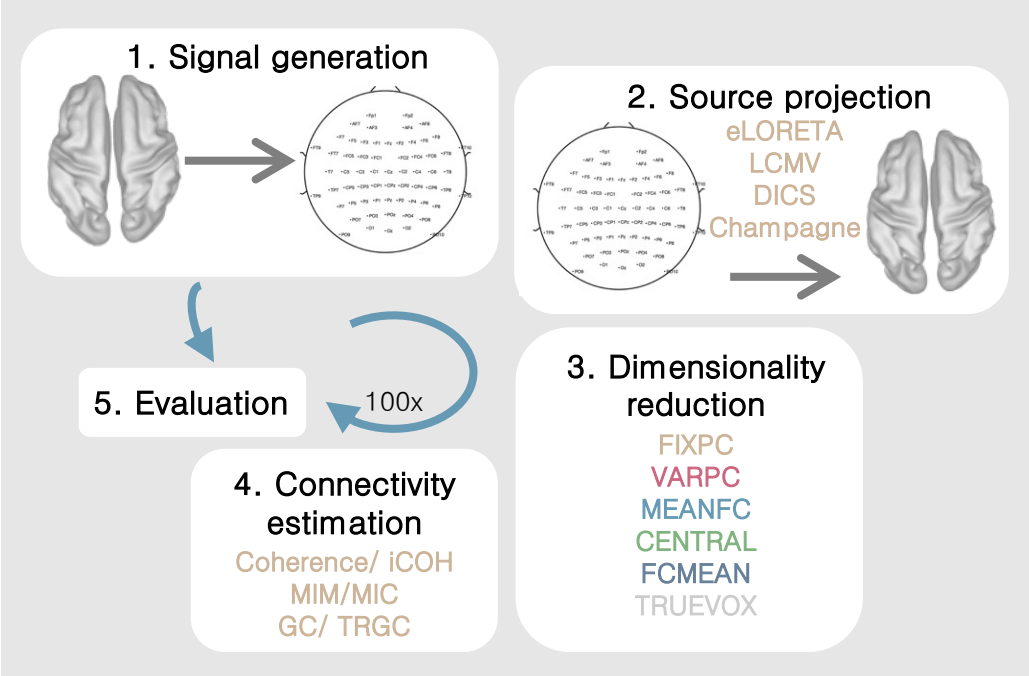


Figure 2: Experimental setup. Every experiment consisted of five consecutive steps: (1) Signal generation. (2) Source projection. (3) Dimensionality reduction within regions. (4) Functional connectivity estimation. (5) Performance evaluation. Every experiment was carried out 100 times.

3.1. Experiment 1

Experiment 1A

In Experiment 1A, we evaluated the performance of different FC metrics in detecting the ground-truth interactions. The ability to detect FC was tested for coherence, iCOH, MIC, MIM, GC, and TRGC. The ability to detect the correct direction of the interaction was tested for GC and TRGC (see Section 2.4).

In Figure 3, we show the performances of different FC metrics. We see that MIM, MIC and TRGC (detection) all have a mean PR of over 0.97 and clearly outperform the other measures in detecting the ground-truth FC. The non-robust metrics coherence (mean PR = 0.59) and GC (mean PR = 0.95) detect the ground-truth interactions less reliably ($p_{\text{coherence, MIM}} < 10^{-4}$; $p_{\text{GC,MIM}} = 0.0040$). When comparing GC and TRGC in their ability to infer the direction of the interaction, TRGC (mean PR = 0.98) outperforms GC (mean PR = 0.96; $p_{\text{GC,TRGC}} < 10^{-4}$).

Experiment 1B

In Experiment 1B, we tested the influence of different strategies of dimensionality reduction within regions. In Figure 4, we show the comparison for MIM (interaction detection) and TRGC (directionality determination). For MIM, we observe that the FIXPC pipelines show a better performance than most of the other pipelines. Within the FIXPC pipelines, the pipelines with two, three or four PCs perform best (all mean PR = 0.99, $p_{\text{FIXPC5,FIXPC3}} < 10^{-4}$). Only the TRUEVOX (baseline) pipeline using ground-truth information on voxel locations expectantly shows a higher performance (mean PR = 1.00; $p_{\text{FIXPC3,TRUEVOX}} < 10^{-4}$). The two VARPC pipelines show a

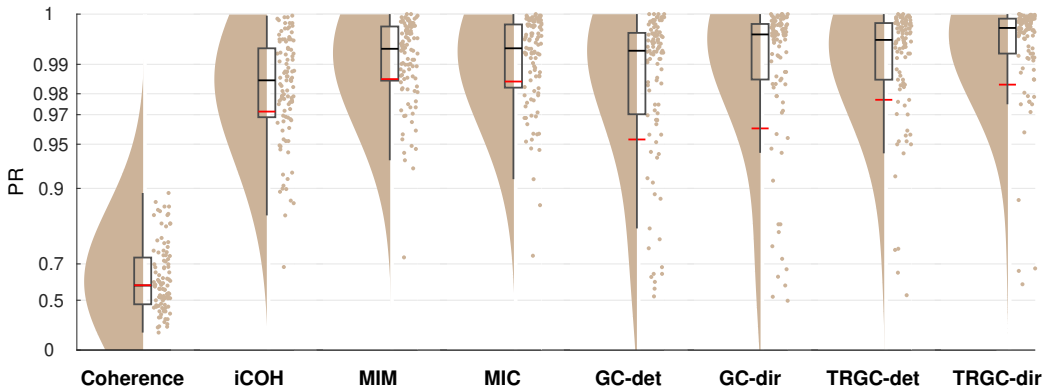
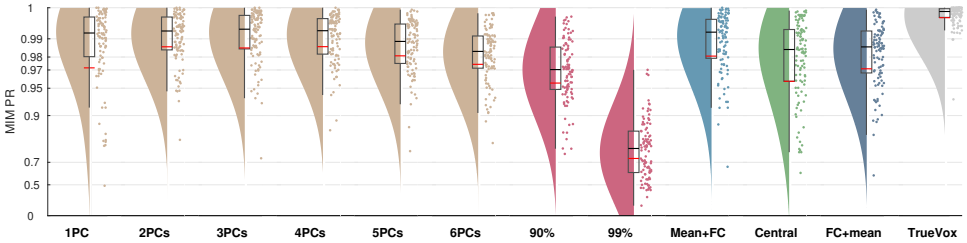


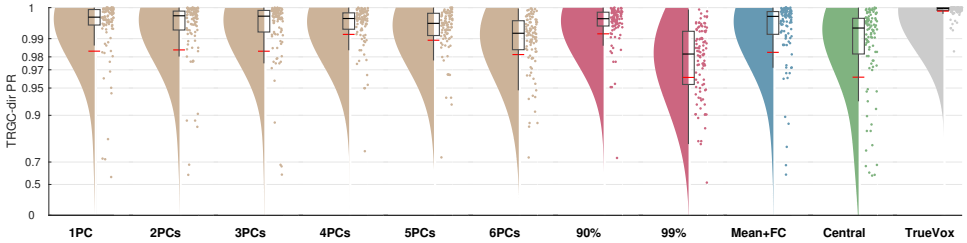
Figure 3: Comparison of different functional connectivity metrics (Experiment 1A). Red and black lines indicate the mean and median percentile rank (PR), respectively. The boxcar marks the 2.5th and 97.5th percentiles.

substantially reduced performance (mean PR = 0.96 and mean PR = 0.73, respectively; both $p_{\text{VARPC,FIXPC3}} < 10^{-4}$). The MEANFC and CENTRAL pipelines (mean PR = 0.98 and mean PR = 0.96, respectively) also show reduced performance in comparison to the FIXPC3 pipeline (both $p < 10^{-4}$). The FCMEAN pipeline (mean PR = 0.97) also did not perform as well as the FIXPC3 pipeline ($p < 10^{-4}$) while taking much longer to compute (FIXPC3 < 1 h, FCMEAN = 32 h, single core, allocated memory: 16 GB).

In terms of directionality estimation using TRGC, the outcome is similar. Again, the TRUEVOX pipeline shows perfect performance (mean PR = 1.00). The FIXPC pipelines also exhibit very high performances (FIXPC4: mean PR = 0.99). Notably, in contrast to the results obtained with MIM, the VARPC90 also achieves competitive performance (mean PR = 0.99, $p_{\text{VARPC90,FIXPC3}} = 0.0235$). Please see Figure S1 to compare computation



(a) Undirected FC reconstruction performance using MIM.



(b) Directed FC reconstruction performance using TRGC.

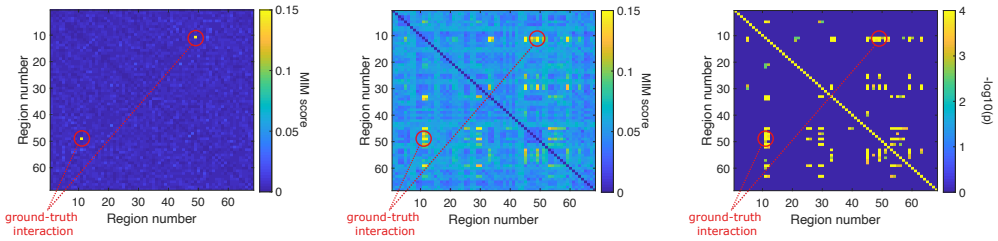
Figure 4: Comparison of different pipelines (Experiment 1B). (a) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median percentile rank (PR), respectively. The boxcar marks the 2.5th and 97.5th percentile.

times of all pipelines.

We show the full matrix of all combinations of FC metrics and dimensionality reduction pipelines in Supplementary Figure S2. However, for all further experiments, we report performances only for MIM (interaction detection) and TRGC (directionality determination) since they performed best in Experiment 1A, and we focus on the FIXPC3 pipeline due the high performance observed in Experiment 1B.

Experiment 1C

To explore how to statistically assess the presence of FC, we performed an additional experiment for a specific setting (SNR = 3.5 dB, one interaction between region 11 and region 49, BSR = 0 dB, LCMV filter, dimensionality reduction to 3PCs, FC metric = MIM). Here, we obtained p-values by testing against a surrogate distribution consistent with the null hypothesis of zero interaction between all region pairs. In Figure 5, we contrast the ground-truth ROI-to-ROI connectome with the estimated FC per region combination as well as the -log10(p) values "surviving" the FDR-correction for this experiment. While in the ground-truth connectome only the ground-truth region combination shows a high MIM score, there are also some high MIM scores in other region combinations than the ground truth in the reconstructed source-level connectome. Still, the ground-truth region combination in this setting still achieves the second-highest MIM score (PR = 0.9996). However, in Figure 5c, we see that testing the statistical significance with a shuffling test results in a substantial number of significant false positive interactions in the vicinity of the simulated interacting region pair. We discuss this result in Section 5.

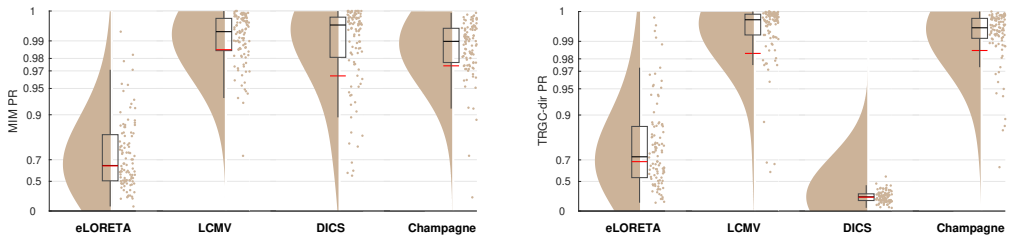


(a) Ground truth source-level connectome measured by MIM.

(b) Reconstructed source-level connectome measured by MIM.

(c) Statistical significance. P-values masked with FDR correction.

Figure 5: Comparison of the ground-truth ROI-to-ROI connectome with the estimated functional connectivity per region combination and the -log10(p) values after FDR-correction for a single experiment. A ground-truth interaction is modeled between region 11 and region 49.



(a) MIM performance.

(b) TRGC performance.

Figure 6: Comparison of different inverse solutions (Experiment 2). (a) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median percentile rank (PR), respectively. The boxcar marks the 2.5th and 97.5th percentile.

3.2. Experiment 2

Experiment 2A

In Experiment 2, we tested the influence of the type of inverse solution on the pipelines' performances. In Figure 6, we show the comparison between eLORETA, LCMV, DICS, and Champagne. We observe that the two beamformer solutions and Champagne clearly outperform eLORETA (mean PR 0.65; Figure 6a) in detecting undirected connectivity (all $p < 10^{-4}$). While DICS, LCMV and Champagne all show very good performances, we see a slight advantage of LCMV (mean PR = 0.99) in comparison to Champagne (mean PR = 0.97, $p_{\text{Champagne,LCMV}} = 0.0013$). We do not observe a significant difference between DICS and LCMV ($p_{\text{DICS,LCMV}} = 0.2805$).

In terms of directionality determination (Figure 6b), the picture is different: while LCMV (mean PR = 0.98) leads to accurate directionality es-

timates, DICS fails to detect the direction of the ground-truth interaction in a high number of experiments (mean PR = 0.28, $p_{\text{DICS,LCMV}} < 10^{-4}$). eLORETA also shows a reduced overall performance (mean PR = 0.69, $p_{\text{eLORETA, LCMV}} < 10^{-4}$). Champagne shows decent performance (mean PR = 0.99), which is, however, lower than that of LCMV ($p_{\text{Champagne,LCMV}} < 10^{-4}$).

The differences in computation times of the different inverse solutions are also remarkable. While LCMV (2 sec) and DICS (178 sec) are fast to compute, eLORETA (388 sec) and Champagne (3747 sec) take much longer to compute as a cross-validation scheme to set the regularization parameter is implemented for both. Setting the regularization parameter to a default value would drastically reduce computation time for eLORETA and Champagne, but would also decrease performance (results not shown).

Experiment 2B

To investigate further why eLORETA performs considerably less well than LCMV in our experiments, we generated ground-truth activity with an interaction between one seed voxel in the left frontal cortex and one target voxel in the left precentral cortex. We then again generated sensor data as described in Section 2.1 and applied pipeline FIXPC1 to calculate regional MIM scores. In Supplementary Figure S3, we show the resulting power maps, as well as seed MIM scores and target MIM scores for data projected with eLORETA and MIM, respectively. We see clearly the advantage of LCMV: while both power and MIM in the eLORETA condition are spread out to other regions, LCMV is able to localize the ground-truth power and connectivity very precisely.

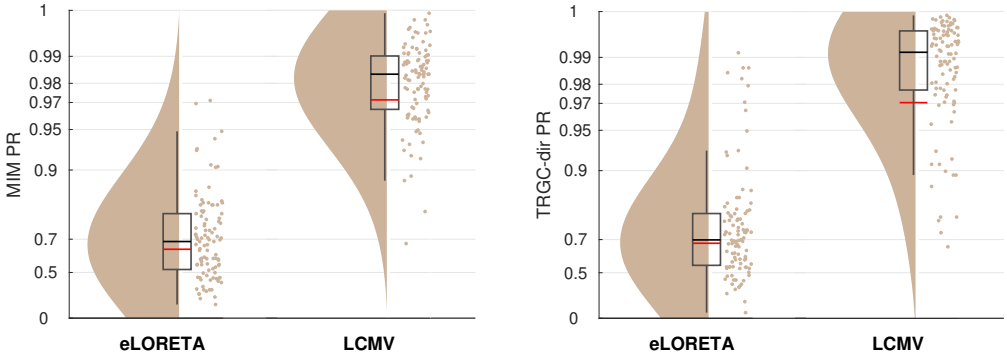
Experiment 2C

Does LCMV only perform so well in our experiment because our experimental setup artificially favors it? In the following additional analysis, we investigated whether LCMV still has an advantage over eLORETA when multiple pairs of correlated sources are present. More specifically, we here simulated two pairs of interacting sources where the time courses of the second source pair were identical to those of the first source pair. Results are presented in Figure 7. Please note that in this case, also the cross-interactions between the seed and target regions were evaluated as ground-truth interactions. We see that, while eLORETA is not much affected by the correlated sources setup, LCMV has a decreased reconstruction performance according to both MIM and TRGC. However, LCMV still performs better than eLORETA even in this setup ($p_{\text{eLORETA,LCMV}} < 10^{-4}$).

3.3. Experiment 3

In real-world EEG measurements, data are to a certain extent corrupted by noise, e.g. from irrelevant brain sources, or by noise sources from the outside. In Experiment 3, we investigated the effect of SNR and BSR on FC estimation performance. In Figure 8a and 8b, we show the performance of the FIXPC3 pipeline for SNRs of -7.4 dB, 3.5 dB and 19.1 dB. For both MIM (Figure 8a) and TRGC (Figure 8b), we observe decreased performances for decreased SNRs, as expected. For an SNR of 19.1 dB, nearly all experiments show a perfect detection of ground-truth interactions (mean PR > 0.99).

Is FC detection more impaired by pink brain noise or white sensor noise? In Experiment 3B, we tested the performance for BSR environments of 100% sensor noise, 25% brain noise, 50 % brain noise, 75% brain noise, and 100%



(a) MIM performance.

(b) TRGC performance.

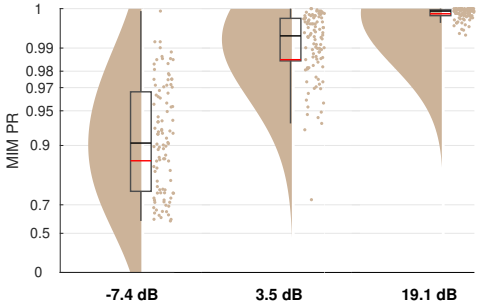
Figure 7: Performance observed for two perfectly correlated source pairs. (a) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median, respectively. The boxcar marks the 2.5th and 97.5th percentile.

brain noise. In Figure 8c and 8d, we show the performances for different BSRs. We observe a slightly better performance for signals more strongly contaminated by correlated brain noise than white sensor noise (mean MIM PR 100% brain noise > 0.99) compared to the opposite case (mean MIM PR 0% brain noise $= 0.97$).

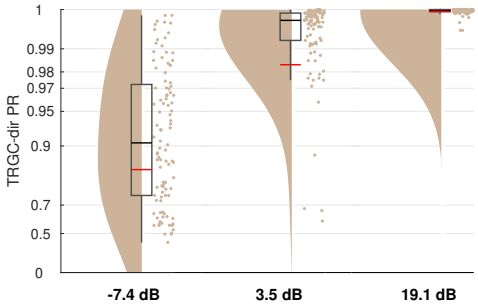
Note that in Experiments 1 to 3, for better comparison between the experimental conditions and to avoid variation due to random factors besides the experimental variation, we used the same generated data within an iteration in every experiment and only varied the tested condition.

3.4. Experiment 4

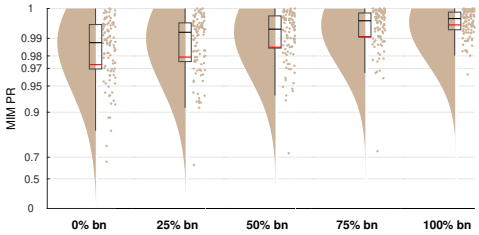
While we focused on a very simple scenario with only two interacting region pairs so far, real brain activity likely involves multiple interacting sources. To increase the complexity in our setup, we compared performances for different numbers of interacting region pairs in Experiment 4. As expected, Figure 9 clearly shows that more simultaneous true interactions lead to decreased ability to reliably detect them. While the detection is nearly perfect for one interaction (mean MIM PR > 0.99; mean TRGC PR > 0.99), the performance is much reduced for 5 interactions (mean MIM PR $= 0.91$; mean TRGC PR $= 0.93$). This applies for both MIM and TRGC. Please note however, that despite using a normalized version of the PR (see Section 2.6), the PR metric is not perfectly comparable for different numbers of true interactions. That is, when calculating the PR on randomly drawn data, the PR distribution is close to uniform when only one interaction is assumed, but shows a normal distribution with increasing kurtosis for higher numbers of interactions. However, the mean of the distribution equals to 0.5 for all



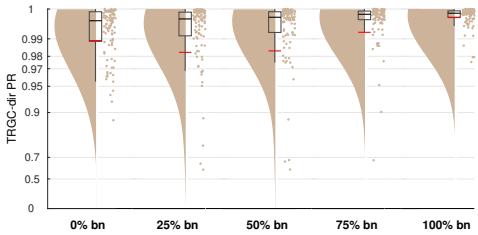
(a) MIM performance increases with SNR.



(b) TRGC performance increases with SNR.

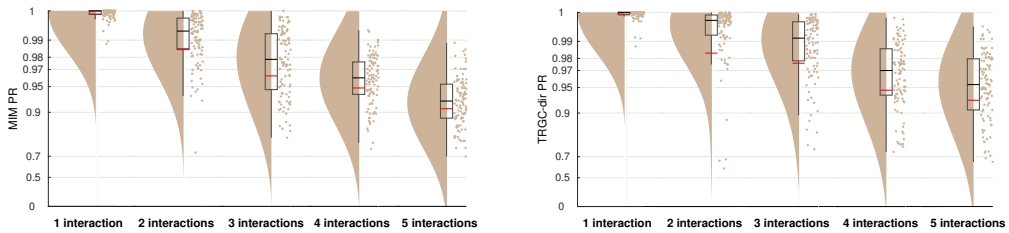


(c) MIM performance increases with BSR.



(d) TRGC performance increases with BSR.

Figure 8: FC estimation performance depends on the signal-to-noise ratio and brain noise-to-sensor noise ratio (Experiment 3). (a/c) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b/d) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median percentile rank (PR), respectively. The boxcar marks the 2.5th and 97.5th percentile.



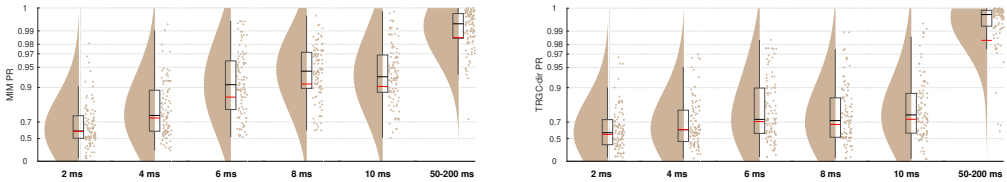
(a) MIM performance. **(b)** TRGC performance.

Figure 9: FC reconstruction performance depends on the number of true interactions (Experiment 4). (a) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median percentile rank (PR), respectively. The boxcar marks the 2.5th and 97.5th percentile.

assumed interactions.

3.5. Experiment 5

While it is not entirely clear how large interaction delays in the brain can be, they likely range between 2 and 100 ms, depending not only on physical wiring, but also on cognitive factors (see Section 5). In Experiment 5, we evaluated to which degree the performance drops when regions interact with shorter time delays of 2, 4, 6, 8, and 10 ms. While the performance for the MIM metric is already quite impaired for a delay of 10 ms (mean PR = 0.90), performance drops drastically for 4 ms (mean PR = 0.73) and 2 ms (mean PR = 0.60) (Figure 10a). Detecting the direction of the interaction with TRGC is already much more difficult at a true delay of 10 ms (mean PR = 0.73) and is further reduced for a delay of 2 ms (mean PR = 0.56;



(a) MIM performance.

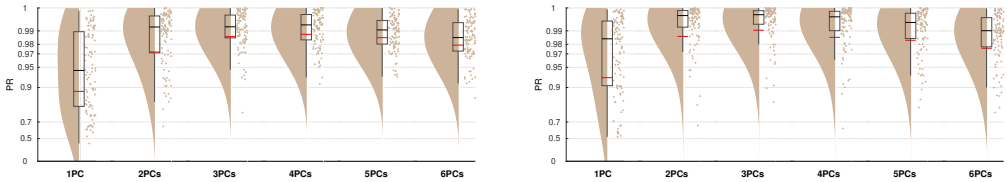
(b) TRGC performance.

Figure 10: Performance for very small interaction delays and the default delay (Experiment 5). (a) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median percentile rank (PR), respectively. The boxcar marks the 2.5th and 97.5th percentile.

Figure 10b).

3.6. Experiment 6

In our previous experiments, the FIXPC pipelines with two to four PCs showed the best performance. But the 'optimal' number of PCs likely depends on the number of (interacting and non-interacting) signals in the brain as well as their relative strengths. To verify that the optimal number of PCs depends on the number of true sources, we increased the number of active voxels per region to two in Experiment 6. We then simulated two bivariate interactions between two different source pairs originating from the same regions.We show the results for pipelines FIXPC1 to FIXPC6. Interestingly, we here see that pipelines FIXPC3 (mean MIM PR = 0.99; mean TRGC PR = 0.99) and FIXPC4 (mean MIM PR = 0.99; mean TRGC PR = 0.99) perform clearly better than FIXPC1 (mean MIM PR = 0.89; mean TRGC



(a) MIM performance.

(b) TRGC performance.

Figure 11: Performance when two active sources per region are simulated (Experiment 6). (a) Undirected FC reconstruction performance achieved using the multivariate interaction measure (MIM). (b) Directed FC reconstruction performance achieved using time-reversed Granger causality. Red and black lines indicate the mean and median, respectively. The boxcar marks the 2.5th and 97.5th percentile.

PR = 0.93) or FIXPC6 (mean MIM PR = 0.98; mean TRGC PR = 0.98). Based on these results, we confirm that the choice of the optimal number of fixed PCs increases with the number of independently active processes within one region (see Section 5 for further discussion).

4. Exploratory analysis of functional connectivity in left vs right motor imagery

To illustrate how the recommended analysis pipeline can be used to analyse real EEG data, we show an exploratory analysis of power and FC in left and right motor imagery. In the Berlin arm of the so-called VitalBCI study (Blankertz et al., 2010; Sannelli et al., 2019), 39 subjects conducted an experiment in which they imagined a movement with either the left or the right hand (Motor Imagery Calibration set; MI-Cb 1-3). Each trial consisted of a visual stimulus showing a fixation cross imposed with an arrow

#Exp.	Tested parameter	Result
1A	FC metric	MIM/TRGC yield best performance.
1B	pipelines	Fixed PC+FC yield best performance.
2	Inverse solution	LCMV yields best performance.
3A	SNR	The higher the better.
3B	BSR	The less sensor noise the better.
4	#Interactions	The lower the better.
5	Short interaction delays	Longer delays yield better performance.
6	Two active sources	Overall lower performance.
		Peak performance at three to four PCs.

Table 1: Summary of the results of experiment one to six. A pipeline including robust multivariate FC metrics like MIM or TRGC, a PCA with fixed number of selected components, and LCMV source reconstruction yields the best performance.

indicating the task for the trial (i.e., left or right motor imagery). After 4 sec, the stimulus disappeared, and the screen stayed black for 2 sec. Every subject conducted 75 left and 75 right motor imagery trials. During the experiment, EEG data were recorded with a 119-channel whole-head EEG system with a sampling rate of 1000 Hz. For this study, we used a 90-channel whole head standard subset of them. For our analysis, we selected only the 26 subjects for which previous studies have reported that the left vs. right motor imagery conditions could be well separated using statistical and machine learning techniques ('Category I' in Sannelli et al., 2019). Further experimental details are provided in Blankertz et al. (2010); Sannelli et al. (2019).

We filtered the data (1 Hz high-pass filter, 48-52 Hz notch filter, and 45 Hz low-pass filter, all zero-phase forward and reverse second-order digital high-pass Butterworth filters), and then sub-sampled them to 100 Hz. We then rejected artifactual channels based on visual inspection of the power spectrum and the topographical distribution of alpha power (between zero and five per participant, mean 1.19 channels) and interpolated them (spherical scalp spline interpolation). A leadfield was computed using the template head model Colin27_5003_Standard-10-5-Cap339 that is already part of the EEGLAB toolbox. We then epoched the data from 1 to 3 sec post-stimulus presentation start and separated left from right motor imagery trials.

We used the `pop_roi_activity` function of the newly developed ROIconnect plugin for EEGLAB to calculate an LCMV source projection filter, apply it to the sensor data, and calculate region-wise power (see Appendix A for a more detailed description). We then normalized the power with respect to

the total power between 3 and 7 Hz as well as 15 and 40 Hz, and averaged it across frequencies between 8 and 13 Hz. The statistical significance of the differences between right and left hand motor imagery power was assessed with a paired t-test in every region. In Figure S4, we show the negative log10-transformed p-values, multiplied with the sign of the t-statistic. As expected, the results show a clear lateralization for the activation of the motor areas.

To estimate inter-regional FC, we used the `pop_roi_connect` function to calculate MIM based on the three strongest PCs of every region. Again, MIM was averaged across frequencies between 8 and 13 Hz. To reduce the region-by-region MIM matrix to a vector of net MIM scores, we summed up all MIM estimates across one region dimension.

Analogous to our statistical evaluation of simulated data, described in Experiment 1C, we assessed the statistical significance of the net FC of each region against the null hypothesis of zero net interaction separately for each of the two motor imagery conditions. Specifically, we first calculated the true MIM score between all region pairs in all subjects. Then, we generated a null distribution of 1000 shuffled MIM scores for every region combination in every subject. Subsequently, the true and shuffled net MIM scores were calculated by averaging across one of the region dimensions. To obtain p-values, we compared the true MIM of every region and subject to the respective null distribution. To aggregate the p-values across subjects, we applied Stouffer's method (see, e.g., Dowding and Haufe, 2018). Finally, FDR-correction (α-level = 0.05) was used to correct for multiple comparisons. We show the negative log10-transformed p-values in Figures 12a and 12b.

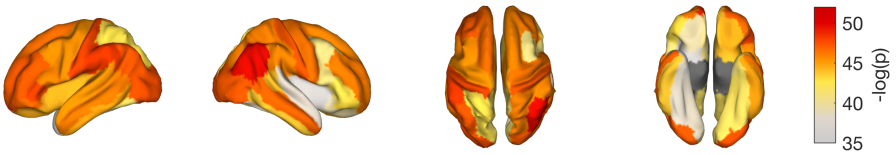
Additionally, we assessed the statistical difference between the net MIM scores of the left vs. right hand motor imagery condition by again using a paired t-test for every region. In Figure 12c, we show the negative log10-transformed p-values, multiplied with the sign of the t-statistic. Again, as expected, the results show a lateralization for the undirected net FC of the motor areas.

Matlab code of the analyses presented in this section is provided under [3].

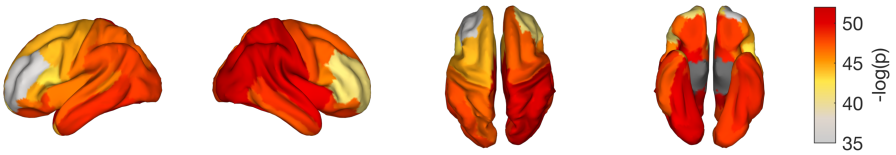
5. Discussion

Estimating functional connectivity between brain regions from reconstructed EEG sources is a promising research area that has generated a number of important results (e.g. Hipp et al., 2011; Schoffelen et al., 2017; Babiloni et al., 2018). However, respective analysis pipelines consist of a number of subsequent steps for which multiple modeling choices exist and can typically be justified. In order to identify accurate and reliable analysis pipelines, simulation studies with ground-truth data can be highly informative. However, most existing simulation studies do not evaluate complete pipelines but focus on single steps. In particular, various published studies assume the locations of the interacting sources to be known a-priori, while, in practice, they have to be estimated as well. To this end, it has become widespread to aggregate voxel-level source activity within regions of an atlas before conducting FC analyses across regions. Multiple ways to conduct this dimensionality reduction step have been proposed, which have not yet been

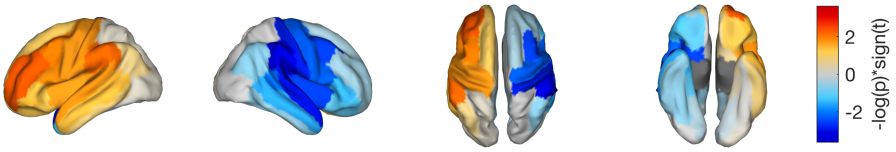
[3] https://github.com/fpellegrini/MotorImag



(a) Statistical significance of the net MIM functional connectivity scores during left motor imagery. Depicted are negative log10-transformed p-values.



(b) Statistical significance of the net MIM functional connectivity scores during right motor imagery. Depicted are negative log10-transformed p-values.



(c) Comparison of net MIM functional connectivity scores in left vs. right hand motor imagery tasks. Depicted are negative log10-transformed p-values, multiplied with the sign of the t-statistic. Positive values indicate higher FC in the left hand motor imagery trials.

Figure 12: Results of the exploratory analysis of functional connectivity in left and right hand motor imagery tasks.

systematically compared using simulations. The main focus of our study was thus to identify those EEG processing pipelines from a set of common approaches that can detect ground-truth inter-regional FC most accurately. For the scenario modelled in this study, we observe that a pipeline consisting of an LCMV source projection, PCA dimensionality reduction, the selection of a fixed number of principal components for each ROI, and a robust FC metric like MIM or TRGC results in the most reliable detection of ground-truth FC (see Table 1). Consistent with results reported in Anzolin et al. (2019), LCMV consistently yielded higher FC reconstruction performance than eLORETA. Thus, we here answer the question that Mahjoory et al. (2017) left open, namely which source reconstruction technique is most suitable for EEG FC estimation. Our results are also in line with a larger body of studies that highlighted the advantages of robust FC metrics compared to non-robust ones (e.g. Nolte et al., 2004; Haufe et al., 2013; Vinck et al., 2015; Winkler et al., 2016; Schoffelen and Gross, 2019).

Inverse solutions

For some inverse solutions, the choice of the regularization parameter has been shown to influence the accuracy of source reconstruction (Hincapié et al., 2016; Hashemi et al., 2021). While the parameter is of little importance for methods like LCMV and DICS, which are fitted separately to each source and thus solve low-dimensional optimization problems, it should be carefully chosen for full inverse solutions like Champagne and eLORETA, which estimate the activity at each source voxel within a single model. To avoid a performance drop due to unsuitable regularization parameter choice in eLORETA and Champagne, we used the spatial cross-validation method

described in (Habermehl et al., 2014; Hashemi et al., 2021). This method automatically sets the parameter based on the data at hand and has been shown to improve the source reconstruction (Hashemi et al., 2021).

As hypothesized, DICS resulted in poor directionality determination performance, while LCMV and TRGC performed well. This can be explained by the difference between LCMV and DICS: while LCMV estimates the inverse solution in the time domain, DICS estimates the source projection for every frequency separately (Gross et al., 2001). This can lead to inconsistencies across frequencies. Since directionality estimation requires the aggregation of phase information across multiple frequencies, such inconsistencies may lead to failure of detecting true interactions and their directionalities. Therefore, we recommend to avoid using DICS source reconstruction when analysing directed FC. For undirected FC measures, this seems to be less of a problem. Still, in our simulation, LCMV consistently performed (even if only slightly) better than DICS. This can be explained by the lower effective number of data samples that are available to DICS at each individual frequency compared to LCMV, which uses data from the entire frequency spectrum. However, there may be cases when using DICS could result in more accurate localization. For example, this could be the case when the noise has a dominant frequency that is different from the signal.

Robust functional connectivity metrics

In this study, we observed a strong benefit of using robust FC metrics over non-robust metrics in detecting genuine neuronal interactions. Overall, the performance of coherence is highly impaired by the volume conduction effect (see Figure 3, c.f. Nolte et al., 2004). The TRGC metric performed well

for the investigation of the interaction direction, but also satisfyingly well for the interaction detection. However, the computation time for calculating TRGC exceeds that of MIM by far. Thus, we recommend using MIM to detect undirected FC in case the direction of the effect is not of relevance. If TRGC is calculated for estimating the direction of interactions, the absolute value of TRGC can be used to detect interactions as well.

Interestingly, GC without time reversal did not perform much worse than TRGC. This is in line with previous results (Winkler et al., 2016) demonstrating that the calculation of net GC values already provides a certain robustification against volume conduction artifacts. Concretely, it has been shown that net GC is more robust to mixed noise than the standard GC; however not as robust as TRGC (Winkler et al., 2016). We generally recommend using robust FC connectivity metrics like iCOH, MIM/MIC, or TRGC.

Aggregation within regions

When comparing different processing pipelines, we found that employing an SVD/PCA and selecting a fixed number of components for further processing performs better than selecting a variable number of components in every ROI. When further investigating this effect, we found that, for MIM and MIC, the final connectivity score of the VARPC pipelines was positively correlated with the number of voxels of the two concerning ROIs (90%: MIM: $r = 0.50$, MIC: $r = 0.32$; 99%: MIM: $r = 0.70$, MIC: $r = 0.41$). This indicates that the flexible number of PCs leads to a bias in MIM and MIC depending on the size of the two involved ROIs. This could be expected, as the degrees of freedom for fitting MIM and MIC scale linearly with the number of voxels within a pair of regions. These in- or explicit model parameters can be tuned

to maximize the FC of the projected data, which may lead to over-fitting. For finite data, this leads to a systematic overestimation of FC, to the degree of which it correlates with the number of voxels. Although representing a multivariate technique as well, similar behavior was not observed for TRGC. Here it is likely that a potential bias of the signal dimensionalities would cancel out when taking differences between the two interaction directions as well as between original and time-reversed data.

An interesting and so far unsolved question is how many fixed components should be chosen for further processing. In Experiment 6, we observed a clear performance peak around three to four components (Figure 11). In the default version with only one active source per ROI, we saw a similar pattern, but not as pronounced as in Experiment 6. This points towards a data-dependent optimal number of components. Future work should investigate how this parameter can be optimized based on the data at hand.

Short time delays

In Experiment 5, we investigated to what extent the performance drops when the true interaction occurs with a very small time delay of 2 to 10 msec, which might be a realistic range for a number of neural interaction phenomena in the brain. Precise data on the typical order of the times within which macroscopic neural ensembles exchange information are, however, hard to obtain, as these transmission times depend not only on the physical wiring but also on cognitive factors that are not straightforward to model. Previous work has shown that delays can range from 2 to 100 msec, depending on the distance and number of synapses between two nodes (e.g. Fries, 2005; Oswal et al., 2016; Shouno et al., 2017; Miocinovic et al., 2018). For example, Oswal

et al. (2016) studied interaction delays between the subthalamic nucleus and the motor cortex and found interaction delays of 20 to 46 msec. The satisfactory performance observed in our study for undirected FC at delays of 8 and 10 msec may therefore be of particular importance for clinical scientists that aim at investigating such long-range interactions. Note that the range of delays that can be detected with robust connectivity metrics strongly depends on the frequency band in which the interaction takes place. If the delay is very short compared to the base frequency of the interaction, then the phase difference it induces is close to either 0 or $\pm\pi$, making it less and less distinguishable from a pure volume conduction effect as it approaches these limits. In addition, the directionality of an interaction can only be resolved by analyzing multiple frequencies. Here, wider interaction bands lead to better reconstructions of the directionality of interactions with shorter delays, whereas higher frequency resolutions (that is, longer data segments) lead to better reconstructions of the directionality of interactions with longer delays. Here, we have demonstrated that alpha-band interactions with physiologically plausible transmission delays can be detected at 0.5 Hz frequency resolution, depending on the underlying SNR as well as additional modeling assumptions (see Limitations below).

Statistical assessment

The goal of this study was to evaluate data analysis pipelines to assess FC. However, we excluded the assessment of any subsequent statistical evaluation of FC, which is not straightforward to investigate in simulation studies. In a simulation setting, we are free to choose the two factors that influence the statistical power of a test—SNR and sample size. Determining realistic

ranges for both in the context of EEG FC estimation is challenging but critical. Second, due to source leakage, we must expect (tiny) spill-over effects from interacting to non-interacting region pairs, an effect termed "ghost interactions" (Palva et al., 2018). As a result, these ghost interactions will inevitably become statistically significant for any source pair at high enough SNRs and sample sizes—an effect that can also be seen in Figure 5c. For these reasons, we here assessed the effect sizes of FC metrics instead of than their statistical significance, and focused on evaluating the performance of different FC estimation pipelines relative to one another rather than on their absolute performance. However, future studies should go one step further by systematically assessing statistical maps derived from connectomes using our results as building blocks.

Limitations

While this study investigates a large range of processing pipelines, phase-to-phase FC metrics, and data parameters, it is far from being exhaustive. Other works have shown that many other parameters like channel density (Song et al., 2015), the location of interacting sources (Anzolin et al., 2019), data length (Astolfi et al., 2007; Van Diessen et al., 2015; Liuzzi et al., 2017; Sommariva et al., 2019), referencing (Van Diessen et al., 2015; Chella et al., 2016; Huang et al., 2017), and co-registration (Liuzzi et al., 2017) can influence FC detection. Besides, we here used the same head model for generating the sensor data and estimating the inverse solution. However, we expect worse performance when the head model has to be estimated, and previous work has shown that the quality of head model estimation also influences FC detection (Mahjoory et al., 2017). Likewise, there exist many other in-

verse solutions, like MNE, wMNE, LORETA, sLORETA, and MSP, just to name a few. Further, there also exist other types of dimensionality reduction techniques. For example, some works selected the source with the highest power within a region or the source that showed the highest correlation to the time series of other sources in the ROI to be representative for all time series of the ROI (Hillebrand et al., 2012; Ghumare et al., 2018). Others have presented a procedure of optimizing a weighting scheme before averaging all time series within a ROI (Palva et al., 2010, 2011).

We also did not investigate the effect of the number of epochs and the epoch length in this study. It has been shown that the number of epochs can introduce a bias for certain connectivity metrics (Vinck et al., 2010). This is the case for connectivity metrics that yield positive values only, like (absolute) coherence, the absolute value of the imaginary part of coherency, MIM, or MIC. For these metrics, for a fixed epoch length, a lower number of epochs will systematically lead to higher values of estimated connectivity, even under the null hypothesis of no interaction. This is due to the higher variance of the estimates for lower samples sizes, which turns into a positive bias when the absolute value is taken. Further, Fraschini et al. (2016) argued that also the epoch length may have an influence on FC estimation, where shorter epochs were found to introduce a positive bias on FC when the number of epochs was held constant. As a result, we recommend to use fixed numbers of epochs throughout a single experiment. This is of particular importance when the goal is to compare different groups or experimental conditions.

As the set of coupling mechanism and corresponding FC metrics that have been proposed is huge, we deliberately constrained our analysis here to phase-

phase coupling using a selection of metrics that have previously been shown to be robust to mixing artifacts (Nolte et al., 2004; Haufe et al., 2013; Ewald et al., 2012). In contrast, non-robust metrics have been shown to be prone to the spurious discovery of interactions (Nolte et al., 2004; Haufe et al., 2013; Bastos and Schoffelen, 2016; Brunner et al., 2016; Van de Steen et al., 2019). This was confirmed here again for absolute coherence and GC. For a detailed overview of the taxonomy of FC metrics we refer to the works of Bastos and Schoffelen (2016); Schoffelen and Gross (2019); Marzetti et al. (2019). Our results are obtained for intra-frequency phase–phase coupling, and make no claims about non-linear interaction metrics quantifying phase–amplitude or amplitude–amplitude coupling within or across frequencies (De Pasquale et al., 2010; Hipp et al., 2012; Colclough et al., 2015). Nevertheless, we expect that robust-to-volume conduction measures for these FC types would be required to obtain optimal performance.

A further limitation of simulation studies in general is that assumptions need to be made that are hard, if not impossible, to confirm. Here, our goal was to generate pseudo-EEG data comprising realistic effects of volume conduction using a physical model of a human head. In terms of the generated time series, we focused on alpha-band oscillations as carriers of the modeled interactions. By adding pink brain noise, uniformly distributed across the entire brain, as well as white sensor noise, we obtained simulated sensor-space EEG data that resemble real data in crucial aspects such as spectral peaks and the general 1/f shape of the power spectrum. On the other hand, numerous additional assumptions were made regarding the linear dynamics of the interacting sources, the conception of the interaction as a pure and fixed time

delay, the focus on an interaction in the alpha band, the number of interactions, the signal-to-noise ratio, and the stationarity of all signal and noise sources. Several of these experimental variables were systematically varied to provide a comprehensive picture of the performance of each pipeline in a wide range of scenarios. The ranking of the pipelines' performances was robust in all tested scenarios. However, a remaining question is how realistic the individual studied parameter choices are. Our simulated environment resembles a setting of task-related (ongoing) activity with few dominant active and interacting sources, as opposed to a resting-state setting with numerous equally active and interacting sources. Hincapié et al. (2017) showed that connectivity estimation pipelines including beamformers perform well for point-like sources, whereas for extended cortical patches, MNE source estimation was found to be more accurate. In this study, we simulated point-like sources, which could lead to an overestimation of beamformer performance. Considering that FC analyses are predominantly performed on ongoing (including resting-state) activity, the assumption of having only a few interacting source pairs standing out against non-interacting background sources may be challenged. However, this assumption was made here for the practical purpose of enabling a comparison between approaches. Considering that FC analyses are predominantly performed on ongoing (e.g., resting-state) activity rather than averaged data, the assumptions of only few interacting source pairs standing out against non-interacting background sources with relatively high SNR can certainly be questioned. However, these assumptions were made here for the practical purpose of enabling a comparison between approaches rather than with the ambition of claiming real-world validity.

Future simulation studies should nevertheless strive to further increase the realism of the generated pseudo-EEG signals. In this regard, Anzolin et al. (2021) presented a toolbox that mimics typical EEG artifacts like eye blinks. We restricted ourselves here to using artificial time series designed to exhibit the specific properties assessed by the studied FC metrics; that is, time-delayed linear dynamics. In contrast, biologically inspired models such as the models implemented within the virtual brain toolbox (Sanz Leon et al., 2013) provide a richer portfolio of non-linear dynamics and thus are alternative ground truth models specifically when the goal is to validate non-linear FC metrics. The COALIA model (Bensaid et al., 2019), for example, has been used to mimick network activity in epilepsy for the purpose of validating FC estimates (Allouch et al., 2022). Further studies used the same model family to study the effect of parameters such as electrode density on FC estimates (Tabbal et al., 2022; Allouch et al., 2023). Similarly, Jirsa and Müller (2013) have used TVB to evaluate metrics of cross-frequency coupling. Overall, these studies provide complementary evidence that is largely aligned with our results, for example with respect to the superiority of robust connectivity metrics. The plausibility of several assumptions made by neural mass models has also recently been questioned (Pathak et al., 2022). Nevertheless, such models hold great promise as validation tools in the future.

Note in this respect that it was not our intention to propose a realistic model of EEG data or even the whole brain but simply to generate data that would allow us to test how well ROI-level FC can be reconstructed in the presence of volume conduction/source leakage. The types of FC we are

interested here (directed and undirected linear FC) have been widely studied and popular metrics to infer these types of FC are known to be heavily affected by volume conduction (Nolte et al., 2004; Haufe et al., 2013). Hence, it was our intention to identify metrics and pipelines that have a high chance of reconstructing FC on the ROI level when signals are mapped to the EEG and back by realistic forward and inverse models. We deliberately do not address the question whether networks estimated using FC metrics provide a correct depiction of actual brain networks.

As a further limitation, our simulations are to some extent restricted to EEG data. However, it can be expected that, qualitatively, the results of this paper could be transferred to MEG data. MEG analyses also suffer from the source leakage problem (Pizzella et al., 2014; Colclough et al., 2016) and benefit from disentangling signal sources with source reconstruction (Marzetti et al., 2019; Schoffelen and Gross, 2019). Moreover, the same FC metrics are typically used in EEG and MEG analyses (Schoffelen and Gross, 2009, 2019). Nevertheless, differences exist, which would be worth studying. In contrast to EEG, which records secondary neuronal return currents, MEG records the magnetic field that is induced by electrical activity and arises in a circular field around an electric current (Hämäläinen et al., 1993). Therefore, MEG cannot record radial neuronal currents (Huang et al., 2007). This must be taken into account when estimating the inverse solution from the leadfield, i.e. it is advised to reduce the rank of the forward model from three to two by applying an SVD at each source location (Westner et al., 2021).

We here provide a simulation framework that is openly accessible by the community. Individual pipeline steps, but also simulated data can easily be

replaced by other variants, following a plug-and-play principle. A such, we encourage readers to test aspects of the pipelines, other data, and other FC metrics not considered here.

6. Conclusion

This work compared an extensive set of data analysis pipelines for the purpose of extracting directed and undirected functional connectivity between predefined brain regions from simulated EEG data. While several individual steps of such pipelines have been benchmarked in previous studies, we focused specifically on the problem of aggregating source-reconstructed data into region-level time courses and, ultimately, region-to-region connectivity matrices. Thereby, we close a gap in the current literature evaluating FC estimation approaches. We show that using non-robust FC metrics greatly reduces the ability to correctly detect ground-truth FC. Further, in our simulated pseudo-EEG data, the use of the eLORETA inverse solution also leads to worse FC detection performance than beamformers. Moreover, the use of inverse solutions that are frequency-specific, such as DICS, may hamper the correct identification of the directionality of interactions. Finally, unequal dimensionalities of signals at different ROIs may bias certain connectivity measures, such as MIC and MIM, degrading their ability to identify true interactions from a noise floor. Thus, dimensionality reduction techniques should be applied such that the number of retained signal components is the same for all regions. We expect that avoiding these pitfalls may enhance the correct interpretation and comparability of results of future connectivity investigations. FC pipelines that show promising results with our simulated

pseudo-EEG data consist of beamformer or champagne source reconstruction, aggregation of time series within ROIs using a fixed number of strongest PCs, and using a robust FC metric like MIM or TRGC. To which scenarios these results can be generalized remains to be shown in further studies. In practice, low SNR, high numbers of interactions, and small interaction delays may, however, reduce the performance even of the best performing pipelines.

Acknowledgements

This project was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 758985) and through Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 424778381 – TRR 295. We thank Tien Dung Nguyen for contributing to the development of the ROIconnect plugin. The computations for this work were partly run on the open Neuroscience Gateway cluster (Sivagnanam et al., 2013).

Data and code availability

The code for the simulation can be found here: `https://github.com/fpellegrini/FCsim`. The code for the ROIconnect plugin can be found here: `https://github.com/sccn/roiconnect`. And the code for the minimal real data example here: `https://github.com/fpellegrini/MotorImag`. Data of the real data example are available upon request.

Appendix A. ROIconnect toolbox

ROIconnect is a freely available open-source plugin to the popular MATLAB-based open-source toolbox EEGLAB for EEG data analysis. It adds the functionality of calculating region-wise power and inter-regional FC on the source level. Moreover, it provides functions to visualize power and FC. All functions can be accessed by the EEGLAB GUI or the command line. ROIconnect uses core EEGLAB functions for importing and preprocessing EEG data, and calculating the leadfield and source model: we refer users to other EEGLAB functions to preprocess data before applying ROIconnect functions. The ROIconnect plugin can be downloaded through github [4] or installed via the EEGLAB GUI extension manager.

Key features

The features of ROIconnect are implemented in three main functions: `pop_roi_activity`, `pop_roi_connect`, and `pop_roi_connectplot`.

`pop_roi_activity` takes an EEG struct containing EEG sensor activity, a pointer to a headmodel and a source model, the atlas name, and the number of PCs for dimensionality reduction as input. It then calculates a source projection filter (default: LCMV) and applies it to the sensor data. Power is then calculated with the Welch method for every frequency on the voxel time series and then summed across voxels within regions. The result is saved in `EEG.roi.source_roi_power`. To estimate region-wise FC, the `pop_roi_activity` function reduces the dimensionality of the time series of every region by employing a PCA and selecting the strongest PCs (as defined

[4] https://github.com/arnodelorme/roiconnect

in the input) for every region. The resulting time series are then stored in `EEG.roi.source_roi_data`.

`pop_roi_connect` calculates FC between regions. It builds on the output of `pop_roi_activity`. That is, it takes the EEG struct as input, as well as the name of the FC metrics that should be calculated. The function calculates all FC metrics in a frequency-resolved way. That is, the output contains FC scores for every region–region–frequency combination. To avoid biases due to different data lengths, `pop_roi_connect` estimates FC for time windows ('snippets') of 60 sec length (default), which subsequently can be averaged (default) or used as input for later statistical analyses. The snippet length can be flexibly adjusted by the user. The output of this function is stored under the name of the respective FC metric under `EEG.roi`.

The `pop_roi_connectplot` function enables visualizing power and FC in the following modes:

- Power as region-wise bar plot.
- Power as source-level cortical surface topography.
- FC as region-by-region matrix.
- Net FC, that is, the mean FC from all regions to all regions, as cortical surface topography.
- Seed FC, that is, the FC of a seed region to all other regions, as cortical surface topography.

For plotting, a specific frequency or frequency band can be chosen by the user. For matrix representations, it is also possible to just plot one of the

hemispheres or only regions belonging to specific brain lobes.

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., Kievit, R. A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome open research 4.

Allouch, S., Kabbara, A., Duprez, J., Khalil, M., Modolo, J., Hassan, M., 2023. Effect of channel density, inverse solutions and connectivity measures on eeg resting-state networks reconstruction: A simulation study. NeuroImage 271, 120006.

Allouch, S., Yochum, M., Kabbara, A., Duprez, J., Khalil, M., Wendling, F., Hassan, M., Modolo, J., 2022. Mean-field modeling of brain-scale dynamics for the evaluation of EEG source-space networks. Brain Topography 35 (1), 54–65.

Anzolin, A., Presti, P., Van De Steen, F., Astolfi, L., Haufe, S., Marinazzo, D., 2019. Quantifying the effect of demixing approaches on directed connectivity estimated between reconstructed EEG sources. Brain topography 32 (4), 655–674.

Anzolin, A., Toppi, J., Petti, M., Cincotti, F., Astolfi, L., 2021. SEED-G: Simulated EEG data generator for testing connectivity algorithms. Sensors 21 (11), 3632.

Aoki, M., Havenner, A., 1991. State space modeling of multiple time series. Econometric Reviews 10 (1), 1–59.

Astolfi, L., Cincotti, F., Mattia, D., Marciani, M. G., Baccala, L. A., de Vico Fallani, F., Salinari, S., Ursino, M., Zavaglia, M., Ding, L., et al.,

2007. Comparison of different cortical connectivity estimators for high-resolution EEG recordings. Human brain mapping 28 (2), 143–157.

Babiloni, C., Del Percio, C., Lizio, R., Noce, G., Lopez, S., Soricelli, A., Ferri, R., Nobili, F., Arnaldi, D., Famà, F., et al., 2018. Abnormalities of resting-state functional cortical connectivity in patients with dementia due to alzheimer's and lewy body diseases: an EEG study. Neurobiology of aging 65, 18–40.

Barnett, L., Barrett, A. B., Seth, A. K., 2018. Solved problems for granger causality in neuroscience: A response to stokes and purdon. NeuroImage 178, 744–748.

Barnett, L., Seth, A. K., 2014. The MVGC multivariate granger causality toolbox: a new approach to granger-causal inference. Journal of neuroscience methods 223, 50–68.

Barnett, L., Seth, A. K., 2015. Granger causality for state-space models. Physical Review E 91 (4), 040101.

Barrett, A. B., Barnett, L., Seth, A. K., 2010. Multivariate granger causality and generalized variance. Physical Review E 81 (4), 041907.

Basti, A., Nili, H., Hauk, O., Marzetti, L., Henson, R. N., 2020. Multi-dimensional connectivity: a conceptual and mathematical review. NeuroImage, 117179.

Bastos, A. M., Schoffelen, J.-M., 2016. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Frontiers in systems neuroscience 9, 175.

Bensaid, S., Modolo, J., Merlet, I., Wendling, F., Benquet, P., 2019. Coalia: A computational model of human EEG for consciousness research. Frontiers in Systems Neuroscience 13, 59.

Blankertz, B., Sannelli, C., Halder, S., Hammer, E. M., Kübler, A., Müller, K.-R., Curio, G., Dickhaus, T., 2010. Neurophysiological predictor of SMR-based BCI performance. Neuroimage 51 (4), 1303–1309.

Bradley, A., Yao, J., Dewald, J., Richter, C.-P., 2016. Evaluation of electroencephalography source localization algorithms with multiple cortical sources. PloS one 11 (1), e0147266.

Bressler, S. L., Seth, A. K., 2011. Wiener–granger causality: a well established methodology. Neuroimage 58 (2), 323–329.

Brunner, C., Billinger, M., Seeber, M., Mullen, T. R., Makeig, S., 2016. Volume conduction influences scalp-based connectivity estimates. Frontiers in computational neuroscience 10, 121.

Castaño-Candamil, S., Höhne, J., Martínez-Vargas, J.-D., An, X.-W., Castellanos-Domínguez, G., Haufe, S., 2015. Solving the EEG inverse problem based on space–time–frequency structured sparsity constraints. NeuroImage 118, 598–612.

Chella, F., Pizzella, V., Zappasodi, F., Marzetti, L., 2016. Impact of the reference choice on scalp EEG connectivity estimation. Journal of neural engineering 13 (3), 036016.

Colclough, G. L., Brookes, M. J., Smith, S. M., Woolrich, M. W., 2015.

A symmetric multivariate leakage correction for MEG connectomes. Neuroimage 117, 439–448.

Colclough, G. L., Woolrich, M. W., Tewarie, P., Brookes, M. J., Quinn, A. J., Smith, S. M., 2016. How reliable are MEG resting-state connectivity metrics? Neuroimage 138, 284–293.

D'Andrea, A., Chella, F., Marshall, T. R., Pizzella, V., Romani, G. L., Jensen, O., Marzetti, L., 2019. Alpha and alpha-beta phase synchronization mediate the recruitment of the visuospatial attention network through the superior longitudinal fasciculus. NeuroImage 188, 722–732.

De Pasquale, F., Della Penna, S., Snyder, A. Z., Lewis, C., Mantini, D., Marzetti, L., Belardinelli, P., Ciancetta, L., Pizzella, V., Romani, G. L., et al., 2010. Temporal dynamics of spontaneous MEG activity in brain networks. Proceedings of the National Academy of Sciences 107 (13), 6040–6045.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31 (3), 968–980.

Dowding, I., Haufe, S., 2018. Powerful statistical inference for nested data using sufficient summary statistics. Frontiers in human neuroscience 12, 103.

Ewald, A., Marzetti, L., Zappasodi, F., Meinecke, F. C., Nolte, G., 2012. Estimating true brain connectivity from EEG/MEG data invariant to linear and static transformations in sensor space. Neuroimage 60 (1), 476–488.

Faes, L., Stramaglia, S., Marinazzo, D., 2017. On the interpretability and computational reliability of frequency-domain granger causality. F1000Research 6.

Fraschini, M., Demuru, M., Crobe, A., Marrosu, F., Stam, C. J., Hillebrand, A., 2016. The effect of epoch length on estimated eeg functional connectivity and brain network organisation. Journal of neural engineering 13 (3), 036015.

Fries, P., 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in cognitive sciences 9 (10), 474–480.

Fries, P., 2015. Rhythms for cognition: communication through coherence. Neuron 88 (1), 220–235.

Friston, K. J., 2011. Functional and effective connectivity: a review. Brain connectivity 1 (1), 13–36.

Friston, K. J., Rotshtein, P., Geng, J. J., Sterzer, P., Henson, R. N., 2006. A critique of functional localisers. Neuroimage 30 (4), 1077–1087.

Geweke, J., 1982. Measurement of linear dependence and feedback between multiple time series. Journal of the American statistical association 77 (378), 304–313.

Ghumare, E. G., Schrooten, M., Vandenberghe, R., Dupont, P., 2018. A time-varying connectivity analysis from distributed EEG sources: A simulation study. Brain topography 31 (5), 721–737.

Gómez-Herrero, G., Atienza, M., Egiazarian, K., Cantero, J. L., 2008. Measuring directional coupling between EEG sources. Neuroimage 43 (3), 497–508.

Gramfort, A., Papadopoulo, T., Olivi, E., Clerc, M., 2010. OpenMEEG: opensource software for quasistatic bioelectromagnetics. Biomedical engineering online 9 (1), 1–20.

Granger, C. W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, 424–438.

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., Salmelin, R., 2001. Dynamic imaging of coherent sources: studying neural interactions in the human brain. Proceedings of the National Academy of Sciences 98 (2), 694–699.

Grova, C., Daunizeau, J., Lina, J.-M., Bénar, C. G., Benali, H., Gotman, J., 2006. Evaluation of EEG localization methods using realistic simulations of interictal spikes. Neuroimage 29 (3), 734–753.

Habermehl, C., Steinbrink, J. M., Müller, K.-R., Haufe, S., 2014. Optimizing the regularization for image reconstruction of cerebral diffuse optical tomography. Journal of Biomedical Optics 19 (9), 096006.

Halder, T., Talwar, S., Jaiswal, A. K., Banerjee, A., 2019. Quantitative evaluation in estimating sources underlying brain oscillations using current source density methods and beamformer approaches. Eneuro 6 (4).

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., Lounasmaa, O. V., 1993. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of modern Physics 65 (2), 413.

Hashemi, A., Cai, C., Kutyniok, G., Müller, K.-R., Nagarajan, S. S., Haufe, S., 2021. Unification of sparse bayesian learning algorithms for electromagnetic brain imaging with the majorization minimization framework. bioRxiv, 2020–08.

Haufe, S., Ewald, A., 2019. A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies. Brain topography 32 (4), 625–642.

Haufe, S., Nikulin, V. V., Müller, K.-R., Nolte, G., 2013. A critical assessment of connectivity measures for EEG data: a simulation study. Neuroimage 64, 120–133.

Haufe, S., Nikulin, V. V., Nolte, G., 2012. Alleviating the influence of weak data asymmetries on granger-causal analyses. In: International Conference on Latent Variable Analysis and Signal Separation. Springer, pp. 25–33.

Haufe, S., Nikulin, V. V., Ziehe, A., Müller, K.-R., Nolte, G., 2008. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. NeuroImage 42 (2), 726–738.

Haufe, S., Tomioka, R., Dickhaus, T., Sannelli, C., Blankertz, B., Nolte, G., Müller, K.-R., 2011. Large-scale EEG/MEG source localization with spatial flexibility. NeuroImage 54 (2), 851–859.

Hillebrand, A., Barnes, G. R., Bosboom, J. L., Berendse, H. W., Stam, C. J., 2012. Frequency-dependent functional connectivity within resting-state networks: an atlas-based MEG beamformer solution. Neuroimage 59 (4), 3909–3921.

Hincapié, A.-S., Kujala, J., Mattout, J., Daligault, S., Delpuech, C., Mery, D., Cosmelli, D., Jerbi, K., 2016. MEG connectivity and power detections with minimum norm estimates require different regularization parameters. Computational intelligence and neuroscience 2016.

Hincapié, A.-S., Kujala, J., Mattout, J., Pascarella, A., Daligault, S., Delpuech, C., Mery, D., Cosmelli, D., Jerbi, K., 2017. The impact of MEG source reconstruction method on source-space connectivity estimation: a comparison between minimum-norm solution and beamforming. Neuroimage 156, 29–42.

Hipp, J. F., Engel, A. K., Siegel, M., 2011. Oscillatory synchronization in large-scale cortical networks predicts perception. Neuron 69 (2), 387–396.

Hipp, J. F., Hawellek, D. J., Corbetta, M., Siegel, M., Engel, A. K., 2012. Large-scale cortical correlation structure of spontaneous oscillatory activity. Nature neuroscience 15 (6), 884–890.

Huang, M.-X., Song, T., Hagler Jr, D. J., Podgorny, I., Jousmaki, V., Cui, L., Gaa, K., Harrington, D. L., Dale, A. M., Lee, R. R., et al., 2007.

A novel integrated MEG and EEG analysis method for dipolar sources. Neuroimage 37 (3), 731–748.

Huang, Y., Zhang, J., Cui, Y., Yang, G., He, L., Liu, Q., Yin, G., 2017. How different EEG references influence sensor level functional connectivity graphs. Frontiers in neuroscience 11, 368.

Idaji, M. J., Zhang, J., Stephani, T., Nolte, G., Mueller, K.-R., Villringer, A., Nikulin, V., 2021. Harmoni: a method for eliminating spurious interactions due to the harmonic components in neuronal data. bioRxiv.

Jaiswal, A., Nenonen, J., Stenroos, M., Gramfort, A., Dalal, S. S., Westner, B. U., Litvak, V., Mosher, J. C., Schoffelen, J.-M., Witton, C., et al., 2020. Comparison of beamformer implementations for MEG source localization. NeuroImage 216, 116797.

Jirsa, V., Müller, V., 2013. Cross-frequency coupling in real and virtual brain networks. Frontiers in computational neuroscience 7, 78.

Korhonen, O., Palva, S., Palva, J. M., 2014. Sparse weightings for collapsing inverse solutions to cortical parcellations optimize M/EEG source reconstruction accuracy. Journal of neuroscience methods 226, 147–160.

Liuzzi, L., Gascoyne, L. E., Tewarie, P. K., Barratt, E. L., Boto, E., Brookes, M. J., 2017. Optimising experimental design for MEG resting state functional connectivity measurement. Neuroimage 155, 565–576.

Mahjoory, K., Nikulin, V. V., Botrel, L., Linkenkaer-Hansen, K., Fato, M. M., Haufe, S., 2017. Consistency of EEG source localization and connectivity estimates. Neuroimage 152, 590–601.

Marzetti, L., Basti, A., Chella, F., D'Andrea, A., Syrjälä, J., Pizzella, V., 2019. Brain functional connectivity through phase coupling of neuronal oscillations: a perspective from magnetoencephalography. Frontiers in neuroscience 13, 964.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., Lancaster, J., et al., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. Neuroimage 2 (2), 89–101.

Miocinovic, S., de Hemptinne, C., Chen, W., Isbaine, F., Willie, J. T., Ostrem, J. L., Starr, P. A., 2018. Cortical potentials evoked by subthalamic stimulation demonstrate a short latency hyperdirect pathway in humans. Journal of Neuroscience 38 (43), 9129–9141.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M., 2004. Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical neurophysiology 115 (10), 2292–2307.

Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., Müller, K.-R., 2008. Robustly estimating the flow direction of information in complex physical systems. Physical review letters 100 (23), 234101.

Nunez, P. L., Srinivasan, R., Westdorp, A. F., Wijesinghe, R. S., Tucker, D. M., Silberstein, R. B., Cadusch, P. J., 1997. EEG coherency: I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. Electroencephalography and clinical neurophysiology 103 (5), 499–515.

Oswal, A., Beudel, M., Zrinzo, L., Limousin, P., Hariz, M., Foltynie, T., Litvak, V., Brown, P., 2016. Deep brain stimulation modulates synchrony within spatially and spectrally distinct resting state networks in parkinson's disease. Brain 139 (5), 1482–1496.

Palva, J. M., Monto, S., Kulashekhar, S., Palva, S., 2010. Neuronal synchrony reveals working memory networks and predicts individual memory capacity. Proceedings of the National Academy of Sciences 107 (16), 7580–7585.

Palva, J. M., Wang, S. H., Palva, S., Zhigalov, A., Monto, S., Brookes, M. J., Schoffelen, J.-M., Jerbi, K., 2018. Ghost interactions in meg/eeg source space: A note of caution on inter-areal coupling measures. Neuroimage 173, 632–643.

Palva, S., Kulashekhar, S., Hämäläinen, M., Palva, J. M., 2011. Localization of cortical phase and amplitude dynamics during visual working memory encoding and retention. Journal of Neuroscience 31 (13), 5013–5025.

Pascual-Marqui, R. D., 2007. Discrete, 3d distributed, linear imaging methods of electric neuronal activity. part 1: exact, zero error localization. arXiv preprint arXiv:0710.3341.

Pascual-Marqui, R. D., Lehmann, D., Koukkou, M., Kochi, K., Anderer, P., Saletu, B., Tanaka, H., Hirata, K., John, E. R., Prichep, L., Biscay-Lirio, R., Kinoshita, T., 2011. Assessing interactions in the brain with exact low-resolution electromagnetic tomography. Philos Trans A Math Phys Eng Sci 369, 3768–3784.

Pathak, A., Roy, D., Banerjee, A., 2022. Whole-brain network models: From physics to bedside. Frontiers in Computational Neuroscience 16.

Perinelli, A., Assecondi, S., Tagliabue, C. F., Mazza, V., 2022. Power shift and connectivity changes in healthy aging during resting-state EEG. NeuroImage, 119247.

Pizzella, V., Marzetti, L., Della Penna, S., de Pasquale, F., Zappasodi, F., Romani, G. L., 2014. Magnetoencephalography in the study of brain dynamics. Functional neurology 29 (4), 241.

Rubega, M., Carboni, M., Seeber, M., Pascucci, D., Tourbier, S., Toscano, G., Van Mierlo, P., Hagmann, P., Plomp, G., Vulliemoz, S., et al., 2019. Estimating EEG source dipole orientation based on singular-value decomposition for connectivity analysis. Brain topography 32 (4), 704–719.

Sannelli, C., Vidaurre, C., Müller, K.-R., Blankertz, B., 2019. A large scale screening study with a SMR-based BCI: Categorization of BCI users and differences in their SMR activity. PLoS One 14 (1), e0207351.

Sanz Leon, P., Knock, S. A., Woodman, M. M., Domide, L., Mersmann, J., McIntosh, A. R., Jirsa, V., 2013. The virtual brain: a simulator of primate brain network dynamics. Frontiers in neuroinformatics 7, 10.

Schaworonkow, N., Nikulin, V. V., 2021. Is sensor space analysis good enough? spatial patterns as a tool for assessing spatial mixing of EEG/MEG rhythms. bioRxiv.

Schoffelen, J.-M., Gross, J., 2009. Source connectivity analysis with MEG and EEG. Human brain mapping 30 (6), 1857–1865.

Schoffelen, J.-M., Gross, J., 2019. Studying dynamic neural interactions with MEG. Magnetoencephalography: from signals to dynamic cortical networks, 519–541.

Schoffelen, J.-M., Hultén, A., Lam, N., Marquand, A. F., Uddén, J., Hagoort, P., 2017. Frequency-specific directed interactions in the human brain network for language. Proceedings of the National Academy of Sciences 114 (30), 8083–8088.

Shouno, O., Tachibana, Y., Nambu, A., Doya, K., 2017. Computational model of recurrent subthalamo-pallidal circuit for generation of parkinsonian oscillations. Frontiers in neuroanatomy 11, 21.

Silfverhuth, M. J., Hintsala, H., Kortelainen, J., Seppänen, T., 2012. Experimental comparison of connectivity measures with simulated EEG signals. Medical & biological engineering & computing 50 (7), 683–688.

Sivagnanam, S., Majumdar, A., Yoshimoto, K., Astakhov, V., Bandrowski, A. E., Martone, M. E., Carnevale, N. T., et al., 2013. Introducing the neuroscience gateway. IWSG 993, 0.

Sommariva, S., Sorrentino, A., Piana, M., Pizzella, V., Marzetti, L., 2019. A comparative study of the robustness of frequency-domain connectivity measures to finite data length. Brain topography 32 (4), 675–695.

Song, J., Davey, C., Poulsen, C., Luu, P., Turovets, S., Anderson, E., Li, K., Tucker, D., 2015. EEG source localization: sensor density and head surface coverage. Journal of neuroscience methods 256, 9–21.

Supp, G. G., Schlögl, A., Trujillo-Barreto, N., Müller, M. M., Gruber, T., 2007. Directed cortical information flow during human object recognition: analyzing induced EEG gamma-band responses in brain's source space. PloS one 2 (8), e684.

Tabbal, J., Kabbara, A., Yochum, M., Khalil, M., Hassan, M., Benquet, P., 2022. Assessing hd-eeg functional connectivity states using a human brain computational model. Journal of Neural Engineering 19 (5), 056032.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., Leahy, R. M., 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. Computational intelligence and neuroscience 2011.

Van de Steen, F., Faes, L., Karahan, E., Songsiri, J., Valdes-Sosa, P. A., Marinazzo, D., 2019. Critical comments on eeg sensor space dynamical connectivity analysis. Brain topography 32, 643–654.

Van Diessen, E., Numan, T., Van Dellen, E., Van Der Kooi, A., Boersma, M., Hofman, D., Van Lutterveld, R., Van Dijk, B., Van Straaten, E., Hillebrand, A., et al., 2015. Opportunities and methodological challenges in EEG and MEG resting state functional brain network research. Clinical Neurophysiology 126 (8), 1468–1481.

Van Veen, B. D., Van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Transactions on biomedical engineering 44 (9), 867–880.

Vinck, M., Huurdeman, L., Bosman, C. A., Fries, P., Battaglia, F. P., Pennartz, C. M., Tiesinga, P. H., 2015. How to detect the granger-causal flow direction in the presence of additive noise? Neuroimage 108, 301–318.

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., Pennartz, C. M., 2010. The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. Neuroimage 51 (1), 112–122.

Wall, M. E., Rechtsteiner, A., Rocha, L. M., 2003. Singular value decomposition and principal component analysis. In: A practical approach to microarray data analysis. Springer, pp. 91–109.

Wang, H. E., Bénar, C. G., Quilichini, P. P., Friston, K. J., Jirsa, V. K., Bernard, C., 2014. A systematic framework for functional connectivity measures. Frontiers in neuroscience 8, 405.

Wang, S. H., Lobier, M., Siebenhühner, F., Puoliväli, T., Palva, S., Palva, J. M., 2018. Hyperedge bundling: A practical solution to spurious interactions in MEG/EEG source connectivity analyses. NeuroImage 173, 610–622.

Westner, B. U., Dalal, S. S., Gramfort, A., Litvak, V., Mosher, J. C., Oostenveld, R., Schoffelen, J.-M., 2021. A unified view on beamformers for m/EEG source reconstruction. NeuroImage, 118789.

Whittle, P., 1963. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. Biometrika 50 (1-2), 129–134.

Winkler, I., Panknin, D., Bartz, D., Müller, K.-R., Haufe, S., 2016. Validity of time reversal for testing granger causality. IEEE Transactions on Signal Processing 64 (11), 2746–2760.

Wipf, D. P., Owen, J. P., Attias, H. T., Sekihara, K., Nagarajan, S. S., 2010. Robust bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using MEG. NeuroImage 49 (1), 641–655.