Measuring Inter-Brain Synchrony: Methods and Pitfalls

Marten de Vries

Graduation Project

Computational Cognitive Science

University of Groningen

Supervisors: Dr. Marieke van Vugt & Lionel Newman MSc

Abstract

Collecting EEG data for two participants simultaneously during a task (i.e., hyperscanning) allows us to study their social interaction. Of particular interest is their inter-brain synchrony (IBS), i.e. how functionally similar their neural oscillations are. Using a full IBS analysis of a tacit coordination experiment and simulations, we study the effect of different methodological choices in an IBS measurement pipeline. Three ways to quantify IBS are studied: the phase-locking value (PLV), the circular correlation (CCorr) and the imaginary part of coherency (ImagCoh). Each measures functional similarity in its own way and has its advantages and disadvantages. We find the CCorr measure to be less stable than the others, but still recommend its use along with the PLV measure because of its natural interpretation and good performance on simulated data. We present a robust version of the circular correlation measure and make recommendations for how to best perform an IBS analysis.

Keywords: Hyperscanning, Inter-Brain Synchrony, methodology, simulation, Tacit Coordination, Theory of Mind, phase locking value, circular correlation, imaginary part of coherency

Measuring Inter-Brain Synchrony: Methods and Pitfalls

Contents

General Introduction

What happens in the brain when we work together? It is a worthwhile question, as social coordination plays a central role in our everyday lives. A better understanding of it could help us compete and cooperate more efficiently, e.g. during negotiations, pair programming or construction projects. Social coordination is also key in more mundane joint actions. For example, carrying a heavy object together (Sebanz et al., 2006). If we can model what happens in the brain during social interaction mathematically, as proposed by Koike et al. (2015), it could even help us build Human-Computer Interaction systems that better anticipate the user's needs (Zander et al., 2010).

The classical way of researching social coordination is to study a single participant in a lab environment (Babiloni & Astolfi, 2014; Hasson et al., 2012). A good example of this is the false belief task, in which the participant is told a story in which the environment changes without the knowledge of an observer (Postle, 2020, p. 458). Afterwards, the participant is asked about the observer's beliefs on the current state of the environment. It is commonly used to study theory of mind[1], i.e. the ability to anticipate other people's behaviour (Postle, 2020, p. 457). Within this research approach, brain imaging is often used to find the cerebral regions that are involved in performing the task (Babiloni & Astolfi, 2014).

While this classical approach has been very succesful, it has its limits too. Humans are known to behave differently when not interacting with an actual person (Babiloni & Astolfi, 2014; Rilling & Sanfey, 2011; Rilling et al., 2004). The approach also cannot be used to study reactions that arise dynamically, i.e. spontaneously, as a result of information exchanged during the social interaction of interest (Babiloni & Astolfi, 2014; Czeszumski, 2020).

Hyperscanning

More recently, an alternate approach has become popular that solves these issues. Montague (2002) named it 'hyperscanning'. As of May 2022, there are over 3530

[1] See Appendix A for more information on theory of mind.

publications on hyperscanning[2]. Most of those were written very recently: 2040 of them were published in 2018 or later. Hyperscanning is defined as recording brain imaging data for two (or more) participants simultaneously. These two participants are also called a dyad. This allows us to treat the dyad's brains as a single entity 'coupled' through their respective perceptual and motor systems (Hasson et al., 2012).

Social interaction is studied with hyperscanning in a large variety of ways (Czeszumski, 2020). Often, studies are performed in the lab where conditions can be precisely controlled. Some of those only allow interaction through a computer interface, as in the prisoner's dilemma studies of De Vico Fallani et al. (2010) and Hu et al. (2018). Other studies strictly control the task but allow participants to see each other either through video links (Dumas et al., 2010; Schippers et al., 2010) or directly while interacting using gestures (Yun et al., 2012). On the other hand, some studies record participants in a more naturalistic setting for the activities they perform, like in the classroom (Dikker et al., 2017) or the monastery (van Vugt et al., 2020). While having complete control allows for more precise conclusions, Konvalinka and Roepstorff (2012) argue that emergent patterns could arise in the brain as a result of social interaction, leading to a difference in experiments where participants are to some extent observers (Schippers et al., 2010, is a good example, it is difficult to avoid in fMRI studies) compared to where they actively interact. The same can be said about being able to interact in person or through a computer (Konvalinka & Roepstorff, 2012). T. Liu and Pelowski (2014) categorize hyperscanning tasks along three dimensions: whether they require concurrent body movement or the participants interact only in a turn-based fashion, whether participants compete or cooperate and whether the participants can influence each other while the task is ongoing or not. If they can influence each other the task is called interdependent, otherwise it is independent.

Inter-brain synchrony

When analysing brain data acquired using hyperscanning, the most common strategy is to look for inter-brain synchrony (IBS; Ayrolles et al., 2021). IBS occurs

[2] As determined by a Google Scholar search for the term 'hyperscanning'.

when there are functional similarities in the brain activity of individuals. IBS is often found in the brain data of such individuals when they socially interact (Konvalinka & Roepstorff, 2012). While the hyperscanning approach is most often used to collect brain data, synchrony has also been found in other physiological signals including "heart rate[3], pupil size, gaze position and saccade rate" (Madsen & Parra, 2022). Novembre and Iannetti (2021) argue hyperscanning alone cannot tell us whether IBS is required for social interaction or if it just co-occurs with it, but extending the paradigm to include multi-brain stimulation could make that possible.

What exactly causes IBS has not yet been firmly established (D. Liu et al., 2018), but suggested causes include common cognitive processing (Hamilton, 2021; Madsen & Parra, 2022), shared observations (Hamilton, 2021), and more generally shared attention (Dikker et al., 2017; Sebanz et al., 2006). How these processes in turn result in synchronized oscillations is also still mostly unknown (D. Liu et al., 2018), but this too is an area of active research (Hamilton, 2021; Koike et al., 2015).

We measure IBS using different functional connectivity measures, all of which calculate the similarity between the brain signals recorded for both (or more) participants in a specific way (Czeszumski, 2020). These measures were originally developed to study connectivity within a single system or brain (Babiloni & Astolfi, 2014). A nice overview of them from that perspective is given by M. X. Cohen (2014, section 5). As inter-brain data has different properties than intra-brain data, their interpretation when used to calculate IBS instead of intra-brain synchrony is different and complex (Ayrolles et al., 2021). For example, when interpreting intra-brain synchrony, you need to be careful to not interpret a single signal measured at multiple points due to volume conduction as synchrony (Czeszumski, 2020). That is not an issue when the signals are coming from different participants. On the other hand, while intra-brain synchrony is driven by the anatomical structure of the brain (Ayrolles et al., 2021; Dumas et al., 2012), IBS can only occur through "an indirect chain of events" (Babiloni & Astolfi, 2014), as (of course) no direct communication can occur between

[3] See also McCraty (2017).

brains as opposed to brain regions (Babiloni & Astolfi, 2014). IBS is driven by sensorimotor coupling instead, which is a less reliable mechanism (Dumas et al., 2012).

Different measures focus on different kinds of similarities (i.e. different aspects of oscillations) in the brain signals. We will see examples of this in the simulation section. Because of that, Czeszumski (2020) argues that it is misleading to refer to them all with the umbrella term 'inter-brain synchrony'.

It is important to keep in mind that many factors can influence IBS before interpreting its results. Cheng et al. (2015) found an effect of gender: more synchrony was found in male-male dyads than female-male dyads, which in turn had higher IBS than female-female ones. The relationship between the participants is also important. Dikker et al. (2021) found a positive correlation between relationship duration and IBS, and Pan et al. (2017) found more IBS between lovers than friends or strangers. Less IBS has been found in individuals with autism spectrum disorder (ASD; Salmi et al., 2013; Valencia & Froese, 2020).

Due to the properties of the signals of different brain imaging methods, they each have their own classes of IBS measures that are often used alongside them (Babiloni & Astolfi, 2014). For example, when working with EEG hyperscanning data frequency domain-based measures are often used, while temporal correlations are more suited to functional magnetic resonance imaging (fMRI) data (Babiloni & Astolfi, 2014). This is due to the lower temporal resolution of the latter (Czeszumski, 2020).

IBS is often found in interacting partners in "prefrontal and centro-parietal brain areas [...] across a wide range of frequencies, including delta, theta, alpha, beta and gamma" (Konvalinka & Roepstorff, 2012). In *prisoner's dilemma* studies, less synchrony is found in the alpha and theta bands when participants defect than when they cooperate (De Vico Fallani et al., 2010; Hu et al., 2018; Valencia & Froese, 2020). De Vico Fallani et al. additionally found the same effect in the beta and gamma bands, and were able to succesfully predict whether a user will defect in an iterated prisoner dilemma task based on IBS data.

Analysis methodologies

Little research has gone into which methodology to adopt when researching IBS. Simple connectivity measures like the phase locking value (PLV; Lachaux et al., 1999) have been most popular (Burgess, 2013; Czeszumski, 2020). The first systematical comparison of the performance of a number of measures in a hyperscanning context was done by Burgess (2013). Burgess found that a number of measures, including PLV, suffered from detecting spurious connections in simulations. Instead, Burgess recommends using the more robust circular correlation (CCorr) and Kraskov mutual information measures.

Burgess (2013) concludes that “different people presented with the same conditions will produce similar EEG responses”, regardless of whether they were interacting. This can be somewhat mitigated by using measures like the imaginary part of coherency (ImagCoh; Nolte et al., 2004; Yoshinaga et al., 2020), which will ignore signals that are in phase (or anti-phase) with each other (M. X. Cohen, 2014, p. 346). An example of such a signal would be brain activity in the sensory cortex caused by a (strong) stimulus (Dikker et al., 2021). The ImagCoh measure was originally developed to counteract volume conduction, but as this is not an issue with hyperscanning it is useless in that respect (Ayrolles et al., 2021).

Ayrolles et al. (2021) recently made a push for standardization in IBS calculation by making a complete hyperscanning analysis pipeline available. Ayrolles et al. also advise to use amplitude based measures like power correlation when interested in neural states and phase-based measures when studying more fine-grained cognitive processes.

Research questions

The hard work of Ayrolles et al. (2021) and Burgess (2013) notwithstanding, it is clear that only a little is known about the consequences of varying parts of an IBS analysis. The same is true for the interpretation of IBS measures in a hyperscanning context. Because of this gap in the literature, the aim of this graduation project is to investigate the sensitivity of IBS calculations in a social coordination task to different connectivity measures and other methodological choices.

To clarify the interpretation of different IBS measures, we calculate and compare their values on (simple) artificial data. This allows us to see what kind of patterns in the data they respond to. Additionally, we develop a method that generates synthetic data for a given IBS value. This method can be used to perform power analyses for IBS experiments. We focus on the PLV, ImagCoh and CCorr measures.

To investigate the sensitivity to methodological choices under realistic conditions, we perform a number of IBS analyses on a cooperative, turn-based and interdependent task. First, we analyse the effect of different time windows of interest and frequency analysis calculation methods on these values. Second, we analyse whether significant IBS is present in the emperical data using different permutation tests. Third, we inspect IBS values over time. Finally, we attempt to predict peformance in the task based on the IBS values. We vary the prediction scenarios and classification methods.

As most of the variations we make should still lead to the same result, we hypothesize that the analysis is robust to such changes in methodology. Based on Burgess (2013)'s previous work, we expect PLV to perhaps return more spurious results than the CCorr measure. Our expectations regarding the ImagCoh measure are more nuanced. It is both a phase- and amplitude based measure, allowing it to potentially pick up on effects that the other phase-based measures might miss. But it might also miss in-phase IBS detected by the other measures.

General Methods

Data set

The EEG hyperscanning data used for this study was collected by Newman et al. (2021) using two daisy-chained BioSemi ActiveTwo EEG systems. Electrodes were placed according to the international 10–20 system. Additionally, four electrodes were placed surrounding the eyes to monitor eye movements and two were placed on the mastoids to serve as linked reference electrodes. The technical details of the EEG hyperscanning setup were as described by Barraza et al. (2019). Data was collected for 42 sessions, but 38 of those are analyzed here as during four sessions recording issues were encountered.

EEG pre-processing

For each of the participants in the dyads, data was re-referenced to the average of the mastoid electrodes and band-pass filtered to remove parts of the signals with a frequency lower than 0.1 Hz and higher than 50 Hz. To prevent edge artifacts one minute of padding consisting of mirrored data was added for the duration of the filtering process. Next, the recorded data was split up into trials which start one second before and end 1.5 seconds after stimulus presentation. The pre-stimulus period was used for baseline correction. At this stage, any linear trends were also removed for each trial. This removed slow low-frequency drifts from the data, which can otherwise show up as artifacts during frequency analysis (Schoffelen, 2010).

Table 1

Trial counts after pre-processing and how often they occur. For most sessions, a maximum of 10 trials were removed. The outlier of 84 trials is session 25.

trial count (1)	84	144	152	157	163	164	165	166	169	170
frequency (1)	1	1	1	1	1	1	1	1	1	2
trial count (2)	171	172	173	174	175	177	178	179	180	
frequency (2)	1	3	1	4	4	1	2	3	8	

Each recorded trial was manually inspected. When electrodes did not make a good connection to the skin or otherwise regularly produced unusable data, they were removed from the data set and reconstructed using spline interpolation from neighbouring electrodes. If an electrode drifted or was very noisy for only one or a few trials, it was interpolated in these trials only. If too many neighbouring electrodes were affected, interpolation became impossible and instead the whole trial was rejected. Trials were also rejected when they contained more than four electrode signals that required interpolation. See for how many trials remained Table 1. Eye blinks, muscle activity and other localized artifacts were left alone at this stage. Instead, they were handled by subtracting highly localized and artifactual components from the data as obtained using independent component analysis (ICA).

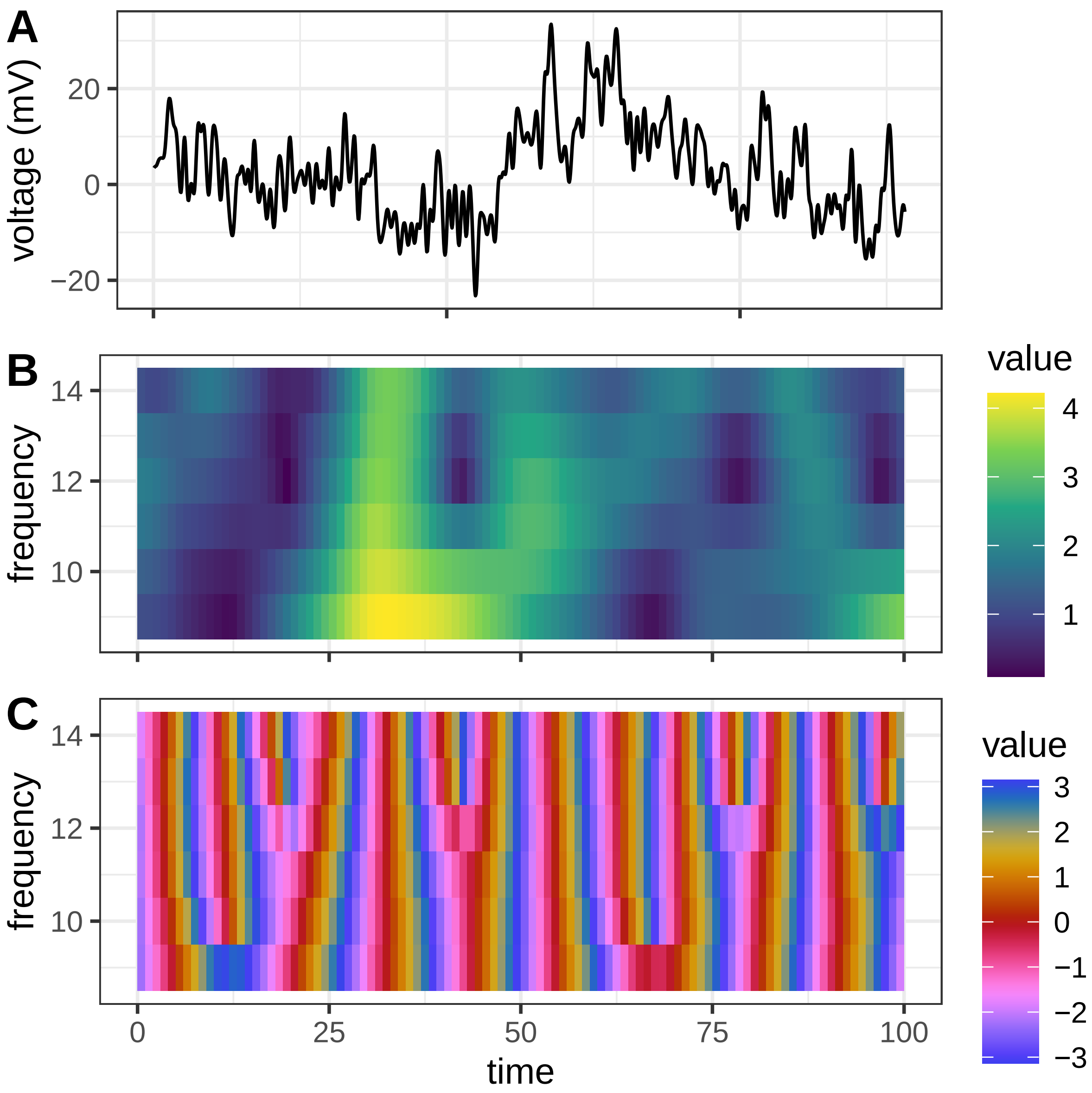


Figure 1

Time-frequency plot showing (B) amplitude and (C) phase values. Example frequency domain representation of session 2, participant 1, trial 1, Pz electrode, alpha band (9–14 Hz). The raw data used for the frequency analysis of this trial is shown in (A).

Time-frequency analysis

The first step to calculating the phase locking value (PLV), circular correlation (CCorr) and imaginary part of coherency (ImagCoh) measures consists of transforming the data into the frequency domain. We do so once every 10 milliseconds for both the alpha band (9–14 Hz) and theta band (4–7 Hz). For most of the analysis, we use a Hann taper with a frequency dependent window length of four cycles per window. At least four cycles are recommended by Ayrolles et al. (2021). For the lowest frequency of interest (4 Hz), this results in a window of exactly one second. As a result, the amplitude and phase of a signal can only be estimated for moments during the trial where half a second of extra data is available before and after. Because of that, we narrow the duration of a trial for the purposes of inter-brain synchrony (IBS) calculation from zero to one second after stimulus presentation exactly.

The result of the frequency analysis is a complex valued Fourier spectrum x_i for each combination of participant, electrode and frequency in the frequency band. From this spectrum, the amplitudes r_i and phases ϕ_i of the input signal can be extracted by representing the complex values in polar coordinates. See Figure 1 for a visualization of r_i and ϕ_i of an example signal.

Measure definitions

We compare the signals of homologous electrodes between participants, e.g. we only compare the signal of the Fz electrode of participant 1 with participant 2's Fz electrode, not with other electrodes. While it is technically possible to do otherwise, we would lose the ability to conveniently interpret high IBS as suggestive of similar mental processes in both participants.

The PLV measures whether the difference in the phase of two signals is kept constant. The PLV measure is defined by Lachaux et al. (1999) as

$$\mathrm{PLV} = \frac{1}{T}\left|\sum_{t=1}^{T} e^{i(\phi_i - \phi_j}\right|, \tag{1}$$

where ϕ_i and ϕ_j are the phases of input frequency spectra x_i and x_j of the different participants. As you can see, the PLV measure is calculated by averaging along a dimension of size T. For this analysis, we will be averaging over time resulting in one measurement per trial and frequency. It is also possible to calculate PLV and the other measures discussed in the current study over trials instead, resulting in a measure of within-trial IBS. But within the context of Newman et al. (2021)'s experiment, how IBS develops over trials is much more interesting. (See Appendix B for more information about Newman et al. (2021)'s task.)

While we calculate measures (including PLV) for each whole number frequency within the frequency band of interest, we are not interested in their differences within the same band. Instead, we average these values resulting in a single measure per trial and frequency band. This should contribute to a more stable estimate. Our PLV implementation was validated against Fieldtrip's implementation.

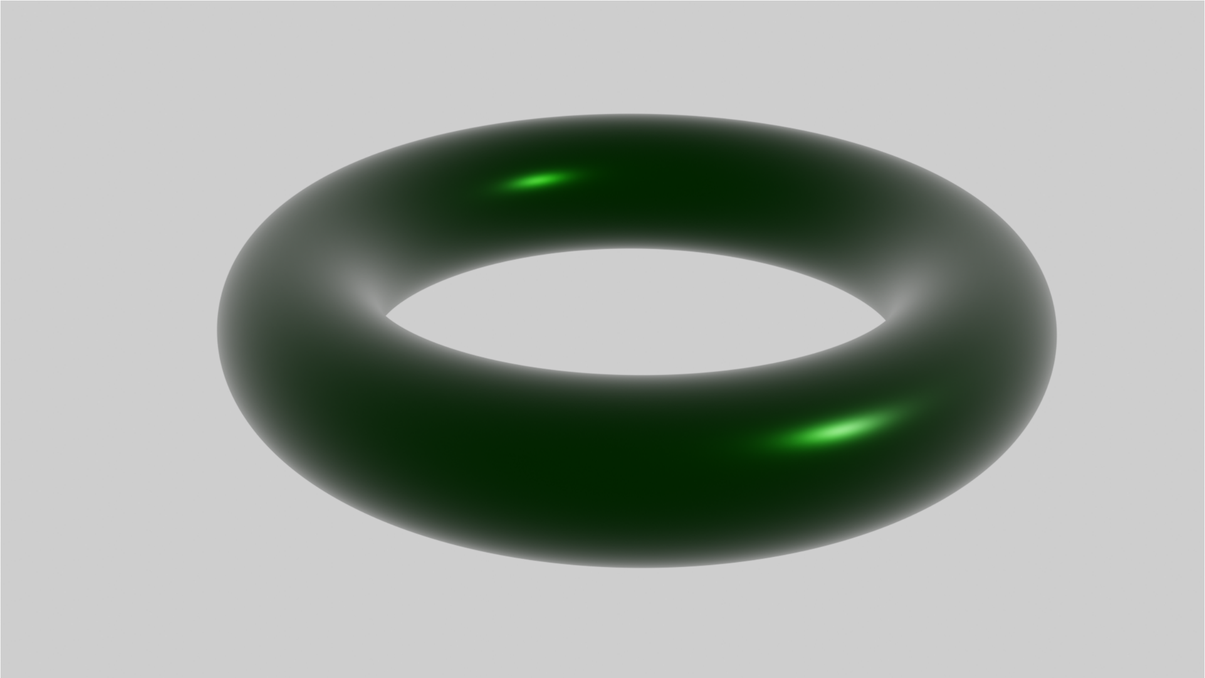


Figure 2

A torus. Bivariate circular data can be thought of as points on a torus. Rendered using Blender (Blender Online Community, 2022).

CCorr is an analogue of Pearson correlation for angular values like a phase. It seems to have been first derived by N. I. Fisher and Lee (1983), but most recent implementations are based on the definition by S. Jammalamadaka and Sarma (1988) which has more recently been republished in a book (S. R. Jammalamadaka &

Sengupta, 2001):

$$\mathrm{CCorr} = \frac{\sum_{t=1}^{T} \sin\left(\phi_i - \bar{\phi_i}\right) \sin\left(\phi_j - \bar{\phi_j}\right)}{\sqrt{\sum_{t=1}^{T} \sin^2\left(\phi_i - \bar{\phi_i}\right) \sin^2\left(\phi_j - \bar{\phi_j}\right)}}. \tag{2}$$

Within this equation, $\bar{\phi_i}$ is the circular mean which can be defined as

$$\bar{\phi_i} = \arg \sum_{t=1}^{T} e^{\phi_t i}, \tag{3}$$

where 'arg' gives us the angle we get when converting the sum to polar form. When interpreting normal Pearson correlation coefficients, I often imagine plotting the two variables of interest against each other. The coefficient then tells us how close the data points are to lying on a line. With CCorr values, it is possible to do the same, but instead of a normal plot you should imagine the data points existing on a torus (Lee, 2010, see also Figure 2). Our implementation of the CCorr measure was inspired by and validated against the implementation in the CircStat MATLAB toolbox (Berens, 2009).

Finally, the ImagCoh measure looks not just at the phase of signals but also takes into account the amplitude. It is defined by Nolte et al. (2004) as

$$\mathrm{ImagCoh} = \mathrm{Im}\left(\frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}\right), \tag{4}$$

where S_{ij} is the crossspectral density of the signals and S_{ii} and S_{jj} are the autospectral densities of each of the signals. These spectral densities are estimated directly from the Fourier transformed data x_i and x_j.

$$S_{ij} = \frac{1}{T} \sum_{t=1}^{T} x_i {x_j}^* \text{ (Schoffelen, 2011)}, \tag{5}$$

where ${x_j}^*$ is the complex conjugate of x_j. Our implementation of the ImagCoh measure was validated against Fieldtrip's.

Statistics

We use linear mixed effect models with random intercepts over sessions and electrodes. More complex random effect structures are not supported by the data, and in some cases including a random intercept for electrodes is not either when the data is too homogeneous across electrodes. In that case, said random intercept is left out. Next to any fixed effects of interest, fixed effects of working memory load, trial and stimulus type were included if they significantly contributed to the model to account for possibly confounding effects of those. The models underlying the model comparisons referred to in the text are reproduced in Appendix C.

Software

All manipulations of the EEG data were performed with Fieldtrip version 20211102 (Oostenveld et al., 2011) running on MATLAB R2020b. All IBS measures were implemented from scratch in both MATLAB for use in the empirical study and R 4.2.0 (R Core Team, 2022) for use in the simulation study. Graphs were generated in R using tidyverse 1.3.2 (Wickham et al., 2019), eegUtils 0.7.0 (Craddock, 2022), gganimate 1.0.7 (Pedersen & Robinson, 2020), ggh4x 0.2.3 (van den Brand, 2022), ggpubr 0.4.0 (Kassambara, 2020), ggvoronoi 0.8.5 (Garrett et al., 2022) and pals 1.7 (Wright, 2021). All statistical tests were performed using R as well, using lme4 1.1.29 (Bates et al., 2015) for the linear mixed effect models. For the generalized additive mixed effect models mgcv 1.8.41 (Wood, 2006) was used alongside itsadug 2.4 (van Rij et al., 2020) for plotting those models. To generate simulated correlated data, faux 1.1.0 (DeBruine, 2021) was used. Finally, Python 3.10.8 (Python Software Foundation, 2021) was used to train and evaluate classifiers, along with imbalanced-learn 0.1.9 (Lemaître et al., 2017), NumPy 1.23.3 (Harris et al., 2020), pandas 1.5.0 (McKinney, 2010), scikit-learn 1.1.2 (Pedregosa et al., 2011) and SciPy 1.9.1 (Virtanen et al., 2020).

Parts of the R and MATLAB code used in this study are available at https://doi.org/10.5281/zenodo.7469929.

Simulation study

Introduction

While the phase locking value (PLV), circular correlation (CCorr) and imaginary part of coherency (ImagCoh) measures have been introduced mathematically, it can be difficult to understand what a certain inter-brain synchrony (IBS) measure value says about the underlying data. We attempt to shine some light on the matter using a simulation study.

We introduce a visualization that shows the relation between the phase components of two EEG signals. For our purposes, one signal from each participant in the dyad. We then apply this visualization method to simulated phase data examples. The examples have been chosen such that they result in a large range of IBS values. This allows us to see what patterns in the phase data the IBS measures detect, and as a result what relations between the two signals they are sensitive to (or not).

In these simulations we ignore the amplitude component. This has two reasons: it is only actually used by the ImagCoh measure, and it would make it harder to draw any conclusions as it would require more complex visualizations. Finally, we explain how our approach can be used to perform a power analysis for IBS experiments.

Methods

As discussed in the general introduction, comparing the relation between two phase signals requires a space that wraps around in two dimensions. One solution to this problem is to plot data on a torus (see Figure 2), but that is not practical for a two-dimensional medium like this report. It would also result in deformed distances. Instead, we use normal plots, but visualize the data for 1.5 periods, resulting in a repeated (and differently shaded) area at the visualization's edges. This makes it easier to imagine the circular repetition of the data and to spot patterns that would otherwise span the edges of the plotting area.

We focus on a single session, trial, electrode and frequency at a time. As we have previusly seen in Figure 1C, this results in a one-dimensional phase signal over time for each participant (ϕ and ψ respectively). We reproduce these signals in Figure 3A, using the y-axis instead of colour to show their values. If we then get rid of the time dimension, as all discussed IBS measures do, we can give each phase signal its own axis in Figure 3B to make it easier to see the relation between the two. The resulting plot demonstrates the primary visualization method used in this study.

The easiest way to simulate this (empirical) phase data is to sample 100 points from a two-dimensional uniform random distribution with a range of $[-\pi, \pi)$. We can then calculate the IBS measures on each of these samples. For the imaginary part of

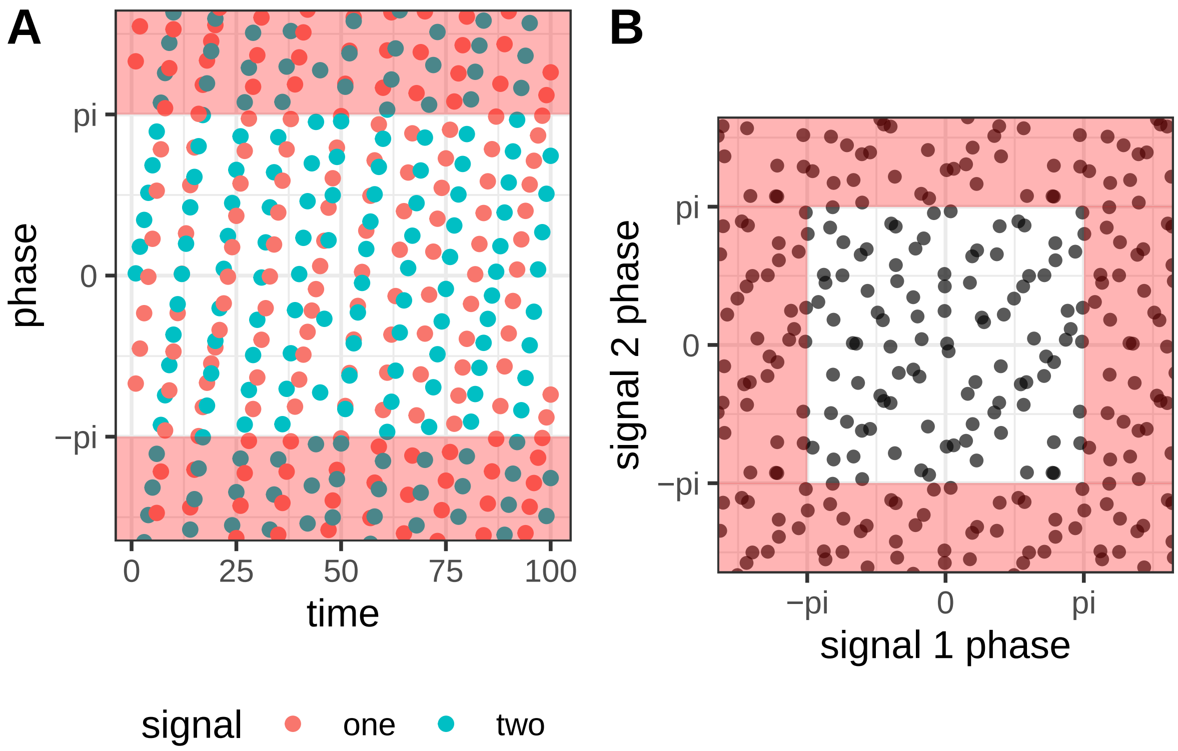


Figure 3

Phase values of both participants in session 2, trial 1 for the Pz electrode at 11Hz. (The trial from Figure 1.) For this trial, circular correlation = -0.01, phase locking value = 0.033 and imaginary part of coherency = 0.012 (ignoring amplitude values). Shaded areas contain repetitions of the (circular) values. (A): Phase values over time. (B): The same two signals plotted against each other, without time, to show the relation between them. This type of visualization is used throughout this simulation study.

coherency measure, we use a fixed amplitude of one. When we repeat this process 10 000 times, it results in a distribution of values for each IBS measure created under the assumption that there is no relation between the two signals.

While that is useful, completely random signals are unlikely to elicit the full range of IBS values. To be able to generate signals for any measure value, we created Algorithm 1.

Algorithm 1 finds example signals that minimize an evaluation function $f(\phi, \psi)$. It uses a combination of global and local search. The global search part of the process generates multiple random signals and picks the ones that minimizes the evaluation

Algorithm 1 Generates random phase data examples that minimize an evaluation function f using a combination of global and local search.

```
Require: $f(\phi, \psi) \to \mathbb{R}$                    ▷ the evaluation function to minimize
  repetitions ← the amount of global search iterations
  start ← the initial amount of data points in the sample (should be $\geq$ end)
  end ← the amount of data points in the sample after local search finishes

  best ← $\infty$
  result ← [ ]
  for 'repetitions' amount of iterations do                    ▷ the global search part
      $\phi, \psi$ ← a 2D uniform random sample of length ‘start’ and range $[-\pi, \pi)$
      while cur > end do                                       ▷ the local search part
          for $i \in 1 \ldots$ length of $\phi$ do
              $\phi'_i, \psi_i$ ← $\phi, \psi$ without the $i$th values
              $e_i \leftarrow f(\phi'_i, \psi'_i)$
          end for
          $i \leftarrow \text{argmin}(e_i)$
          $\phi, \psi \leftarrow \phi'_i, \psi'_i$
      end while
      if $f(\phi, \psi) <$ best then
          best ← $f(\phi, \psi)$
          result ← $[\phi\ \psi]$
      end if
  end for
  return result
```

function. The local search part of the process optimizes each candidate before evaluation by removing a number of 'outliers' (as determined by the evaluation function), one at a time. By varying the input parameters, it is possible to trade-off between the (unbiased, but unlikely to cover the whole range) global search process and the (biased, but more flexible) local search process.

We apply Algorithm 1 in two tasks.

First, we use it to generate examples that have CCorr and ImagCoh values close to $-1, -0.75, \ldots, 0.75$ and 1. And the same for the PLV values $0, 0.125, \ldots, 0.875$ and 1. To approximate values, we use an L2 loss function as the evaluation function. E.g. to get a CCorr value of -0.75 the evaluation function is

$$f(\phi, \psi) = \left(CCorr(\phi, \psi) - (-0.75)\right)^2, \tag{6}$$

where $CCorr$ is as defined in Equation 2.

Secondly, to further contrain the examples and to see where the IBS measures differ, we use evaluation functions that constrain the measures in different ways simultaneously. For example, when minimizing the evaluation function

$$f(\phi, \psi) = 1 - CCorr(\phi, \psi) + |ImagCoh(\phi, \psi)| + PLV(\phi, \psi), \tag{7}$$

we obtain example signals with a positive CCorr value, a (close to) zero ImagCoh value and a low PLV value. We generate example signals for all possible constraint permutations. The PLV and $ImaghCoh$ definitions are given in Equations 1 and 4.

Results

In Figure 4, we see the distribution of IBS values when the underlying signals are completely random. We see that the distributions for the CCorr and ImagCoh values are symmetric and centered around their middle value of zero. We observe a skewed distribution for the PLV measure values, most likely because the values are all close to the minimum value the measure can take (i.e. zero). It will be virtually impossible to distinguish an effect that has an IBS value in the high density part of the shown distributions from noise.

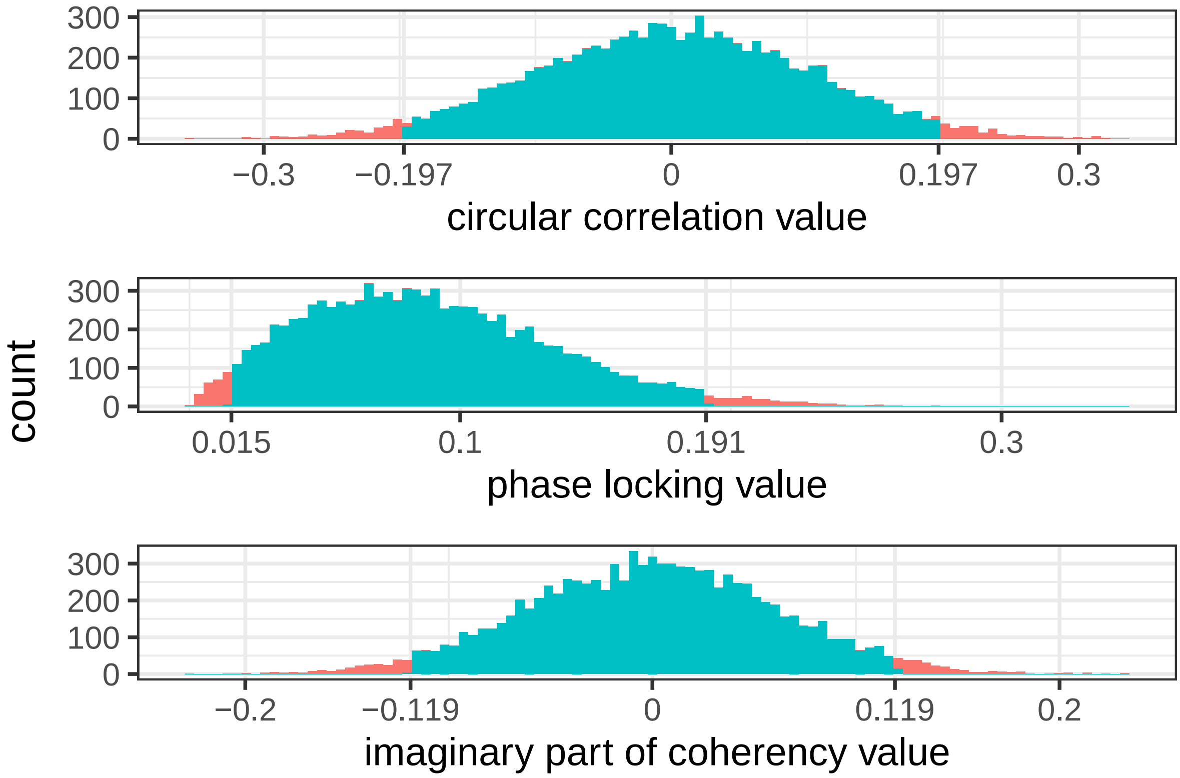


Figure 4

A histogram of inter-brain synchrony values calculated on 10 000 pairs of uniform random signals. It shows inter-brain synchrony values typical for when the underlying signals are unrelated. The central 95% of the data is shown in blue.

In Figure 5, we see the phase components of example signals for the whole range of CCorr, ImagCoh and PLV measure values. Finding ImagCoh examples was harder than finding examples for the other measures, resulting in different input parameters for Algorithm 1 being required to cover the whole range. These parameters can be found in Table 2.

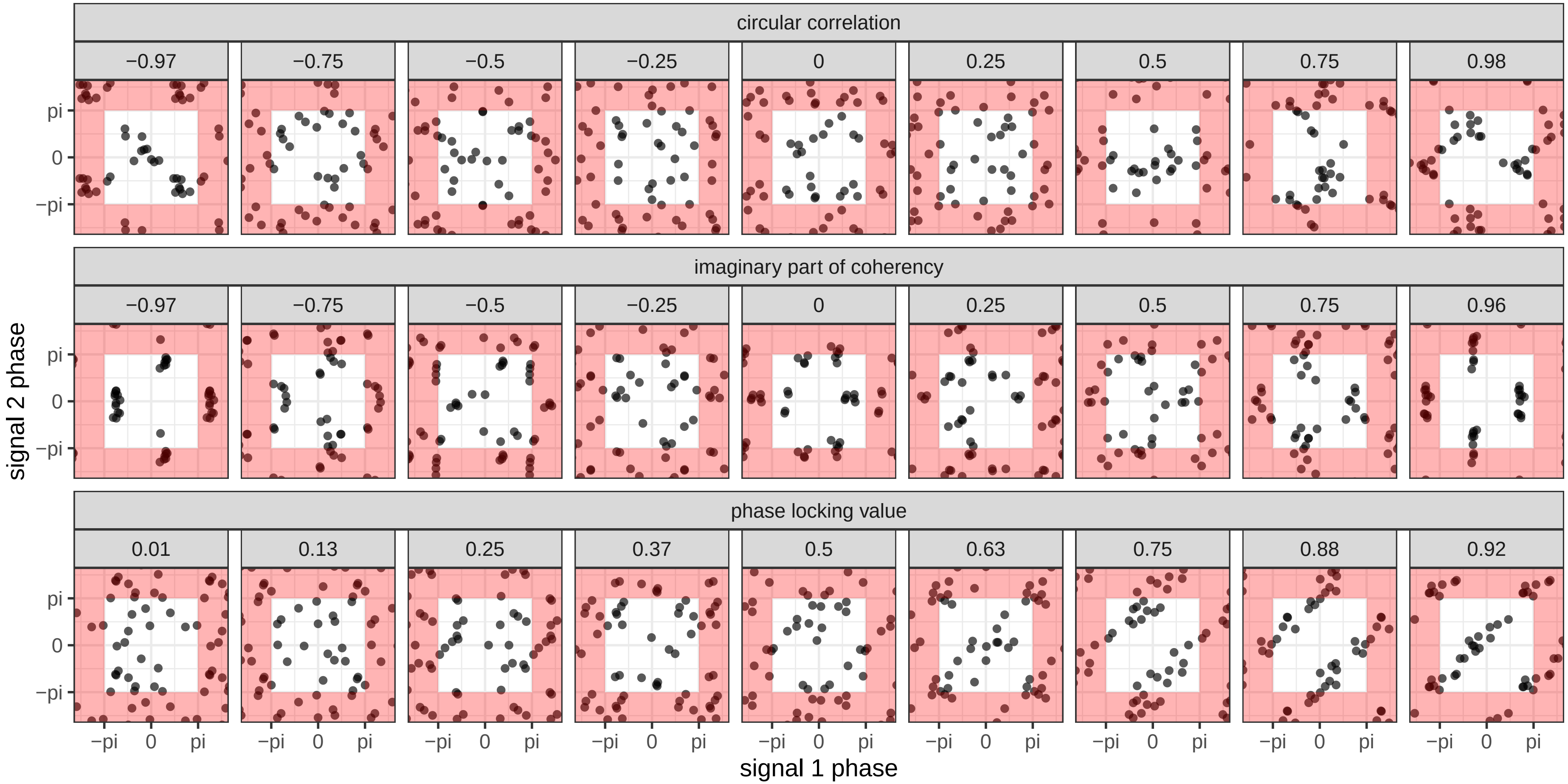


Figure 5

Simulated examples for the whole range of inter-brain synchrony measure values. As in Figure 3, shaded areas contain repetitions of the (circular) values. A high phase locking value requires a positive linear relation between signals. For a high circular correlation or imaginary part of coherency, multiple clumps suffice. But interestingly, while the simulation finds random noise examples for PLV and CCorr values of zero, the ImagCoh is assigned a clumped example instead.

Table 2

Parameter values of Algorithm 1 used to generate Figure 5.

Parameter	circular correlation & phase locking value	imaginary part of coherency
repetitions	1000	20
start	40	200
end	(/2 =) 20	(/10 =) 20

The examples show a few clear trends. First of all, an increase in PLV seems to lead to a more positive and linear relation between the phase components of the example signals. For the other measures, higher (and lower) IBS values seem to lead to the formation of clumps, i.e. patterns where a lot of the phase components are approximately constant. Surprisingly, while the other measures seem to converge on a (at first glance) random noise example for IBS values of zero, which is in line with our findings in Figure 4, this is not the case for the ImagCoh measure. There, the central example is clumped just like the examples at the tails.

To see whether these examples are typical or just the first configuration Algorithm 1 finds, we further constrain the examples by forcing them into configurations that contrast the IBS measure values. Figure 6 shows these examples. As this optimization problem is harder, an optimal solution is not always found. Because of that, the size of the error is shown as well, which is at the same scale as the IBS values themselves. While the total error occasionally surpasses 0.5, the error is in practise divided up among measures. So while the examples might not match the target IBS values exactly, they are never so far off as to become misleading.

To generate Figure 6, 100 global search repetitions were used. The local search started with 100 samples, and reduced that to 20 samples.

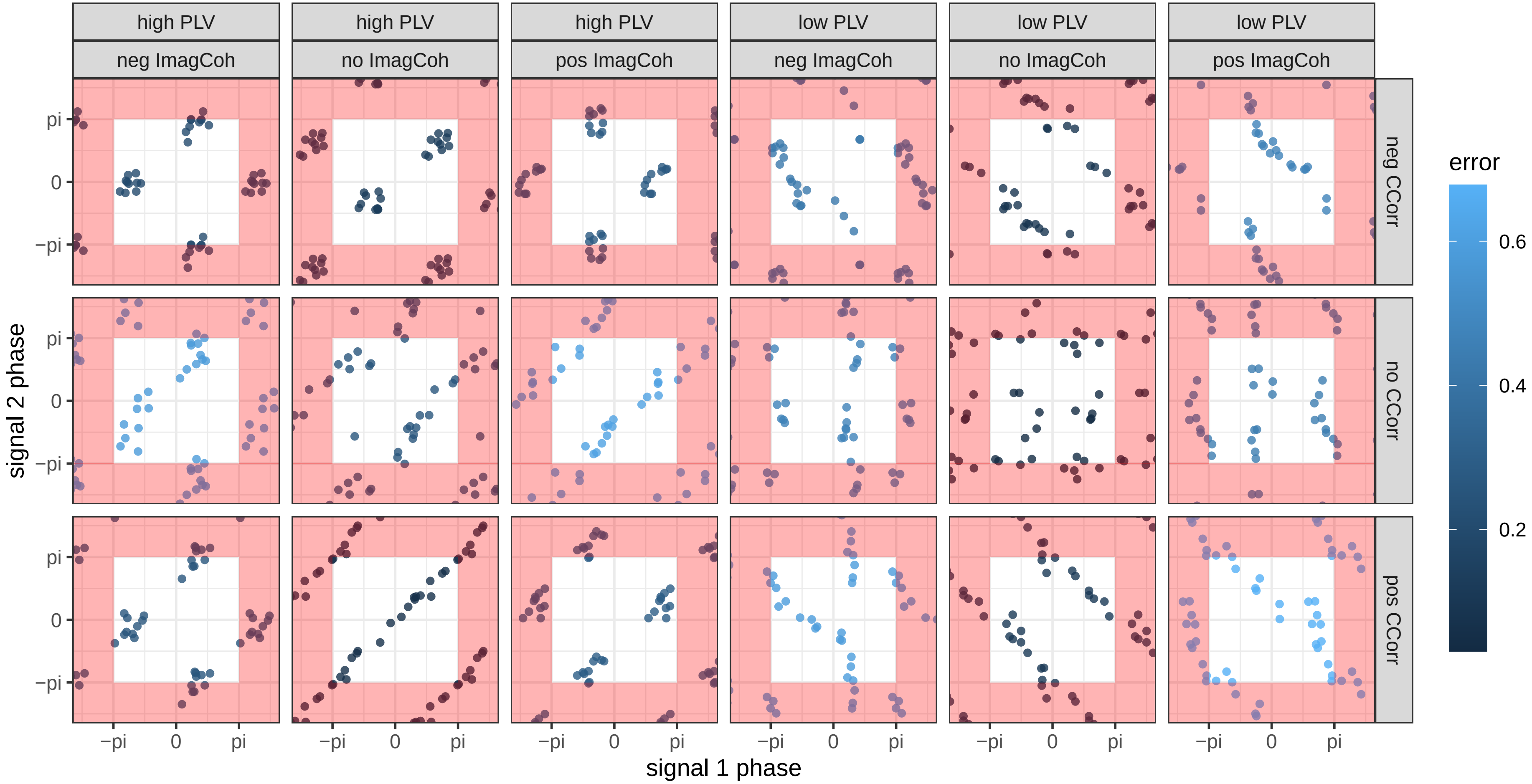


Figure 6

Simulated examples that minimize (low phase locking value, negative imaginary part of coherency, negative circular correlation), maximize (high PLV, positive ImagCoh, positive CCorr) or zero out (no CCorr, no ImagCoh) IBS values simultaneuously. Light blue dots indicate the example is not perfect (fulfilling the ImagCoh requirement is often most difficult), while black dots indicate all constraints were met perfectly.

We see that PLV is not sensitive to negative linear trends (as opposed to positive ones) at all. This is exploited in the right side of Figure 6, as depending on the exact configuration the CCorr and ImagCoh measures are sensitive to negative trends. We further see confirmation that the imaginary part of coherency can reach a value of zero just fine with an (apparently) random example (the 'low PLV, no ImagCoh, no CCorr' condition). Also, it is interesting to point out that the ImagCoh is sensitive to the phase components of signals being constant in one dimension but not in the other, contrary to the CCorr. (See e.g. the 'low PLV, no CCorr' conditions with negative or positive ImagCoh values). Next, it is worth noting that a configuration with a seemingly negative trend ('low PLV, no ImagCoh, pos CCorr') results in a positive CCorr value. Finally, to the eye immediately apparent trends are not always picked up on by the ImagCoh measure. (E.g. the two 'pos CCorr, no ImagCoh' conditions.)

Discussion

The lack of response of the PLV measure to a clear negative relation between the phase components of the input signals is not unexpected, as it only reports whether the phases are directly coupled, not if one of them can be used to predict the other (Burgess, 2013). But it is a downside, as you would most likely want to detect such effects in IBS experiments.

Similarly, we saw that the ImagCoh measure failed to sometimes detect trends. That might be because it is designed not to detect signals that are perfectly in phase, as a way to (originally) prevent spurious effects due to volume conduction (Nolte et al., 2004). But these simulations suggest that the cost of that might be too high. On the other hand, it is important to keep in mind that these simulations are not a level playing field for the ImagCoh measure: amplitude components of the signal on which it is normally dependent are held constant.

In the end, the CCorr values seem the least surprising given the studied examples, although the direction of any relations (i.e. whether the correlation coefficient is positive or negative) should probably not be relied on.

Power analysis for inter-brain synchrony experiments

Taking a step back, it is worth pointing out that Algorithm 1 is a very flexible method to generate phase component data for a target IBS value. It could potentially be used to perform an up-front power analysis for a test in an IBS experiment. The steps would be as follows:

1. Choose a target effect size. That is, what IBS value would you expect your experiment to find? The simulations in this section give you some guidance on what would be reasonable values, but ideally the target value would be decided based on what other similar studies found.

2. For a range of sample sizes, repeatedly simulate the outcome of your test using the Monte Carlo method (P. R. Cohen, 1995, p. 150 gives a nice introduction) as follows:

   (a) Use Algorithm 1 to generate fake trials for your IBS value of choice. By varying the trade-off between global- and local search part of the algorithm, you have some control over the variation around your target IBS value. Ideally, you would again use this to match the variation found in other similar studies.

   (b) Perform your test on the simulated data, recording the outcome.

3. Use the collected outcomes to estimate the power of your test for each sample size.

A downside of this method, and in fact of this simulation study in general, is that the local search part of Algorithm 1 introduces a bias due to the way it removes outliers. After all, removing them one at a time is just one of many possible approaches. While the results seem reasonable looking at the graphs, it could be that some examples we observe are in fact not typical but artifacts of the process used to generate them. This could for example perhaps explain the clumps in the ImagCoh plot in the very middle of Figure 5.

Varying time-frequency analysis methods

Introduction

Inter-brain synchrony (IBS) values are calculated on a frequency domain representation of the original signals. Obtaining this frequency domain representation requires making some methodological choices: you need to choose a window of interest, a calculation method and a resolution. We assess how these choices affect the final IBS values.

Ideally, the IBS values should be robust to slight changes in these parameters. For the window of interest case, IBS values could presumably vary a bit as the window could include or exclude cognitive processes that take a while to start after the stimulus. But the other parameters are just technical details of the time-frequency analysis process and should not have a big effect on the IBS values when reasonably chosen.

Methods

To assess the effect of different windows of interest on the results, we repeat the main frequency analysis using a window of half a second before to one second after stimulus presentation, making full use of each available pre-processed data point. To determine the effect of different frequency analysis methods on the final results, we also repeat the (alpha band) analysis using multitapers instead of a Hann taper. To find whether there is an effect of resolution, we perform the frequency analysis both more often (for each original data point, i.e. 512 times per second or approximately every 2 ms) and less often (once every 20 ms). None of these variations are extreme, and all could have been reasonably chosen for the main experiment instead.

For each of these variations, we first compare the averaged IBS values by plotting the data and assess the significance using a linear mixed effect model. If those do not show a difference at first glance, we further assess possible differences in the underlying data structure by calculating correlations between the values for different conditions and (where necessary) by plotting the data. Correlations are calculated for each session, Fisher transformed (R. A. Fisher, 1915), averaged and transformed back.

Results

Frequency analysis window size

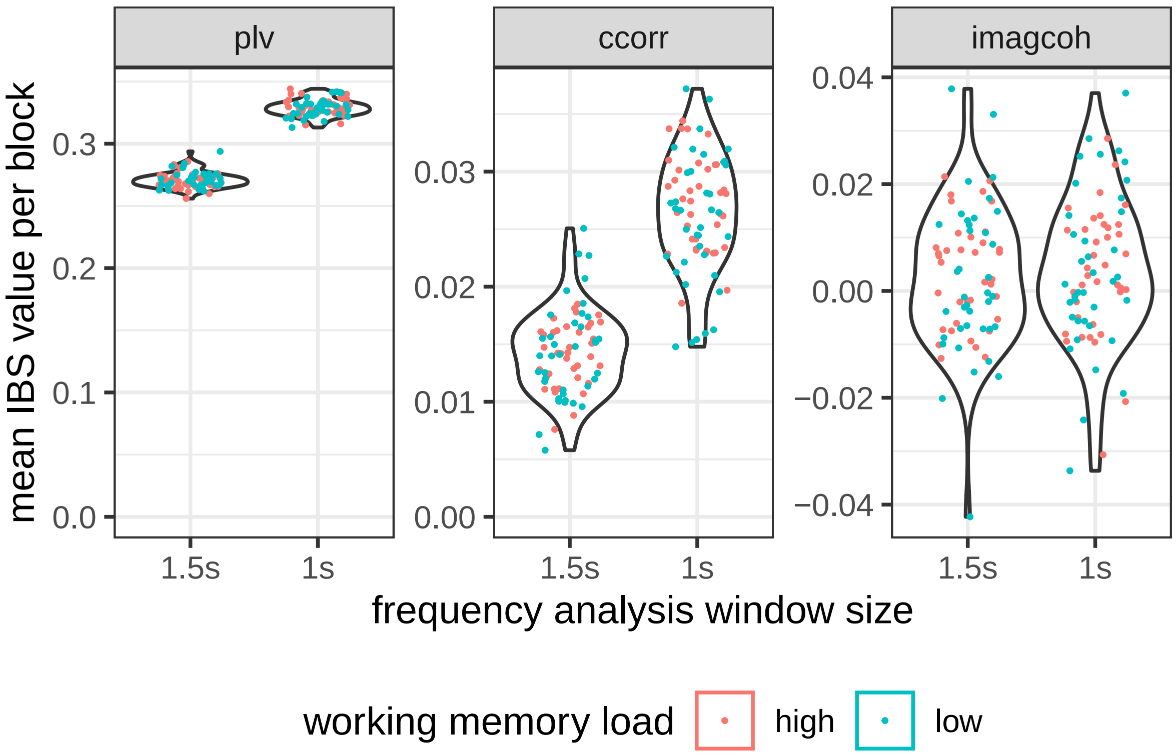


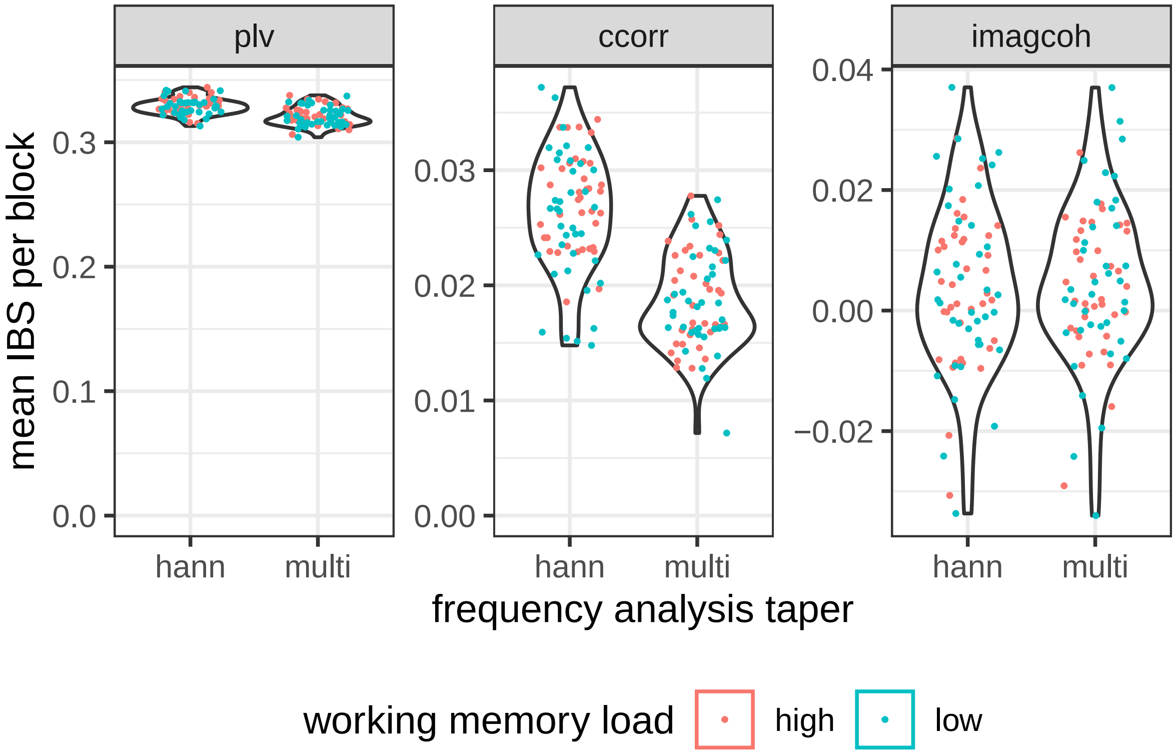
Figure 7

(Mean) alpha band inter-brain synchrony values are sensitive to different time windows of interest within a trial. The 1s window starts at the presentation of the stimuli, while the 1.5s window starts half a second earlier during fixation. Working memory load has no effect. Finally, the (mean) circular correlation and imaginary part of coherency values are both very close to zero considering they are on a scale from -1 to 1. The phase locking value is on a 0-1 scale.

We tested the effect of window size on IBS values by calculating the IBS measures on overlapping windows of 1s and 1.5s respectively. Contrary to our initial expectations, we found IBS values to be sensitive to changes in frequency analysis window size. See Figure 7. The phase locking value (PLV) significantly decreased when the larger window was used ($\chi^2(1) = 29984$, p $<$ 0.001, ΔAIC = 29982, ΔBIC = 29971) and so did the circular correlation (CCorr; $\chi^2(1) = 1071$, p $<$ 0.001, ΔAIC = 1069,

ΔBIC = 1058), but no significant effect was found for the imaginary part of coherency values (ImagCoh; $\chi^2(1) = 2.27$, n.s., ΔAIC = 0.27, ΔBIC = 10.66).

Frequency analysis taper



s

Figure 8

Choice of taper matters when calculating (mean) 'phase locking value', 'circular correlation' and 'imaginary part of coherency' inter-brain synchrony values for the alpha band. Working memory load has no influence.

To see the effect of taper choice on the frequency analysis, we compared the IBS values obtained from spectra generated using a Hann taper and a multitaper. As you can see in Figure 8, there is an effect of taper choice on PLV synchrony values ($\chi^2(1) = 626$, $p < 0.001$, ΔAIC = 624, ΔBIC = 614) and CCorr synchony values. ($\chi^2(1) = 452$, $p < 0.001$, ΔAIC = 450, ΔBIC = 440). In both cases, the IBS value decreases a bit when multitapers are used. Again, there is no significant effect on ImagCoh synchrony values ($\chi^2(1) = 0.51$, n.s., ΔAIC = 1.49, ΔBIC = 12.4). When we compare how values calculated using the different methods correlate (Figure 9), we again see that IBS value calculation is sensitive to choice of taper contrary to our hypothesis. The CCorr measure

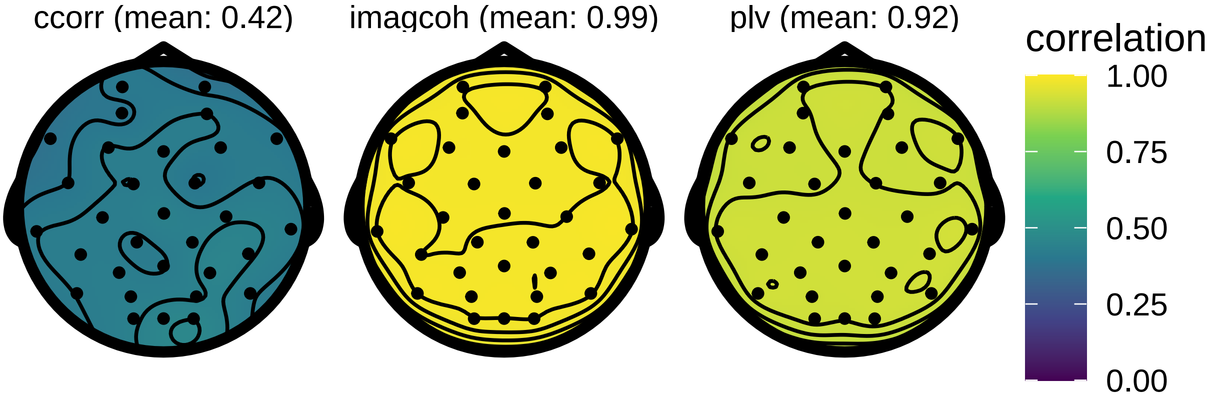


Figure 9

When calculating phase locking value and circular correlation measures, the choice of taper alone can lead to inter-brain synchrony values different enough that they do not perfectly correlate with each other. Imaginary part of coherency values are less affected than circular correlations and phase locking values.

is especially affected with a mean correlation of 0.42 across sessions and electrodes.

Frequency analysis resolution

We now turn to a less discussed parameter of the frequency analysis: the resolution of the resulting spectrum. At first sight, there does not appear to be an effect of resolution on IBS values (Figure 10). When using the resolution as a continuous parameter ($\frac{1000}{512}$ ms, 10 ms or 20 ms), statistics confirm this for the PLV ($\chi^2(1) = 0.145$, n.s., ΔAIC = 1.85, ΔBIC = 13.2) and ImagCoh ($\chi^2(1) = 0$, n.s., ΔAIC = 2.0, ΔBIC = 13.3) measures, but report a significant positive (though small) effect of resolution on CCorr synchrony values ($\chi^2(1) = 7.4$, $p < 0.01$, ΔAIC = 5.4, ΔBIC = -5.9).

Looking into it further, we see that CCorr values in fact correlate much worse across resolutions than the other measures (see Figure 11). This is unexpected when varying such a ‘boring’ parameter as resolution, which you normally do not think twice about when choosing it.

Discussion

An effect of window size was found for the CCorr and PLV measures, but not for the ImagCoh measure. In hindsight, the effect of window size is not that surprising, as

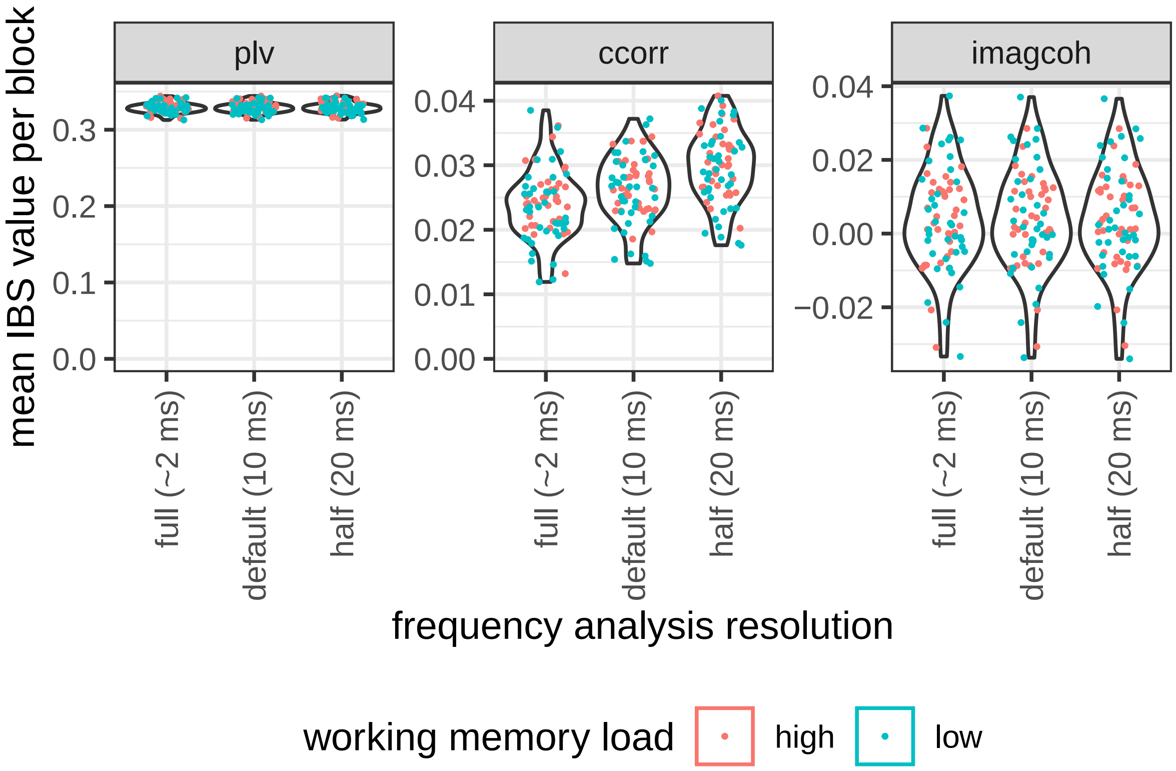


Figure 10

Mean inter-brain synchrony values do not appear to vary much with resolution, with the exception of a slight effect on circular correlation values.

all measure definitions weigh each data point equally and one third of the data is new for the 1.5s window of interest. Still, as all IBS present in the short window is also present in the longer window, we would not expect any effects to change direction, especially as the extra half second is time in which the participants are ‘only’ watching the fixation point. This indeed seems to be the case.

The lack of an effect on the ImagCoh measure could mean that the underlying functional (dis)similarities that the other measures now pick up on are in phase in both signals. But it could also just indicate a lack of sensitivity of the ImagCoh measure.

Using multitapers also changed CCorr and PLV values. This could be because multitapers are ill suited to performing a frequency analysis of low-frequency data (M. X. Cohen, 2014, p. 203). (Which the alpha band (9–14 Hz) data used for this experiment is.) But that does not explain why the CCorr and PLV measures are again more affected than the ImagCoh measure. Especially the CCorr value with a mean

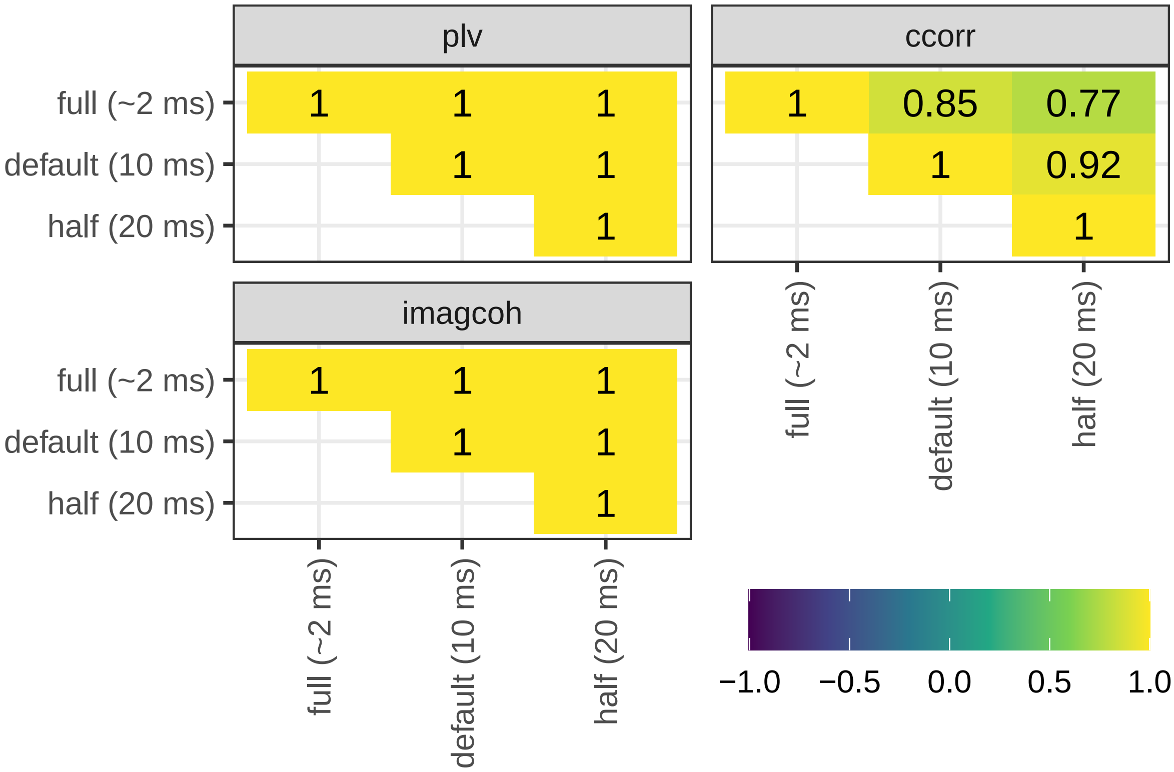


Figure 11

The phase locking value and imaginary part of coherency measures are robust to being calculated on frequency spectra of different resolutions. Circular correlation values, in contrast, will not correlate perfectly when comparing values across resolutions. Figure D1 shows the full underlying data for the CCorr case where r = 0.77.

correlation when comparing tapers of only 0.42 across sessions and electrodes (Figure 9).

As expected, no effect of resolution was found for the PLV and ImagCoh measures. In contrast, the CCorr values are not stable when calculated for different resolutions. While you could still argue that the variance was reasonable for the multitaper case, it seems conceptually bad for a change in sampling rate to have a big effect on an IBS value when the underlying data has not changed. So it is worth discussing the big variance introduced by the CCorr measure in more detail.

Stability of circular correlation synchrony values

We found the CCorr values to change quite a bit when only small changes to the frequency analysis process where made. Apparently, contrary to PLV and ImagCoh, the values do not converge to a single stable value. This is surpising, because Burgess

(2013) previously found CCorr to be more robust than other measures. Pauen and Ivanova (2013) also found it to at least not perform worse than PLV.

Algorithm 2 Calculates a robust circular correlation coefficient. Based on Mahmood (2022)'s work, using the (univariate) dispersion measure from Pewsey et al. (2013, p. 28).

```
Require: $\phi, \psi$                                   ▷ The input signals (phases).
  $n \leftarrow$ length of $\phi$
  $n' \leftarrow n \cdot 0.95$                          ▷ How many data points to keep?
  while $n > n'$ do
      for $i \in 1 \ldots n$ do
          $d_i \leftarrow \sum_{j=0}^{n} \mathrm{dist}(\phi_i, \phi_j) + \mathrm{dist}(\psi_i, \psi_j)$
      end for
      $i \leftarrow \mathrm{argmax}(d)$                 ▷ The point that maximizes the distances.
      remove $\phi_i$ and $\psi_i$ from $\phi$ and $\psi$ respectively
      $n \leftarrow n - 1$
  end while
  return CCorr$(\phi, \psi)$                            ▷ As defined in Equation 2.
```

When investigating why the CCorr varies this much, we hypothesised it could be overly influenced by outliers. Mahmood (2022) proposes a robust version of the CCorr measure which removes values that "lie far away from the majority of the circular data based on the circular geometry theory". Mahmood shows using simulations that his 'trimmed robust circular correlation' measure succesfully reduces the influence of outliers. That makes it ideal to test our hypothesis. But sadly, not enough information is provided to unambiguously reproduce Mahmood's method. While 'lying far away' is well-defined for univariate circular data using the dispersion measure (Pewsey et al., 2013, p. 26), robust correlations need to have a way of detecting bivariate outliers (Maronna, 2019, p. 12) as points can be outliers while not being extreme in any dimension by itself. To the best of my knowledge, such methods have only been published for normal correlations (Bebbington, 1978; Maronna, 2019; Shevlyakov & Smirnov, 2010, chapter 6), not circular ones. As a result, it is likely that Mahmood instead only considered outliers that are isolated in a single dimension. Working with this assumption, we defined the robust circular correlation as given in Algorithm 2.

The resulting algorithm is slow, so it was only applied to the data sets obtained using the standard frequency analysis resolution (10 ms) and half the resolution (20 ms). It results in circular correlation values higher (around 0.11) than those found previously (around 0.03; Figure 10). When correlating the IBS values for the different resolutions, we get a (mean) correlation of 0.87, which is close to the value of 0.92 we got for normal CCorr (Figure 11). But as it is still not ‘1’, our robust circular correlation measure clearly did nothing to resolve the circular correlation stability issue. Apparently, outliers are not the problem.

As sampling at half the resolution is equivalent to just leaving out every second data point, the issue is also unlikely to be caused by more sophisticated frequency analysis issues like spectral leakage. Instead, after further investigation, the problem seems to be inherent to correlation measures. This is most easily demonstrated with a simulation. We replace the CCorr measure with a Pearson correlation, and the underlying phase signals by data drawn from a multi-variate normal distribution while making sure the signals are somewhat correlated. We also calculate correlations after downsampling our ‘signals’ by half. The result can be seen in Figure 12. Clearly, the correlation ($r = 0.72$) between CCorrs of different resolutions is not perfect when using normal correlations to simulate them either.

On the one hand, this is good news. Normal correlations are not fundamentally flawed, so there is no reason not to use the CCorr measure either. On the other hand, it cannot be denied that CCorr values are less stable than PLV or ImagCoh values. As such, more care is required when interpreting raw values, as we will do in the time course analysis section. Permutation tests are an elegant solution to the problem: as they encounter the variation issue also during null distribution construction, it is automatically taken into account when determining the final p-value.

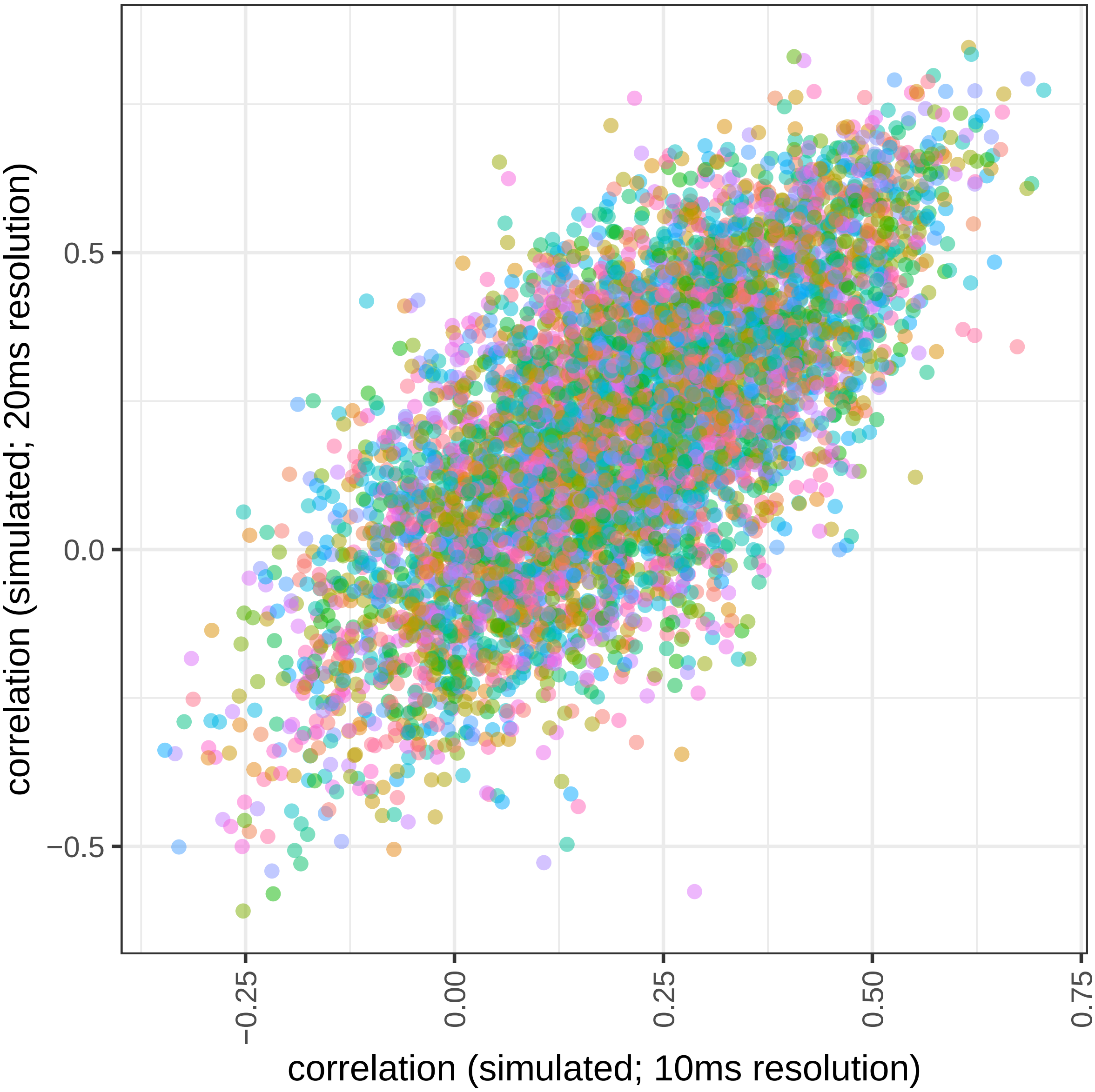


Figure 12

The correlation of simulated (correlated) normal data before and after downsampling. Simulates a single session with 32 electrodes (each represented by a different color) for 180 trials. The result matches the emperical data in Figure D1 quite well. This illustrates that the circular correlation stability issue also exists for normal correlations.

Permutation test analysis

Introduction

At this point it is clear how we calculate the raw inter-brain synchrony (IBS) values, and we have some ideas on how to interprete them. But to explore how they integrate in a full analysis, we need some task-related analysis goals to cut our teeth on.

Task-related research questions

We decided to determine whether there is an effect of cooperation in Newman et al. (2021)'s coordination task on IBS. Assuming there is an effect, we want to know:

1. Is the effect of cooperation merely task-dependent (e.g. due to stimuli or motor responses), or due to actual interaction within the dyad?

2. Is there also an effect of the study's manipulation (i.e. varying working memory load) on IBS, and how does it develop over time?

3. Is it possible to predict for new EEG data whether cooperation was succesful using just the IBS values?

Based on the hyperscanning studies discussed in the introduction, we hypothesize regarding our task-related research questions that more cooperation will lead to higher synchrony. We also expect such an effect to not just be caused by the task but also by the interaction itself. And as a consequence, we expect prediction of accuracy (i.e. succesful cooperation) on the basis of newly collected EEG data to also be possible. We expect IBS to vary over time, as at some point we expect participants to stumble upon a cooperation strategy. Finally, we hypothesize high working memory load to be detrimental to IBS because of Maehara and Saito (2011) and Newman et al. (2021)'s behavioural results.

While most of the task-related questions we plan to answer have an exploratory nature, we also have one directly testable hypothesis: we expect IBS in frontal and temperoparietal areas in the alpha band (Newman et al., 2021). This hypothesis is based on the findings of van Vugt et al. (2020). They found "frontal alpha oscillations" during "moments of agreement" in monastic debate. And also on the findings of Hu et al. (2018), who found higher (phase locking value) synchrony in the alpha band in centro-parietal regions during high cooperation than low cooperation.

Methods

The task-dependent and dyad-dependent effects on IBS are tested separately, the former by running a permutation test against shuffled samples and the latter by running a permutation test against virtual dyads assembled from random participants that never performed the task together.

By shuffling samples, we force the recording of one participant to be independent from the recording of the other participant as they no longer match up in time (Lachaux et al., 1999). This also destroys any effects of (temporal) task structure. There is a problem though: there are multiple ways to shuffle samples. We can shuffle the original signal or the frequency spectrum. If we could convert between the time and frequency domain at any resolution, both approaches would be equivalent. But as we have previously seen in this study, phase and amplitude information is in practice estimated over time windows of up to a second. We run both tests to determine the differences in practise. The permutation tests use 200 repetitions.

By shuffling dyads, we can determine whether there is something that makes the IBS values of actual dyads different compared to randomly assembled dyads that were never actually cooperating. As each participant saw the same stimuli in Newman et al. (2021)'s experiment (albeit in a different order), it is possible to construct virtual dyads such that they still saw the same stimuli. This prevents the permutation test from detecting effects that are actually due to the stimuli instead of the participants themselves. For this permutation test, all possible combinations of virtual dyads are generated. As the amount of participants is limited, this is computationally feasible.

The result of a permutation test is a large data set that has an IBS value for each IBS measure, session, electrode, trial and permutation test repetition. We first average out trial, then session resulting in a distribution for each measure and electrode. Using the same averaging on the actually observed data, we get a single value to compare each distribution against. We calculate a p-value from this by looking how extreme this value is compared to the distribution (see Phipson & Smyth, 2010, for a

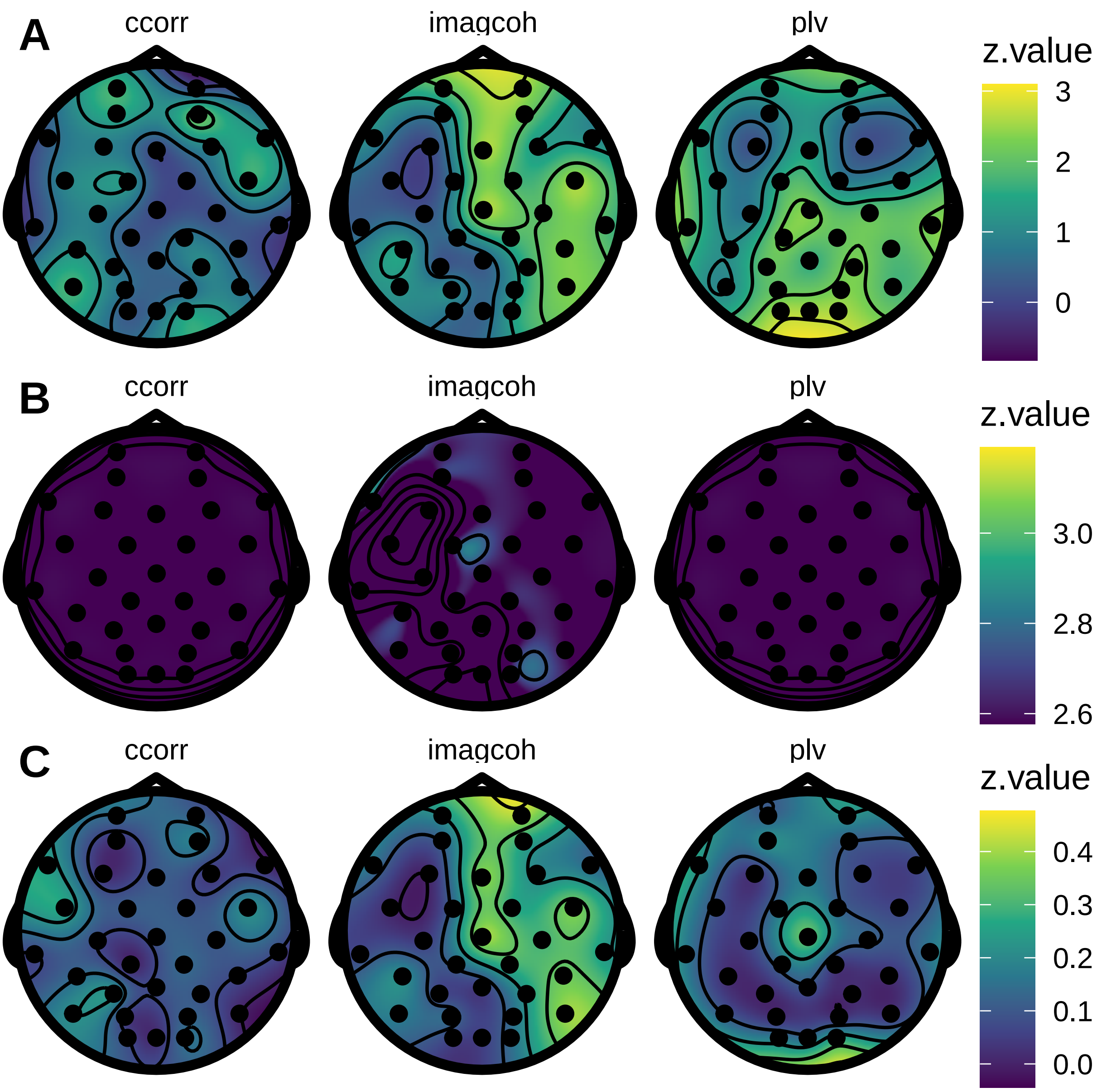


Figure 13

Is there an effect of the task (A, B) or the within-dyad interaction (C) on IBS in the alpha (9–14 Hz) band? After FDR correction, most values in (B) meet the significance threshold, which in that case lies at 2.58. But (B), which shuffles the spectrum instead of the original data (A), does not visualize a valid permutation test (see text). Other tests do not meet the significance threshold.

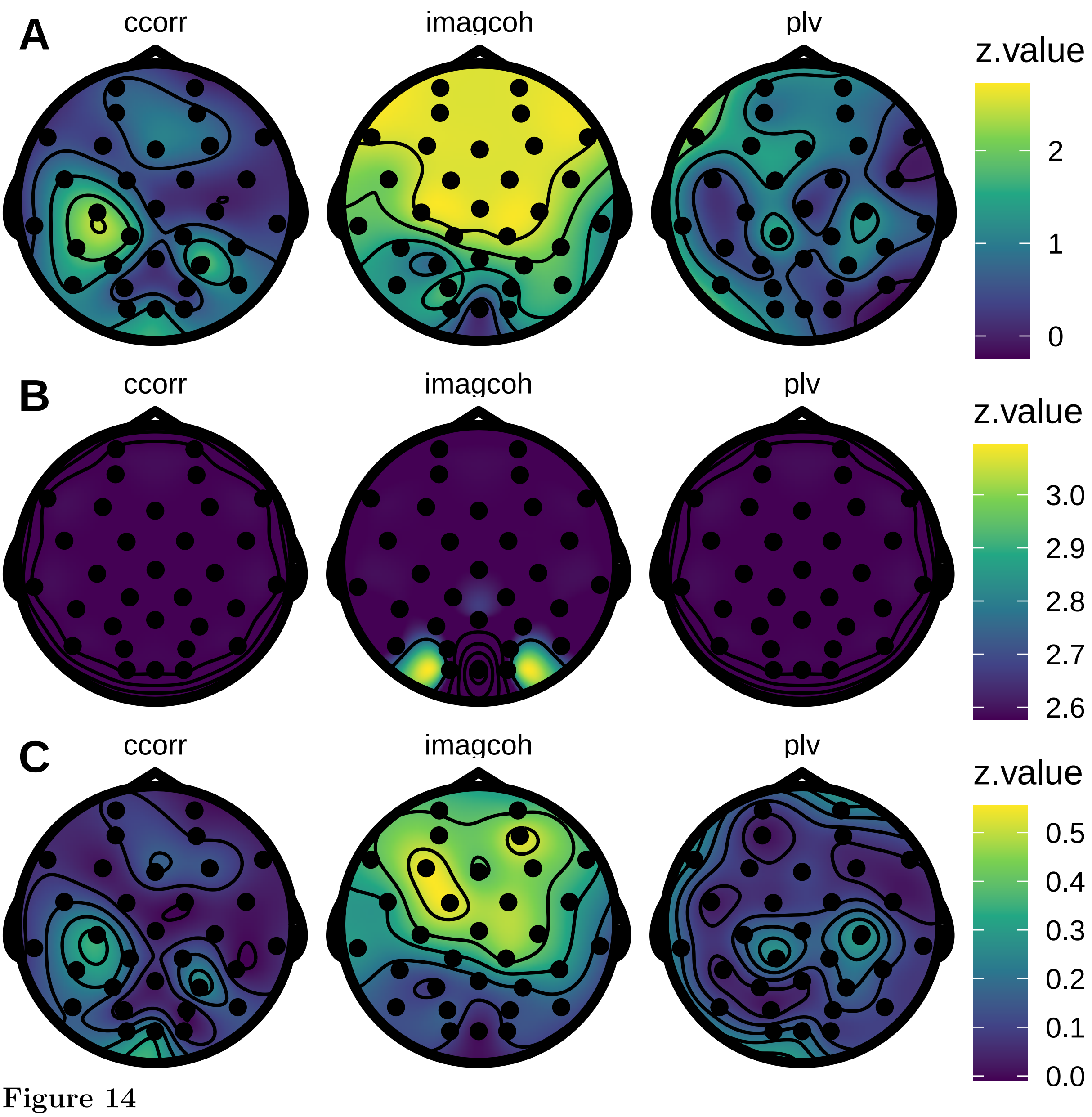


Figure 14

Is there an effect of the task (A, B) or the within-dyad interaction (C) on IBS in the theta (4–7 Hz) band? After FDR correction, most values in (B) meet the significance threshold, which in that case lies at 2.58. But (B), which shuffles the spectrum instead of the original data (A), does not visualize a valid permutation test (see text). Other tests do not meet the significance threshold.

robust method). We convert these p-values to z-scores when visualizing the results, to make any colour transitions more gradual.

Because we perform comparisons for each measure and electrode, we control the false discovery rate (FDR) using a Benjamini and Hochberg (1995) procedure and report the resulting threshold.

Results

To assess whether there is an effect of the task, we perform a permutation test where we shuffle the EEG timeseries data within trials before calculating IBS. When we shuffle in the time domain (panel (A) in Figures 13 & 14), we find no significant effect. When we instead shuffle in the frequency domain, we appear to find a significant effect almost everywhere on the scalp for all the measures (panel (B) in Figures 13 & 14). The only exceptions are for the imaginary part of coherency measure, which is not significant after FDR correction for the Oz electrode in the theta band and the F3, FC5, T7, C3, P3, Pz, O1 and Oz electrodes in the alpha band.

Considering we could not find a task-related effect on IBS when shuffling the time series data, it is not surprising that the tests for a within-dyad interaction effect are also insignificant for all electrodes and measures (panel (C) in Figures 13 & 14). After all, it tests a more specific claim: whether there is an effect of working together.

Discussion

The difference in significance in panels (A) and (B) of Figures 13 & 14) is because the permutation test null distributions differ depending on the shuffling method (see Figure 15). Shuffling the spectrum results in a less conservative test. When we look at the difference between the spectra (Figure 16), it becomes clear that the frequency analysis process normally results in a smoothed spectrum. But also, that this is not the case when the spectrum is shuffled. As a result, this method of generating a permutation test null distribution should not be used. It results in an invalid test.

Contrary to our expectations, we found neither a task-dependent nor a dyad-dependent effect of on IBS. As a result, testing whether this effect varies by level of cooperation (i.e. presumably accuracy), working memory load or over time does not

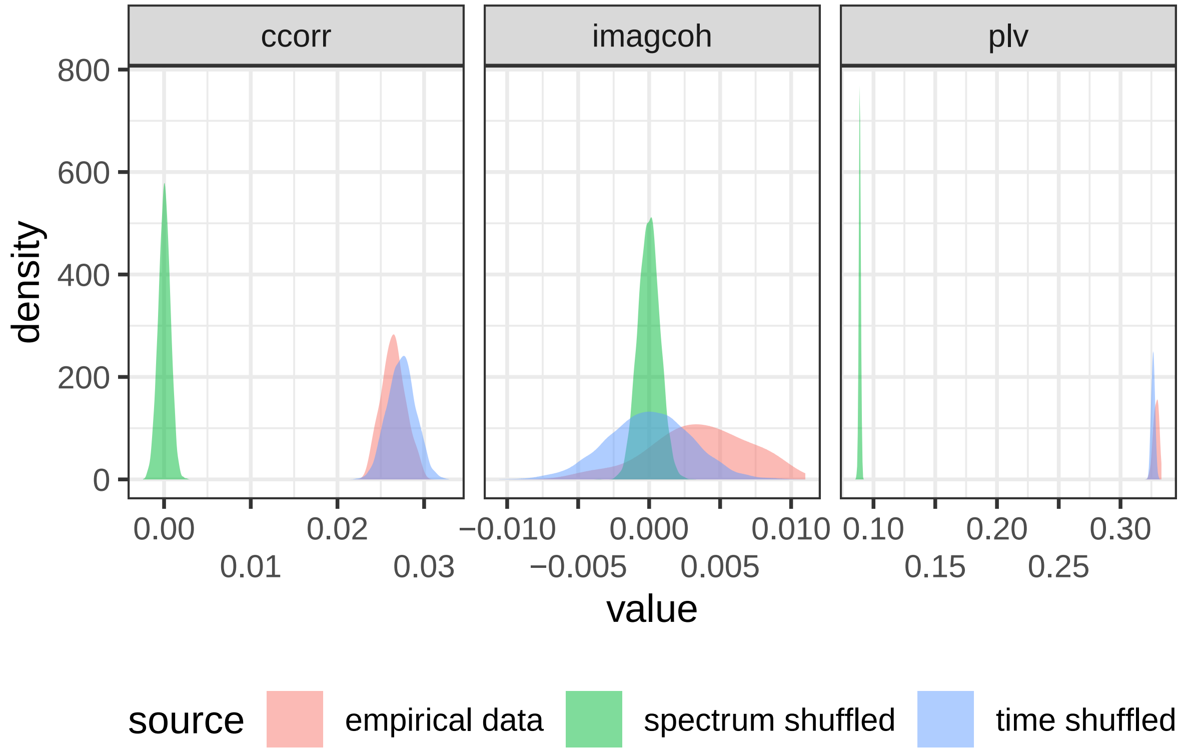


Figure 15

Shuffling the frequency spectrum is not equivalent to shuffling the underlying time series and then estimating the spectrum. Clearly, the shortcut of shuffling the spectrum results in a less conservative test.

make much sense. For the purpose of exploring the full IBS pipeline, the next two sections will make an attempt regardless by analysing the time course and trying to predict accuracy from the IBS values.

We also expected IBS in frontal and temperoparietal areas for the alpha band. No such effect was found.

Inter-brain synchrony over time

Introduction

In Newman et al. (2021)'s experiment, participants need to converge on a strategy to pick the same image or shape. As the task consists of two blocks, they need to do so twice. We hypothesize more inter-brain synchrony (IBS) at the start of a block, when participants need to figure out what the other is doing, and less IBS towards the

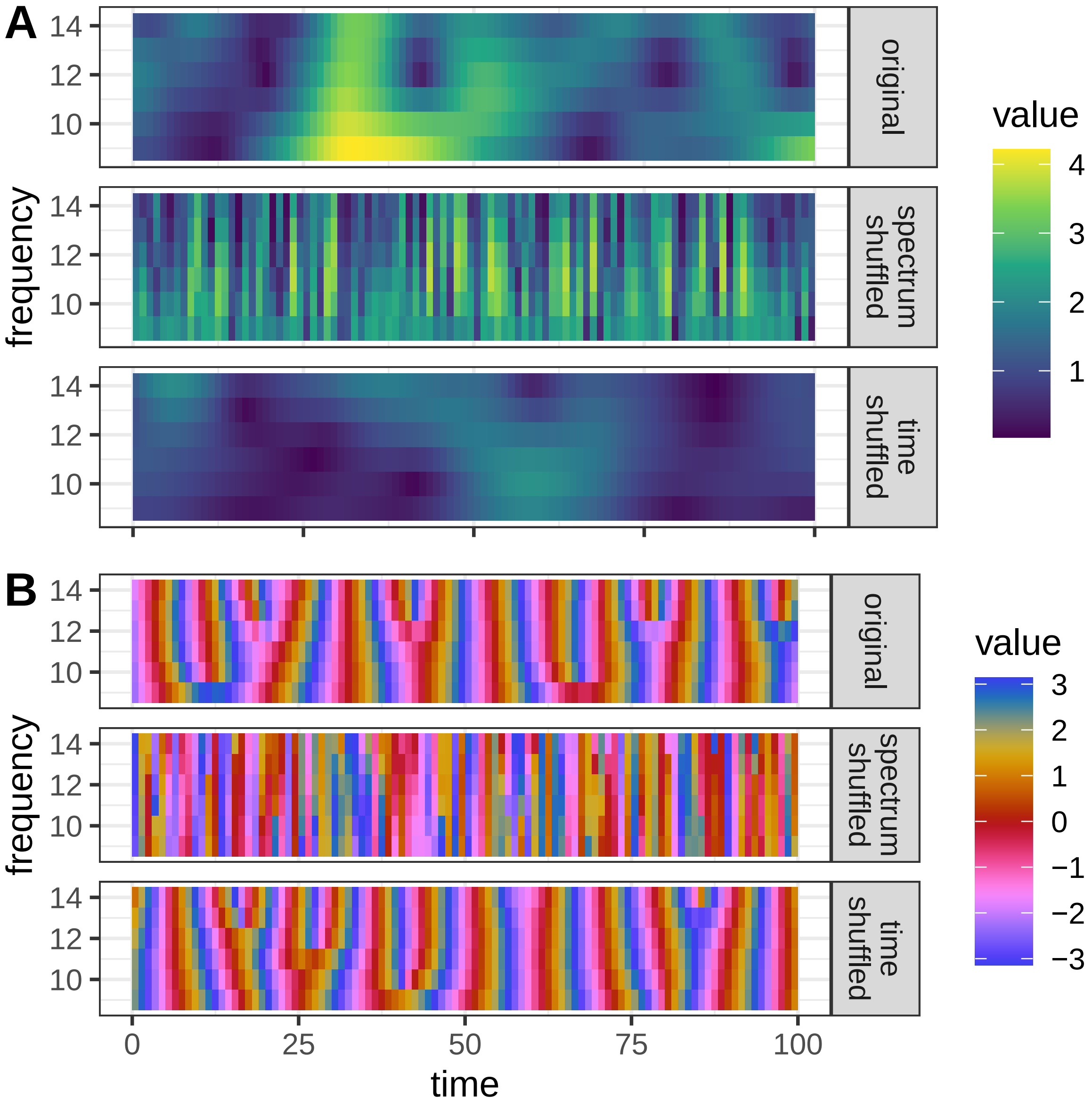


Figure 16

An example of an original spectrum (identical to Figure 1B & C), the same spectrum but shuffled, and a spectrum generated from the same data but shuffled before frequency analysis. Spectrum amplitudes (A) and phases (B) are shown.

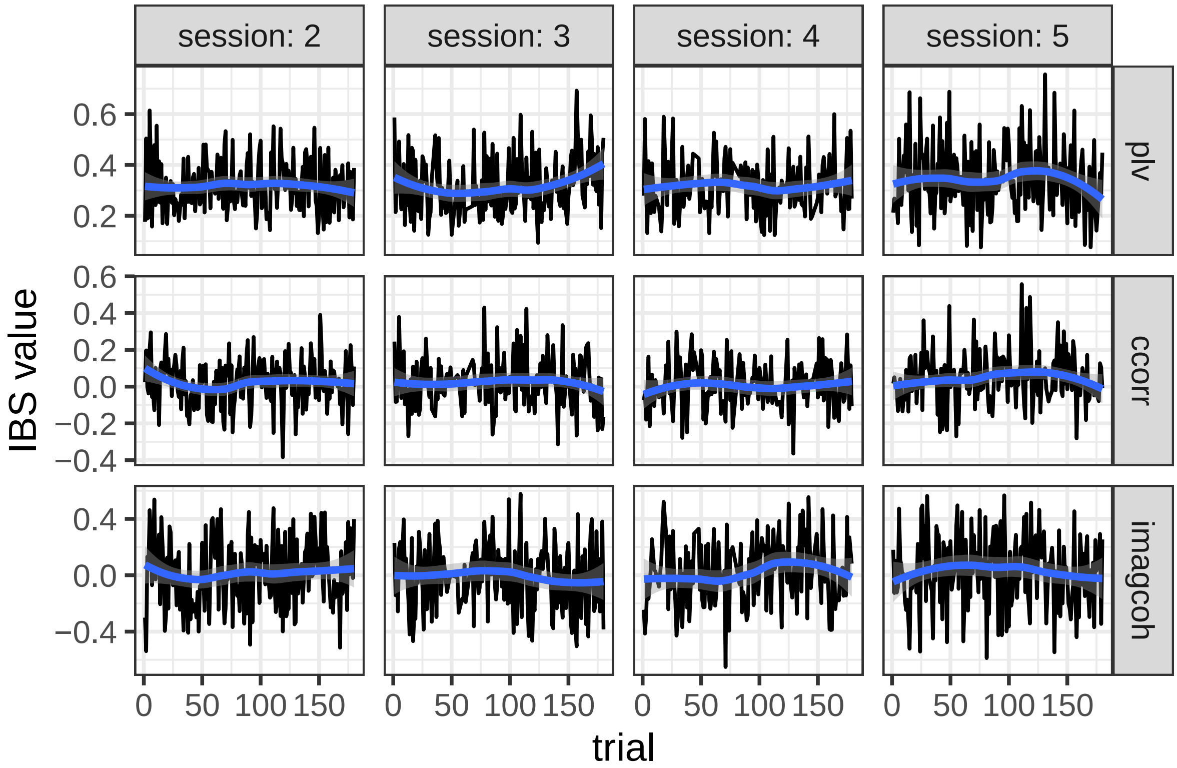


Figure 17

Development of inter-brain synchrony during the task. Variance is high, and there are no clear trends. (First 4 sessions only; alpha band; Pz electrode.)

end, when participants will have switched to exploiting a by then fixed strategy.

Methods

We first visualize IBS for a couple of representative sessions. Because we are interested in effects over time that are potentially non-linear, we assess their significance by comparing Generalized Additive Mixed Effect Models (Wood, 2006, GAMMs) for each IBS measure. These models contain two random effects: a factor smooth of trial by subject, and a factor smooth of trial by electrode. This allows the model to generalize over session- and electrode-specific trends in IBS values. GAMMs can deal with the structure in the data caused by having multiple data points per dyad without requiring averaging. To determine whether a given effect is significant, we add it as a (smooth) fixed effect to one model, then compare both models.

Because it is potentially possible for effects to only show up in a couple electrodes of interest, we additionally fit a (simpler) linear mixed effect model with a

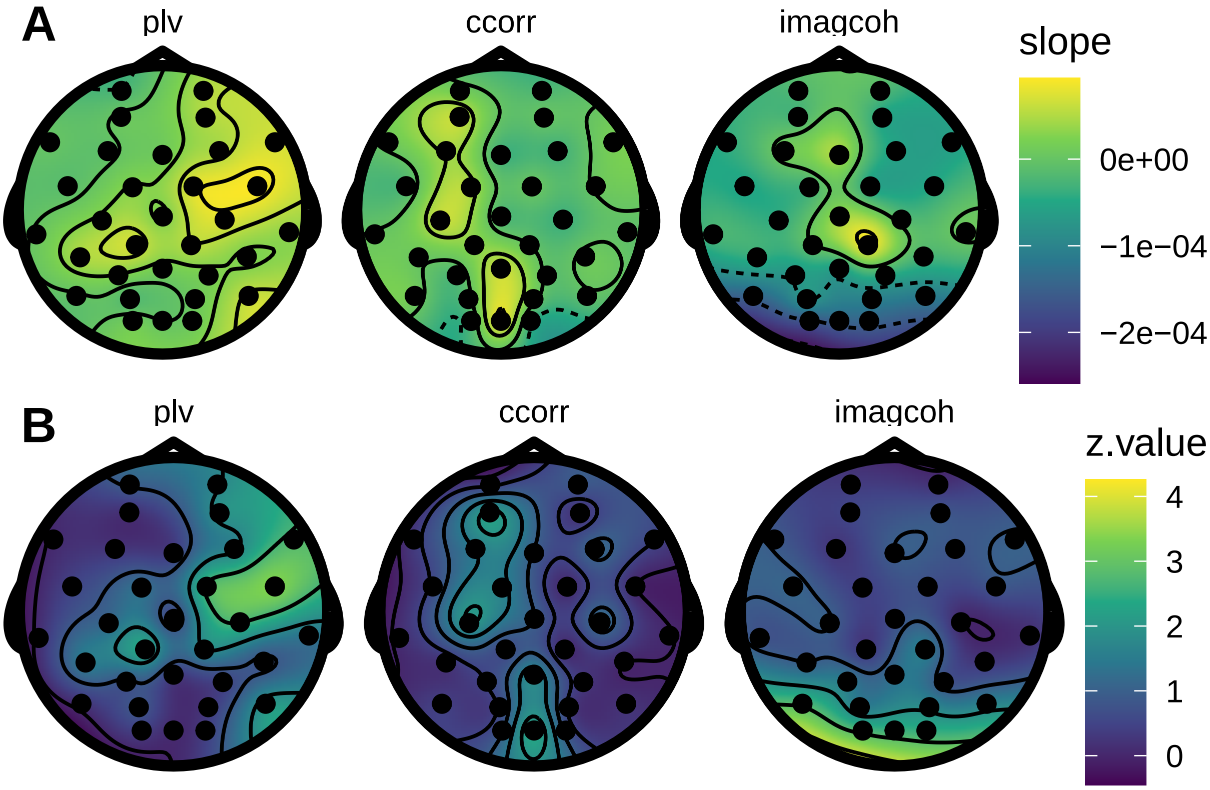


Figure 18

One inter-brain synchrony value is calculated per trial in the alpha band. (A) shows their (average) slope when we fit a line through them. (B) shows none of these slopes are significantly different from zero after FDR correction by comparing a linear mixed effect model that includes the slope to one that does not for each electrode.

random intercept per session on a subset of the data for each combination of a measure and electrode. This allows us to see if there is a linear effect of time within a session for each electrode and measure.

Results

Some representative timecourses of IBS data during the task are shown in Figure 17. Just like with raw EEG data, there is high variance and it can be hard to spot any trends without statistics or averaging.

No significant (smooth) effect of time (i.e. trial) was found for phase locking value (PLV; $\chi^2(2) = 5.199, n.s.$), circular correlation (CCorr;$\chi^2(2) = 5.862, n.s.$) or imaginary part of coherency (ImagCoh; $\chi^2(2) = 3.734, n.s.$) in the alpha band. The same was true for the theta band in the case of PLV ($\chi^2(2) = 4.982, n.s.$) and CCorr

($\chi^2(2) = 4.732, n.s.$), but not in the case of ImagCoh: $\chi^2(2) = 5.805, p = 0.003$). This is due to an increase in ImagCoh values towards the end of the second block. (A plot of the predicted values can be found in the appendix, Figure D2.)

When looking at the electrode level in the alpha band, we see that the linear effect of trial on IBS values is always close to zero (Figure 18A). Unsurprisingly, none of these are significant after FDR correction (Figure 18B). It might appear as if that is not the case for the bottom left of the ImagCoh plot, but this is an interpolation artifact: z-values are only defined at the electrode positions. No effect of trial on IBS was found at the electrode level in the theta band as well. (See Figure D3 in the appendix).

Discussion

We found that contrary to our hypothesis, most IBS values in Newman et al. (2021)'s experiment do not change over time, with the possible exception of the ImagCoh values in the theta band. Interestingly, in that case the effect was in a different direction than expected, with IBS increasing towards the end of the block instead of going down.

One possible explanation is that performing the chosen strategy results in similar brain activity in both participants, even if having theory of mind is at that point no longer necessary. This could be due to performing the same strategy, or simply due to shared environmental stimuli, like the end of the experiment approaching. Alternatively, the assumption that towards the end of a block participants will have converged on a strategy could be incorrect. Finally, it is important to consider that the effect is not that big, especially when taking into account only one out of six tests came out significant. It could be a spurious result, especially as a robust result would presumably be detectable by more than a single IBS measure. On the other hand, ImagCoh is the only measure that includes amplitude information, so it could be that it really found something the others are unable to. Ayrolles et al. (2021) argue that amplitude information reflects cognitive states better than phase information because of its larger timescale.

Prediction of task performance

Introduction

We attempt to predict the performance of the dyads in Newman et al. (2021)'s cooperation task based on inter-brain synchrony (IBS) values and the amplitude of the P3 event-related potential (ERP) component. There are many potential mechanisms that could cause IBS and simultaneously be predictive of task performance. For example, functional similarities could arise in the neural oscillations as a result of participants placing themselves in their partner's shoes (i.e., theory of mind). Alternatively, performing the same strategy could lead to similar brain activity, as could focussing on the same (if chosen automatically correct) stimulus. In the end, while the exact mechanism is interesting from a theoretical point of view, it does not matter when the goal is to predict task performance. Instead, it would be a possible follow-up question.

The P3 ERP component, also known as the P300 component (Luck, 2014, p. 5), is a positive deflection in an EEG signal about 300ms after a stimulus is shown (Sutton et al., 1965). It generally occurs as a response to infrequent but task-related stimuli (Polich, 2011). The mechanism underlying the P3 is unclear, but the most popular theory is the *context updating model* (Luck, 2014, p. 96). It explains the P3 as a consequence of updating the neural representation of the environment when the stimulus (unexpectedly) changes (Polich, 2011). The P3 decreases during mind-wandering (Jin et al., 2019). In Newman et al. (2021)'s task the context is a bit more complex than frequent or infrequent stimuli being shown: participants instead need to reason about how the other participant is choosing one of the four stimuli. But this still requires keeping track of choices, feedback and hypotheses. It is not out of the question this also would evoke a context updating P3 ERP. Alternatively, negative feedback could be surprising in itself when the participant believes to have hit upon a 'shared rule' on how to perform the task.

We consider a number of classification methods: logistic regression (Goodfellow et al., 2016, p.137), support vector machines (Goodfellow et al., 2016, p.137–139), random decision forests (Tin Kam Ho, 1995) and multi-layer perceptrons (Rumelhart & McClelland, 1987). We attempt both across-session and within-session prediction.

Methods

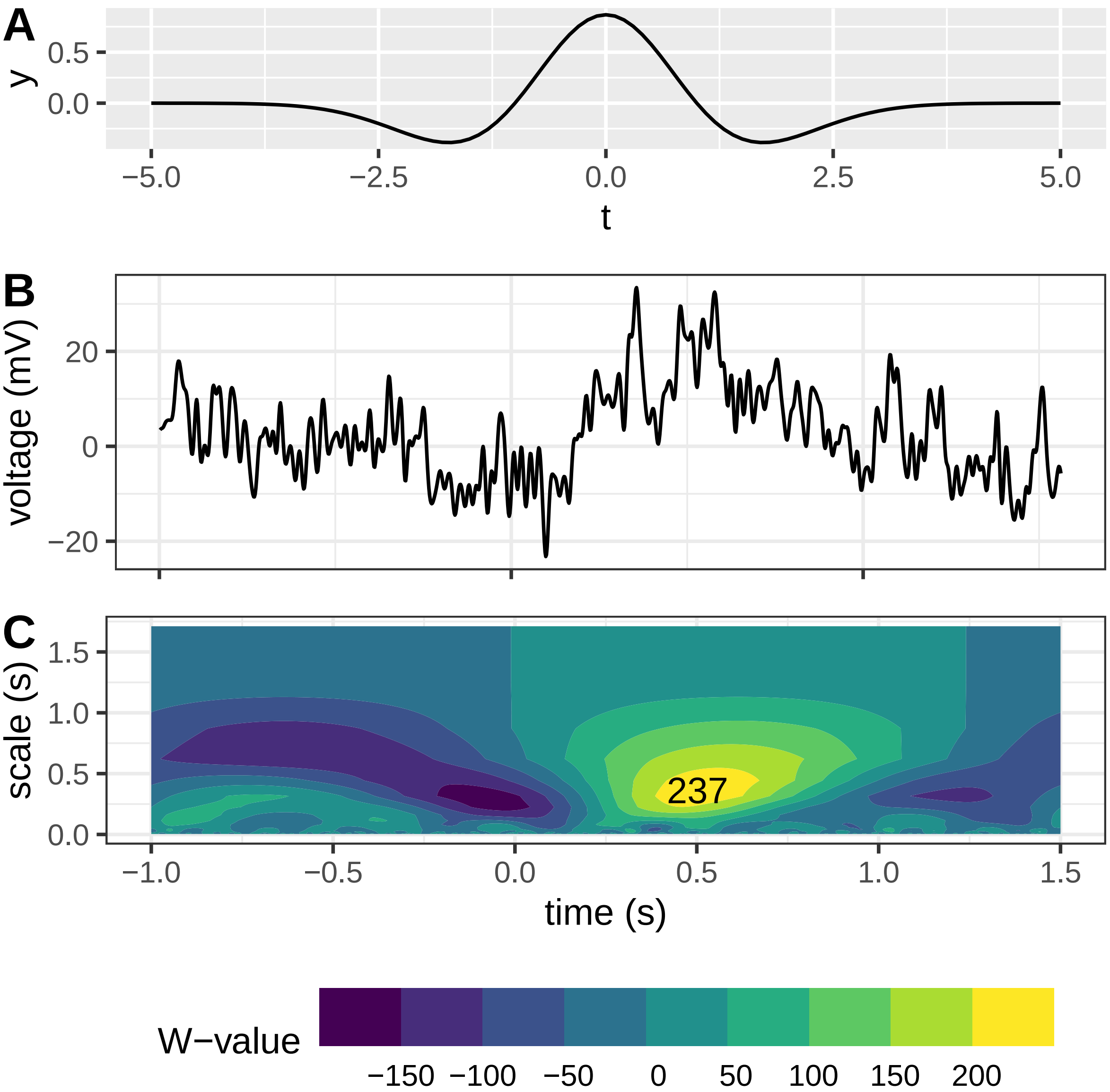


Figure 19

When applying the continuous wavelet transform using the mexican hat template $\psi(t)$ in (A) to an example trial $f(t)$ in (B) we obtain (C). The local maximum in (C) is shown and is our single-trial ERP measure. (B) is identical to Figure 1A.

The methods of the prediction task are based on those of a study by Jin et al. (2019).

Because EEG data contains a lot of noise, ERP components are normally identified by averaging over multiple trials (Luck, 2014, p. 259). This is not feasible when predicting task performance, as we need to predict whether the dyad guessed correctly for each trial. Instead, we use a method that attempts to match each trial's EEG signal with the shape of a template. The template $\psi(t)$ takes the form of an idealized ERP component (see Figure 19A). This function, sometimes called 'mexican hat', is defined as follows (Bostanov & Kotchoubey, 2006):

$$\psi(t) = (1 - 16t^2)e^{-8t^2}. \tag{8}$$

The method we use, which is devised by Bostanov and Kotchoubey (2004), uses a continuous wavelet transform (CWT) to calculate the covariance between the signal and the template at different time points and for different template scales. The CWT is defined as (Bostanov & Kotchoubey, 2006):

$$W(s,t) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(\tau) \cdot \psi\left(\frac{\tau - t}{s}\right) d\tau \tag{9}$$

where $f(t)$ is the signal that is transformed, s scales the template $\psi(t)$ and τ shifts the template. See for an example signal Figure 19B, and for a corresponding example CWT output Figure 19C.

The single-trial P3 ERP is defined as the local maximum in $W(s,t)$ between $t = 250$ms and $t = 600$ms (Jin et al., 2019).

Table 3

Thirty sessions were randomly assigned to the train set, eight to the test set.

assignment	session numbers
train set	2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 16, 18, 19, 21, 22, 23, 24, 25, 27, 29, 31, 33, 35, 36, 37, 38, 39, 41, 42
test set	5, 7, 15, 17, 20, 28, 30, 40

The data set was randomly split into a train- and a test set (see Table 3). The latter was not accessed during training and only used for final model evaluation. Such a test set is sometimes also called a lock box (Hosseini et al., 2020).

Because on average dyads are correct a bit more than they are incorrect, we use random oversampling during training to account for this imbalance in the data set (Chawla, 2005). Otherwise, a model that classifies every example as correct would result in an accuracy higher than 50%, which is not helpful when determining whether IBS measures and the P3 component can predict task performance.

For each trial, the three IBS measures where calculated in both the alpha and theta band. Additionally, single-trial P3 ERP components were calculated for both participants. These eight calculations were all repeated 32 times for each electrode, resulting in a total of 256 features.

Phase locking value values were normalized using the inverse cumulative density function of the normal distribution. Circular correlation and imaginary part of coherency values were Fisher-transformed. All single-trial ERP trials were log-transformed. This was impossible for (a trivial amount of) negative values, which were capped at 0.05 before the transformation. All the resulting values were additionally z-transformed.

We report sensitivity, specificity and balanced accuracy. Sensitivity looks at correct trials. It tells us what proportion of those the classifier predicted to be correct (Yerushalmy, 1947). Specificity looks at the incorrect trials. It tells us what proportion of those the classifier predicted to be incorrect (Yerushalmy, 1947). Balanced accuracy is the mean of the two. Classifiers were trained to maximize balanced accuracy.

During training, 10-fold cross validation was used. Folds were chosen such that data of a single session did not leak into both a train and validation set, as this could potentially lead to overoptimistic accuracy estimates.

Random hyperparameter search was used (Bergstra & Bengio, 2012) for 30 iterations, with hyperparameters sampled from log uniform distributions with the exception of the random forest integer hyperparameter values, were a discrete uniform distribution was used instead.

For the logistic regression classifier, which served as a baseline, the L1 norm was used as it forces unused features to be dropped entirely. We optimized the regularization strength hyperparameter C (also known as *capacity*; Goodfellow et al., 2016, p. 117).

For the support vector machine, a radial basis function was used. We optimized its radius (γ) and regularization (C) hyperparameters. As the SVM performed best in earlier EEG prediction tasks (Jin et al., 2019; Lotte et al., 2007), we used it for two variations on the experiment as well. An SVM was trained without P3 ERP component data (i.e. taking 192 features as its input), and 256 SVMs were trained that only took a single feature each to determine the relative importance of features. For these variations, hyperparameter values of the main experiment were used.

Two hyperparameters were optimized for the random forest classifier. Amount of trees (1–250) and maximum amount of features (1–30). For the multi-layer perceptron, a fixed architecture consisting of a single hidden layer of 10 neurons was used. The ReLU function, i.e.

$$f(x) = \max(0, x) \tag{10}$$

was used as activation function. The learning rate (alpha) was optimized.

Finally, some within-session and within-condition classification was attempted using SVM classifiers. This results in small data sets of $~90$ rows, which were split in train sets containing $~75\%$ of the rows and test sets containing $~25\%$. Hyperparameter optimization was the same as in the 'main' prediction experiment, except for the choice of cross-validation folds. There were no groups to take into account, instead folds were kept balanced such as to have both 'correct' and 'incorrect' examples. A dimension reduction step using principal component analysis (PCA) was added to better cope with the small amount of available data. The amount of components k was optimized using cross-validation.

Results

The outcome of the cross-validation procedure used to determine the hyperparameters was visualized in a number of parameter vs. test performance plots. These plots can be found in Appendix D for logistic regression (Figure D4), for the

SVMs (Figure D5), for the Random Forest classifier (Figure D6), for the multi-layer perceptron (Figure D7) and finally for the within-session SVM (Figure D8).

Evaluation

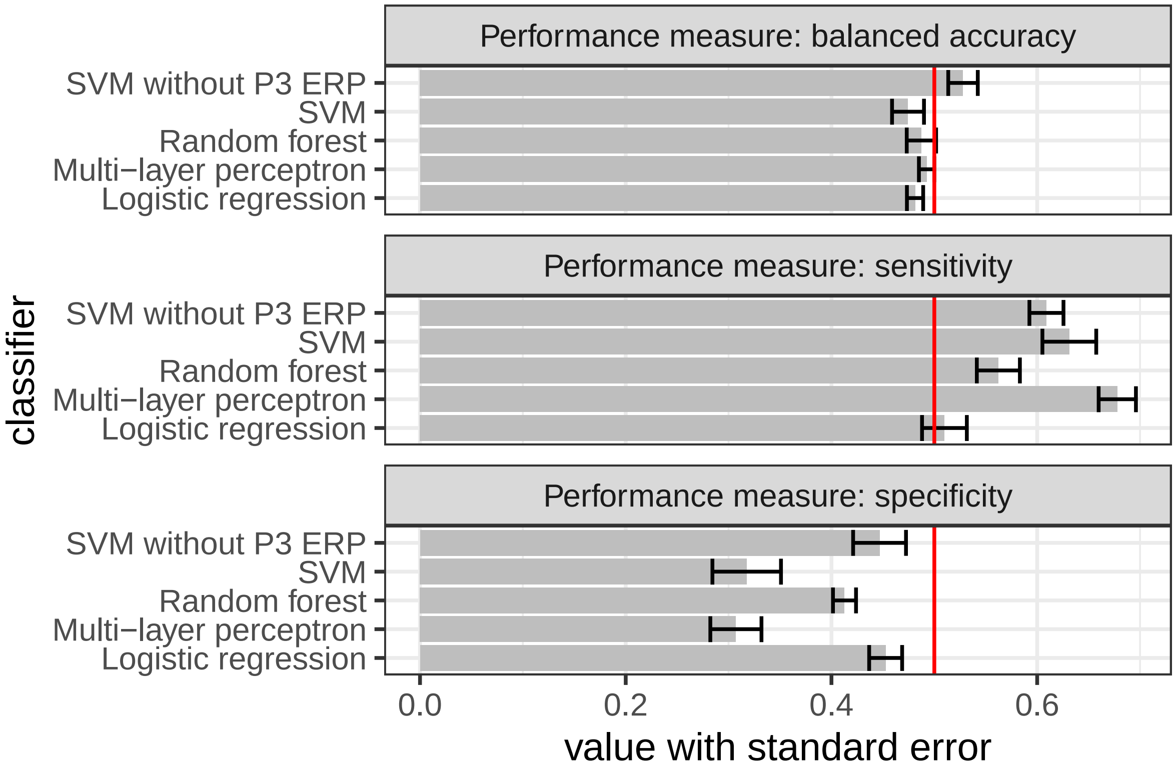


Figure 20

Mean classification performance on the test set for different performance metrics.

Figure 20 shows that classification performance, as measured using the balanced accuracy, is at chance level (i.e. 0.5) for all classifiers. Although it might seem like some error bars do not overlap with 0.5, this would be the case if confidence intervals were shown instead of standard errors, as those are almost two times as big. Figure 20 also suggests that an SVM classifier that was not trained on P3 ERP component-based features outperforms an SVM classifier that was.

We see that all classifiers have a higher sensitivity than specificity. In other words, the classifiers are better at predicting correct trials as correct than incorrect trials as incorrect. This would make sense if we had not corrected for the imbalance in the data, but between the balanced accuracy performance measure and the random oversampling process, that is not the case. Apparently, the classifiers converge on a

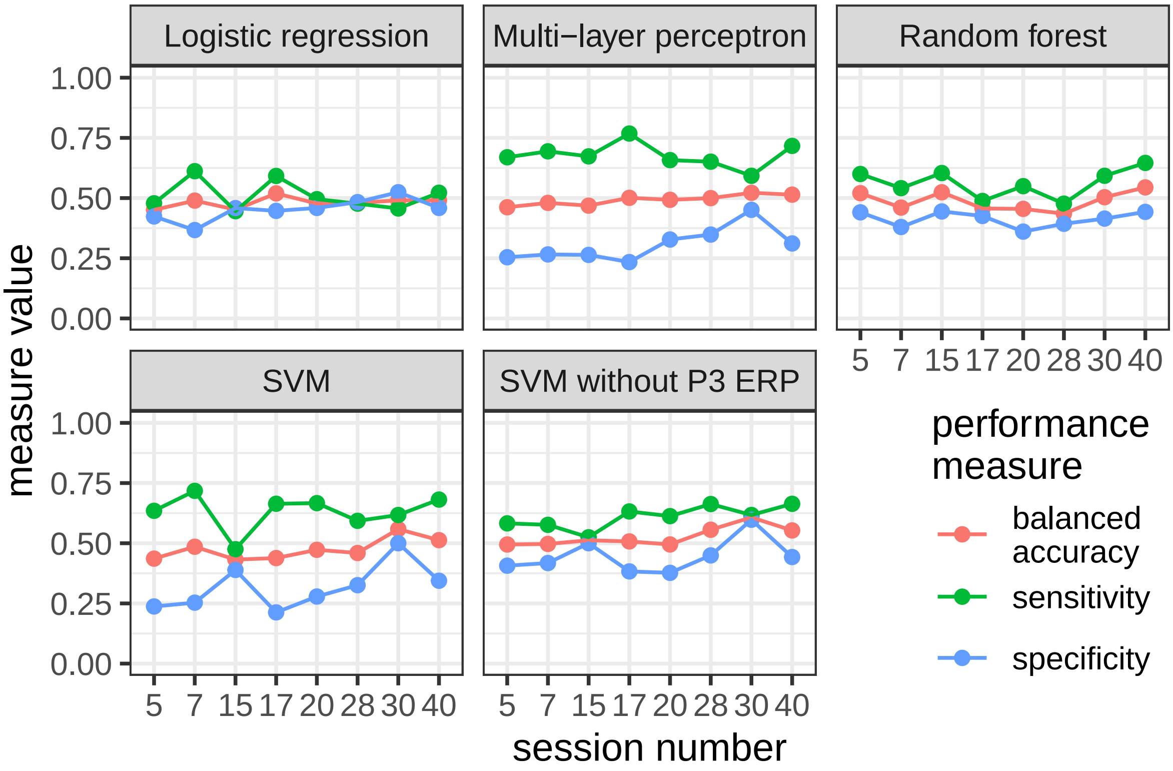


Figure 21

Classification performance for each test set session. (Detailed version of Figure 20).

(slight) bias to predict trials to be correct regardless. This is not the case for all sessions, as can be seen in Figure 21. Also, the logistic regression classifier seems to show this phenomenon less strongly.

Important features

The final logistic regression model drops most predictors, and puts the highest importance on P3 ERP component features (see Figure 22). This is likely to be a fluke, as we would expect an actual pattern to be duplicated among P3 ERP components for both subjects. Otherwise, no pattern is discernible, which is what we would expect for a model that predicts at chance level.

Another way of assessing the importance of individual features in predicting task performance is to look at the balanced accuracy of the SVMs that were trained on single features (see Figure 23). In general, these classifiers also perform at around chance level. There is a bit more variation in perfomance of the classifiers that were trained on P3 ERP components of the first participant compared to the other classifiers

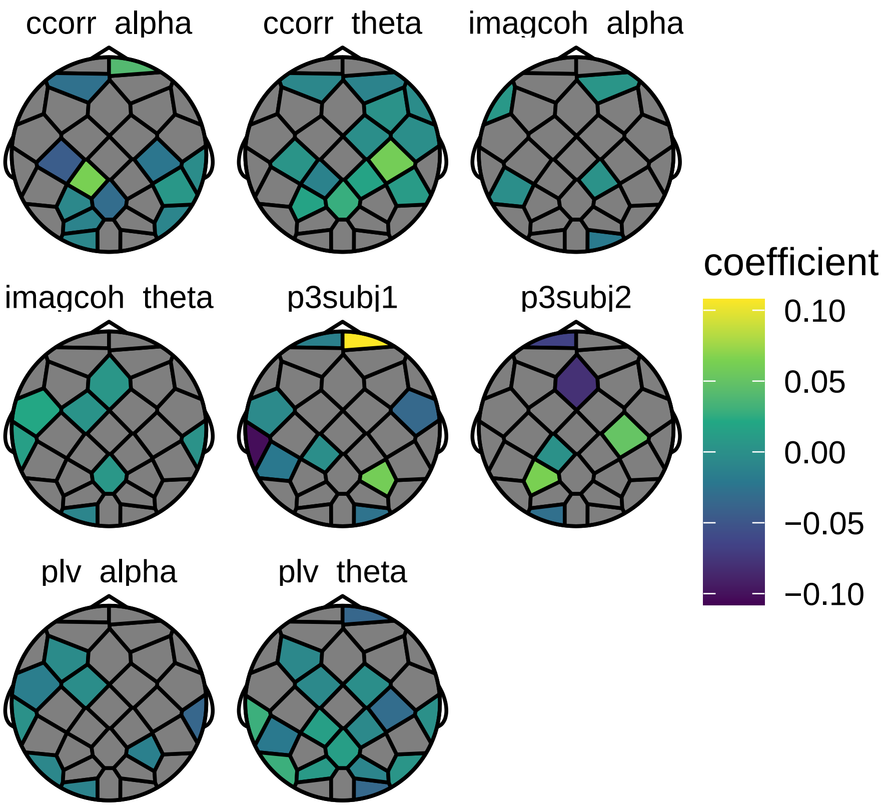


Figure 22

Logistic regression (L1 norm): coefficients for each feature. Missing data (grey background) means the feature was dropped by the classifier completely. Features with coefficients further from zero have a greater influence on the final prediction. Synchrony measures (ccorr = circular correlation, imagcoh = imaginary part of coherency, plv = phase locking value) were calculated for both the alpha and theta band. Both subjects contribute a P3 single-trial ERP value.

(see Figure 24). As this matches our findings using the logistic regression classifier with L1 norm, it is likely to be caused by a pattern in the P3 ERP component data and not just by a classifier induced artifact. But as the pattern is again not reproduced for the second participant, it is unlikely it contributes to predicting task performance.

Within-session classification

When predicting task performance within sessions for both low and high working memory load, we see the same pattern as for the between-session classifiers. Performance is still at chance level, and classifiers have on average higher sensitivity than specificity (see Figure 25).

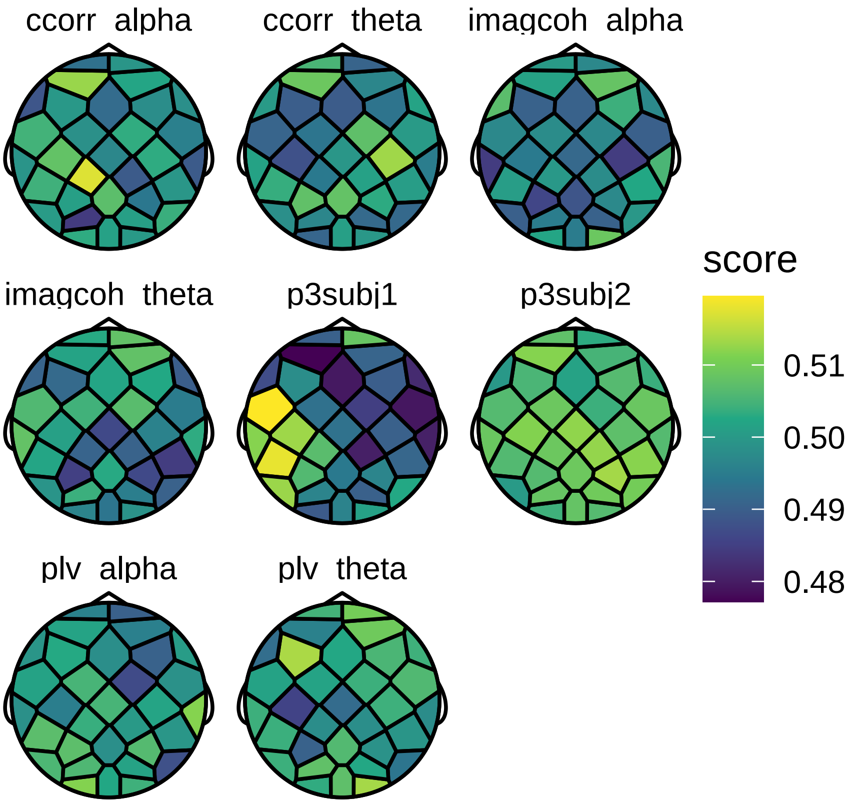


Figure 23

Balanced accuracy of SVMs trained on single features, averaged over sessions. Synchrony measures (ccorr = circular correlation, imagcoh = imaginary part of coherency, plv = phase locking value) were calculated for both the alpha and theta band. Both subjects contribute a P3 single-trial ERP value. Outliers lie further than 1.5 inter-quartile ranges from the hinge.

When looking at the raw data in Figure 26, we see more variation (cf. Figure 21). But this is to be expected as the test sets are much smaller.

Discussion

Prediction of task performance based on IBS values and single-trial P3 ERP component values failed. There are two possible explanations for this. It could be that another classification method would perform better. But as different classifiers, classification scenarios and hyperparameters were tried, another method is unlikely to yield wildly different results. The more likely explanation is that there is simply not enough information in IBS values and single-trial P3 ERP components to be able to predict task performance. That would also be in line with the null results found in the

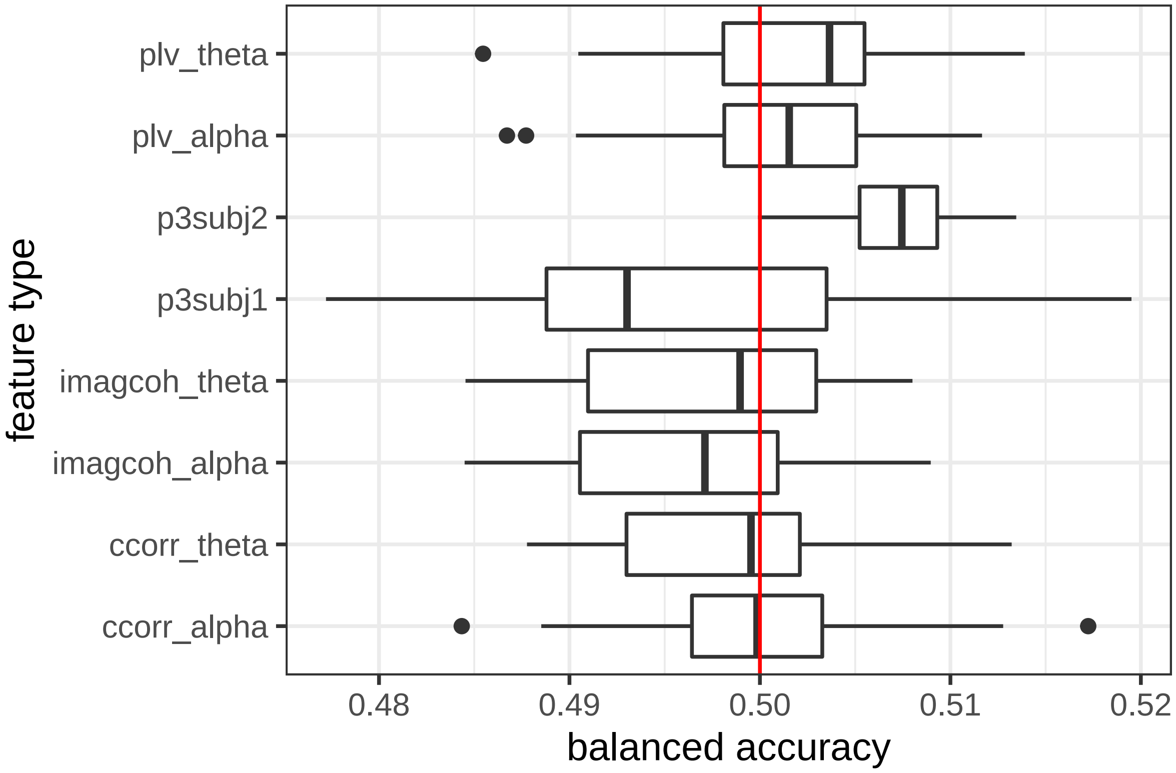


Figure 24

Balanced accuracy of SVMs trained on single features, averaged over electrodes and sessions. A box plot summary of the data shown in Figure 23.

time course analysis and permutation test analysis described in this report.

While most classifiers had higher sensitivity than specificity, this was not the case for the logistic regression classifier (see Figure 21). Possibly, this is because it is one of the more constrained models from a theoretical point of view, having a smaller representational capacity (Goodfellow et al., 2016, p. 110).

Interestingly, an SVM trained only on IBS values seems to perform slightly better than one trained also on P3 ERP components. The difference is small, so it could just be due to variation in the data. But an alternative explanation worth considering is the *curse of dimensionality*: because the data set is relatively small compared to the amount of features, models are not constrained all that much by the training examples (Goodfellow et al., 2016, p. 151–152). Leaving out features that do not contribute much is helpful as a result. But there are other ways to constrain classifiers. One is to force them to model the underlying distribution smoothly, e.g. using regularization. This was

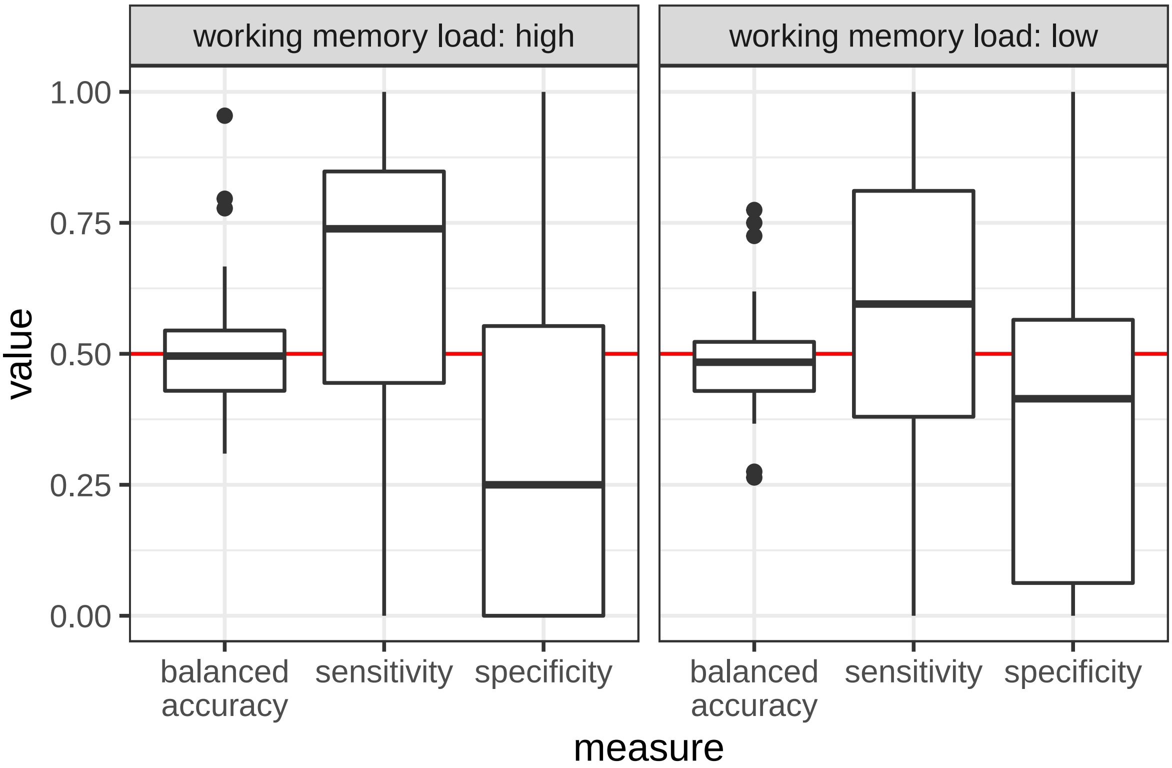


Figure 25

Within-session classification performance on the test set (box plot, outliers lie further than 1.5 inter-quartile ranges from the hinge).

the case for all discussed classifiers. Another is to pre-process the data using a dimension reduction technique. This approach was used for the within-session classifiers, which included a PCA step. But that did not yield better classifiers.

The attempt to identify the most influential values in the classification process was largely stymied by the lack of classifiers performing above chance level. It suggests the first participant's P3 ERP component values might be more influential. One possible explanation for that could be that the ERP component features stand out because their distribution is the furthest from a normal distribution. This could lead them to have an oversized effect on the models. Negative ERP values being set to a fixed value, especially, introduces a few (rare) outliers. Negative single-trial ERP values are rare and suggest errors in the data cleaning, but in practise they are hard to avoid as getting rid of them all would also throw out a lot of good data of other electrodes.

It is unlikely the first participant's P3 features are influential because they

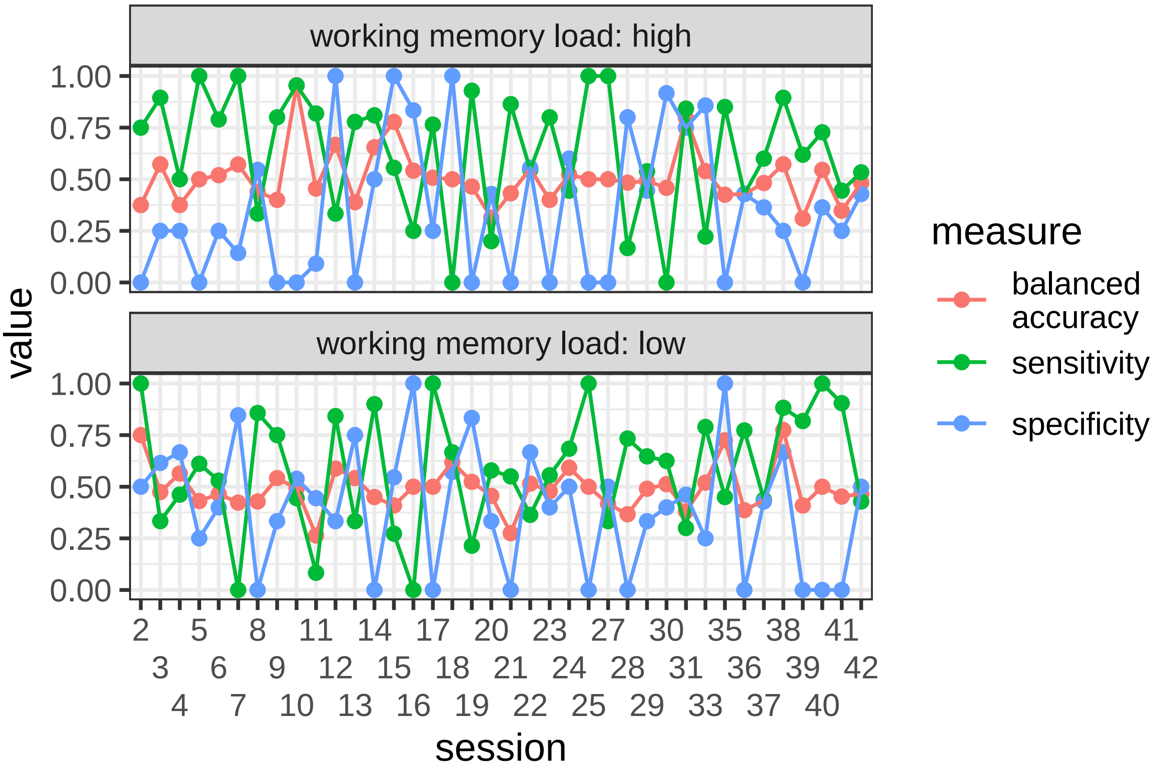


Figure 26

Within-session performance on the test set (raw metrics, see Figure 25) for summarized numbers.

actually help classification. In that case, we would expect them to also show up in the P3 features for the second participant. Also, we would expect features close to the midline to be more influential, as that is where the P3 effect is strongest (Polich, 2011). Neither is the case (see Figures 22 and 23).

Finally, a few notes. The use of balanced accuracy instead of normal accuracy is important, as random oversampling is only used during training, not during testing. Originally, I overlooked this, and in this case, it lead to models that only predict a single outcome. This only became clear after looking at the sensitivity and specificity measures. Random oversampling works well, but it can lead to overfitting (Chawla, 2005). Especially if the imbalance in the data is big, it is worth considering more sophisticated methods to construct balanced samples (Chawla, 2005, e.g. SMOTE).

While performing within-session classification can profit from dyad-specific signals that predict task performance in IBS and P3 values, it comes at a cost of having only very little data available. In some test sets, no examples of both correct and incorrect trials were available. This made it impossible to train a classifier in a couple of cross-validation folds, and is also a cause of the large variance in Figure 26. The lack of data also means that we have no choice but to randomly sample cross-validation folds from the train set instead of respecting their causal ordering in time. When the underlying time series is autocorrelated, as is not unlikely in EEG-derived data, this could lead to overoptimistic predictions.

General Discussion

We investigated the sensitivity of a hyperscanning data analysis to different methodological choices by performing an analysis of inter-brain synchrony (IBS) data recorded during Newman et al. (2021)'s tacit coordination task. We built a complete analysis pipeline that tested three IBS measures: the phase locking value (PLV), the circular correlation coefficient (CCorr) and the imaginary part of coherency (ImagCoh). Contrary to our expectations, we found the analysis outcome to be sensitive to relatively minor changes to this pipeline.

All studied measures of IBS rely on a frequency analysis step to transform the raw EEG data into the frequency domain. We found that varying the resolution of the output or the exact tapering method used to control spectral leakage resulted in different IBS values. The CCorr measure was especially sensitive to such changes. As long as you are comparing apples to apples, i.e. only values that have been calculated with the same methodology, this variation should not be a problem. But it is a reason to caution against comparing raw IBS values across experiments or analyses. Using statistical methods that can take this into account, like permutation tests that will make the same assumptions when generating a null distribution, is recommended.

Burgess (2013) found the CCorr measure to be less sensitive to detecting spurious IBS than other measures. Our study did not encounter this issue, because the permutation tests did not detect any IBS. On the other hand, our simulation study clearly illustrates Kayhan et al. (2022)'s observation that PLV only measures the consistency of the phase components of the EEG signals coming from each participant, not whether they co-vary. Most strikingly, we see it completely ignore a strong negative linear relation between the two phase components (see Figure 6). The ImagCoh measure is hardest to evaluate. It seems to be less sensitive in general to changes in the data it is calculated upon. For example, in the simulation study, finding examples for different ImagCoh values was harder than for the other measures. Also, it only responded little to changes in the frequency analysis process. If it still picks up on 'real' effects, it would be the best measure tested. But the fact that it is so insensitive, makes me doubtful about whether it would quantify such effects. In the end, weighing all the evidence, I would prefer using the CCorr measure for measuring IBS. But the PLV measure is also worth considering considering. While it has its flaws, its ubiquitousness in the hyperscanning literature makes it more familiar to the average reader.

Contributions

Next to the research project's results and pipeline description, we make available validated implementations of the PLV, CCorr and ImagCoh measures for both MATLAB and R. During the project, we also developed a MATLAB implementation of Mahmood (2022)'s robust circular correlation measure (Algorithm 2), although the implementation is slow and as discussed previously the measure itself is not well-defined from a theoretical point of view. Finally, in the end of the simulation study section, we describe a way to perform a power analysis for tests used in IBS studies. It reuses the method the simulation study uses to generate fake data for a given IBS value (Algorithm 1). As a consequence, the test will only have access to the phase component of the signal, as the simulation study ignored amplitude components. But that can still be useful for power analyses of tests that target phase-based measures only.

Limitations

This research project, especially the simulation study part, has been heavily focused on phase-based IBS measures. The only exception is the ImagCoh measure. Ayrolles et al. (2021) suggest phase-based measures are better at measuring "ongoing cognitive processing", while amplitude-based measures are better for measuring "cognitive state". It would be interesting to also consider other amplitude-based measures, like the 'power envelope correlation between orthogonalized signals' measure described by Hipp et al. (2012). That measure is also used by Dikker et al. (2021), who call it 'projected power correlation' instead. Another measure that was considered for inclusion in this study is the Kraskov mutual information measure (Kraskov et al., 2004). Burgess (2013) recommends it alongside the CCorr measure. But while Burgess's work seems to have single-handedly popularized the latter[4], the former seems to be have much less uptake. Perhaps it is due to the lack of implementations being available [5], or the more complex (information-theoretic) definitions. At least, that is the reason why it has not been included in the present project.

Figures 13, 14, 18 and D3 use topographical plots of the scalp that are a common sight in EEG research. While actual values are only available for the electrode sites, the visualization fits a surface to them to present a continuous image. While interpreting some of these figures during this project, this lead me to the wrong conclusions at times. For example, it is common to see extreme values around the edges of the scalp because the surface continues on outside the data's range for a bit. As a result, I switched to drawing Voronoi cells around the electrodes instead when analysing the prediction data (see Figures 22 and 23). Of course, this approach also has its downsides. It will be less familiar to researchers in the field, and the discrete nature of the visualization is unrealistic.

IBS permutation tests that generate their null hypothesis distribution by shuffling dyads, as we did in the permutation test analysis section, are a nice way to determine whether synchrony is just task-related, or due to cooperation within the dyad. That said, if such a test yields a significant result, there are other possible explanations. For example, if the two participants both have a faster response time than other dyads, this could lead to the test finding synchrony between them that is

[4] Most discussions of the CCorr measure I have seen can be traced back to Burgess (2013)'s work (Chen et al., 2021; Farahzadi & Kekecs, 2021; Goldstein et al., 2018; Kingsbury & Hong, 2020; Kurihara et al., 2022; Wikström et al., 2022, to name just a few).

[5] https://github.com/otoolej/mutual_info_kNN/blob/master/mi_cont_cont.m comes the closest, but it is does not match Burgess's definition exactly. For one, it does not use an angular distance metric.

‘just’ due to their early motor response. Such a response would be solely task-related, not due to the participants working together or interacting otherwise. It is something to keep in mind when designing IBS experiments.

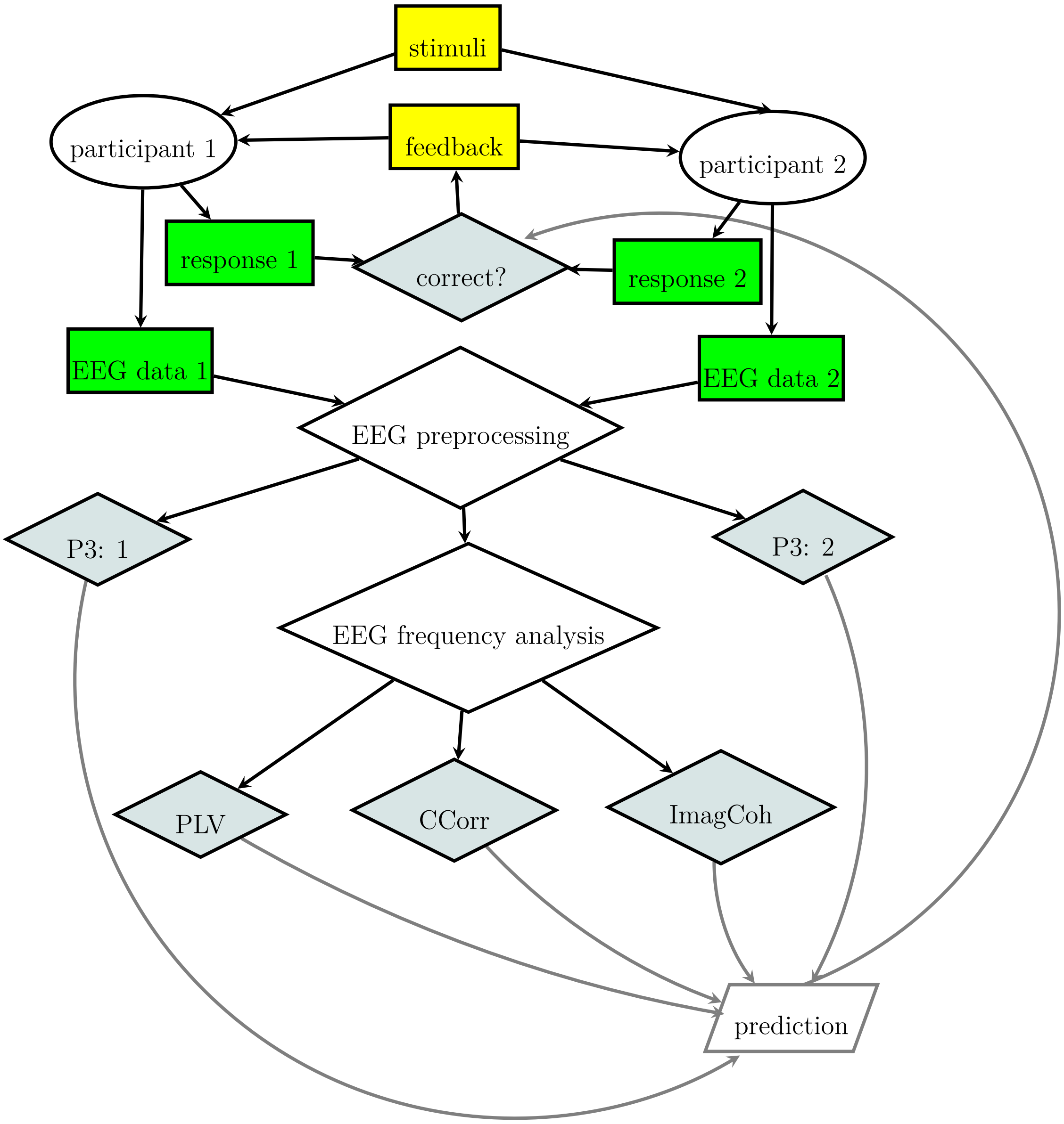


Figure 27

A diagram of the causal structure of Newman et al. (2021)’s experimental setup, and its interaction with the accuracy classifiers described in this thesis. Stimuli are shown in yellow, while recorded measurements are shown in green. Calculated values are shown in grey.

Finally, it is worth reflecting a bit on the prediction task. Considering that we did not find significant IBS previously, it always was a long shot. But even if we had, it is worth mapping out the causal path that would lead to a correct prediction in Newman et al. (2021)'s task. See Appendix B for more information about the task. Figure 27 does exactly that. When a stimulus comes in, both participants give a response, and get feedback based on if they both picked the same image or shape. They use that feedback to adjust their mental model of what the other is doing, which they will use in future trials. We record their brain activity while that is going on, run it through the IBS pipeline, and get out IBS values (PLV, CCorr & ImagCoh in the diagram) and normal EEG values (P3 single trial ERP values). These are then in turn used by the classifier to make a prediction of the accuracy in the current trial. Now, what would be the mechanism that increases the odds of predicting whether the dyad chose the same image or shape?

There are multiple possible ways. Theoretically, a group of images or shapes dissimilar to previous examples could lead to a P3 ERP, and would most likely decrease their chances of picking the same image or shape. But as the images are similar switching only their colours, such an advantage would be unlikely to last long. Alternatively, one of the participants could 'simulate' what the other is doing, thereby mirroring the other's brain activity. The IBS measures could then pick up on this, which the classifier could use to predict a correct response. This is the 'theory of mind' explanation. Personally, I think it unlikely that the functional activity would (1) occur simultaneously enough for the IBS measures to pick up on and (2) would result in a strong, identifiable EEG signal considering that these seem to me relatively abstract, high-level and complex thoughts. Yet another way combines the two. In this case, we assume that integrating (unexpected) feedback causes a P3, or some other neural activity that the IBS measures pick up on due to it presumably being shared across participants. The problem with this explanation is that the activity would need to last into the start of the next trial. There might be other hypotheses, but it is clear that it is not a trivial exercise to find a mechanism that explains why predicting performance would be possible in the first place. Considering our results, perhaps it is not possible.

On the other hand, you could make similar arguments for De Vico Fallani et al. (2010)'s prediction task, which did succeed. Still, considering the causal structure of the problem is probably a worthwhile exercise when attempting prediction using IBS data.

Conclusion

While a lot has been written about the mathematical definitions of different IBS measures, it would be very nice if more intuitive descriptions or visualizations became available. Figure 6 is my own attempt at this, but it has its limitations. It is still my favourite figure in this thesis, though!

It is my hope this research project can contribute to the design of future IBS studies using EEG, by showing the consequences and pitfalls of different methodological choices. As mentioned in the introduction, the standardization of IBS research methods has only just started. But it is encouraging to see that early contributions, like Burgess (2013)'s recommendation to use the CCorr measure, are being taken into account in a lot of studies now appearing.

Appendix A

Theory of Mind

A good introduction about Theory of Mind is given by Postle (2020, p. 455–467). It defines the key process of mentalizing as “engaging in mentation about the thoughts, motivations, and knowledge of another”. It also lists a number of brain regions associated with theory of mind: the right posterior temporal sulcus, the temporal poles, the anterior paracingulate cortex and/or the medial pre-frontal cortex, and finally (to a lesser degree and with lots of caveats) the temperoparietal junction. Theory of Mind co-occurs with the development of executive control, but the mechanism behind that is still an active area of research (Bradford et al., 2015; Perner & Lang, 1999). A high working memory load will disrupt Theory of Mind ability even in adults (Maehara & Saito, 2011), causing them to (incorrectly) fall back on their own beliefs. Finally, impairments in theory of mind may underlie ASD (Baron-Cohen et al., 1985; Frith & Frith, 2005; Postle, 2020, p. 457).

Appendix B

Newman et al.'s tacit coordination task

Tacit coordination tasks, i.e. tasks in which participants have to silently work together, are widely studied: de Weerd et al. (2015) found in a simulation study that higher-order theory of mind ('I know that she knows that I know. . . ') is only useful up to a certain point in such tasks. De Kwaadsteniet and van Dijk (2012) review different coordination rules people use in tacit coordination tasks.

To study the effect of working memory on theory of mind, Newman et al. (2021) developed a tacit coordination experiment in which two participants need to look at four images, and pick the same one. They get to see the other participant's choice after each trial. The underlying idea is that both participants need to apply theory of mind to determine how the other participant makes their choice, so both can converge on a shared strategy and perform at a better than chance level. Newman et al. (2021) tested the effect of working memory load on theory of mind by alternating trials with either a 2-back task (Kirchner, 1958) or a 0-back task (i.e. even/odd classification). During the experiment, EEG data was collected for both participants. That data is analysed in the current project.

One trial in the experiment consists of a fixation cross screen shown between 1000–3000ms to prevent anticipation effects, a self-paced screen where the participant sees the images and chooses one and a feedback screen which is shown for 4000ms. Then, the working memory task takes over with three screens with the same purpose but different timings: the answering screen in the n-back task is always shown 3000ms and the feedback screen is shown for only 1500ms.

Two different abstract stimulus image types are used. One which varies colors, and one which varies shapes. The stimuli were taken from a game-theoretic study by Alberti et al. (2012). This study focuses on modeling why participants tend to prefer certain (more 'salient') images. Luckily, Alberti et al. (2012) found that for the abstract image set the experiment borrows, these preferences tend not to be structural across participants (see also Figure B1).

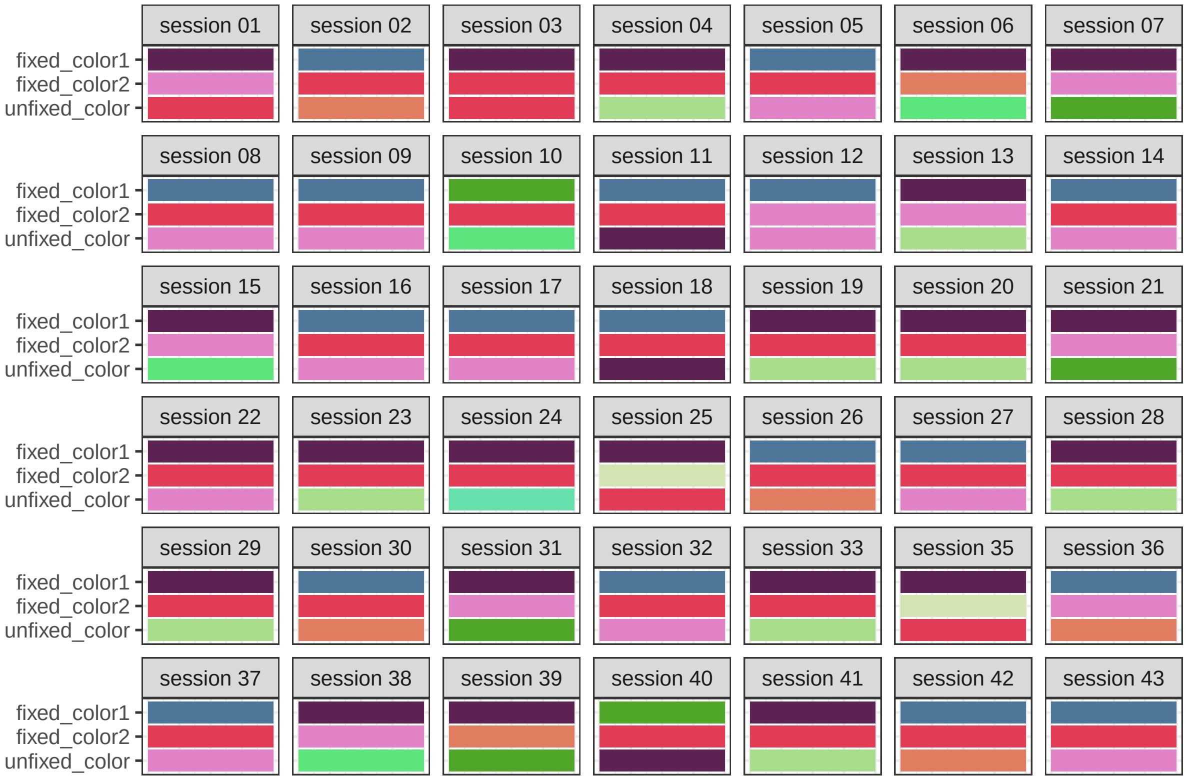


Figure B1

Each dyad's favourite colors for the three parts of the stimulus images.

Finally, it is worth mentioning participants in a dyad were matched for gender and all participants filled in three self-report questionnaires (Christodoulou, 2021): the Interaction Anxiousness Scale, the Interpersonal Reactivity Index and the Autism Spectrum Quotient. This made it possible to check for confounding effects of social anxiety, empathy and ASD (Akcay, 2021).

Appendix C

Linear mixed effect models

purpose	repetition	base model	extra term
effect of window size	by measure	value ∼ trial + (1 | session) + (1 | electrode)	winsize
effect of taper on PLV	none	plv ∼ trial + (1 | session) + (1 | electrode)	taper
effect of taper on CCorr	none	ccorr ∼ wm_load + (1 | session)	taper
effect of taper on ImagCoh	none	imagcoh ∼ trial + (1 | session) + (1 | electrode)	taper
effect of resolution on PLV	none	plv ∼ trial + wm_load + (1 | session) + (1 | electrode)	resolution
effect of resolution on CCorr	none	ccorr ∼ wm_load + (1 | session) + (1 | electrode)	resolution
effect of resolution on ImagCoh	none	imagcoh ∼ trial + stim_type + (1 | session) + (1 | electrode)	resolution
effect of trial	by measure, electrode, band	value ∼ 1 + (1 | session)	trial

Note that the random effect structure of the final model is not always supported by the data, but we decided it better to sometimes have a ‘singular fit’ error than to have a model that sometimes does not account for the structure within sessions.

Appendix D

Supplementary figures

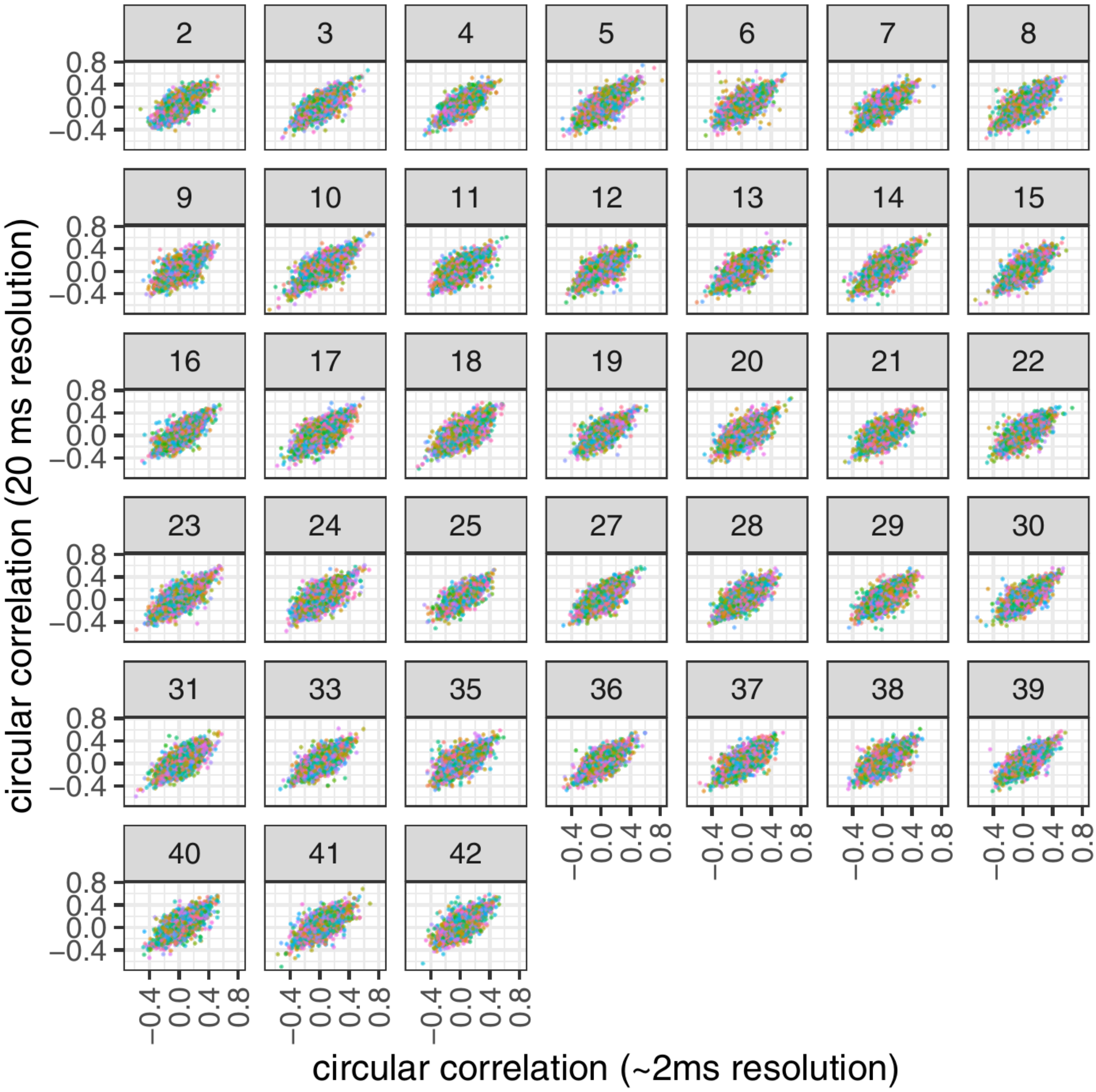


Figure D1

Circular correlation values do not correlate perfectly across different frequency analysis resolutions, contrary to phase locking values and imaginary part of coherency values (not shown here). Each subplot represents a single session. Each dot represents the data for a single timepoint. Colours are assigned based on electrode.

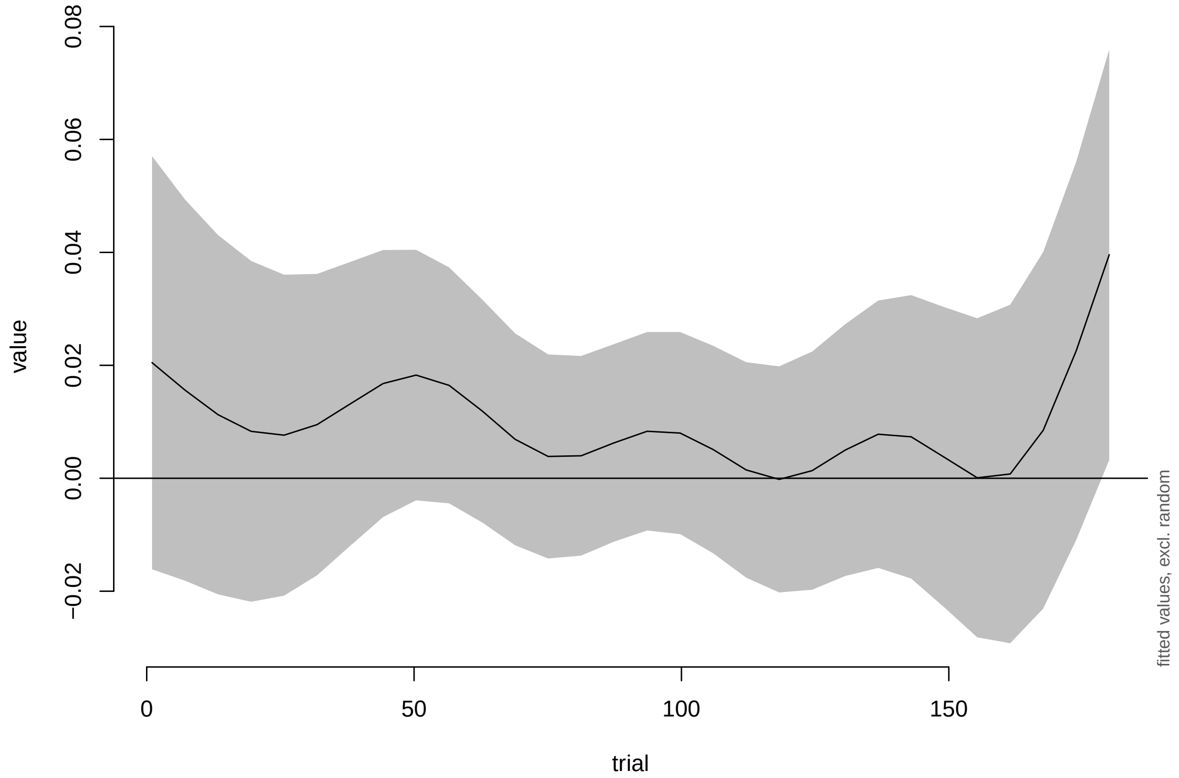


Figure D2

Predicted imaginary part of coherency without random effects in the theta band.

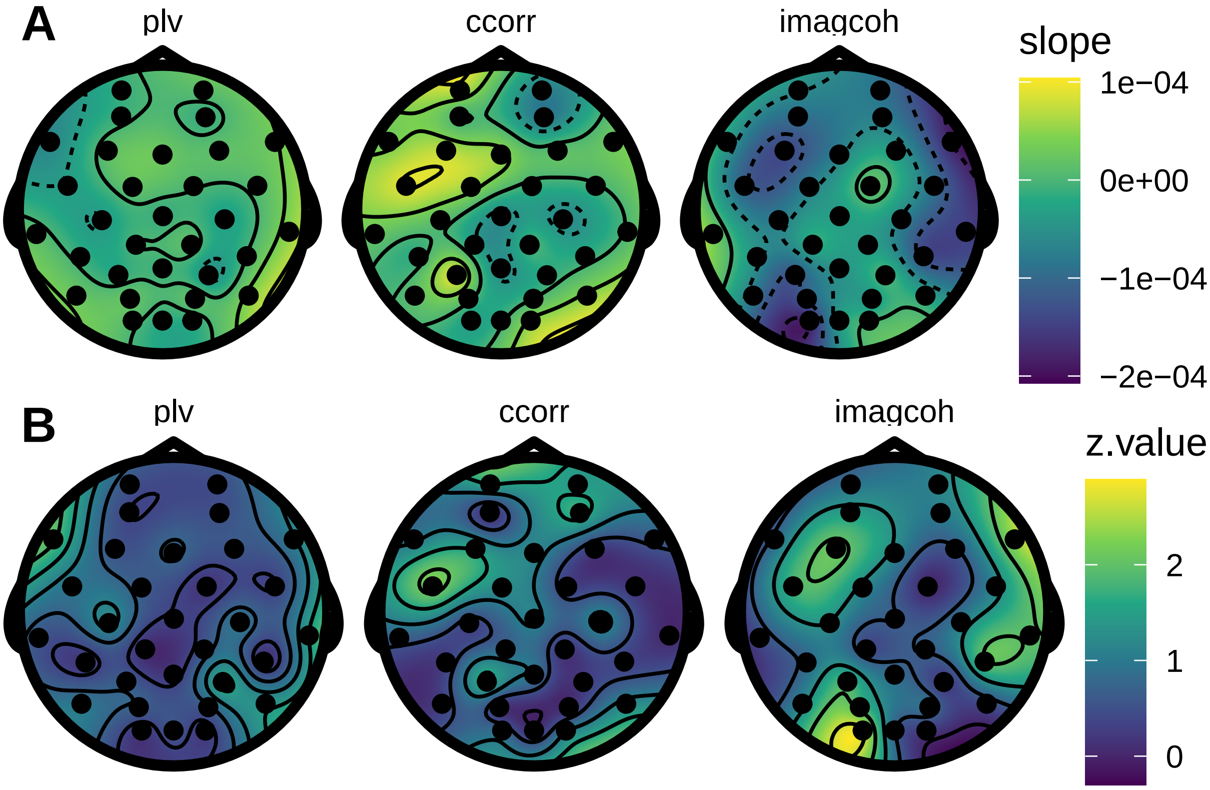


Figure D3

One inter-brain synhrony value is calculated per trial in the theta band. (A) shows their (average) slope when we fit a line through them. (B) shows none of these slopes are significantly different from zero after FDR correction by comparing a linear mixed effect model that includes the slope to one that does not for each electrode.

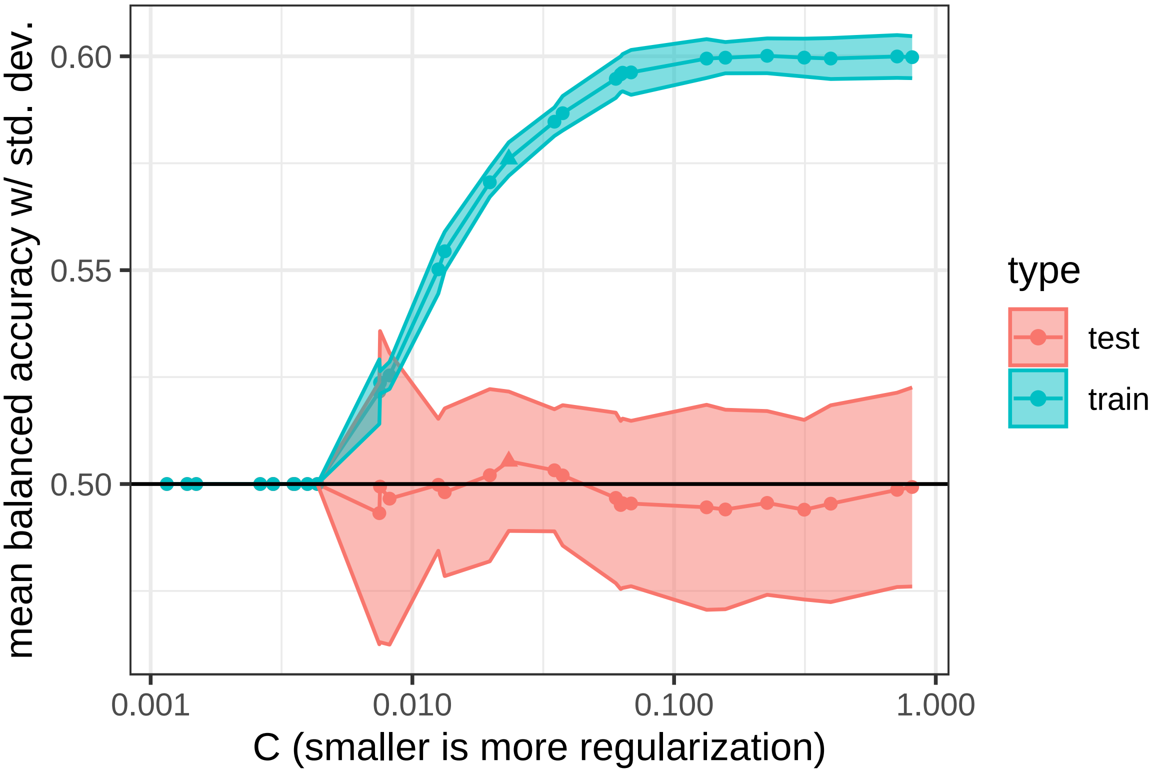


Figure D4

Logistic regression: performance on the train and test set during cross-validation for different regularization parameters. The triangle shows the parameter used for final evaluation.

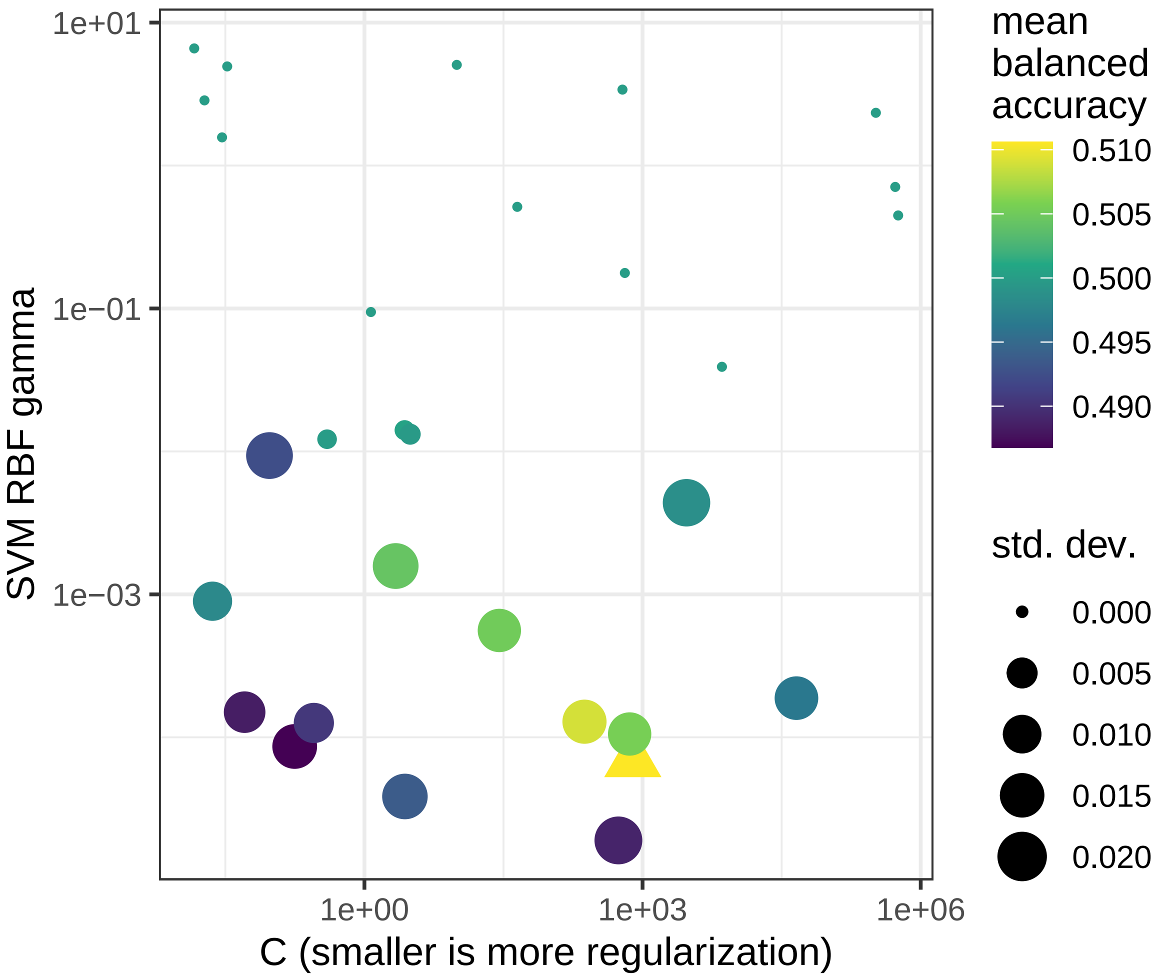


Figure D5

SVM: performance on the test set during cross-validation for different regularization and radial basis function size parameters. The triangle shows the parameters used for final evaluation.

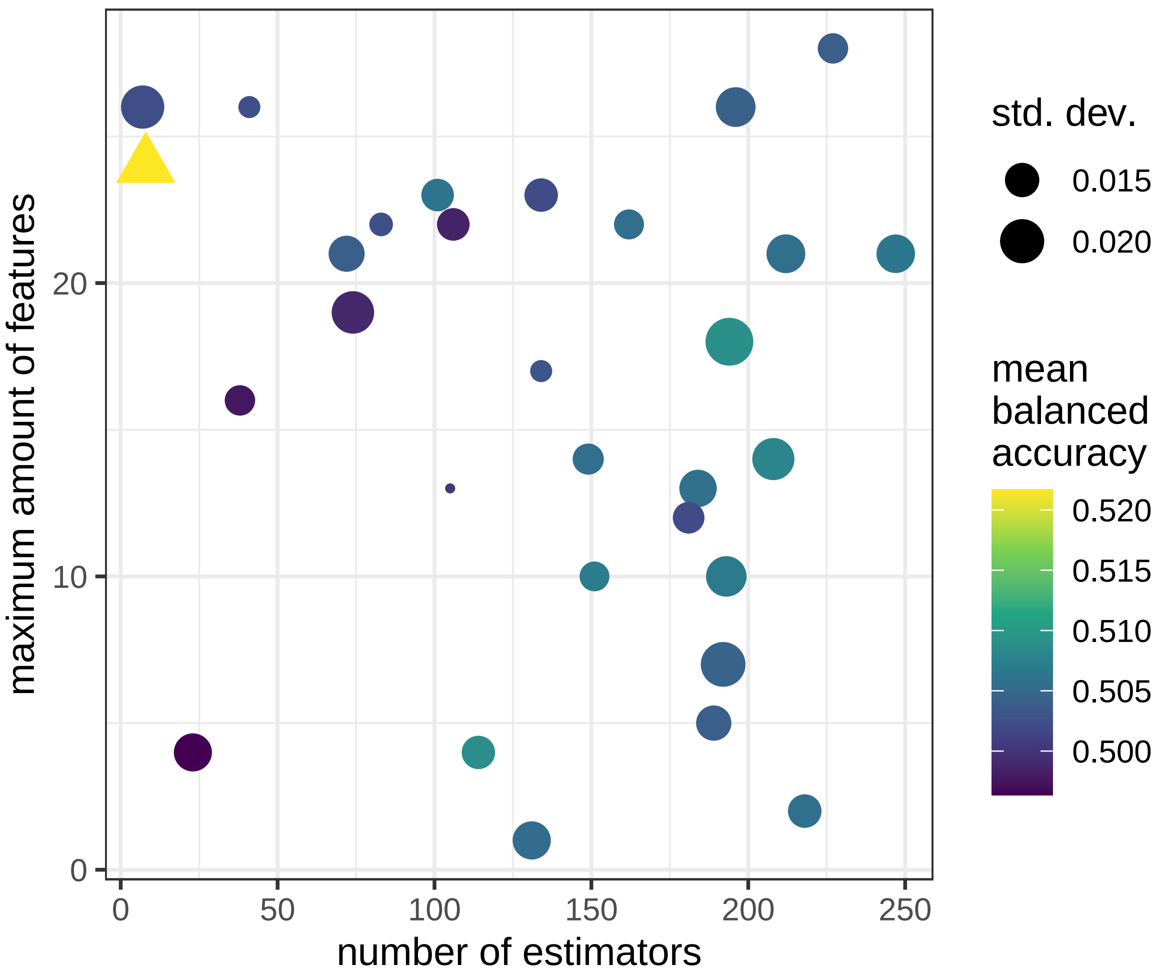


Figure D6

Random Forest: performance on the test set during cross-validation for different number of estimators and maximum amounts of features. The triangle shows the parameters used for final evaluation.

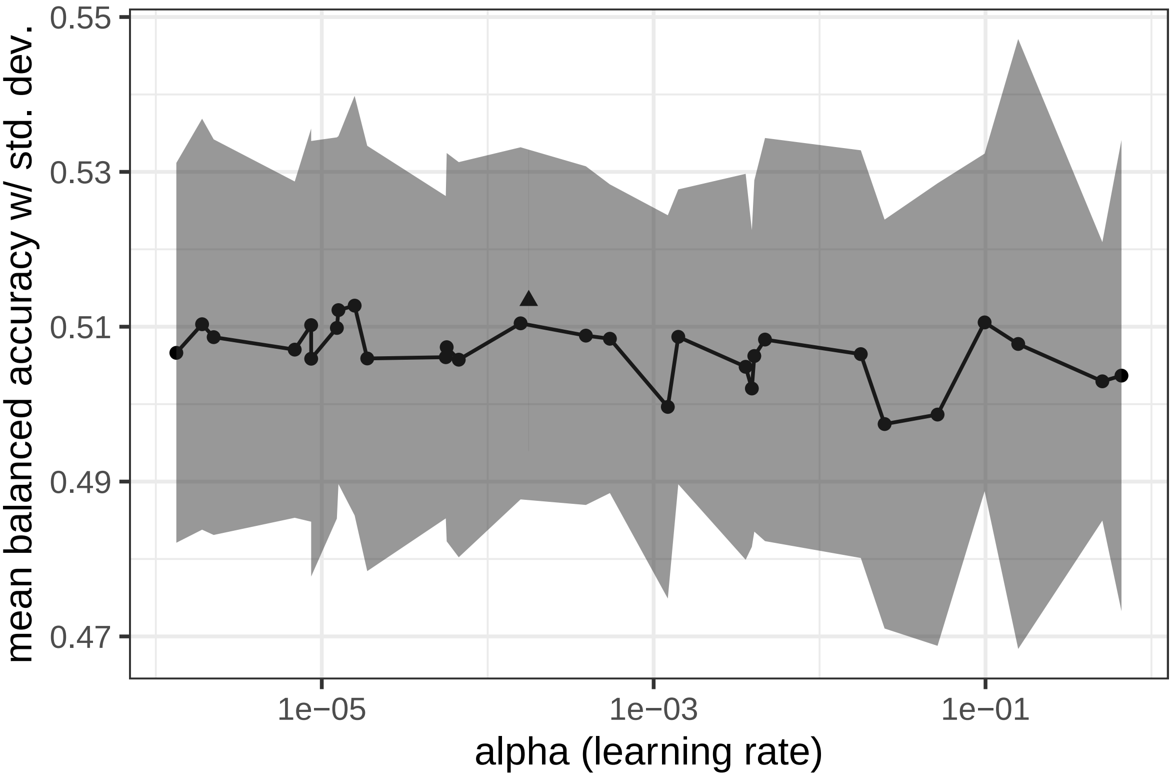


Figure D7

Multi-layer perceptron: performance on the test set during cross-validation for different learning rates. The triangle shows the rate used for final evaluation.

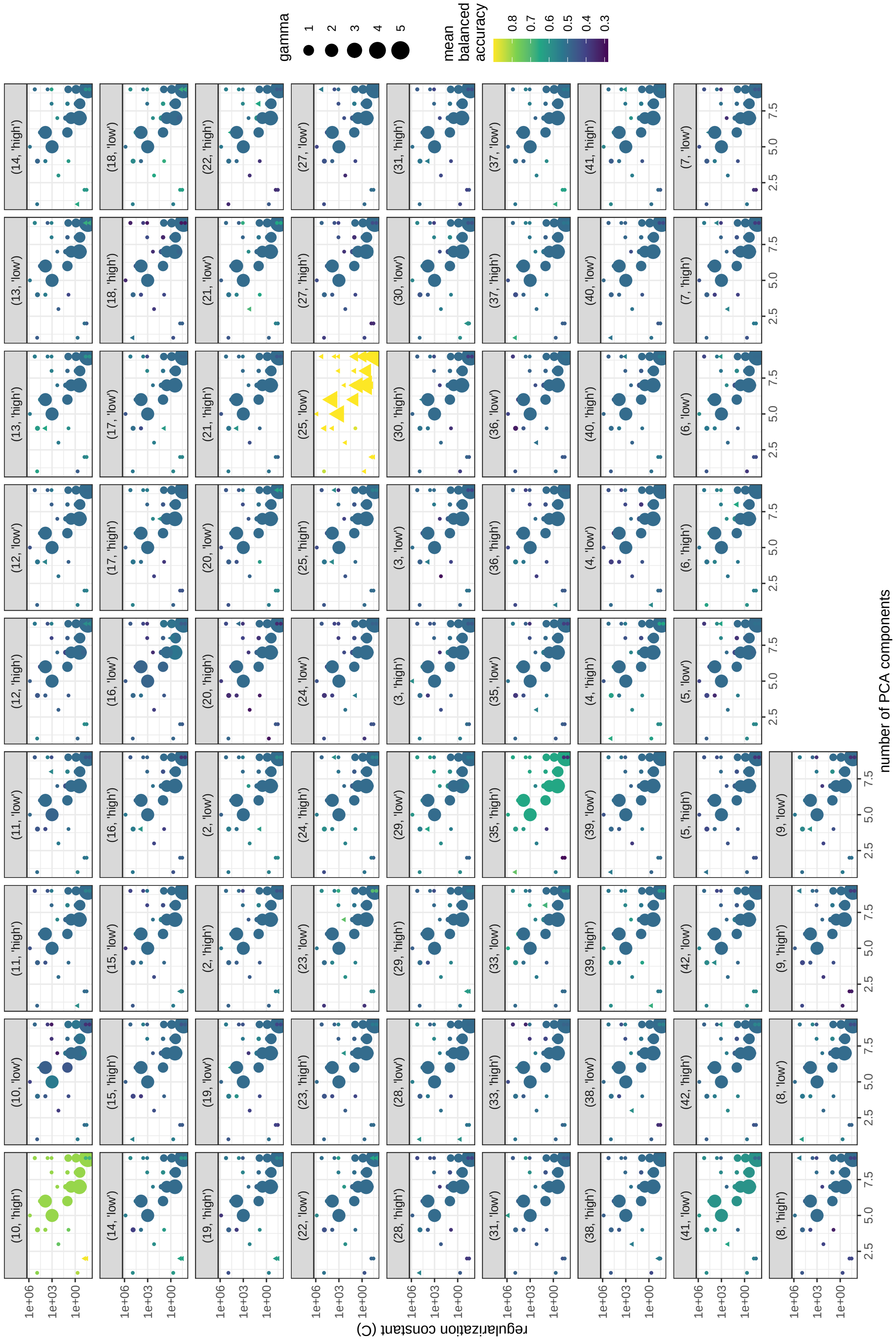


Figure D8

Within-dyad SVM classifiers: performance on the test set during cross-validation for different regularization and SVM RBF gamma parameters. A triangle shows the parameters used for final evaluation. One plot for each session and WM manipulation.

Appendix E

*

References

Akcay, E. B. (2021). The role of empathy, social anxiety and autistic traits in theory of mind: A behavioral study of tacit coordination, 8.

Alberti, F., Sugden, R., & Tsutsui, K. (2012). Salience as an emergent property. *Journal of Economic Behavior & Organization*, *82*(2), 379–394. https://doi.org/10.1016/j.jebo.2011.10.016

Ayrolles, A., Brun, F., Chen, P., Djalovski, A., Beauxis, Y., Delorme, R., Bourgeron, T., Dikker, S., & Dumas, G. (2021). HyPyP: A hyperscanning python pipeline for inter-brain connectivity analysis. *Social Cognitive and Affective Neuroscience*, *16*(1), 72–83. https://doi.org/10.1093/scan/nsaa141

Babiloni, F., & Astolfi, L. (2014). Social neuroscience and hyperscanning techniques: Past, present and future. *Neuroscience & Biobehavioral Reviews*, *44*, 76–93. https://doi.org/10.1016/j.neubiorev.2012.07.006

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Barraza, P., Dumas, G., Liu, H., Blanco-Gomez, G., van den Heuvel, M. I., Baart, M., & Pérez, A. (2019). Implementing EEG hyperscanning setups. *MethodsX*, *6*, 428–436. https://doi.org/10.1016/j.mex.2019.02.021

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bebbington, A. C. (1978). A method of bivariate trimming for robust estimation of the correlation coefficient. *Applied Statistics*, *27*(3), 221. https://doi.org/10.2307/2347156

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical*

Society: Series B (Methodological), *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*(10). https://doi.org/10.18637/jss.v031.i10

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305. http://jmlr.org/papers/v13/bergstra12a.html

Blender Online Community. (2022). *Blender - a 3d modelling and rendering package* (Version 3.3.1). Amsterdam, Stichting Blender Foundation. http://www.blender.org

Bostanov, V., & Kotchoubey, B. (2006). The t-CWT: A new ERP detection and quantification method based on the continuous wavelet transform and student's t-statistics. *Clinical Neurophysiology*, *117*(12), 2627–2644. https://doi.org/10.1016/j.clinph.2006.08.012

Bostanov, V., & Kotchoubey, B. (2004). Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations. *Psychophysiology*, *41*(2), 259–268. https://doi.org/10.1111/j.1469-8986.2003.00142.x

Bradford, E. E., Jentzsch, I., & Gomez, J.-C. (2015). From self to social cognition: Theory of mind mechanisms and their relation to executive functioning. *Cognition*, *138*, 21–34. https://doi.org/10.1016/j.cognition.2015.02.001

Burgess, A. P. (2013). On the interpretation of synchronization in EEG hyperscanning studies: A cautionary note. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00881

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 853–867). Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_40

Chen, P., Kirk, U., & Dikker, S. (2021, June 29). *Trait mindfulness predicts inter-brain coupling during naturalistic face-to-face interactions* (preprint). Neuroscience. https://doi.org/10.1101/2021.06.28.448432

Cheng, X., Li, X., & Hu, Y. (2015). Synchronous brain activity during cooperative exchange depends on gender of partner: A fNIRS-based hyperscanning study: Synchronous brain activities. *Human Brain Mapping*, *36*(6), 2039–2048. https://doi.org/10.1002/hbm.22754

Christodoulou, K. (2021). Effects of working memory load on tacit coordination and inter-brain synchrony, 62. https://fse.studenttheses.ub.rug.nl/25859/

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. The MIT Press.

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. MIT Press.

Craddock, M. (2022). *eegUtils: Utilities for electroencephalographic (EEG) analysis*. https://craddm.github.io/eegUtils/

Czeszumski, A. (2020). Hyperscanning: A valid method to study neural inter-brain underpinnings of social interaction. *Frontiers in Human Neuroscience*, *14*, 17. https://doi.org/10.3389/fnhum.2020.00039

De Vico Fallani, F., Nicosia, V., Sinatra, R., Astolfi, L., Cincotti, F., Mattia, D., Wilke, C., Doud, A., Latora, V., He, B., & Babiloni, F. (2010). Defecting or not defecting: How to "read" human behavior during cooperative games by EEG measurements (O. Sporns, Ed.). *PLoS ONE*, *5*(12), e14187. https://doi.org/10.1371/journal.pone.0014187

DeBruine, L. (2021). *Faux: Simulation for factorial designs*. Zenodo. https://doi.org/10.5281/zenodo.2669586

de Kwaadsteniet, E. W., & van Dijk, E. (2012). A social-psychological perspective on tacit coordination: How it works, when it works, (and when it does not). *European Review of Social Psychology*, *23*(1), 187–223. https://doi.org/10.1080/10463283.2012.718136 de Vries, M. (2022, December). *Marten-de-vries/measuring-inter-brain-synchrony: Final* (Version final). Zenodo. https://doi.org/10.5281/zenodo.7469929

de Weerd, H., Verbrugge, R., & Verheij, B. (2015). Higher-order theory of mind in tacit communication game, 19. https://doi.org/10.1016/j.bica.2014.11.010

Dikker, S., Michalareas, G., Oostrik, M., Serafimaki, A., Kahraman, H. M., Struiksma, M. E., & Poeppel, D. (2021). Crowdsourcing neuroscience: Inter-brain coupling during face-to-face interactions outside the laboratory. *NeuroImage*, *227*, 117436. https://doi.org/10.1016/j.neuroimage.2020.117436

Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J., Michalareas, G., Van Bavel, J. J., Ding, M., & Poeppel, D. (2017). Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current Biology*, *27*(9), 1375–1380. https://doi.org/10.1016/j.cub.2017.04.002

Dumas, G., Chavez, M., Nadel, J., & Martinerie, J. (2012). Anatomical connectivity influences both intra- and inter-brain synchronizations (S. Boccaletti, Ed.). *PLoS ONE*, *7*(5), e36414. https://doi.org/10.1371/journal.pone.0036414

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Inter-brain synchronization during social interaction (J. Lauwereyns, Ed.). *PLoS ONE*, *5*(8), e12166. https://doi.org/10.1371/journal.pone.0012166

Farahzadi, Y., & Kekecs, Z. (2021). Towards a multi-brain framework for hypnosis: A review of quantitative methods. *American Journal of Clinical Hypnosis*, *63*(4), 389–403. https://doi.org/10.1080/00029157.2020.1865129

Fisher, N. I., & Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika*, *70*(2), 327–332. https://doi.org/10.1093/biomet/70.2.327

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507. https://doi.org/10.2307/2331838

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, *15*(17), R644–R645. https://doi.org/10.1016/j.cub.2005.08.041

Garrett, R. C., Nar, A., & Fisher, T. J. (2022). *Ggvoronoi: Voronoi diagrams and heatmaps with 'ggplot2'*. https://CRAN.R-project.org/package=ggvoronoi

Goldstein, P., Weissman-Fogel, I., Dumas, G., & Shamay-Tsoory, S. G. (2018). Brain-to-brain coupling during handholding is associated with pain reduction. *Proceedings of the National Academy of Sciences*, *115*(11). https://doi.org/10.1073/pnas.1703643115

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Hamilton, A. F. d. C. (2021). Hyperscanning: Beyond the hype. *Neuron*, *109*(3), 404–407. https://doi.org/10.1016/j.neuron.2020.11.008

Harris, C. R., Millman, K. J., Walt, S. J. v. d., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. v., Brett, M., Haldane, A., Río, J. F. d., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy [Publisher: Springer Science and Business Media LLC]. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121. https://doi.org/10.1016/j.tics.2011.12.007

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. https://doi.org/10.1038/nn.3101

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, *119*, 456–467. https://doi.org/10.1016/j.neubiorev.2020.09.036

Hu, Y., Pan, Y., Shi, X., Cai, Q., Li, X., & Cheng, X. (2018). Inter-brain synchrony and cooperation context in interactive decision making. *Biological Psychology*, *133*, 54–62. https://doi.org/10.1016/j.biopsycho.2017.12.005

Jammalamadaka, S., & Sarma, Y. (1988). A correlation coefficient for angular variables. statistical theory and data analysis. *Proceedings of the Second Pacific Area Statistical Conference.*, 349–364.

Jammalamadaka, S. R., & Sengupta, A. (2001). *Topics in circular statistics*. World Scientific.

Jin, C. Y., Borst, J. P., & van Vugt, M. K. (2019). Predicting task-general mind-wandering with EEG. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(4), 1059–1073. https://doi.org/10.3758/s13415-019-00707-1

Kassambara, A. (2020). *Ggpubr: 'ggplot2' based publication ready plots*. https://CRAN.R-project.org/package=ggpubr

Kayhan, E., Matthes, D., Marriott Haresign, I., Bánki, A., Michel, C., Langeloh, M., Wass, S., & Hoehl, S. (2022). DEEP: A dual EEG pipeline for developmental hyperscanning studies. *Developmental Cognitive Neuroscience*, *54*, 101104. https://doi.org/10.1016/j.dcn.2022.101104

Kingsbury, L., & Hong, W. (2020). A multi-brain framework for social interaction. *Trends in Neurosciences*, *43*(9), 651–666. https://doi.org/10.1016/j.tins.2020.06.008

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

Koike, T., Tanabe, H. C., & Sadato, N. (2015). Hyperscanning neuroimaging technique to reveal the "two-in-one" system in social interactions. *Neuroscience Research*, *90*, 25–32. https://doi.org/10.1016/j.neures.2014.11.006

Konvalinka, I., & Roepstorff, A. (2012). The two-brain approach: How can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00215

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138

Kurihara, Y., Takahashi, T., & Osu, R. (2022). The relationship between stability of interpersonal coordination and inter-brain EEG synchronization during anti-phase tapping. *Scientific Reports*, *12*(1), 6164. https://doi.org/10.1038/s41598-022-10049-7

Lachaux, J.-P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, *8*(4), 194–208. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C

Lee, A. (2010). Circular data: Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 477–486. https://doi.org/10.1002/wics.98

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Liu, D., Liu, S., Liu, X., Zhang, C., Li, A., Jin, C., Chen, Y., Wang, H., & Zhang, X. (2018). Interactive brain activity: Review and progress on EEG-based hyperscanning in social interactions. *Frontiers in Psychology*, *9*, 1862. https://doi.org/10.3389/fpsyg.2018.01862

Liu, T., & Pelowski, M. (2014). Clarifying the interaction types in two-person neuroscience research. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00276

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *4*(2), R1–R13. https://doi.org/10.1088/1741-2560/4/2/R01

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second edition). The MIT Press.

Madsen, J., & Parra, L. C. (2022). Cognitive processing of a common stimulus synchronizes brains, hearts, and eyes (K. E. Nelson, Ed.). *PNAS Nexus*, *1*(1), pgac020. https://doi.org/10.1093/pnasnexus/pgac020

Maehara, Y., & Saito, S. (2011). I see into your mind too well: Working memory adjusts the probability judgment of others' mental states. *Acta Psychologica*, *138*(3), 367–376. https://doi.org/10.1016/j.actpsy.2011.09.009

Mahmood, E. A. (2022). Robust circular-circular correlation coefficient. *Communications in Statistics - Theory and Methods*, 1–9. https://doi.org/10.1080/03610926.2022.2117561

Maronna, R. A. (2019). *Robust statistics: Theory and methods (with r)* (Second edition). WIley.

McCraty, R. (2017). New frontiers in heart rate variability and social coherence research: Techniques, technologies, and implications for improving group dynamics and outcomes. *Frontiers in Public Health*, *5*, 267. https://doi.org/10.3389/fpubh.2017.00267

McKinney, W. (2010). Data structures for statistical computing in python. In S. v. d. Walt & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Montague, P. (2002). Hyperscanning: Simultaneous fMRI during linked social interactions. *NeuroImage*, *16*(4), 1159–1164. https://doi.org/10.1006/nimg.2002.1150

Newman, L. A., Cao, M., Täuber, S., & van Vugt, M. (2021, January). *Effects of working memory load on tacit coordination* (Poster) [Poster].

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. https://doi.org/10.1016/j.clinph.2004.04.029

Novembre, G., & Iannetti, G. D. (2021). Hyperscanning alone cannot prove causality. multibrain stimulation can. *Trends in Cognitive Sciences*, *25*(2), 96–99. https://doi.org/10.1016/j.tics.2020.11.003

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Pan, Y., Cheng, X., Zhang, Z., Li, X., & Hu, Y. (2017). Cooperation in lovers: An fNIRS-based hyperscanning study: Cooperation in lovers. *Human Brain Mapping*, *38*(2), 831–841. https://doi.org/10.1002/hbm.23421

Pauen, K., & Ivanova, G. (2013). Circular correlation coefficients versus the phase-locking-value. *Biomedical Engineering / Biomedizinische Technik*. https://doi.org/10.1515/bmt-2013-4162

Pedersen, T. L., & Robinson, D. (2020). *Gganimate: A grammar of animated graphics*. https://CRAN.R-project.org/package=gganimate

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, *3*(9), 337–344. https://doi.org/10.1016/S1364-6613(99)01362-5

Pewsey, A., Neuhäuser, M., & Ruxton, G. D. (2013). *Circular statistics in r* (First edition) [OCLC: ocn865566482]. Oxford University Press.

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, *9*(1). https://doi.org/10.2202/1544-6115.1585

Polich, J. (2011, December 15). *Neuropsychology of p300*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0089

Postle, B. R. (2020). *Essentials of cognitive neuroscience* (Second edition). Wiley.

Python Software Foundation. (2021, October 4). *The python language reference, version 3.10*. https://docs.python.org/3.10/reference/

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, *62*(1), 23–48. https://doi.org/10.1146/annurev.psych.121208.131647

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*(4), 1694–1703. https://doi.org/10.1016/j.neuroimage.2004.04.015

Rumelhart, D. E., & McClelland, J. L. (1987). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (pp. 318–362).

Salmi, J., Roine, U., Glerean, E., Lahnakoski, J., Nieminen-von Wendt, T., Tani, P., Leppämäki, S., Nummenmaa, L., Jääskeläinen, I., Carlson, S., Rintahaka, P., & Sams, M. (2013). The brains of high functioning autistic individuals do not synchronize with those of others. *NeuroImage: Clinical*, *3*, 489–497. https://doi.org/10.1016/j.nicl.2013.10.011

Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences*, *107*(20), 9388–9393. https://doi.org/10.1073/pnas.1001791107

Schoffelen, J.-M. (2010, December 22). *Why does my TFR look strange (part II, detrending)?* Retrieved June 3, 2022, from https://www.fieldtriptoolbox.org/faq/why_does_my_tfr_look_strange_part_ii/

Schoffelen, J.-M. (2011, June 8). *In what way can frequency domain data be represented in FieldTrip?* Retrieved June 3, 2022, from https://www.fieldtriptoolbox.org/faq/in_what_way_can_frequency_domain_data_be_represented_in_fieldtrip/

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Shevlyakov, G. L., & Smirnov, P. O. (2010). Robust estimation of a correlation coefficient: An attempt of survey, 9.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188. https://doi.org/10.1126/science.150.3700.1187

Tin Kam Ho. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Valencia, A. L., & Froese, T. (2020). What binds us? inter-brain neural synchronization and its implications for theories of human consciousness. *Neuroscience of Consciousness*, *2020*(1), niaa010. https://doi.org/10.1093/nc/niaa010

van den Brand, T. (2022). *Ggh4x: Hacks for 'ggplot2'.* https://CRAN.R-project.org/package=ggh4x

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2020). Itsadug: Interpreting time series and autocorrelated data using GAMMs.

van Vugt, M. K., Pollock, J., Johnson, B., Gyatso, K., Norbu, N., Lodroe, T., Gyaltsen, T., Phuntsok, L., Thakchoe, J., Khechok, J., Lobsang, J., Tenzin, L., Gyaltsen, J., Moye, A., & Fresco, D. M. (2020). Inter-brain synchronization in the practice of tibetan monastic debate. *Mindfulness*, *11*(5), 1105–1119. https://doi.org/10.1007/s12671-020-01338-1

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wikström, V., Saarikivi, K., Falcon, M., Makkonen, T., Martikainen, S., Putkinen, V., Cowley, B. U., & Tervaniemi, M. (2022). Inter-brain synchronization occurs without physical co-presence during cooperative online gaming. *Neuropsychologia*, *174*, 108316. https://doi.org/10.1016/j.neuropsychologia.2022.108316

Wood, S. N. (2006, February 27). *Generalized additive models* (1st ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781420010404

Wright, K. (2021). *Pals: Color palettes, colormaps, and tools to evaluate them.* https://CRAN.R-project.org/package=pals

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Reports (1896-1970)*, *62*(40), 1432. https://doi.org/10.2307/4586294

Yoshinaga, K., Matsuhashi, M., Mima, T., Fukuyama, H., Takahashi, R., Hanakawa, T., & Ikeda, A. (2020). Comparison of phase synchronization measures for identifying stimulus-induced functional connectivity in human magnetoencephalographic and simulated data. *Frontiers in Neuroscience*, *14*, 648. https://doi.org/10.3389/fnins.2020.00648

Yun, K., Watanabe, K., & Shimojo, S. (2012). Interpersonal body and neural synchronization as a marker of implicit social interaction. *Scientific Reports*, *2*(1), 959. https://doi.org/10.1038/srep00959

Zander, T. O., Kothe, C., Jatzev, S., & Gaertner, M. (2010). Enhancing human-computer interaction with input from active and passive brain-computer interfaces [Series Title: Human-Computer Interaction Series]. In D. S. Tan &

A. Nijholt (Eds.), *Brain-computer interfaces* (pp. 181–199). Springer London. https://doi.org/10.1007/978-1-84996-272-8_11