Bayesian analysis of phase data in EEG and MEG

Sydney Dimmock1*, Cian O’Donnell1,2, Conor Houghton1

1 Department of Computer Science, University of Bristol, Bristol, England; 2 School of Computing, Engineering & Intelligent Systems, Ulster University, Derry/Londonderry, Northern Ireland

Abstract Electroencephalography and magnetoencephalography recordings are non-invasive and temporally precise, making them invaluable tools in the investigation of neural responses in humans. However, these recordings are noisy, both because the neuronal electrodynamics involved produces a muffled signal and because the neuronal processes of interest compete with numerous other processes, from blinking to day-dreaming. One fruitful response to this noisiness has been to use stimuli with a specific frequency and to look for the signal of interest in the response at that frequency. Typically this signal involves measuring the coherence of response phase: here a Bayesian approach to measuring phase coherence is described. This Bayesian approach is illustrated using an example from neurolinguistics and is more descriptive and more data-efficient than the traditional statistical approaches.

Introduction

In an electroencephalography (EEG) or magnetoencephalography (MEG) frequency-tagged experiment the stimuli are presented at a specific frequency and the neural response is quantified at that frequency. This provides a more robust response than a typical event-related potential (ERP) paradigm because the response the brain makes to the stimuli occurs at the predefined stimulus frequency while noise from other frequencies, which will correspond to other cognitive and neurological processes, does not contaminate the response of interest.

Frequency tagging is a well-established tool in the study of vision, where it is often referred to as the steady state visual evoked potential Regan (1966). At first, it was predominately used to study low-level processing and attention, see Norcia et al. (2015) for a review. Latterly, though, it has been used for more complex cognitive tasks, such as face recognition and discrimination Alonso-Prieto et al. (2013); Farzin et al. (2012); Liu-Shuang et al. (2014), perception of number Guillaume et al. (2018); Van Rinsveld et al. (2020), and the “human quality” of dance movements Alp et al. (2017). It has been applied to other modalities: audition Galambos et al. (1981); Picton et al. (2003); Bharadwaj et al. (2014), somatosensation Tobimatsu et al. (1999); Regan (1989) and nociception Colon et al. (2012, 2014). It has been used to study broad phenomenon like memory, Lewis et al. (2018) and lateralisation, Lochy et al. (2015, 2016) along with more specific types of neurocognitive response, such as visual acuity Barzegaran and Norcia (2020) and the perception of music Nozaradan (2014). Furthermore frequency tagging can be used in the assessment of disorders such as autism Vettori et al. (2020a,b) and schizophrenia Clementz et al. (2008). It has even being used to study neural responses to social interaction Oomen et al. (2022). Beyond EEG and MEG, phase coherence has been proposed as a mechanism for signal routing Abeles (1982); Salinas and Sejnowski (2001); Börgers and Kopell (2008), assembly formation Singer (1999); Buzsáki (2010) and coding O’Keefe and Recce (1993); Panzeri et al. (2010), so the measurement of phase coherence for
electrocorticography, local field potentials and neuronal spiking is important for the neuroscience of neuronal systems.

One striking application of frequency tagging is in neurolinguistics Ding et al. (2016, 2017). Neurolinguistic experiments are difficult; since language experiments inevitably involve humans and target a phenomena whose temporal grain is often too fine for magnetic resonance imaging, the principal neural imaging techniques are EEG and MEG. However, the complexity of the neural processing of language makes the signals recorded in these experiments particularly noisy, difficult to analyse and difficult to disentangle from other cognitive processes. A further difficulty in neurolinguistics is that an ERP is often difficult to obtain because the tens or hundreds of repetitions required would render the stimulus meaningless to the participant, a phenomenon known as the semantic satiation Jakobovits (1962).

Consider, as an example, the frequency-tagged experiment described in Burroughs et al. (2021a) and following the paradigm given in Ding et al. (2017). There, the neural response to phrases was investigated by comparing the response to grammatical adjective-noun (AN) phrases

\[ \text{... old rat sad man ill wife ...} \]

and to ungrammatical adjective-verb (AV) pairs

\[ \text{... rough give ill tell thin chew ...} \]

where care had been taken to have a similar 2-gram frequency for adjacent word pairs in each condition. The words are all presented at 3.125 Hz, however the frequency of interest is the ‘phrase rate’, 1.5625 Hz, corresponding to the phrase structure of the AN stimuli, see Figure 1. In Burroughs et al. (2021a) it is suggested that the strength of the response to AN stimuli relative to AV stimuli at this frequency measures a neural response to the grammatical structure, rather than to lexical category of the words.

This investigation required a quantitative measurement of the strength of the response. The obvious choice: the power at 1.5625 Hz does not work, empirically this proves too noisy a quantity for the stimulus-dependent signal to be easily detected and, indeed, although the frequency tag produces a more robust signal than an ERP, for more high-level or cognitive tasks where frequency-tagging is now proving valuable, the power is not a useful measure. Instead the typical approach to frequency-tagged data for cognitive tasks is to use the inter-trial phase coherence. The inter-trial phase coherence (ITPC) is defined using the mean phase angle:

\[
R(f, \theta) = \frac{1}{K} \sum_k e^{i\theta_{f,k}}
\]

where \( f \) is the frequency, \( k \) is the trial index, \( K \) is the number of trials, \( \theta \) represents other indices such as electrode number or experimental condition and \( \theta_{f,k} \) is the phase of the complex Fourier coefficient for the EEG or MEG trace \((f, k, \theta)\). Across different applications, mean phase angle is often called the mean resultant, a term we will use here. The ITPC is the length of the mean resultant:

\[
R(f, \theta) = |R(f, \theta)|
\]
The ITPC is chosen as a quantitative measure to extract the portion of the response at the frequency of interest that is phase locked to the stimulus and therefore consistent in phase from trial-to-trial. This is another of the denoising strategies required by these noisy data.

In the case of the experiment we are discussing here the hold-all index \( \phi \) is made up of participant, condition and electrode indices. To produce a result, the ITPC is averaged over the 32 electrodes used in the experiment to give \( R(f,p,c) \) where \( p \) labels participants and \( c \) labels conditions. The principal result of Burroughs et al. (2021a) is that, in the language of frequentist statistics, \( R(f = 1.5626 \text{ Hz}, p, c = \text{AN}) \) is significantly larger than \( R(f = 1.5626 \text{ Hz}, p, c = \text{AV}) \).

The result of these experiments analysed using ITPC are summarised in Figure 2. This graphs the ITPC measure for all six experimental conditions, the two, AN and AV, that have already been described and four others; a table of the experimental conditions is provided as Table 2. In Figure 2A it is seen that there is a strong peak in ITPC at the syllable rate, 3.125 Hz, and, in the case of AN, at the phrase rate: 1.5625 Hz. The ITPC at the phrase rate is graphed in Figure 2B where, again, it appears only AN has a response at the phrase rate.

In Burroughs et al. (2021a) all analysis is done for ITPC averaged across electrodes; nonetheless in Figure 2C we show the condition-to-condition difference in ITPC at each electrode, but averaged across participants:

\[ \Delta R_{pc} = \langle \Delta R_{pe,c} \rangle_p \]  

where  

\[ \Delta R_{pe,c} = R(f = 1.5626 \text{ Hz}, e, c = \text{AN}) - R(f = 1.5626 \text{ Hz}, e, c = \text{RR}). \]

Visually these data appear to show a left temporal and right parietal response. However, the data are noisy and deciding the significance of any comparison is complicated: a straightforward calculation of the \( p \)-value using an uncorrected paired two-sided Wilcoxon signed-rank test gives a value less than 0.05 for 54 of the 480 possible comparisons. This includes some comparisons that fit well with the overall picture, for example when comparing AN to AV the P4 electrode shows a difference with \( p = 0.002 \) and the T7 electrode with \( p = 0.044 \). It also includes some more surprising results: for the comparison of RR to RV the CP5 electrode has \( p = 0.0182 \) and the FC1 electrode has \( p = 0.0213 \). If we interpret these as significant difference, the apparent difference between two conditions without an apparent phrase structure is odd and, presumably misleading. However, a naïve Bonferroni correction would use a factor of \( 32 \times 15 = 480 \) and in a manner typical of this very conservative approach, it massively reduces the number of significantly different responses, to one in this case.

There are a number of disadvantages to the ITPC. The most obvious problem is that the item in the statistical analysis of ITPC is a participant, not a trial. In the results described in Burroughs et al. (2021a) the statistical significance relied on a \( t \)-test between conditions with a pair of data points for each participant: there are actually 24 trials for each participant but these are used to calculate the ITPC values. Some of the analysis in Burroughs et al. (2021a) is done using 20 participants, some using 16; sticking to the latter for simplicity, the hypothesis testing is performed using 16 pairs of values, rather than \( 16 \times 24 = 384 \) or even \( 16 \times 24 \times 32 = 12288 \) items if the individual electrodes are included. In short, the ITPC is itself a summary statistic, a circular version of variance, and so it hides the individual items inside a two stage analysis.

\[ \text{items} \rightarrow \text{ITPC} \rightarrow \text{statistical analysis} \]  

However, this is hard to rectify: it is difficult to compare items across participants, or across electrodes, because the mean phase, \( \text{phase}[R(f, \phi)] \) is very variable and not meaningful to the scientific questions of interest. This variability is graphed in Figure 3: this Figure shows the value of

\[ \mu_{pe,c} = \text{phase}[R(f = 1.5625, p, e, c = \text{AN})] \]

the phase of the mean resultant for the AN condition. For illustrative purposes, three example electrodes are picked out and the distribution across participants is plotted. What is clear is how
Figure 2. Summarising the ITPC for different conditions. 

A: Coloured lines show the ITPC for each participant traced across all frequencies. The mean trace is overlaid in black. Vertical lines mark the sentence, phrase, and syllable frequencies as increasing frequencies respectively. 

B: Statistical significance was observed with an uncorrected paired two-sided Wilcoxon signed-rank test (*: 0.05, **: 0.01). 

C: ITPC differences calculated at the phrase frequency and interpolated across the skull.
variable these phases are; this means that individual responses cannot be compared across participants and electrode since $p$ and $e$ have such a strong effect on their value.

Here we provide a Bayesian approach to phase data. We believe this has advantages when compared to the ITPC: it permits a per-item analysis and correspondingly a more statistically efficient and richer use of the data. Furthermore, as a Bayesian approach, it supports a better description of the data, replacing a hypothesis-testing and significance-based account with a narrative phrased in terms of models and their consequence.

A Bayesian account provides a parameterised probabilistic model of the data. The model proposes a distribution for the data given a set of model parameters: this is the likelihood. In our case, the likelihood will be the probability for the phases of the responses, given our model. Our interest is in how the variance of this distribution depends on the condition. In addition to the likelihoods, the full Bayesian model also includes a prior distribution for parameters, for example, it includes priors for the parameters which determine the relationship between the condition and the distribution of phases. The goal is to calculate the posterior distribution, the probability distribution for the parameters given the experimental data, this follows from Bayes’s theorem:

$$P(\Theta | \Delta) = \frac{P(\Delta | \Theta) P(\Theta)}{P(\Delta)}$$

where $\Theta$ are the parameters and $\Delta$ the data. Essentially, $P(\Delta | \Theta)$ is the likelihood, the distribution of the data given some parameters: the goal is to take the data and from them calculate the posterior distribution of the parameters: $P(\Theta | \Delta)$. The denominator $P(\Delta)$ can usually be ignored because it is just a normalising constant that is independent of the parameter values and therefore does not change the shape of the posterior distribution. There are a number of excellent descriptions of the Bayesian approach to data, including the textbook (Gelman et al., 1995) and a recent review van de Schoot et al. (2021); our terminology and notation will often follow conventions established by the textbook McElreath (2018). In many ways Bayesian descriptions are more intuitive and easier to follow than the frequentist approaches that have been favoured over the last century. The impediment to their use has been the difficulty of calculating the posterior distribution. These days, however, powerful computing resources and new insight into how to treat these models mean that there are a variety of approaches to estimating the posterior; one approach, the one used here, is to sample from the posterior without calculating it analytically using Markov chain Monte-Carlo (MCMC) techniques. Probabilistic programming languages such as Stan, (Carpenter et al., 2017), and Turing, (Ge et al., 2018), make it easy to use advanced MCMC sampling methods such as Hamiltonian / Hybrid Monte-Carlo (HMC) and the no U-turn sampler (NUTS) (Duane et al., 1987; Neal, 2011; Betancourt, 2013) making the complexity of a frequentist analysis unnecessary. Here, we report results calculated using Stan though many of the computations were carried out in both Stan and Turing.

Methods and Materials

Circular distributions

Here, the data are a set of phases and so the model for the data is a probability distribution on a circle. The motivation which informs the ITPC is that the phases to a greater or lesser extent have a unimodal distribution around the circle and so the model should be a unimodal distribution on the circle; see Figure 3. One common class of distributions on the circle is given by the wrapped distribution with probability density function

$$p(\theta) = \sum_{n=-\infty}^{\infty} p_{r}(\theta + 2\pi n)$$

where $p_{r}(x)$ is a probability density of a distribution on the real line and $\theta$ is the angle. It might seem that the obvious choice for $p_{r}(x)$ would be the Gaussian distribution. In fact, the wrapped Gaussian
Figure 3. Mean phases are uniform across participants. The left hand panel shows the distribution of phases across electrodes for each participant: each column corresponds to one participant and each dot marks the mean phase $\mu$ for each of the 32 electrodes calculated at the phrase frequency for the AN grammatical condition. To show how a given electrode varies across participant three example electrodes are marked, T7 in teal, P4 in orange and F8 in blue. The right hand panel shows the distribution of mean phases across participants: for each of the three example electrode the violin plots show the distribution $\mu$ across participants.

The distribution is not a very satisfactory example of a wrapped distribution because $p(\theta)$ cannot be calculated in closed form. A much better example is the Cauchy distribution

$$p_c(x) = \frac{1}{\pi \gamma} \frac{\gamma^2}{(x - x_0)^2 + \gamma^2}$$  \hspace{1cm} (9)$$

where $x_0$ is a parameter giving the median of the distribution and $\gamma$ is a scale parameter; the corresponding wrapped distribution has the closed form:

$$p(\theta) = \frac{1}{2\pi \cosh \gamma - \cos(\theta - \mu)}$$  \hspace{1cm} (10)$$

and, in contrast to the Cauchy distribution on the real line, where the moments are not defined, the wrapped distribution has a well-defined and convenient value for the mean resultant:

$$R = e^{i\mu - \gamma}$$  \hspace{1cm} (11)$$

Thus, as illustrated in Figure 4A, a large value of $\gamma$ corresponds to a highly dispersed distribution; a low value to a concentrated one. With this explicit relationship between parameter values and the mean resultant, the Cauchy distribution is a convenient choice for our model.

Prior distributions

The next important element is the choice of priors both for the mean of the angular distribution, $\mu$, and for $\gamma$, which determines how dispersed the distribution is. The prior for $\mu$ is the more straightforward: a different value of $\mu$ is required for each participant, condition and electrode. This prior should be uniform over the unit circle. Although there is likely to be correlations in $\mu$ values for the same electrode across participants and for the electrodes for a given participant, since the value of $\mu$ is not of interest it is convenient to ignore this and pick an independent value $\mu_{pc}$ for each triplet of participant-condition-electrode values. Future studies that aim to extend this model could consider adding correlations to $\mu$.

Since $\mu$ has a uniform prior over the unit circle it would seem that the correct prior is

$$\mu_{pc} \sim \text{Uniform}(-\pi, \pi)$$  \hspace{1cm} (12)$$

This is, however, wrong: the ideal distribution is a uniform distribution on the circle, not on the interval and while the uniform distribution on an interval has the same numerical probability values, it has a different topology. This matters when sampling using a MCMC method. In MCMC, to create the list of samples, referred to as the chain, the sampler moves from sample to sample, exploring
Figure 4. Model construction and geometry. A: Example wrapped Cauchy density functions for different values of the scale parameter. B: The Bundt distribution has a shape reminiscent of a cake made in a Bundt tin. C: The mean phase is sampled from an axially symmetric prior distribution with a soft constraint on the radius. Highlighted pairs \((x, y)\) give the location of example points; these points correspond to the mean angle for a wrapped Cauchy distribution using \(\text{angle}(x, y)\). D: An example distribution for \(S = 1 - R\) which determines the \(\gamma\) parameter in the wrapped Cauchy distribution. \(S\) is related to other parameters such as condition and participant number through a logistic regression, as in Eq. 17, the priors for the slopes in the regression are used to produce the distribution shown here; as before, two example points are chosen, each will correspond to a different value for \(\gamma\) in the corresponding wrapped Cauchy distribution. E: Example wrapped Cauchy distributions are plotted in correspondence with the numbered prior proposals in C and D.
the parameter space. In this exploration for $\mu$, if the posterior value is, for example, close to $\pi$ then the chain should explore the region near to $\pi$, which includes values near $-\pi$ in $[-\pi, \pi]$. A small change should move the sampler from $\pi$ to $-\pi$. However, dynamics on the interval $[-\pi, \pi]$ can only get from one to the other by traversing the potentially low-likelihood interval in between. Nothing in the mathematical description of the common MCMC samplers, such as NUTS, prevents the prior from being defined on a circle or other compact region. However, there is a problem: the current high-quality implementations of these methods in Stan and Turing do not allow priors over circles.

**Sampling from a circular prior distribution**

As a practical approach to avoiding this difficulty, we introduce a two-dimensional distribution which, in polar coordinates, is uniform in the angle coordinate and in the radial part restricts sampling to a ring around the origin. Because its probability density function resembles the Bundt cake tin, used to make kugelhopf Hudgins (2010), see Figure 4B, this will be referred to as a Bundt distribution. The choice of the radial profile of the Bundt distribution is not critical, its purpose is to restrict the samples to a ring: we sample points $(x, y)$ on a plane so that their radius $r = \sqrt{x^2 + y^2}$ is drawn from a gamma distribution

$$
\rho \sim \text{Gamma}(10, 0.1). \quad (13)
$$

giving what we will call a Bundt-gamma distribution. This distribution has mean 1 and standard deviation 0.1 giving the golden ring of likely $(x, y)$ values seen in Figure 4C. In fact, the radial values are not used in the model; what is used is the angle:

$$
\mu_{pce} = \text{angle}(x_{pce}, y_{pce}) \quad (14)
$$

**A linear model for the scale of the wrapped Cauchy distribution**

The final element of the model is the prior for $\gamma$; obviously the intention is have this depend on the condition. To make our priors easier to interpret it is convenient to use a link function, first converting from $\gamma$ to the circular variance $S$:

$$
\gamma_{pce} = -\log(1 - S_{pce}) \quad (15)
$$

$S$ is bound between zero and one so a second link function is applied

$$
S_{pce} = \sigma(\gamma_{pce}) \quad (16)
$$

where $\sigma(\cdot)$ is the logistic function. The quantity $\gamma_{pce}$ quantifies the effect of participant, condition and electrode on response. In this model it is linear

$$
\gamma_{pce} = \alpha_{c} + \beta_{p} + \delta_{e} \quad (17)
$$

so $\alpha_{c}$ is understood as quantifying the effect of condition, $\beta_{p}$ the effect of participant and $\delta_{e}$ the effect of electrode. In the language of regression, these are slopes. In the case of $\beta_p$ and $\delta_e$, experimenting with different models has demonstrated a better fit when these are interaction terms, allowing the effect of respectively participant and electrode to be condition dependent.

Thus, the main objects of interest are $\alpha_{c}$, $\beta_{p}$, and $\delta_{e}$ and our result is calculated by sampling the posterior distribution for these quantities. Of course, these quantities also require priors; these are described as part of a full description of the model in the supporting information, see Appendix Full model. The prior for the slope $\beta_{p}$ has a hierarchical structure, allowing correlations across conditions: $\beta_{p}$ models the participant response and, roughly speaking, modelling the idea that a participant who is not paying attention in one condition is likely to be inattentive for all of them.

The posteriors were sampled using the NUTS algorithm implementation in Stan. Four chains were run for 4000 iterations, with half dedicated to a warm up or calibration epoch. Details of the software packages and libraries used can be found in the Appendix Software.
Results

Comparison with ITPC results

The posterior distributions are described in Figure 5. This figure reprises, using our analysis, the ITPC analysis exhibited as Figure 2. Figure 5A show a point estimate of the mean resultant length across all frequencies estimated using the optimise function within RStan; as in the earlier figure, Figure 2A, there is a phase peak visible at the phrase frequency 1.5626 Hz for AN, but not for the other conditions.

At the bottom of Figure 5B we see the posterior distributions over the mean resultant length for each condition. These are obtained by transforming posterior samples \( \alpha_c \) that describe the effect of condition on the response within the regression to circular variance, as described in Eq. 16, then subtracting one. It appears that the AN condition has a higher value of the mean resultant length than the other five conditions. To examine this further the upper panel in Figure 5B also shows the 90% highest density intervals (HDIs) and posterior medians of the posterior distribution over the differences between the mean resultant length of all condition pairings. The HDI provides a summary of the full posterior distribution: a HDI is the smallest width interval that contains a specified proportion of the total probability density and here, above the violin plot for each \( \alpha_c \), we have plotted the HDI for that condition relative to the other four: this could be considered as a Bayesian equivalent to the confidence brackets common in frequentist plots like Figure 2. Here, the only HDIs which do not overlap zero are the ones corresponding to the difference between AN and another condition: this clearly shows that in our model there is a neural response at the phrase stimulus frequency for AN but not for the other conditions. It appears, for example, that although the MP condition consists of grammatical phrases, the fact that these phrases are of different types means that there does not appear to be a response. This suggests that the neuronal response observed for AN is a response to a specific type of phrase, not to any phrase.

In Figure 5C we see the electrode-by-electrode comparisons across conditions. These graphs show a clearer structure than the corresponding ITPC analysis in Figure 2C; there is a left temporal and right parietal response for AN and nothing else.

Participant effects

In Figure 6A we plot the 90% HDIs for the participant slopes, \( \beta_p \), for \( c = \text{AN} \); more positive values of \( \beta \) correspond to less attentive participants, more negative values correspond to more attentive. These have been arranged in increasing order of \( \beta \) with the participant number \( p \) given on the \( x \)-axis. From an experimental point-of-view, this plot gives some reassurance that there is no systematic trend, with participation becoming better or worse as the experiment progressed through participants. Our model includes a condition-dependent standard deviation for the participant response, see Appendix Full model; posterior distributions for these standard deviations are plotted in Figure 6B. This appears to indicate there is more across-participant variation in responses to the MP and ML conditions, where there is a structure but a complicated or confusing one, than to either the highly structured and grammatical AN condition or the RV and RR conditions, with little or no structure at the phrase rate.

Electrode effects

To investigate the electrode dependent response Figure 6C is an enlarged version of of the first headcap plot from Figure 5C: the difference in mean resultant between AN and AV. The heat map colour-scale is recalibrated since here it refers only to this one plot. The localisation of the response is seen very clearly. It is difficult to combine a headcap plot and information about the posterior distribution, so the HDI for

\[
\Delta R_e = R_{c_1e} - R_{c_2e}
\]

where \( c_1 = \text{AN}, c_2 = \text{AV} \) and

\[
R_{ce} = 1 - \sigma(\alpha_c + \delta_{ce})
\]
Figure 5. Posterior distributions. A: The traces show point estimates of the mean resultant length calculated across all 58 frequencies using the optimisation procedure. B: The marginal posterior distributions for each transformed condition effect $\alpha_i$ are shown with a violin plot. Posteriors over condition differences are given directly above, the colour of which represents the condition against which the comparison is made. Posterior differences and marginal intervals are all given as 90% highest density intervals marked with posterior medians. C: Posterior medians are interpolated across the skull for all condition comparisons.
is plotted for three example electrodes, one electrode from each of the two active areas and one from an area that shows little activity. The response for P4 and T7 is clearly different from zero, indicating that there is a stronger response to the AN condition than to the AV condition at these two electrode. The same HDI analysis for RR versus RN does not show any electrodes whose HDI does not overlap zero; the presumably misleading results for CPS and FC1 noted in the discussion of ITPC results do not appear here.

**Sampler diagnostics**

When calculating posteriors using MCMC it is necessary to check the success of sampling; sometimes it can become stuck in one part of the parameter space Gelman et al. (1995); McElreath (2018). Figure 7 plots standard diagnostic measures for our MCMC sampling quality. There does not appear to have been any problems: the most commonly used measure of the success of sampling is $\hat{R}$, often referred to as R-hat. This is a measure of sampling correlation along and across chains; ideally it would be 1.0, but typically a value of less than 1.05 is considered acceptable and less that 1.02 desirable. Here, all values of R-hat are less than 1.006 indicating good mixing; values are plotted in Figure 7A; Figure 7C plots the chains for the parameter with the largest R-hat value for each parameter type; none of these plots appear to show the sort of pathological behaviour associated with poor sampling, chains are both stationary and convergent. Another measure of sampling success, the comparison of marginal and transitional probabilities, is exhibited in Figure 7B; this also indicates good sampling. See Appendix Tree depth warnings for a note on tree depth warnings.

**Discussion and Conclusions**

Here, we have presented a Bayesian description of phase data using a particular example from neurolinguistics. Our approach reprises the original conclusions of the statistical analysis of these data, but, we believe, does so in a more expressive and more natural way. Our account focuses on neurolinguistics, where frequency tagging is common and we use a specific neurolinguistic example, an example with which we are familiar and for which data is openly available Burroughs et al. (2021b). However, we believe that our approach has broad application across the multiple applications of frequency tagging. Bayesian analysis is, essentially, a more modest and honest approach to data than the more commonly used frequentist analysis: where a frequentist approach seeks to establish with significant certainty whether a hypothesis is true or false, in a Bayesian analysis we restrict ourselves to the more achievable goal of estimating the values of parameters in a model of the data and calculating our certainty or uncertainty in making those estimates.

**Model design choices**

Our model resembles a logistic regression with the circular variance $S_{\rho\omega}$ related to condition, participant and electrode through parameters $\alpha_i$, $\beta_p$, and $\delta_e$, Eq. 17. These parameters all need priors; the obvious place to start is the condition effects $\alpha_i$, because effects are weak in these data our prior belief is that for any condition the circular variance should be reasonably large, likely bigger than a half. Conversely, the parameters $\beta_p$ and $\delta_e$ correspond to deviations about the baseline level $\alpha$, which can be represented easily using unbounded symmetric distributions. The participants slopes, $\beta_p$ were assigned a multivariate t-distribution, the t-distribution was chosen because its heavy tails give a more robust estimation in the presence of ‘unusual’ participants: exceptional strong or exceptionally worse, probably due to lack of attention. The multivariate parameterisation allows for a simultaneous two-way regularisation process due to information sharing both within conditions and across conditions. This idea of self-regularising priors is common in hierarchical Bayesian models, and is often referred to as partial pooling, see Gelman et al. (1995) for a review. A similar approach was adopted for the electrode slopes however these were not treated using partial pooling because testing showed that this was not useful on this dataset.
Figure 6. Participant attentiveness and localised electrode effects. A: The intervals show participant effects for the grammatical AN condition given as 50/90% HDI's and posterior medians. B: The posterior distributions over the standard deviation of participant slopes for each condition. Outer vertical lines mark the 90% posterior HDI's, inner lines marks the posterior median. C: The skull plot from Fig.5C for the AN-AV difference with electrode names marked. D: Posterior distributions over electrode differences for those positions on the skull where the grammatical condition shows a higher coherence of phases at the average participant in (C). Intervals give 50/90% HDI's and the posterior medians.
Figure 7. Sampler performance and diagnostics. 

A: The performance of the sampler is illustrated by plotting \( \hat{R} \) (R-hat) against the ratio of the effective number of samples for each parameter in the model. Points represent individual model parameters grouped by colour with a different colour for each parameter type. For convenience the dot sizes are scaled so the more numerous parameters have smaller dots, the less numerous, fewer, so, for example, \( \alpha \) with only six examples, is large.

B: A histogram comparing the marginal energy distribution \( E \), and the transitional energy distribution \( \Delta E \) of the Hamiltonian.

C: Post-warmup trace plots. All four chains for the poorest performing parameter within each parameter group are overlaid. Corresponding points in A are marked with a black border and zero transparency.
Figure 8. Efficiency of the frequentist and Bayesian approaches. A: The \( p \)-value is plotted for \( \Delta R \), the comparison between the ITPC for the AN and RR conditions: this is calculated using an uncorrected paired Wilcoxon signed-rank test with decreasing numbers of participants. B: The 90/95/99.16\% posterior HDI’s for the AN-RR comparison. The Bayesian model is re-fitted for each of the same reduced data sets as A.

Data efficiency

The Bayesian approach also appears to make more efficient use of the data. In order to investigate the data efficiency of the frequentist and Bayesian approaches, the result has been simulated for the data we would have had if the experiment had been stopped early, down to a minimum participant number of four. It can be misleading to directly compare frequentist and Bayesian results; the aims of the two approaches are different; nonetheless we have done just that in Figure 8. In Figure 8A we plot the uncorrected \( p \)-value for a two-sided uncorrected paired Wilcoxon signed-rank test for the experimental data with smaller and smaller participant numbers produced by removing the participants starting with the last tested. We also mark the lines for \( \alpha = 0.1 \), \( \alpha = 0.05 \) and \( \alpha = 0.0083 \); these would correspond to significant differences in an uncorrected one-sided test, an uncorrected two-sided test and a two-sided test in which a Bonferroni correction of six is used to correct for multiple comparisons across the four phrase conditions. We are not advocating for any of these \( \alpha \) values and the uncertainty in deciding an appropriate value of \( \alpha \) plagues frequentist approaches; however, for comparison, in Figure 8B we plot the HDI of 90\%, 95\% and 99.16\% for the equivalent \( \Delta R \). This shows that the posterior points to a difference from zero in cases where the low participant number has pushed the \( p \)-value over one or other of the \( \alpha \) values.

Interestingly, when both participants 15 and 16 are removed from the data, leaving 14 participants, the \( p \)-value increases by a factor of approximately two (n=15 : 0.005, n=14 : 0.011). Figure 6A can explain this result since the posteriors for \( \beta \) shows that participant 15 performs better on the task than participant 16 so removing this participant from the analysis weakens the result more than participant 16. For this particular analysis the degrees of freedom parameter \( \nu \) was fixed to 30 to address divergent transitions arising for small participant numbers.

It may appear that our motivation is contradictory; we first explain that frequency-tagging produces robust encephalography results, but then explain that a new framework is required to analyse these results because they are often too noisy to study using a naive power analysis. Of course, there is no contradiction; the encephalographic study of cognitive phenomena like language demands both a robust experimental paradigm and a cutting edge analysis pipeline!

EEG data can benefit from a Bayesian analysis

The Bayesian approach we have advanced in this paper is undoubtedly much more computationally demanding than a frequentist approach; it also demands some thought and experiment in the formulation of the model and its priors. Frequency tagging is, in this regard, a particularly demanding application of the approach. However, we believe that the clarity of a Bayesian description and the complete way it presents the model and its evidence, along with the great data efficiency it
provides, makes it superior. Some of the complexity of our approach derives from the difficulty of sampling a circle and we hope this example will be helpful in incorporating compact distributions into the standard probabilistic packages such as Stan and Turing.

In general, Bayesian models become worth the effort in scenarios with two properties: 1) where the data are limited and noisy so statistical uncertainty is high and therefore worth representing explicitly; 2) where the dataset has a strong structure, which the Bayesian model can be designed to match and therefore share information across parameters. For these reasons, we also believe that similar Bayesian approaches will have broad application to EEG data. The nature of EEG data, its noisiness high-dimension and the tendency to small participant numbers, make it likely that Bayesian methods will be helpful. This certainly is evident in the preliminary work report in Turco and Houghton (2022).

Acknowledgements
CH is a Leverhulme Research Fellow (RF-2021-533). We would also like to acknowledge funding from the MRC (MR/S026630/1 to COD), and an EPSRC Doctoral Training Partnership award to SD.

We would like to recognise the following discussion\(^1\) that highlighted some of the difficulties involved with sampling directional statistics and the potential ways to ameliorate them in our model.

Competing interests
No competing interests.

References


\(^1\)https://discourse.mc-stan.org/t/divergence-treedepth-issues-with-unit-vector/8059


Regan D. Human brain electrophysiology. Evoked potentials and evoked magnetic fields in science and medicine. 1989; .


Full model
In the Bayesian model the individual phases are modeled as draws from a wrapped Cauchy distribution:

$$\theta_{pcek} \sim \text{Wrapped-Cauchy}(\mu_{pce}, \gamma_{pce})$$

(20)

where, as above, $p$, $c$ and $e$ are participant, condition and electrode number and $k$ is the trial number. The mean phase is derived from the Bundt-gamma distribution:

$$(x_{pce}, y_{pce}) \sim \text{Bundt-Gamma}(10, 0.1)$$

(21)

The probability density function for the Bundt-gamma distribution can be derived through a Jacobian adjustment from polar to Cartesian coordinates. Our assumptions in polar coordinates are a uniform angle, and a gamma distributed radius:

$$\rho_{pce} \sim \text{Gamma}(10, 0.1)$$

(22)

$$\mu_{pce} \sim \text{Uniform}(\pi, -\pi)$$

(23)

To represent these assumptions in Cartesian coordinates we multiply by $1/\rho$:

$$p(x_{pce}, y_{pce}) = \frac{1}{\rho_{pce}} p(\rho_{pce}, \mu_{pce})$$

(24)

$$= \frac{1}{2\pi \rho_{pce}} \text{Gamma}(\rho_{pce}, 100, 0.1)$$

(25)

This gives an angle uniform on the circle, not on the interval:

$$\mu_{pce} = \text{angle}(x_{pce}, y_{pce}).$$

(26)

As described above, the model for $\gamma$ uses a pair of link functions so

$$\gamma_{pce} = -\log(1 - S_{pce})$$

(27)

and

$$S_{pce} = \sigma(v_{pce})$$

(28)

with a linear model for $v_{pce}$:

$$v_{pce} = c_e + p_{vc} + e_c$$

(29)

We have priors for each of $a_e, \beta_p$ and $\delta_{vc}$, what in linear regression are referred to as slopes. The prior for $a_e$ is induced through placing a prior over $\sigma(a_e)$ which represents the baseline circular variance for each condition

$$\sigma(a_e) \sim \text{Beta}(3, 2)$$

(30)

By applying the change of variables formula we can work out the pdf for the prior induced on $a_e$:

$$p(a_e) = \text{Beta}(\sigma(a_e)|3, 2) \frac{e^{a_e}}{(1 + e^{a_e})^2}$$

(31)

As discussed above, for $\beta_p$ we have a hierarchical structure modelling covariance of participant responses across conditions, thus:

$$\beta_p \sim \text{MvT}(v, \mathbf{0}, \Sigma)$$

(32)
where \( \beta_e \) is a vector over the \( e \) index. With six conditions, the scale matrix \( \Sigma \) is a \( 6 \times 6 \) matrix. It is made up of a correlation matrix \( \Omega \) and a set of scales, \( \sigma_1 \) to \( \sigma_6 \).

\[
\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_6) \cdot \Omega \cdot \text{diag}(\sigma_1, \ldots, \sigma_6)
\]  

(33)

To facilitate the interpretation as a covariance matrix this scale matrix needs to be multiplied by \( \nu/(\nu - 2) \). The correlation matrix has a Lewandowski-Kurowicka-Joe prior \cite{Lewandowski2009, Gelman1995}:

\[
\Omega \sim \text{LKJ}(2)
\]  

(34)

The prior for the degrees of freedom parameter \( \nu \) is given a gamma prior:

\[
\nu \sim \text{Gamma}(2, 10)
\]  

(35)

and the scales have half-normal priors:

\[
\sigma_e \sim \text{Half-Normal}(0, 0.5)
\]  

(36)

Finally, for \( \delta_e \), we partially pool electrodes within condition only:

\[
\delta_e \sim \text{Normal}(0, \tau_e)
\]  

(37)

\[
\tau_e \sim \text{Half-Normal}(0, 0.5)
\]  

(38)

To attempt a standard notation we have followed the conventions set by the \texttt{julia} library \texttt{Distribution.jl} by writing the distributions as words and using the same arguments as are found there: in particular the two parameters for the Gamma distribution correspond to shape and scale.

The prior distributions for \( \beta \) and \( \delta \) were implemented using a reparameterisation known as non-centring \cite{Papaspiliopoulos2007}. This is a commonly adopted technique in hierarchical Bayesian modelling to help alleviate funnels, a class of pathological feature in the target distribution that cause slow and biased sampling. This reparamaterisation does not change the mathematical model; its sole purpose is to help the numerical computation. See \cite{McElreath2018}, and \cite{Betancourt2015}, for an introduction to this approach.
Software
Posters were sampled using rstan v2.21.5, and cmdstanr v0.5.2. Data and posters were analysed using R v4.2.1; tidyverse v1.3.1; reshape2 v1.4.4, and HDInterval v0.2.2. All graphs were plotted in ggplot2 v3.3.6. Figure 2B used ggsignif v0.6.3 for hypothesis testing and additional plotting functionality; Figure 4B used viridis v0.6.2 for heatmap colours; headcaps were interpolated using mgcv v1.8-40 for Figure 2C, Figure 5C and Figure 6C; ridgeplots were created for Figure 6B with ggridges v0.5.3; Hamiltonian energy distributions were plotted in figure Figure 7B using bayesplot v1.9.0 Gabry and Mahr (2022); Gabry et al. (2019). All panels were composed using inkscape v1.1.1.

Code and data
The data used here are from the open dataset Burroughs et al. (2021b); all code is available at https://github.com/conorhoughton/NeuralProcessingOfPhrases.

Tree depth warnings
The sampler has been observed to produce a low number (< 1%) of max_treedepth warnings. This does not imply biased computation like those arising from divergences, but it is a warning about efficiency. A higher tree depth comes at the cost of doubling the number of gradient evaluations required at the previous depth Hoffman et al. (2014), adding a penalty to the run time.

Computing resources
Posters were sampled on a locally on a Dell XPS 13 7390 laptop (Intel i7-10510U @ 1.80GHz, 16GB of RAM) running under Ubuntu 20.04.4 LTS.
Appendix 3

Table of experimental conditions

The six experimental conditions were:

<table>
<thead>
<tr>
<th>c</th>
<th>description</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN</td>
<td>adjective noun pairs</td>
<td>...old rat sad man...</td>
</tr>
<tr>
<td>AV</td>
<td>adjective verb pairs</td>
<td>...rough give ill tell...</td>
</tr>
<tr>
<td>ML</td>
<td>adjective pronoun verb preposition</td>
<td>...old this ask in...</td>
</tr>
<tr>
<td>MP</td>
<td>mixed grammatical phrases</td>
<td>...not full more green...</td>
</tr>
<tr>
<td>RV</td>
<td>random words with every fourth a verb</td>
<td>...his old from think...</td>
</tr>
<tr>
<td>RR</td>
<td>random words</td>
<td>...large out fetch her...</td>
</tr>
</tbody>
</table>

Of these four are “phrase conditions”, AN, AV, MP and RR, and were analysed in Burroughs et al. (2021a); the other two, ML and RV, were “sentence conditions” which formed part of the experiment and were used to investigate phenomena which proved to be absent. ML stands for “mixed lexical” and provides a four-word analogue of the AV condition, repeating lexical category but avoiding grammatical structure. All stimuli are available, see the data and code availability list at Appendix Code and data.