UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

The relationship between frequency content and representational dynamics in the decoding of neurophysiological data

Higgins, Cameron; van Es, Mats W.J.; Quinn, Andrew J.; Vidaurre, Diego; Woolrich, Mark W.

DOI:
10.1016/j.neuroimage.2022.119462

License:
Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Higgins, C, van Es, MWJ, Quinn, AJ, Vidaurre, D & Woolrich, MW 2022, 'The relationship between frequency content and representational dynamics in the decoding of neurophysiological data', *NeuroImage*, vol. 260, 119462. https://doi.org/10.1016/j.neuroimage.2022.119462

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 16. May. 2024

NeuroImage 260 (2022) 119462

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

The relationship between frequency content and representational dynamics in the decoding of neurophysiological data

Cameron Higgins[a], Mats W.J. van Es[a,*], Andrew J. Quinn[a], Diego Vidaurre[a,b], Mark W. Woolrich[a]

[a] Oxford Centre for Human Brain Activity, Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, Oxford, UK
[b] Center of Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

ARTICLE INFO

Keywords:
Representational dynamics
Decoding
Aliasing
Complex spectrum decoding

ABSTRACT

Decoding of high temporal resolution, stimulus-evoked neurophysiological data is increasingly used to test theories about how the brain processes information. However, a fundamental relationship between the frequency spectra of the neural signal and the subsequent decoding accuracy timecourse is not widely recognised. We show that, in commonly used instantaneous signal decoding paradigms, each sinusoidal component of the evoked response is translated to double its original frequency in the subsequent decoding accuracy timecourses. We therefore recommend, where researchers use instantaneous signal decoding paradigms, that more aggressive low pass filtering is applied with a cut-off at one quarter of the sampling rate, to eliminate representational alias artefacts. However, this does not negate the accompanying interpretational challenges. We show that these can be resolved by decoding paradigms that utilise both a signal's instantaneous magnitude and its local gradient information as features for decoding. On a publicly available MEG dataset, this results in decoding accuracy metrics that are higher, more stable over time, and free of the technical and interpretational challenges previously characterised. We anticipate that a broader awareness of these fundamental relationships will enable stronger interpretations of decoding results by linking them more clearly to the underlying signal characteristics that drive them.

1. Introduction

The field of representational dynamics uses temporal patterns in decoding accuracy timecourses to test hypotheses about how the brain processes information (Carlson et al., 2013; Cichy et al., 2014; Kietzmann et al., 2019; King & Dehaene, 2014). By decoding different experimental stimuli from recorded brain activity at high temporal resolution, researchers use information theoretic measures to quantify what features of a stimulus are explicitly represented in neural data as a function of time from stimulus onset (Carlson et al., 2011; Cichy et al., 2016; Ince et al., 2017). An emerging question in neuroscience is how these representational dynamics relate to the brain's underlying neurophysiology (Gross et al., 2013; Jafarpoura et al., 2013; Kriegeskorte & Kievit, 2013; Schyns et al., 2011). Such analyses seek to go beyond merely answering *what* is represented in recorded brain activity, by also characterising the neural mechanisms explaining *how* that information is represented (Higgins, Vidaurre, et al., 2021; Kikumoto & Mayr, 2018; Valentin et al., 2020; van de Nieuwenhuijzen et al., 2013; Zhan et al., 2019).

This commonly involves a decoding paradigm we will refer to as *instantaneous signal decoding*, where classifiers are trained and tested on the raw broadband signal recorded over all sensors at each timepoint following a stimulus (Carlson et al., 2013; Cichy & Pantazis, 2017; Grootswagers et al., 2017), and the representational dynamics interpreted (often with reference to activity in canonical frequency bands). This can be used for example to study the phase-locking of information content to canonical oscillations (Kerrén et al., 2018; Kunz et al., 2019; van Es et al., 2020), the dynamics of memory (Higgins, Liu, et al., 2021; LaRocque et al., 2013; Wolff et al., 2015), or the direction of information flow (Cichy et al., 2014; Dijkstra et al., 2020; Goddard et al., 2016; Linde-Domingo et al., 2019). A closely related paradigm, we will refer to as *narrowband signal decoding*, applies the same procedure after filtering the data into a narrowband of interest. This explicitly links observed patterns with canonical frequency bands (Samaha et al., 2016; Xie et al., 2020).

Unfortunately, however, the fundamental relationship between the frequency content of the stimulus evoked signal and the inferred information content is not widely recognised. Whilst many decoding approaches aim to be agnostic about the specific data characteristics over time that drive their results, there is a considerable risk of misinterpretation when this relationship is not considered. In this paper we draw attention to this relationship, highlighting that the spectral con-

* Corresponding author.
E-mail address: mats.vanes@psych.ox.ac.uk (M.W.J. van Es).

https://doi.org/10.1016/j.neuroimage.2022.119462.
Received 3 March 2022; Received in revised form 4 July 2022; Accepted 8 July 2022
Available online 22 July 2022.
1053-8119/© 2022 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

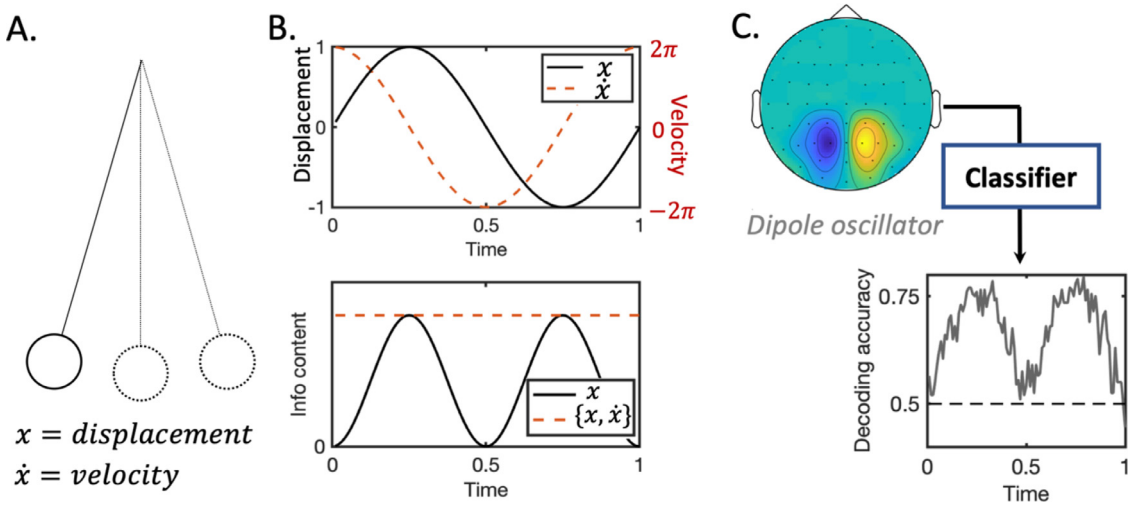

Fig. 1. a pendulum analogy for decoding oscillatory neural signals. A. Commonly used *instantaneous signal decoding* pipelines can only offer a partial view of the brain's representational dynamics, as they only use the instantaneous data values and cannot detect information stored in the gradient or higher moments of the dynamic signal trajectory. When dealing with a dynamical system such as the brain, this is like trying to predict the behaviour of a pendulum given only its displacement at a single instant in time – not its velocity or momentum, which would fully characterise the dynamic system. **B.** Suppose we wish to classify if a pendulum is moving or stationary given noisy estimates of its displacement and velocity over time. The information content associated with only the displacement readout peaks at the pendulum's extrema and drops to zero in between. Including the velocity information instead achieves stable information content over time. **C.** Evoked neural data with strong oscillatory components behaves in the same way as the pendulum. When researchers apply *instantaneous signal decoding* to such data, classifiers should perform well at the peaks and troughs of sinusoidal components in the evoked response, and poorly in between. This problem would be overcome if the local gradient information was included as features for classification, resulting in information content metrics that are stable over time.

tent of the evoked response is translated to double its original frequency in associated decoding accuracy metrics when using the *instantaneous signal decoding* or *narrowband signal decoding* paradigms most typically used in the literature (Carlson et al., 2013; Cichy & Pantazis, 2017; Grootswagers et al., 2017). From this, we identify two problems: the first is the presence of artefacts due to representational aliasing; the second is the broader challenge of how we should interpret information theoretic metrics that systematically oscillate at double the frequency of the evoked response spectrum.

We argue that these problems arise from a narrow focus on information content in the instantaneous signal at a single moment in time, which ignores information stored in the signal's gradient or higher moments. Conceptually, this is analogous to analysing a simple pendulum by measuring only its displacement at a single instant in time – not its velocity or acceleration, which would together fully define the dynamic system. As illustrated in Fig. 1, such a narrow focus only on the pendulum's displacement leads to inferred information content that peaks at the pendulum's extrema (i.e. the peaks and troughs of the oscillation); taking a broader view of the information contained in both the displacement and velocity leads to a measure of information content that is stable over time.

We extend the same logic to the dynamic trajectory of neural activity evoked by a stimulus. This motivates a third decoding paradigm that we refer to as *complex spectrum decoding* (Angjelichinoski et al., 2019; Ince et al., 2017), which is one way of including such temporal gradient information. Returning to our example above, if we applied a Fourier decomposition to the pendulum's displacement over time, we would obtain a single complex frequency component with a real part (tracking the displacement) and an imaginary part (tracking the velocity). This concept generalises to neural activity, where we would expect more complex Fourier dynamics played out simultaneously over multiple frequency bands and spatial channels. When this complex spectrum information is included as features to a classifier, we show that this results in inferred representational dynamic patterns that have higher accuracy, are more stable over time, and which we believe to provide a better characterisation of the brain's representational architecture.

2. How the spectrum of the evoked response determines the signal information content

We first ask: what is the fundamental relationship between frequency-specific features of the stimulus-evoked response and the resulting timecourse of decoding accuracy? We address this question using a generative modelling approach, where we model the neural data recorded on individual trials as a Fourier series with bandlimited Gaussian noise. From a probabilistic modelling perspective, the *mutual information* is the theoretical quantity analogous to decoding accuracy that we can then derive. This allows us to characterise how the information content of a signal varies as a function of time and frequency (Table 1).

2.1. Generative model of stimulus evoked responses

We wish to model epoched electrophysiological data recorded from P channels under two different experimental conditions. Let us denote by $x_{n,t}$ the $[P \times 1]$ vector of data recorded at time $t \in \{1, 2, \dots T\}$ on trial $n \in \{1, 2, \dots N\}$, where $y_n \in \{1, -1\}$ denotes the experimental condition for that trial. We model $x_{n,t}$ as comprising a condition-independent evoked response term μ_t of dimension $[P \times 1]$, and residual terms that are decomposed into a sum of $[P \times 1]$ Fourier components $z_{n,t,\omega}$:

$$x_{n,t} = \mu_t + \sum_{\omega=0}^{\Omega} z_{n,t,\omega} \tag{1}$$

We henceforth refer to $x_{n,t}$ as the ***'broadband signal'***, and the multiple $z_{n,t,\omega}$ terms as the ***'narrowband signals'***. If we assume each narrowband signal $z_{n,t,\omega}$ has a multivariate Gaussian distribution with mean conditioned on the stimulus (see Appendix A for full details), we obtain the following expression for the distribution of the broadband signal:

$$P(X_t|Y) \sim N\left(\mu_t + Y \sum_{\omega=0}^{\Omega} A_\omega \cos\left(\omega t + \phi_\omega\right), \sum_{\omega=0}^{\Omega} \Sigma_\omega\right) \tag{2}$$

Each A_ω term is a diagonal $[P \times P]$ matrix, where the ith diagonal entry, denoted by $a_{\omega,i}$, reflects the magnitude of the component at frequency ω on channel i. Both ω and t are scalar indices reflecting the frequency and time respectively; ϕ_ω is a $[P \times 1]$ vector, each entry of which

Table 1
Overview of random variables modelled in this paper.

Random Variable	Observation on nth trial	Domain and dimension	Used to model:
X_t	$x_{n,t}$	$\mathbb{R}^{1\times P}$	Recorded broadband signal at time t; input used for *instantaneous signal decoding*
Y	y_n	$\{1,-1\}$	Stimulus class
$Z_{t,\omega}$	$z_{n,t,\omega}$	$\mathbb{R}^{1\times P}$	Narrowband signal at time t in frequency band ω; input used for *narrowband signal decoding*
$W_{t,\omega}$	$w_{n,t,\omega}$	$\mathbb{C}^{1\times P}$	Complex spectrum signal at time t in frequency band ω; input used for *complex spectrum decoding*

contains the phase offset of the oscillation at frequency ω across the P channels. Finally, we model induced effects (i.e. narrowband power that is not phase aligned to the stimulus) independently in each frequency band, where Σ_ω is the $[P \times P]$ covariance matrix modelling the spatial variance and correlations expressed at frequency band ω. Note that this corresponds to an assumption that only the evoked response, not the induced response, differs over the two conditions – this is a simplifying assumption that we later relax in Section 2.4.

We can now characterise the mutual information between the broadband signal X_t or its constituent narrowband signals $Z_{t,\omega}$ and the class labels Y.

2.2. Information content available to narrowband signal decoding

We wish to explore how the spectrum of the evoked response determines the representational dynamics inferred from the decoding paradigms that are most typically used in the literature (T. Carlson et al., 2013; Cichy & Pantazis, 2017; Grootswagers et al., 2017). We start by considering instantaneous decoding of narrowband signals $Z_{t,\omega}$, which we refer to as *narrowband signal decoding*.

Given a probabilistic model, we can calculate the *mutual information* $I(Z_{t,\omega}, Y)$, which expresses the amount of information shared between the signal and the condition label time courses. This measure of information content in the signal that pertains to the condition labels can be thought of as a surrogate measure of decoding accuracy were one to do *narrowband signal decoding*. Starting with a single Fourier component of the evoked response at frequency ω, the mutual information is itself a sinusoidal function that has been translated to double the original frequency, 2ω:

$$I\left(Z_{t,\omega}, Y\right) = f\left(c_\omega + \mathrm{r}_\omega \cos\left(2\omega t + \xi_\omega\right)\right) \tag{3}$$

Where f is a monotonic, concave function (see Appendix C for proof and Fig. A1 for plot of the function); and $c_\omega, \mathrm{r}_\omega$ and ξ_ω are all scalar values that are constant with respect to time (see Appendix D for their exact values, and Appendix B and D for proof of the above result). The intuition for this is based on what was discussed in Fig. 1: if $Z_{t,\omega}$ were the displacement of a pendulum oscillating at frequency ω, a decoder will perform best at the peaks and troughs of that oscillation and poorly in between these points.

We illustrate this relationship in example 1 (Fig. 2), where we simulate an evoked response under two conditions. Suppose that one condition (in blue) contains information content at 10Hz across both channels, and the second condition (in black) does not. The information content associated with this narrowband component is itself a sinusoidal function oscillating at 20Hz.

2.3. Information content available to instantaneous signal decoding

Realistic neural signals are not expressed in a single component frequency across all spatial areas, but are rather comprised of a number of spatially distinct components at multiple frequencies. How then does the entire frequency spectrum of the evoked response determine the frequency spectrum of the associated information content? This equates to the paradigm of *instantaneous signal decoding* that is most widely performed in the literature (Carlson et al., 2013; Cichy & Pantazis, 2017; Grootswagers et al., 2017). For the broadband signal X_t given in our model, the information content is given by:

$$I\left(X_t, Y\right) = f\left(c_B + \sum_{\omega}^{\Omega} r_{B,\omega} \cos\left(2\omega t + \xi_{B,\omega}\right) + \mathrm{h(t)}\right) \tag{4}$$

Where $c_B, r_{B,\omega}$ and $\xi_{B,\omega}$ are scalar values that are constant over time, and h(t) refers to additional sinusoidal harmonics distributed across the frequency spectrum between zero and 2Ω (see Appendix E for their exact values along with proof of this result).

Importantly, if the highest frequency component of the evoked response on any channel is Ω, it follows that the highest frequency in the associated information spectrum will be2Ω. We illustrate this point with example 2 in Fig. 2; for simplicity we simulate an evoked response comprising just 2 spectral components under each condition at 10Hz and 15Hz; the associated information content displays multiple peaks over time, represented in its Fourier spectrum by frequency components distributed between 0Hz and 30Hz. As we will explore further in Section 3.1, this also means that commonly used anti-aliasing filters are insufficient to stop representational aliasing, i.e. alias artefacts in the inferred information content dynamics. In order to further illustrate the representational dynamics of instantaneous signal decoding, we created *Representational Dynamics Simulator*, a web application analogous to Fig. 2, where the user can interactively change the parameters of the evoked spectrum and see the resulting information content (Van Es et al, 2022; hosted at https://representational-dynamics.herokuapp.com).

2.4. Modelling induced effects

It is important to consider the degree to which these findings are specific to our chosen modelling assumptions. We have specifically limited our discussion to that of evoked effects by assuming the noise distribution was invariant over conditions. In the frequency domain, this means that we have limited our analysis to the part of the signal that is phase aligned to stimulus onset. When we introduce condition-specific induced effects – i.e. to model the case where one condition induces an increase in bandlimited power that has random phase alignment with the stimulus onset – we can no longer derive an exact analytic expression for the mutual information; however, we can derive an upper bound on the information content. This upper bound is a function of components at the same frequencies specified in equations 3 (for the narrowband case; see Appendix G and H for proof) and 5 (for the instantaneous signal case; see Appendix I for proof). This result is not mathematically trivial, but may nonetheless be intuitive to some readers on the basis that the information content of a signal containing both evoked and induced effects must be less than the combined information content of each of those effects assessed independently; and the information content of induced effects assessed independently is constant with respect to time (owing to the uniform phase distribution that defines induced effects). Thus, we are able to generalise our findings to the case where induced effects are present.

3. Technical and Interpretational issues raised

The relationship we have characterised above between the stimulus-evoked signal spectrum and the spectrum of the information content raises several issues with commonly used *instantaneous signal decoding*

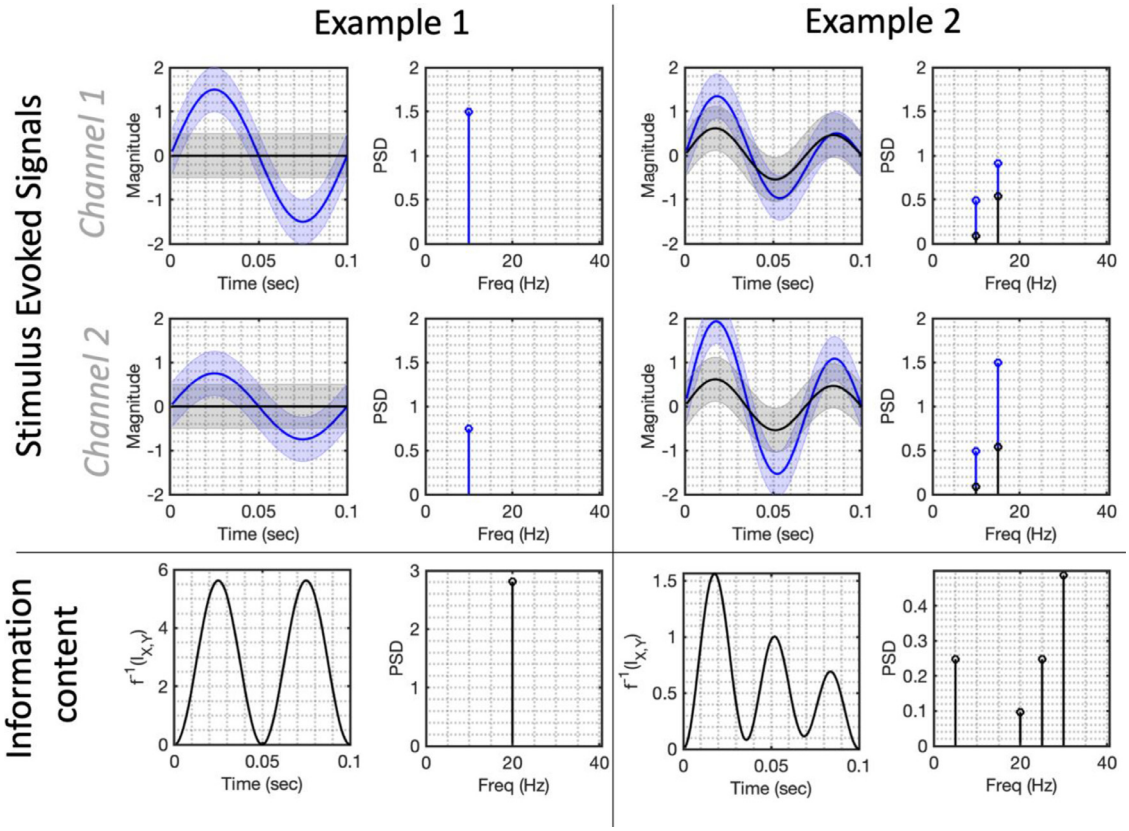

Fig. 2. How the stimulus evoked spectrum determines the spectrum of information content when *instantaneous signal decoding* is used. In example 1 on the left, we simulate two conditions across two channels, with the upper panel showing the two conditions' trial-averaged evoked responses (one in blue and one in black). The first condition evokes a phase-locked 10Hz oscillation on both channels, the second condition is a null condition in which there is no evoked response. The information content (i.e. the mutual information $I(X_t, Y)$ between the broadband data X_t and the stimulus labels Y) is plotted in the lower panel, and can be thought of as a surrogate measure of decoding accuracy were one to do *instantaneous signal decoding*. Although the only oscillation in the evoked response is at 10Hz, the information content is a 20Hz sinusoidal signal, reaching maxima at each peak and trough of the evoked response. In example 2 on the right, we simulate a signal comprised of 2 Fourier components at 10Hz and 15Hz in both conditions, with slightly different amplitudes. The associated information content is a signal with three distinct peaks, with Fourier components at 20Hz and 30Hz and additional harmonics at 5Hz and 25Hz. For illustrative purposes, we created a web application where these parameters can be changed interactively (Van Es et al., 2022; hosted at https://representational-dynamics.herokuapp.com).

pipelines. On a technical level, there is a risk of high frequency artefacts which we refer to as representational aliasing. On a broader level, this raises questions about how certain features of decoding accuracy timecourses should be interpreted.

3.1. Representational aliasing

The Nyquist frequency defines the highest frequency component that can be correctly resolved from data that has been digitally sampled at a specified sampling rate. It is standard practice to apply a low pass anti-aliasing filter prior to sampling which ensures no *signal* components are above the Nyquist frequency and that all signal components can therefore be correctly resolved. However, this only applies to the signal components, not their associated information spectrum, which we have shown contains spectral contents at double the highest frequency of the signal spectrum.

It follows that representational aliasing artefacts will be present in *instantaneous signal decoding* accuracy metric unless the following condition is met:

$$F_S \geq 4\Omega \tag{5}$$

Where Ω is the highest frequency component of the evoked response and F_s is the sampling rate. Thus, *instantaneous signal decoding* pipelines need to use low pass filters with cut-off no higher than one quarter of the sampling rate – before training classifiers – in order to eliminate representational aliasing effects. Fig. 3 illustrates this graphically.

3.2. How should we interpret oscillatory information content?

The oscillatory nature of information content associated with sinusoidal components of the evoked response is, we argue, interpretationally problematic. Features resembling multiple successive peaks in the timecourse of classification accuracy are quite commonly reported in the literature (Gennari et al., 2021; Hogendoorn & Burkitt, 2018; Mohsenzadeh et al., 2018; Robinson et al., 2020); in some cases, the dynamics of these successive peaks have been interpreted as evidence for complex cognitive phenomena such as phase-locked memory reactivation (Fuentemilla et al., 2010; Kerrén et al., 2018). As we have shown in Fig. 2, successive peaks arise naturally from an evoked response containing sinusoidal components. We argue that a simpler explanation for their common appearance in the literature could merely be that the typical evoked response is characterised by a succession of peaks and troughs (e.g. the N70, P100 and N175) that resemble a transient sinusoidal waveform.

We believe a fuller picture of information content should include the information stored in the dynamic gradient of the signal that is not available using *instantaneous signal decoding* pipelines. In Section 4 we explore a third paradigm that includes such information, and show that this results in narrowband information content that is stable over time.

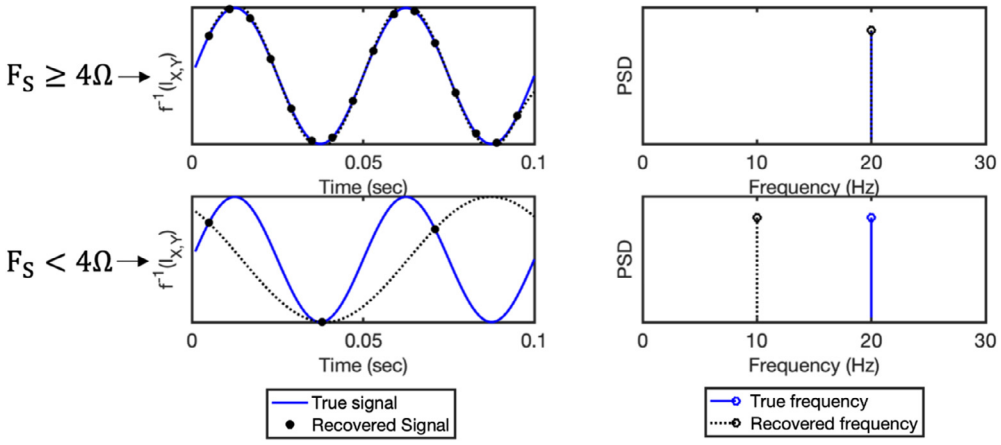

Fig. 3. Demonstration of representational aliasing associated with *instantaneous signal decoding*. Consider the information content associated with example 1 from Fig. 2, where one condition was associated with a stimulus-evoked component at 10Hz. As Fig. 2 shows, the single oscillatory mode at 10 Hz is associated with a true information content that oscillates at a frequency of 20 Hz (plotted in blue above). In the first case on the top row, given a sampling rate of 160 $Hz > 4\ \Omega$, the recovered representational dynamics (plotted with black dashed line showing sinusoidal interpolation between the discrete samples) match the true frequency. In the second case however, given an inadequate sampling rate of 30 $Hz < 4\ \Omega$, the recovered dynamics are subject to representational aliasing, resulting in spurious dynamics at 10Hz rather than the true 20Hz pattern.

However, as these methods will not always be practical for reasons given in the discussion, we would more generally argue that representational dynamics obtained using *instantaneous signal decoding* and representing the 'double peak' feature shown in Fig. 2 (and widely characterised in the literature) should first be assumed to correspond merely to peaks and troughs of an evoked sinusoidal signal, rather than more complex cognitive phenomena.

4. Obtaining measures of sinusoidal information content that are stable over time

We contend that the profile of information content obtained by *instantaneous signal decoding* is potentially misleading, as it suggests the brain's representational dynamics are much faster than the actual evoked spectrum from which they are derived. Whilst *instantaneous signal decoding* pipelines are the most popular way to apply decoding to neural data at high temporal resolution, alternative methods exist that overcome these limitations. We focus our attention on Fourier analysis (for continuity with our modelling approach and because of these methods are well-established in neural data analysis), but emphasise these benefits are not specific to Fourier analysis per se – rather, they arise whenever methods include information in a dynamic signal's higher temporal derivatives (e.g. its gradient and rate of change) as features for classification.

4.1. Complex spectrum decoding

We previously characterised the information content between stimulus labels Y and the narrowband Fourier series components $Z_{t,\omega}$. These narrowband components do not in fact include all the information that is returned by a Fourier signal decomposition; they reflect only the real component of a complex number representation. The imaginary components of these narrowband components reflect the instantaneous gradient information of each narrowband signal; we here characterise the information content associated with the full complex signal representation of each narrowband component, analogous to the decoding accuracy that would be obtained when both the narrowband signal and its local gradient are used as features for classification as in Angjelichinoski et al., 2019; Ince et al., 2017.

4.1.1. Real and complex components of a Fourier decomposition

Fourier decompositions provide a complex representation of the underlying signal that includes both a real signal component and an orthogonal imaginary component, which we omitted from our model outline in Section 2 for simplicity. Including this complex-valued information, the same model can equivalently be written:

$$x_{n,t} = \mu_t + \sum_{\omega=0}^{\Omega} z_{n,t,\omega} \tag{6}$$

$$z_{n,t,\omega} = \frac{w_{n,t,\omega} + w^*_{n,t,\omega}}{2} \tag{7}$$

$$w_{n,t,\omega} = y_n A_\omega e^{i(\omega t + \phi_\omega)} + \varepsilon_{n,\omega} e^{i\omega t} \tag{8}$$

$$\varepsilon_{n,\omega} = N\left(0, \Sigma_\omega\right) + iN\left(0, \Sigma_\omega\right) \tag{9}$$

Where $w^*_{n,t,\omega}$ denotes the complex conjugate of $w_{n,t,\omega}$. This is exactly equivalent to the model of Eq. (2), however it includes the complex spectral representation $w_{n,t,\omega}$ of each narrowband Fourier series component. It includes a condition-dependent evoked term $y_n A_\omega e^{i(\omega t+\phi_\omega)}$ (i.e. the component of the response that is phase-locked to the stimulus), and a condition-independent residual term (i.e. the residual component with randomly drawn phase and amplitude on each trial; note that the values for the phase and amplitude are respectively the angle and magnitude of the complex valued $\varepsilon_{n,\omega}$ converted to polar coordinates).

4.1.2. Information content available to complex spectrum decoding

An alternative to decoding on the raw signal at each point in time is to use both the real and imaginary parts of the complex-valued Fourier coefficients as features/inputs to a classifier (Angjelichinoski et al., 2019; Ince et al., 2017). We will refer to this decoding paradigm as *complex spectrum decoding*. When all this information is included as features for classification, then the resulting information content in each frequency band is given by:

$$I\left(W_{t,\omega}, Y\right) = f\left(2c_\omega\right) \tag{10}$$

Where c_ω is the average value of the sinusoidal expression associated with the real information content in Eq. (4) (see Appendix F for proof). Importantly, this expression is no longer sinusoidal; it is stable over time, and greater or equal to the peak information content that can be obtained using only the real spectrum (see Fig. 4). Consequently, this overcomes both the problematic interpretational issues associated with *instantaneous signal decoding* discussed above, as well as the risk of representational aliasing that would otherwise require low-pass filtering with cut-off one quarter of the sampling rate.

4.2. Practical considerations for non-stationary and non-oscillatory signals

We emphasise the generality of these results, deriving from the fact that *any* arbitrary time series can be mapped into the frequency domain

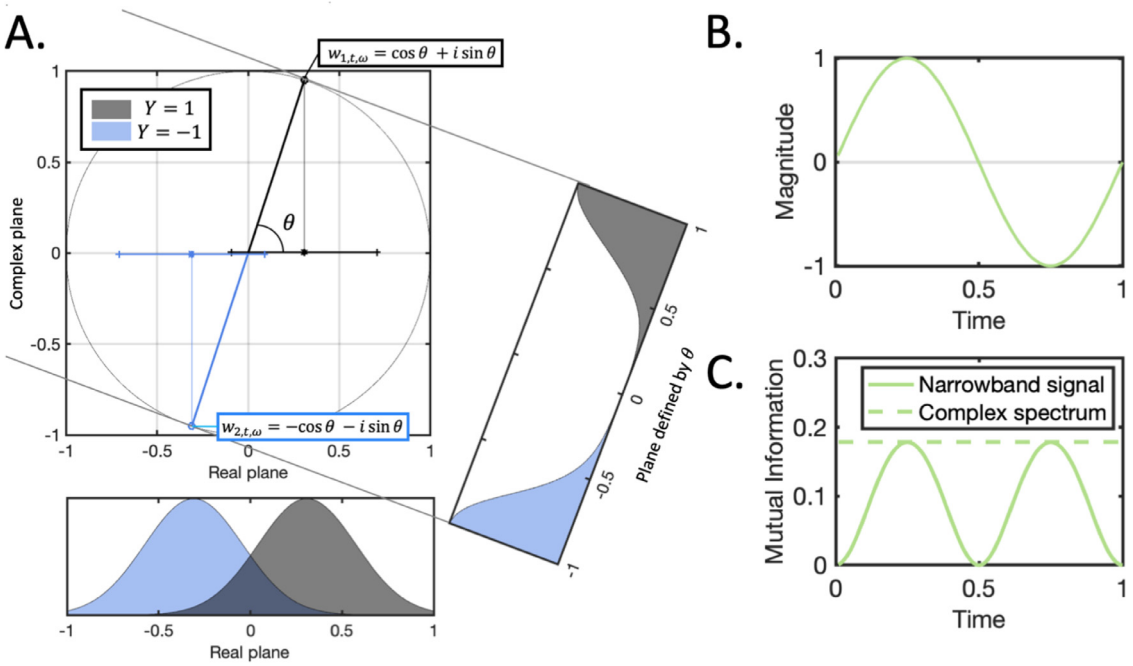

Fig. 4. Motivation for complex spectrum decoding. A. The signal modelled by $w_{n,t,\omega}$ can be visualised as a point rotating around a circle in the complex plane, with the two stimulus conditions shown in grey and blue corresponding to opposite sides of the circle. As the signal crosses the imaginary axis (i.e. as θ approaches $\frac{\pi}{2}$), the separability of the two conditions in the real plane is minimised (corresponding to the troughs in the narrowband information content in Fig. 2); however, at the same point, the two conditions in the complex plane (see the plane defined by θ) are still highly separable. In fact, projecting onto the plane defined by the instantaneous phase θ results in information content that is stable over time and not varying with the phase of the oscillatory signal. B. The real part of the signal (i.e. $z_{n,t,\omega} = Re(w_{n,t,\omega})$) is a sinusoid as previously characterised. C. The corresponding narrowband mutual information (i.e. $I(Z_{t,\omega}, Y)$) drops to zero in the sinusoidal troughs, whereas the complex spectrum information term (i.e. $I(W_{t,\omega}, Y)$) is constant throughout.

by a Fourier decomposition. Whilst we have so far simulated quite simplified evoked responses comprising only a few frequency components, our approach generalises to those that contain non-stationary and/or non-oscillatory components. In this section we demonstrate this with some more complex simulations.

4.2.1. Sliding window Fourier decompositions

Real evoked responses are more complex than the illustrative examples we have simulated so far and in particular do not have spectral profiles that are constant over the whole trial epoch. We therefore anticipate that the methods introduced above will be most informative when combined with sliding window methods, e.g. where separate Fourier decompositions are applied to each window within a trial epoch rather than a single Fourier decomposition applied to the whole epoch.

There are numerous methods for estimating spectral properties over sliding windows, which are typically similar in motivation but different in implementation. Perhaps the most important factor is how the trade-off between time and frequency resolution is handled. Given our focus on characterising representational dynamics over time, we prefer methods that use a fixed temporal resolution, such as the Short-Time Fourier Transform (STFT). This provides complex-valued Fourier coefficients in each frequency band at each timepoint within a trial, allowing decoding accuracy to then be computed timepoint-by-timepoint without the interpretational problems previously discussed.

4.2.2. Non-stationary oscillatory signals

To test these methods on evoked signals characterised by transient spectral properties, we simulated a signal over two channels using a combination of frequency chirp functions and unit step functions (example 1 in Fig. 5). To maintain simplicity only one of the two conditions has this profile, the other is a null condition of stationary Gaussian noise. As shown by the time-frequency diagram on Fig. 5A, the frequency distribution of the signal varies over time and over the two channels. For this signal, we then computed:

(i) The *broadband information content;* This corresponds to the information content available to ***instantaneous signal decoding***, i.e. the timepoint-by-timepoint decoding approaches that are most typically used in the literature (Carlson et al., 2013; Cichy & Pantazis, 2017; Grootswagers et al., 2017).

(ii) The *complex spectrum information content*; this corresponds to the information content available to ***complex spectrum decoding*** as we have proposed. In this case however we have estimated the complex spectral features using a sliding window (specifically using a STFT with 50ms sliding Hamming window).

As shown in Fig. 5B, the broadband information content (analogous to the decoding accuracy obtained by *instantaneous signal decoding*) contains fast dynamics that do not clearly relate to the evoked signal shown in Fig. 5A. Applying a similar STFT analysis to this information content (Fig. 5B, right hand side) shows it reflects components at up to double the frequency of the corresponding signals (i.e. it contains components at up to 100Hz, double the frequencies identified in Fig. 5A).

In contrast, the complex spectrum information content provides frequency band specific measures of information content that more closely reflect the spectral distribution of information at each moment in time over the course of the trial (i.e. Fig. 5B, lower plot reflects the combined contributions of the channel power spectral density plots in Fig. 5A). From the perspective of representational dynamics, such information is at least complementary, and we would argue more informative than that available to *instantaneous signal decoding*.

4.2.3. Non-oscillatory evoked signals

In Section 2 we showed that consecutive peaks in decoding accuracy timecourses could arise due to a simple oscillatory signal, even if this oscillatory signal is itself stable over time. We argued that these peaks

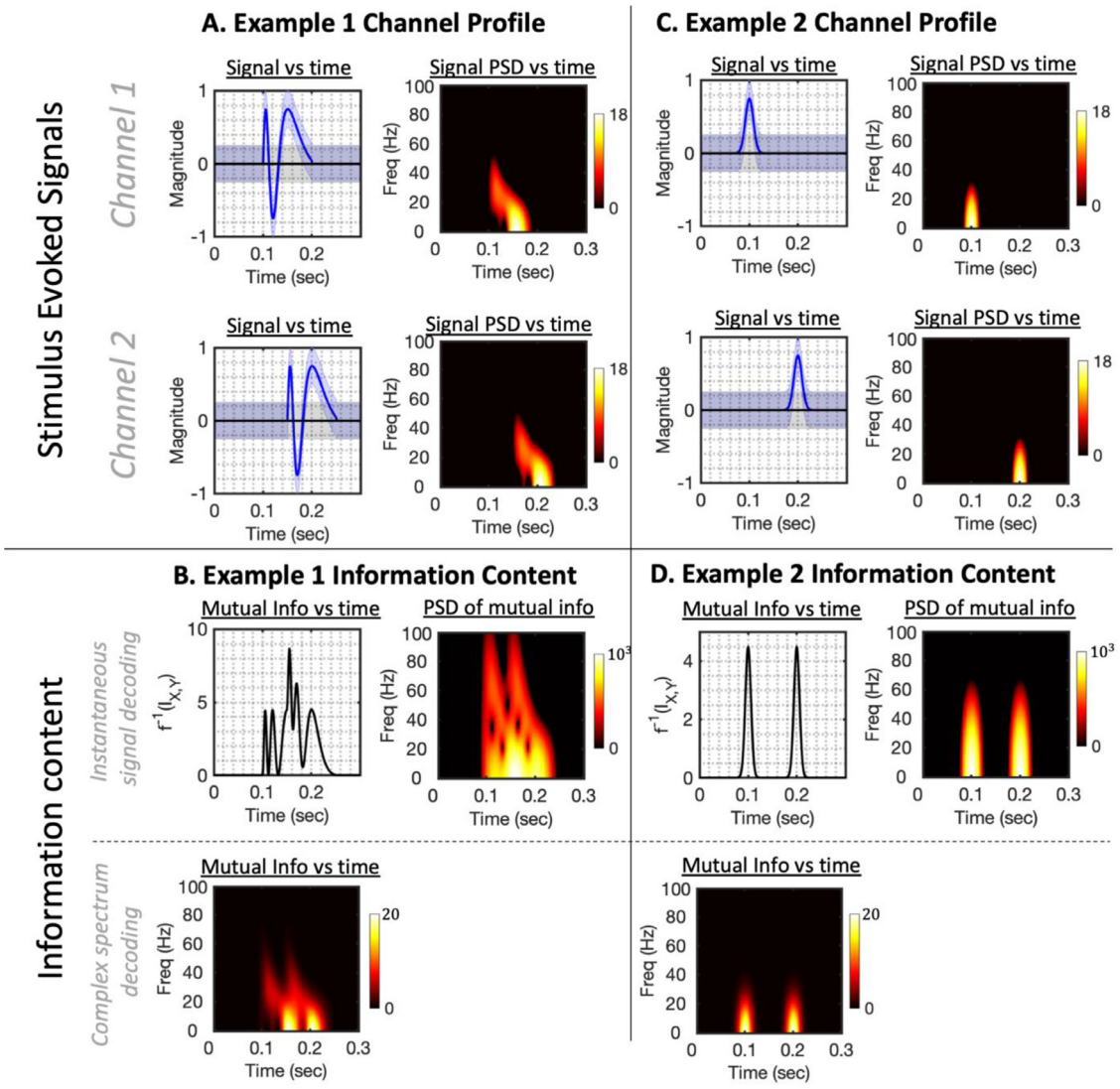

Fig. 5. Complex spectrum decoding remains applicable and informative even when the spectral properties vary over time or are not fundamentally oscillatory. A. In Example 1, we simulate a signal across two channels with a time-varying frequency 'chirp' response, with a different onset time on each of the two channels. Left hand side plots the actual trial-averaged evoked response for each channel, right hand side the PSD as a function of time on each channel, showing the frequency content is transient on each channel and limited to frequencies below 50Hz. **B.** The information content associated with this signal. Top row plots the information content obtained by doing *instantaneous signal decoding*; right hand side plots the frequency profile of this mutual information timecourse, which reflects a mix of the spectrum from the two channels translated to double their original frequency (i.e. up to 80 Hz). Lower plot shows the mutual information obtained by *complex spectrum decoding* in each frequency band, which reflects the true frequencies at which information is present in the original signal. **C.** In Example 2, we simulate a non-oscillatory evoked response comprising two distinct processes occurring at different times; these characterise the signal over each channel in time and frequency. **D.** The information content available to instantaneous signal decoding identifies and separates these processing stages. This profile of two distinct peaks is similarly recovered from the complex spectrum information content (provided a suitable size of sliding window is used), demonstrating that this approach does not obscure such features where they are genuinely reflected in non-sinusoidal activity.

should not be interpreted as representing discrete events or cognitive phenomena. This begs the question, how do our methods perform if the underlying signals *do* derive from discrete temporal events, where the underlying signals cannot be parsimoniously represented using sinusoidal components?

To test this, we simulated an evoked response deriving from two spatially and temporally distinct "activations", and repeated the analysis described above to compare the broadband and narrowband information content. To simulate non-oscillatory signals, each activation was characterised by a Gaussian kernel function (Fig. 5C). As shown in Fig. 5D, the broadband information content (i.e. that available when doing *instantaneous signal decoding*) produces two distinct peaks corresponding to each activation. Notably, this profile is replicated in the complex spectrum information content (Fig. 5D, lower panel) showing that this method does not obscure such phenomena – provided the sliding window width is less than the period between these activations. Wider window lengths progressively include more information from both activations and the peaks become much less pronounced (see Supplementary Information, Section 2 and Figure S2). We therefore conclude that, subject to appropriate sliding window sizes, *complex spectrum decoding* can eliminate the fast dynamics associated with sinusoidal components of the evoked response, whilst not eliminating the structure associated with spatially distinct, potentially non-oscillatory evoked activations.

5. Evidence from MEG data

The results we have presented are fundamentally theoretical and supported by simulated data from models of evoked activity. We therefore wanted to test how these findings extend to real data, and therefore tested our main predictions on a MEG dataset of visual image decoding.

5.1. Methods

We took a publicly available dataset comprising 15 subjects viewing 118 different visual stimuli (Cichy et al., 2016). This data had been acquired on an Elekta Neuromag scanner with 306 channels (204 planar gradiometers and 102 magnetometers) at 1kHz sampling rate, with filtering applied at acquisition with bandpass 0.03Hz to 300Hz. We downsampled the data to 100 samples per second with an anti-aliasing filter with cut-off at 50Hz and extracted the 0.5 second epochs immediately following stimulus presentation. The data was then mapped into a complex time-frequency decomposition using an STFT with Hamming window length of 100ms. The epoched data was then decoded to predict the trial condition labels using the three paradigms:

i ***Instantaneous signal decoding***: decoding the raw broadband signal time-point-by-timepoint as widely performed in the literature (T. Carlson et al., 2013; Cichy & Pantazis, 2017; Grootswagers et al., 2017).
ii ***Narrowband Signal decoding***: sliding window decoding using the time-frequency estimates from the STFT, but only using the real coefficients across all sensors as a set of features. This method is analogous to decoding on data filtered into specific frequency bands of interest.
iii ***Complex spectrum decoding***: sliding window decoding using the time-frequency estimates from the STFT, using both the real and imaginary coefficients across all sensors as a set of features.

Each approach fit linear support vector machine classifiers using three-fold cross validation. This was applied to each pair of the 118 images in a mass pairwise classification paradigm as originally implemented by (Cichy et al., 2016). In cases (ii) and (iii), classifiers were trained separately on each frequency band. The decoding used three-fold cross-validation to obtain independent classification accuracy metrics as a function of time and frequency for each pair of images and each participant.

Finally, to test the hypothesis that different frequency bands contained complementary information, we trained an aggregate classifier to estimate the aggregate information distributed over all frequency bands. We did this through a nested cross validation procedure. An inner cross validation loop simply consisted of the *complex spectrum decoding* estimates described above. The outer cross validation loop then partitioned all of the stimuli into two equally sized groups and applied two-fold cross validation to obtain accuracy estimates. This outer loop consisted of a random forest ensemble classifier with 100 trees, trained to predict the class label from the outputs of the *complex spectrum decoding* classifiers on each trial. This outer loop was run ten times with replacement for each subject, randomly sampling a different subset of stimuli with replacement on each cross validation fold.

5.2. Decoding accuracy vs time in different decoding paradigms

Fig. 6A plots the decode accuracy derived from decoding under the three identified paradigms. As paradigms (ii) and (iii) provide accuracy in each frequency band independently, for ease of visualisation they are each plotted separately against paradigm (i). Averaged over all pairs of stimuli and all subjects, this identifies a systematic variation in the information content at different frequencies as a function of time. The earliest detectable information appears in higher frequencies, but these peak quite transiently at relatively low values and are quickly surpassed by information content in lower frequencies, which rise to higher values and are then sustained for a longer duration. Notably, the information in either the 10Hz or the 0Hz band exceeds that obtained by *instantaneous signal decoding* for nearly the entire period analysed – the accuracy averaged over all timepoints is higher for both measures (paired t-test, $p<0.001$ Bonferroni corrected for multiple comparisons over frequency bands) which correspond to higher accuracy over a majority of timepoints in these frequency bands (see Supplementary Information, Section 5 and Figure S4). From the perspective of representational dynamics, this establishes first and foremost that Fourier decompositions can improve decoding accuracy over *instantaneous signal decoding* methods whilst retaining a profile of how the representational content evolves in both time and frequency.

5.3. Complex spectrum decoding accuracy exceeds narrowband signal decoding accuracy

Fig. 6B compares the average classification accuracy in each frequency band, averaged over all subjects and pairs of stimuli, when either the *complex spectrum decoding* or *narrowband signal decoding* is applied (it follows from the definition of the discrete Fourier transform that the imaginary coefficients in the 0Hz and 50Hz frequency bands are always zero, so in these bands the two paradigms are in fact equivalent). In all cases the classification accuracy obtained using *complex spectrum decoding* exceeds that obtained using *narrowband spectrum decoding*; this information gap can be interpreted as the information stored in the gradient of these sinusoidal components.

5.4. Narrowband signal decoding produces spectral peaks at double their original frequency in inferred decoding accuracy metrics

Our models predict that the information content associated with evoked spectral components at a given frequency is itself oscillatory at double that frequency, unless *complex spectrum decoding* is applied. We have so far plotted the average over all subjects and all comparisons, therefore obscuring some of the temporal dynamics evident in each comparison. For example, in Fig. 6C we plot the timecourse obtained for one subject and one pair of stimuli; the accuracy timecourse obtained from *complex spectrum decoding* appears to follow the envelope of the equivalent timecourse obtained by *narrowband signal decoding* which appears to show sinusoidal dynamics. If we take the PSD of these accuracy timecourses, we observe a peak at double the frequency band being analysed (i.e. the 10Hz and 20Hz bands are associated with a 20Hz and 40Hz spectral peak respectively). If we take the PSD of the timecourse for every pair of stimuli and every subject and average, we see the PSD is significantly higher in the 10Hz and 20Hz bands at approximately 20Hz and 40Hz, respectively.

Given a sampling rate of 100Hz, we expect representational aliasing to occur linked to any evoked spectral content above 25Hz. Specifically, given evoked spectral content at 30Hz or 40Hz, we expect representational aliasing artefacts at 40Hz and 20Hz, respectively (for example, a 30Hz component is translated to 60Hz in the accuracy timecourses; as this is 10Hz above the Nyquist frequency, it is aliased to 10Hz below the Nyquist frequency, i.e. to 40Hz). For both of these narrowband signals, we see peaks at these locations, confirming the presence of representational aliasing. We stress that this aliasing effect must also be present in the *instantaneous signal decoding* results, they just cannot be explicitly resolved as we have no knowledge of the frequencies at which they would be expected.

Finally, in these plots we note that spectra are significantly more weighted towards the lower end of the frequency spectrum for *complex spectrum decoding* vs *narrowband signal decoding*, whilst the opposite relationship is the case towards the upper end of the frequency spectrum. This means that the higher accuracies obtained by *complex spectrum decoding* in Fig. 5B are a result of increased low frequency content, or representational dynamics that are more stable over time.

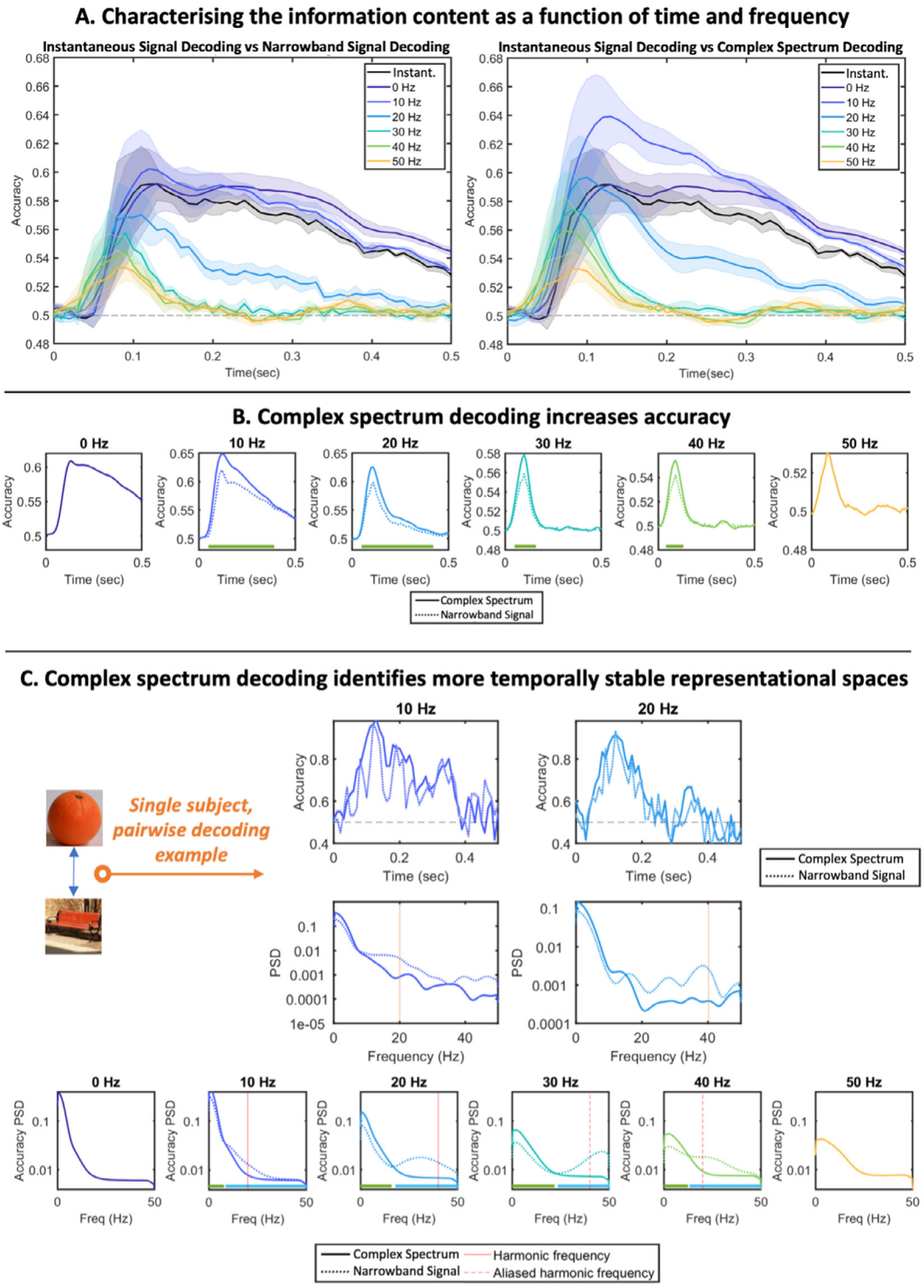

Fig. 6. Characterising information content across the frequency spectrum. A. Decoding accuracy timecourses obtained by *instantaneous signal decoding* vs either *narrowband signal decoding* (left) or *complex spectrum decoding* (right). All plots show mean +/- SE over subjects. **B.** Directly comparing the accuracy vs time in each frequency band of *narrowband signal decoding* and *complex spectrum decoding*. Significance bars denote periods where *complex spectrum decoding* accuracy is significantly greater than *narrowband signal decoding*: $p<0.01$ using cluster permutation tests. **C.** Single subject results. The dynamics, and therefore any attempt to assess the temporal stability of the decode accuracy, are obscured by averaging over all participants and subjects. Top: taking a single subject, single pairwise comparison as an example, we show the decode accuracy obtained by either *narrowband signal decoding* or *complex spectrum decoding* as a function of time (upper) or frequency (lower). These show significant harmonic components at double the fundamental frequency in each band when *narrowband signal decoding* is used. Lower plot: We took all individual accuracy vs time plots (across all subjects and all stimulus comparisons) and computed their power spectral density, then averaged (plots show mean +/- SE over subjects). The power spectrum obtained using *narrowband signal decoding* show strong peaks at harmonic frequencies (for 10Hz and 20Hz bands) and at aliased frequencies (for 30Hz and 40Hz bands). Significance bars denote significance at $p<0.01$ levels using cluster permutation tests; green bars denote *complex spectrum decoding* greater than *narrowband signal decoding*; blue bars denote *narrowband signal decoding* greater than *complex spectrum decoding*. This shows that, in all frequency bands, the increased accuracy obtained by *complex spectrum decoding* is concentrated in lower frequencies, reflecting more temporally stable representational spaces.

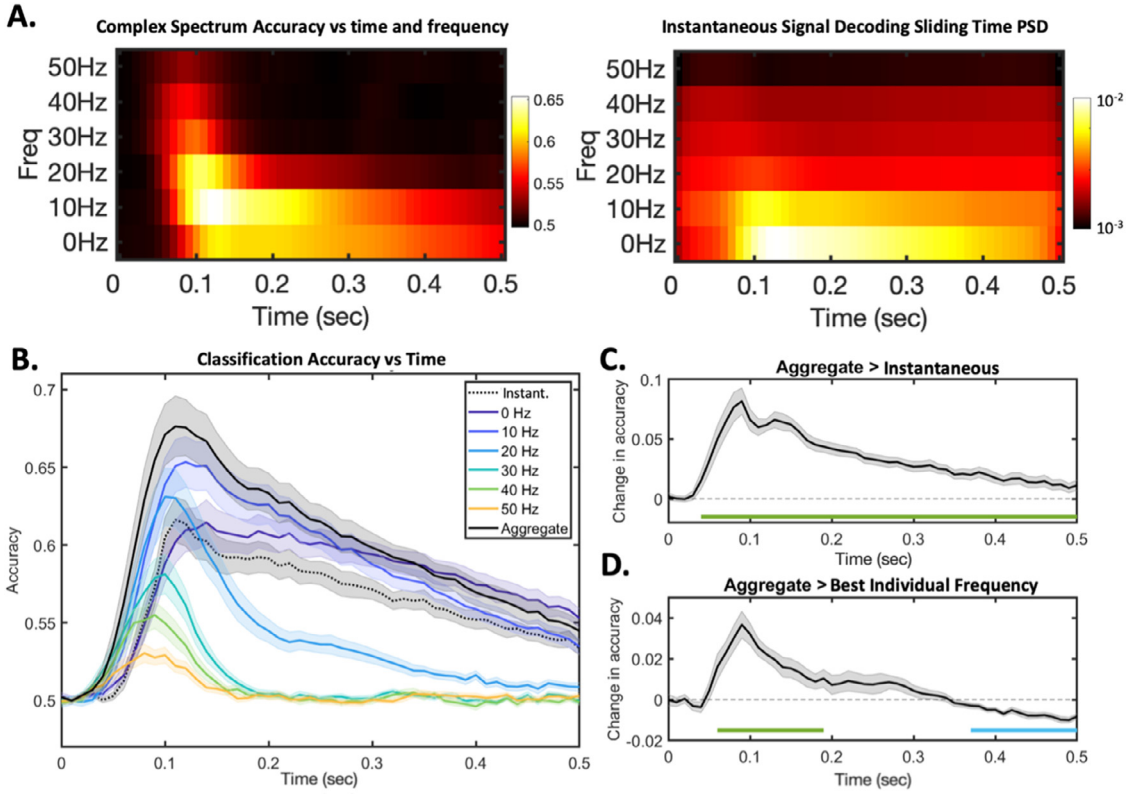

Fig. 7. Interpreting and aggregating information over multiple frequency bands. A. Complex spectrum decoding (left) directly reflects the frequencies at which information is represented in the brain, revealing representational dynamics that have both spectral and temporal structure. In contrast, taking the PSD of the instantaneous signal decoding accuracy timecourse (right) is neither interpretable nor free of aliasing artefacts. **B.** Plotting the classification accuracy achieved vs time for *instantaneous signal decoding,* for each individual frequency using *complex spectrum decoding*, and then for the aggregate classifier. All plots show mean +/- SE over subjects. **C.** Plotting the contrast of the aggregate classifier accuracy minus the instantaneous signal decoding accuracy; significance bars denote $p<0.01$ using cluster permutation tests, green bars denote contrast is significantly positive, blue bars denote the sontrast is significantly negative. **D** Plotting the contrast of aggregate classifier accuracy minus the best individual frequency (i.e. the highest accuracy obtained by any *complex spectrum decoder* at each timepoint); significance bars defined as for C.

In Fig. 5, we made the argument that complex spectrum decoding reflects the true frequencies at which information is present in the original signal. This point is now reinforced with real data. In Fig. 7A, we plot the accuracies per frequency alongside the PSD of the accuracy timecourses obtained using instantaneous signal decoding. The former is interpretable and reveals an information profile with both spectral and temporal structure. In contrast, it follows from Eq. (4) that the PSD of the instantaneous signal decoding timecourse does not correspond to the frequencies of actual information. The representational aliasing effects characterised above, and harmonics created when multiple carrier frequencies are combined, have contaminated the spectral profile such that no prominent structure can be easily observed.

5.5. Complex spectrum decoding accesses information content that is complementary over frequencies

Having established that *complex spectrum decoding* accesses information content that is not available to *instantaneous signal decoding*, one final question arises; is the complex spectral information across different frequencies overlapping, or complementary? That is to say, if we aggregate the information over frequency bands, do we obtain performance that is merely equivalent to the best individual frequency band – or exceeding it?

Fig. 7B plots the performance of the aggregate classifier against the *complex spectrum decoding* accuracies achieved in each frequency band, and that obtained by *instantaneous signal decoding*. The aggregate classifier significantly outperforms the *instaneous signal decoder*, reaching a peak accuracy of 67.6% vs 61.6%. As plotted in Fig. 7C, this difference quantifies the total amount of information that is inadvertently being omitted by the insensitivity of *instantaneous signal decoding* paradigms to information stored in signal gradients. However the aggregate decoder accuracy also peaks at a level higher than that obtained in *any* individual frequency band. As in Fig. 7D, over the period between 70msec and 190msec following stimulus presentation, the aggregate classification accuracy significantly exceeded the information content in any individual frequency. This coincides with the time over which significant information content was distributed across multiple frequency bands, especially higher frequency bands, proving that these different frequency bands contain information content that is complementary. The performance is quite different for timesteps more than 370msec after stimulus onset, with the ensemble classifier underperforming slightly relative to the best narrowband classifiers (albeit still outperforming standard broadband methods). Over this period, the classifiers trained on higher frequencies output chance level predictions, and only lower frequency bands contain meaningful information content. We conclude that over this period, all meaningful information is concentrated in lower frequency bands, and the inclusion of high frequency bands that only contain noise is in fact detrimental to classifier performance.

6. Discussion

We have outlined a widely overlooked problem in decoding pipelines: that frequency components in the evoked response produce corresponding components at double their original frequency content in the resulting accuracy metrics. Where researchers are not aware of this fundamental relationship, there is a considerable risk of misinterpret-

ing results and, in particular, of inferring relationships with canonical frequency bands that are in fact trivial representations of the evoked response spectrum.

We have argued that including a signal's higher temporal derivatives in decoding better reflects the full picture of information content available to downstream brain regions, such that the more stable temporal profiles obtained by *complex spectrum decoding* (and related methods) are a better depiction of the true information content compared to *instantaneous signal decoding*. It is notable that neural circuits – being fundamentally conductance-based at the cellular level – are perfectly placed to compute such higher temporal derivatives, such that this information is readily available for any further computation. This certainly does not mean that the brain does *not* encode information to a particular phase of an underlying oscillation, just that commonly used *instantaneous signal decoding* methods may mistakenly suggest so.

In particular, studies investigating memory reactivation have interpreted the oscillatory dynamics of the classification accuracy as evidence that reactivation is functionally phase-locked to canonical frequency bands. Kerrén et al. (2018) analyse the spectrum of the classification accuracy timecourse, much as we have done in Fig. 6C, and interpreted the peak at 7Hz as evidence for theta phase locking. Our results suggest this may rather be the result of a 3.5Hz sinusoidal component in the evoked response. In a similar vein, Fuentemilla et al. (2010) found that classification accuracy was modulated by theta phase. Crucially, this work derived decoding accuracy by using wavelet power estimates as inputs to the classifier, a measure which is theoretically phase invariant and therefore could circumvent the effects that we have characterised above. Nonetheless, in practice this becomes largely dependent on the parameters controlling the resolution of time and frequency when power is estimated, such that the relationships we have characterised remain a potential confound (see Supplementary Information, Section 4 and Figure S3). We therefore argue broadly for caution in interpreting oscillatory dynamics in classification accuracy timecourses, and – where authors wish to make stronger interpretations such as those discussed here – recommend rigorous parameter testing with suitably defined non-parametric tests of significance (Brookshire, 2022) to prove such characteristics could not arise trivially.

Our results are fundamentally mathematical, and should be interpreted as such; they derive from the expected Fourier spectrum of the evoked response, not from the fundamental frequency of a canonical neural oscillation. For example, a 40Hz Fourier component can be produced by a vast range of underlying neural sources, only a subset of which would be considered 'gamma oscillations'. Our results apply regardless; thus the recommendation to low-pass filter with a cut-off frequency of one quarter of the sampling rate applies to any researcher doing *instantaneous signal decoding*, irrespective of the frequencies of neural activity they may be interested in or expecting.

As an information theoretic result, if our modelling assumptions hold then these results are fundamental and apply to any *instantaneous signal decoding* approach regardless of methodological choices on the part of the researcher; they cannot be overcome by use of nonlinear classifiers, machine learning tools, or by analysing different accuracy metrics. The result similarly applies more broadly beyond our focus of decoding wherever unsigned statistics are used – for example, applying timepoint-by-timepoint F-tests in an ANOVA analysis to ascertain when a univariate sensor signal significantly differs over conditions would exhibit the same behaviour. In our analysis we have derived the spectrum of the information content up to an arbitrary monotonic scaling denoted by the function f. It follows that other widely used metrics to assess decoding accuracy (such as classification accuracy, distance from the classification hyperplane etc.) are each a different monotonic scaling of this quantity (see Table 2 and SI for further details). We therefore argue that our results are universally applicable to *instantaneous signal decoding* pipelines (and indeed many other pipelines that utilise unsigned statistics) regardless of any variations in methodological choices.

We have characterised three major decoding paradigms but do not claim these to be exhaustive with respect to the literature. A very common approach involves the application of classifiers not to a recorded signal itself but to a set of Fourier features derived from a signal, which in most applications will be equivalent to the narrowband or complex spectrum decoding paradigms such that all our results remain applicable. A related area of research uses the recorded signal and its central-difference gradient as features, similarly obtaining enhanced accuracy as a result (Ince et al., 2016). However an emerging area of research infers nonlinear time-domain features, for example through the training of temporal convolutional networks or recurrent neural networks, that are then used as inputs for classification (Kalafatovich et al., 2020; Schirrmeister et al., 2017; Zubarev et al., 2019). These methods typically offer a greater ability to separate conditions, however the accompanying barriers to interpretability have to date limited their direct application in the study of representational dynamics. We hope that such interpretability barriers will be challenged and overcome in future work, and that the relationships we have outlined here may aid this endeavor.

Finally, we have shown that *complex spectrum decoding* overcomes the problem of representational aliasing whilst also presenting other benefits; specifically, leading to higher accuracies that are more stable over time. We are not the first to use complex features for decoding and find they achieve greater classification accuracy (Angjelichinoski et al., 2019; Ince et al., 2017), nor more generally to use gradient information as features for decoding (Zhan et al., 2019), however the theoretical principles for the underlying relationship were not previously established. However, it similarly presents its own challenges. The significant increase in dimensionality associated with a feature vector that varies simultaneously over time, space and frequency may present computational challenges. Furthermore, whilst we see interpretational benefits to having results that are resolved in both frequency and time, in some circumstances (such as the non-sinusoidal signal example simulated in Section 4.2.3) this additional complexity may not harbour any new insights. We have spoken broadly of Fourier analysis, again to stress that these results apply generically to STFTs, wavelet decompositions, or any other such method – however each of these apply different assumptions that mostly result in different trade-offs of time and frequency resolution. These trade-offs are likely to be especially pertinent in the context of high temporal resolution decoding. Nonetheless, the benefits can be quite substantial and well justified by the results.

Table 2
Mutual information results generalise to other common decoding metrics. We have characterised the information content associated with three variables up to an arbitrary monotonic function f. All our results generalise to other commonly used decoding metrics such as classification accuracy and distance from the hyperplane simply by substituting f by another monotonic function, here denoted by F and $\breve{F}$ (see Supplementary Information Section 1 and Figure S1 for plots of these functions).

Signal	Mutual information	Classification accuracy	Distance from hyperplane
X_t	$I(X_t, Y) = f(c_B + \sum_{\omega}^{\Omega} r_{B,\omega} \cos(2\omega t + \xi_{B,\omega}) + \mathrm{h(t)})$	$A(X_t, Y) = F(c_B + \sum_{\omega}^{\Omega} r_{B,\omega} \cos(2\omega t + \xi_{B,\omega}) + \mathrm{h(t)})$	$D(X_t, Y) = \breve{F}(c_B + \sum_{\omega}^{\Omega} r_{B,\omega} \cos(2\omega t + \xi_{B,\omega}) + \mathrm{h(t)})$
$Z_{\omega,t}$	$I(Z_{t,\omega}, Y) = f(c_\omega + \mathrm{r}_\omega \cos(2\omega t + \xi_\omega))$	$A(Z_{t,\omega}, Y) = F(c_\omega + \mathrm{r}_\omega \cos(2\omega t + \xi_\omega))$	$D(Z_{t,\omega}, Y) = \breve{F}(c_\omega + \mathrm{r}_\omega \cos(2\omega t + \xi_\omega))$
$W_{t,\omega}$	$I(W_{t,\omega}, Y) = f(2c_\omega)$	$A(W_{t,\omega}, Y) = F(2c_\omega)$	$D(W_{t,\omega}, Y) = \breve{F}(2c_\omega)$

7. Conclusion

We have characterised the relationship between the stimulus evoked spectrum and the information content spectrum, which is commonly used to investigate the brain's representational dynamics. Understanding how these two quantities relate is crucial to interpreting results obtained via decoding pipelines. By establishing these relationships under three different decoding paradigms, this work opens the door to much stronger interpretation of decoding results by linking the question of *what* is being represented with the neural mechanisms explaining *how* it is being represented. We hope this will enable more targeted scientific enquiry to uncover the true mechanisms by which the brain processes diverse forms of information.

Author contributions

C.H.: conceptualisation, methodology, software, formal analysis, investigation, writing – original draft, project administration; M.V.E.: methodology, software, validation, resources, writing – review and editing, visualisation; A.Q.: validation, writing - review and editing; D.V.: validation, writing - review and editing; M.W.: validation, writing – review and editing, supervision, funding acquisition.

Data and code availability

The data used in Results Section 4 is from a previously published work (Cichy et al., 2016); it is publicly available for download at http://userpage.fu-berlin.de/rmcichy/fusion_project_page/main.html. The code to perform all the analysis and example simulations published in this paper is publicaly available at https://github.com/OHBA-analysis/RepresentationalDynamicsModelling. The interactive web application accompanying Fig. 2 is published at https://doi.org/10.5281/zenodo.6579997 and is hosted at https://representational-dynamics.herokuapp.com/.

Ethics statement

The data used in Section 4 is from a previously published work (Cichy et al., 2016); as established in the original publication, the study was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee (Institutional Review Board of the Massachusetts Institute of Technology).

Declaration of Competing Interest

The authors have no interests to declare.

Acknowledgments

This research was funded by the Wellcome Trust (106183/Z/14/Z, 215573/Z/19/Z), the New Therapeutics in Alzheimer's Diseases (NTAD) study supported by UK MRC and the Dementia Platform UK (RG94383/RG89702) and the EU-project euSNN (MSCA-ITN H2020-860563), and supported by the NIHR Oxford Health Biomedical Research Centre. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). DV is supported by a Novo Nordisk Emerging Investigator Award (NNF19OC-0054895) and by the European Research Council (ERC-StG-2019-850404). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119462.

A. Full model specification

The assumptions expressed in Section 2.1 can be more fully expressed mathematically, with corresponding expressions obtained for the probability distribution of each of the random variables X_t, Y and $Z_{t,\omega}$.

We have assumed that the stimulus class is binary with equal class probabilities. This corresponds to the following distribution for Y:

$$P(Y) = \begin{cases} 1 \text{ with probability } 0.5 \\ -1 \text{ with probability } 0.5 \end{cases} \tag{A1}$$

Following on from Eq. (1), since the μ_t term captures the mean over both conditions, we can then model the expected value of each signal $z_{n,t,\omega}$ on individual trials with polarity determined by the trial condition y_n. We assume these narrowband signals have multivariate Gaussian noise that is independent and identically distributed (over trials and conditions) – this corresponds to an assumption that only the evoked response, and not the induced response, differs over the two conditions. This is a simplifying assumption that we discuss further and ultimately relax in Section 2.4. This allows us to specify the probability distributions of $Z_{t,\omega}$ (of which $z_{n,t,\omega}$ is the sample corresponding to the nth trial) as follows:

$$P\left(Z_{t,\omega}|\mathrm{Y}\right) = N\left(Y A_\omega \cos\left(\omega t + \phi_\omega\right), \Sigma_\omega\right) \tag{A2}$$

Each A_ω term is a diagonal $[P \times P]$ matrix, where the ith diagonal entry, denoted by $a_{\omega,i}$, reflects the magnitude of the component at frequency ω on channel i. Both ω and t are scalar indices reflecting the frequency and time respectively, whereas ϕ_ω is a $[P \times 1]$ vector, each entry of which contains the phase offset of the oscillation at frequency across the P channels. Finally, we model induced effects (i.e. narrowband power that is not phase aligned to the stimulus) independently in each frequency band, where Σ_ω is the $[P \times P]$ covariance matrix modelling the spatial variance and correlations expressed at frequency band ω.

For any set of discretely sampled data recordings with at least P total trials (i.e. more trials than channels), all of the above parameters are fully identifiable. The data for each channel and each trial can be decomposed into a discrete Fourier series representation of the above form where the number of frequency components equals half the number of timepoints in the trial $\Omega = \frac{T}{2}$ (for simplicity we here model a single Fourier decomposition over the trial; this can equivalently be computed over sliding windows as in Section 4, in which case the number of frequency components equals half the number of samples in a window). Following from the uniqueness of the Fourier transform, unique values can be obtained for the diagonal matrix A_ω, the phase offsets ϕ_ω and the patterns of spatial correlations in each frequency Σ_ω.

The distribution of the broadband signal X_t given in Eq. (2) then follows directly from Eq. (1) (by observing that a sum of narrowband Gaussian components is also Gaussian distributed).

B. Mutual information for a Gaussian mixture model with equal covariances

Let us first consider a simpler model and derive a general result that we can then use to prove our claims. Suppose we have a random variable Y distributed as given in the text:

$$\mathrm{P}(Y) = \begin{cases} 1 \text{ with probability } 0.5 \\ -1 \text{ with probability } 0.5 \end{cases}$$

Suppose we then have another random variable B of dimension $P \times 1$ conditioned on Y as follows:

$$P(B|Y) = N(\mu + Y\mathrm{m}, \mathrm{S})$$

This is equivalent to a Gaussian mixture model with two components (corresponding to the cases $Y = 1$ and $Y = -1$) and equal covariances. The marginal distribution can then be expressed as follows:

$$\begin{aligned}
P(B) &= P(Y=1)P(B|Y=1) + P(Y=-1)P(B|Y=-1) \\
&= \frac{1}{2\sqrt{2\pi|S|}}\left(e^{-\frac{1}{2}(B-\mu-m)^T S^{-1}(B-\mu-m)} + e^{-\frac{1}{2}(B-\mu+m)^T S^{-1}(B-\mu+m)}\right) \\
&= \left(\frac{1}{\sqrt{2\pi|S|}} e^{-\frac{1}{2}(B-\mu-m)^T S^{-1}(B-\mu-m)}\right)\left(\frac{1+e^{-2(B-\mu)^T S^{-1} m}}{2}\right) \\
&= N(B|\mu+m,S)\left(\frac{1+e^{-2(B-\mu)^T S^{-1} m}}{2}\right)
\end{aligned}$$

Where we use the notation $N(B|\mu+m,S)$ to denote the Gaussian distribution over B with mean $\mu+m$ and covariance S. We shall furthermore use the notation $\mathbb{E}_{P(x)} f(x)$ to denote the expectation of a function $f(x)$ given the probability distribution $P(X)$. We can now compute the following result for the entropy of B:

$$\begin{aligned}
H(B) &= -\int P(B)\log P(B) dB \\
&= -\int \frac{1}{2\sqrt{2\pi|S|}}\left(e^{-\frac{1}{2}(B-\mu-m)^T S^{-1}(B-\mu-m)} + e^{-\frac{1}{2}(B-\mu+m)^T S^{-1}(B-\mu+m)}\right)\log\left(N(B|\mu+m,S)\left(\frac{1+e^{-2(B-\mu)^T S^{-1} m}}{2}\right)\right) dB \\
&= \log 2 - \frac{1}{2}\mathbb{E}_{N(B|\mu+m,S)}\log N(B|\mu+m,S) - \frac{1}{2}\mathbb{E}_{N(B|\mu+m,S)}\log\left(1+e^{-2(B-\mu)^T S^{-1} m}\right) \\
&\quad - \frac{1}{2}\mathbb{E}_{N(B|\mu-m,S)}\log N(B|\mu-m,S) - \frac{1}{2}\mathbb{E}_{N(B|\mu-m,S)}\log\left(1+e^{2(B-\mu)^T S^{-1} m}\right)
\end{aligned}$$

We then observe that the second and fourth terms correspond to the entropy of a multivariate Gaussian, which has a known solution; we similarly observe that the third and fifth remaining terms are each an expectation of a univariate function in $u = 2(B-\mu)^T S^{-1} m$ and $v = 2(B+\mu)^T S^{-1} m$ respectively. With a substitution of variables this simplifies to:

$$H(B) = \log 2 + \frac{1}{2}\log|S| + \frac{P}{2}(1+\log 2\pi) - \mathbb{E}_{N(u|2m^T S^{-1} m, 4m^T S^{-1} m)}\log(1+e^{-u})$$

Similarly, we find that the conditional entropy is given by:

$$\begin{aligned}
H(B|Y) &= -\frac{1}{2}\mathbb{E}_{N(B|\mu-m,S)}\log N(B|\mu-m,S) - \frac{1}{2}\mathbb{E}_{N(B|\mu+m,S)}\log N(B|\mu+m,S) \\
&= \frac{1}{2}\log|S| + \frac{P}{2}(1+\log 2\pi)
\end{aligned}$$

We can therefore apply the chain rule to derive the mutual information:

$$\begin{aligned}
I(B,Y) &= H(B) - H(B|Y) \\
&= \log 2 - \mathbb{E}_{N(u|\alpha,2\alpha)}\log(1+e^{-u})
\end{aligned}$$

Note that the second term involves an integral that is intractable, but is a function of the scalar product $\alpha = 2m^T S^{-1} m$. We therefore can state equivalently that:

$$I(B,Y) = f(\alpha)$$

Where

$$f(\alpha) = \log 2 - \int \frac{1}{2\sqrt{\pi\alpha}} e^{-\frac{1}{4\alpha}(u-\alpha)^2}\log(1+e^{-u}) du$$

C. Proof that the function f is monotonic and concave

Firstly, let us define the following probability distribution:

$$Q(u) = \frac{1}{C_1 2\sqrt{\pi\alpha}} e^{-\frac{1}{4\alpha}(u-\alpha)^2}\log(1+e^{-u})$$

Where $C_1 = \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi\alpha}} e^{-\frac{1}{4\alpha}(u-\alpha)^2}\log(1+e^{-u})\, du = \mathbb{E}_{N(u|\alpha,2\alpha)}\log(1+e^{-u})$. Our proof below rearranges the first and second derivative of f in terms of the higher moments of this distribution, thus we now seek an expression for these moments. The moment generating function for $Q(u)$ is:

$$\begin{aligned}
\mathbb{E}_{Q(u)} e^{tu} &= \frac{1}{C_1}\int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi\alpha}} e^{-\frac{1}{4\alpha}(u-\alpha)^2}\log(1+e^{-u}) e^{tu}\, du \\
&= \frac{1}{C_1} e^{\alpha t(1+t)} \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi\alpha}} e^{-\frac{1}{4\alpha}(u-\alpha(1+2t))^2}\log(1+e^{-u})\, du
\end{aligned}$$

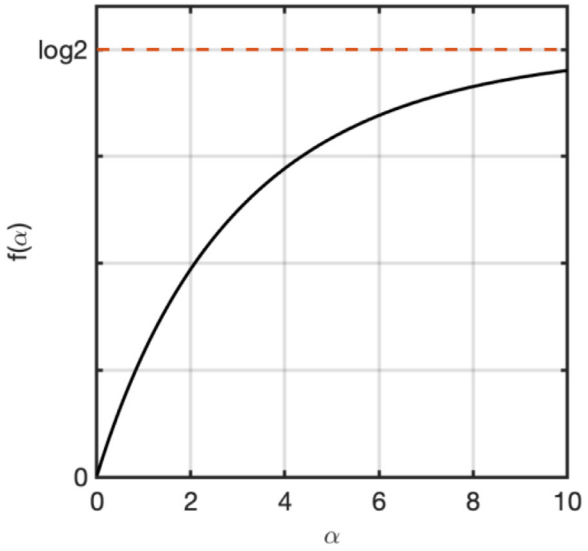

Fig. A1. the function $f(\alpha)$, which uniquely determines the information content of data generated from the Gaussian model specified in the text, is a monotonic concave function.

$$= \frac{\mathbb{E}_{N(u|\alpha(1+2t),2\alpha)}\log(1+e^{-u})}{\mathbb{E}_{N(u|\alpha,2\alpha)}\log(1+e^{-u})} e^{\alpha t(1+t)}$$
$$= \frac{\mathbb{E}_{N(u|\alpha,2\alpha)}\log\left(1+e^{-u-2\alpha t}\right)}{\mathbb{E}_{N(u|\alpha,2\alpha)}\log\left(1+e^{-u}\right)} \mathbb{E}_{N(u|\alpha,2\alpha)} e^{tu}$$

The moment generating function allows us to compute the higher moments of the distribution that are ultimately required for the proof. As the algebra for this is somewhat tedious we refer readers to Supplementary Information Section 3 for full details, where we obtain the following expressions for these moments:

$$\mathbb{E}_{Q(u)}u^2 = \alpha^2 + 2\alpha - \frac{4\alpha^2}{C_1}\mathbb{E}_{N(u|\alpha,2\alpha)}(1-\sigma(u))^2$$
$$\mathbb{E}_{Q(u)}u^4 = \alpha^4 + 12\alpha^3 + 12\alpha^2 - \frac{8\alpha^3}{C_1}\mathbb{E}_{N(u|\alpha,2\alpha)}\left((1-\sigma(u))^2\left(\alpha(2\sigma(u)-1)^2+6\right)\right)$$

Where $\sigma(u) = \frac{1}{1+e^{-u}}$ denotes the logistic sigmoid function. Note that the terms inside the expectations are strictly positive, such that their expectation is always greater than zero.

Now, consider the function $f(\alpha)$, plotted in Fig. A1 and specified as in Appendix B.

The first derivative is:

$$f'(\alpha) = \frac{d}{d\alpha}\mathbb{E}_{N(u|\alpha,2\alpha)}\log\left(\frac{1}{1+e^{-u}}\right)$$

Let us denote by $g(\alpha,u) = \frac{1}{2\sqrt{\pi\alpha}}e^{-\frac{(u-\alpha)^2}{4\alpha}}\log(\frac{1}{1+e^{-u}})$. We can then evaluate by Leibniz rule:

$$\frac{d}{d\alpha}\left(\int g(\alpha,u)du\right) = \int \frac{\partial}{\partial\alpha}g(\alpha,u)du$$

$$f'(\alpha) = \int \log\left(\frac{1}{1+e^{-u}}\right)\left(\frac{-1}{4\sqrt{\pi\alpha^3}}e^{-\frac{(u-\alpha)^2}{4\alpha}} + \frac{1}{2\sqrt{\pi\alpha}}e^{-\frac{(u-\alpha)^2}{4\alpha}}\left(\frac{u^2-\alpha^2}{4\alpha^2}\right)\right)du$$
$$= \int\left(\frac{1}{2\sqrt{\pi\alpha}}e^{-\frac{(u-\alpha)^2}{4\alpha}}\right)\log\left(1+e^{-u}\right)\left(\frac{\alpha^2+2\alpha-u^2}{4\alpha^2}\right)du$$
$$= \frac{C_1}{4\alpha^2}\mathbb{E}_{Q(u)}\left(\alpha^2+2\alpha-u^2\right)$$

Substituting the above expression for $\mathbb{E}_{Q(u)}u^2$, we have:

$$f'(\alpha) = \frac{C_1}{4\alpha^2}(\alpha^2+2\alpha-\alpha^2-2\alpha+\frac{4\alpha^2}{C_1}\mathbb{E}_{N(u|\alpha,2\alpha)}(1-\sigma(u))^2$$
$$= \mathbb{E}_{N(u|\alpha,2\alpha)}(1-\sigma(u))^2$$
$$> 0$$

We conclude that f is monotonic.

A second application of Leibniz' rule gives us the second derivative:

$$
\begin{aligned}
f''(\alpha) &= \int \frac{\partial}{\partial\alpha}\left(\frac{1}{2\sqrt{\pi\alpha}}e^{-\frac{(u-\alpha)^2}{4\alpha}}\right)\left(\frac{u^2-4\alpha^2-2\alpha}{4\alpha^2}\right)\log\left(\frac{1}{1+e^{-u}}\right)du \\
&= \int \log\left(\frac{1}{1+e^{-u}}\right)\left(\frac{\alpha^4+4\alpha^3+\left(12-2u^2\right)\alpha^2-12u^2\alpha+u^4}{32\sqrt{\pi\alpha^9}}\right)e^{-\frac{(u-2\alpha)^2}{4\alpha}}du \\
&= -C_1\int \frac{1}{2C_1\sqrt{\pi\alpha}}e^{-\frac{(u-\alpha)^2}{4\alpha}}\log\left(1+e^{-u}\right)\left(\frac{\alpha^4+4\alpha^3+12\alpha^2-2u^2\left(6\alpha+\alpha^2\right)+u^4}{16\alpha^4}\right)du \\
&= -\frac{C_1}{16\alpha^4}\mathbb{E}_{Q(u)}\left(\alpha^4+4\alpha^3+12\alpha^2-2u^2\left(\alpha^2+6\alpha\right)+u^4\right)
\end{aligned}
$$

Now substituting the above expressions for $\mathbb{E}_{Q(u)}u^2$ and $\mathbb{E}_{Q(u)}u^4$ we again find that most terms cancel out leaving us with:

$$
\begin{aligned}
f''(\alpha) &= -\frac{C_1}{16\alpha^4}\left(2\left(\frac{4\alpha^2}{C_1}\mathbb{E}_{N(u|\alpha,2\alpha)}(1-\sigma(u))^2\right)\left(6\alpha+\alpha^2\right)-\frac{8\alpha^3}{C_1}\mathbb{E}_{N(u|\alpha,2\alpha)}\left((1-\sigma(u))^2\left(\alpha(2\sigma(u)-1)^2+6\right)\right)\right) \\
&= -\frac{1}{2\alpha}\mathbb{E}_{N(u|\alpha,2\alpha)}(1-\sigma(u))^2\left(6+\alpha-\left(\alpha(2\sigma(u)-1)^2-6\right)\right) \\
&= -\frac{1}{2}\mathbb{E}_{N(u|\alpha,2\alpha)}(1-\sigma(u))^2\left(1-(2\sigma(u)-1)^2\right) \\
&= -\mathbb{E}_{N(u|\alpha,2\alpha)}2\sigma(u)(1-\sigma(u))^3 \\
&< 0
\end{aligned}
$$

Which follows from $0<\sigma(u)<1$. We conclude that f is concave.

D. Mutual information of the narrowband signal

Consider the narrowband real signal $Z_{\omega,t}$ as specified in Equation A2. It can be seen that this is a special case of the model specified in Appendix B by substituting the following:

$$m = A_\omega\cos\left(\omega t+\phi_\omega\right)$$

$$S = \Sigma_\omega$$

It therefore follows that the information content is given by:

$$I\left(Z_{t,\omega},Y\right) = f\left(2m^T S^{-1}m\right)$$

Re-writing the mean term in cartesian form, we have:

$$m = \frac{\left(u_\omega+iv_\omega\right)e^{i\omega t}+\left(u_\omega-iv_\omega\right)e^{-i\omega t}}{2}$$

Where $u_\omega = A_\omega\cos(\phi_\omega)$ and $v_\omega = A_\omega\sin(\phi_\omega)$. This allows us to determine the information term:

$$
\begin{aligned}
2m^T S^{-1}m &= \frac{1}{2}\left(\left(u_\omega+iv_\omega\right)e^{i\omega t}+\left(u_\omega+iv_\omega\right)e^{-i\omega t}\right)^T\Sigma_\omega^{-1}\left(\left(u_\omega+iv_\omega\right)e^{i\omega t}+\left(u_\omega+iv_\omega\right)e^{-i\omega t}\right) \\
&= \frac{1}{2}\left[e^{2i\omega t}\left(u_\omega+iv_\omega\right)^T\Sigma_\omega^{-1}\left(u_\omega+iv_\omega\right)+e^{-2i\omega t}\left(u_\omega-iv_\omega\right)^T\Sigma_\omega^{-1}\left(u_\omega-iv_\omega\right)+2u_\omega^T\Sigma_\omega^{-1}u_\omega+2v_\omega^T\Sigma_\omega^{-1}v_\omega\right] \\
&= \left(u_\omega^T\Sigma_\omega^{-1}u_\omega+v_\omega^T\Sigma_\omega^{-1}v_\omega+\left(u_\omega^T\Sigma_\omega^{-1}u_\omega-v_\omega^T\Sigma_\omega^{-1}v_\omega\right)\frac{e^{2i\omega t}+e^{-2i\omega t}}{2}-2u_\omega^T\Sigma_\omega^{-1}v_\omega\frac{e^{2i\omega t}-e^{-2i\omega t}}{2i}\right) \\
&= \left(u_\omega^T\Sigma_\omega^{-1}u_\omega+v_\omega^T\Sigma_\omega^{-1}v_\omega+\left(u_\omega^T\Sigma_\omega^{-1}u_\omega-v_\omega^T\Sigma_\omega^{-1}v_\omega\right)\cos 2\omega t-2u_\omega^T\Sigma_\omega^{-1}v_\omega\sin 2\omega t\right) \\
&= c_\omega+r_\omega\cos\left(2\omega t+\xi_\omega\right)
\end{aligned}
$$

Where $c_\omega = u_\omega^T\Sigma_\omega^{-1}u_\omega+v_\omega^T\Sigma_\omega^{-1}v_\omega$, $\tan\xi_\omega = \frac{2u_\omega^T\Sigma_\omega^{-1}v_\omega}{u_\omega^T\Sigma_\omega^{-1}u_\omega-v_\omega^T\Sigma_\omega^{-1}v_\omega}$ and $r_\omega^2 = \left(u_\omega^T\Sigma_\omega^{-1}u_\omega-v_\omega^T\Sigma_\omega^{-1}v_\omega\right)^2+\left(2u_\omega^T\Sigma_\omega^{-1}v_\omega\right)^2$.

Alternatively, returning to the polar coordinates used throughout the paper, we have:

$$
\begin{aligned}
c_\omega &= \cos\left(\phi_\omega^T\right)A_\omega\Sigma_\omega^{-1}A_\omega\cos\left(\phi_\omega\right)+\sin\left(\phi_\omega^T\right)A_\omega\Sigma_\omega^{-1}A_\omega\sin\left(\phi_\omega\right) \\
&= Tr\left(A_\omega\Sigma_\omega^{-1}A_\omega\left(\cos\phi_\omega\cos\left(\phi_\omega^T\right)+\sin\left(\phi_\omega\right)\sin\left(\phi_\omega^T\right)\right)\right) \\
&= Tr\left(A_\omega\Sigma_\omega^{-1}A_\omega\cos\left[\phi_\omega-\phi_\omega^T\right]\right)
\end{aligned}
$$

And applying the same steps for the remaining variables gives us:

$$r_\omega^2 = Tr^2\left(A_\omega\Sigma_\omega^{-1}A_\omega\cos\left(\left[\phi_\omega+\phi_\omega^T\right]\right)\right)+Tr^2\left(A_\omega\Sigma_\omega^{-1}A_\omega\sin\left(\left[\phi_\omega+\phi_\omega^T\right]\right)\right)$$

$$\xi_\omega = tan^{-1}\left(\frac{Tr\left(A_\omega\Sigma_\omega^{-1}A_\omega\sin\left(\left[\phi_\omega+\phi_\omega^T\right]\right)\right)}{Tr\left(A_\omega\Sigma_\omega^{-1}A_\omega\cos\left(\left[\phi_\omega+\phi_\omega^T\right]\right)\right)}\right)$$

Where we have used the notation $[\phi \pm \phi^T]$ for the $[P \times P]$ matrix constructed from the $[P \times 1]$ vector ϕ_ω such that the i,jth matrix entry is given by $\phi_{\omega,i} \pm \phi_{\omega,j}$.

We conclude that the narrowband signal information content is given by:

$$I\left(Z_{t,\omega}, Y\right) = f\left(c_\omega + \mathrm{r}_\omega \cos\left(2\omega t + \xi_\omega\right)\right)$$

E. Mutual information of the broadband signal

Consider the broadband signal X_t as specified in Eq. (2). It can be seen that this is a special case of the model specified in Appendix B by substituting the following:

$$m = \sum_{\omega=0}^{\Omega} A_\omega \cos\left(\omega t + \phi_\omega\right)$$

$$S = \sum_{\omega=0}^{\Omega} \Sigma_\omega$$

It therefore follows that the information content in the broadband signal is given by

$$I\left(X_t, Y\right) = f\left(2m^T S^{-1} m\right)$$

Substituting the above values:

$$2m^T S^{-1} m = 2\left[\sum_{\omega=0}^{\Omega} A_\omega \cos\left(\omega t + \phi_\omega\right)\right]^T \left[\sum_{\omega=0}^{\Omega} \Sigma_\omega\right]^{-1} \left[\sum_{\omega=0}^{\Omega} A_\omega \cos\left(\omega t + \phi_\omega\right)\right]$$

$$= 2\left(\sum_{\omega=0}^{\Omega} \cos\left(\omega t + \phi_\omega\right)^T A_\omega \Sigma_B^{-1} A_\omega \cos\left(\omega t + \phi_\omega\right) + 2\cos\left(\omega t + \phi_\omega\right)^T A_\omega \Sigma_B^{-1}\left(\sum_{\psi=\omega+1}^{\Omega} A_\psi \cos\left(\psi t + \phi_\psi\right)\right)\right)$$

Where $\Sigma_B = \sum_{\omega=0}^{\Omega} \Sigma_\omega$. Writing in cartesian coordinates where $u_\omega = A_\omega \cos(\phi_\omega)$ and $v_\omega = A_\omega \sin(\phi_\omega)$, such that:

$$A_\omega \cos\left(\omega t + \phi_\omega\right) = \frac{\left(u_\omega + iv_\omega\right)e^{i\omega t} + \left(u_\omega - iv_\omega\right)e^{-i\omega t}}{2}$$

This becomes:

$$2m^T S^{-1} m$$

$$= 2\left(\sum_{\omega=0}^{\Omega} \frac{(u_\omega^T + iv_\omega^T)e^{i\omega t} + (u_\omega^T - iv_\omega^T)e^{-i\omega t}}{2} \Sigma_B^{-1} \frac{(u_\omega + iv_\omega)e^{i\omega t} + (u_\omega - iv_\omega)e^{-i\omega t}}{2}\right.$$

$$\left. + 2\frac{(u_\omega^T + iv_\omega^T)e^{i\omega t} + (u_\omega^T - iv_\omega^T)e^{-i\omega t}}{2} \Sigma_B^{-1}\left(\sum_{\psi=\omega+1}^{\Omega} \frac{\left(u_\psi + iv_\psi\right)e^{i\psi t} + \left(u_\psi - iv_\psi\right)e^{-i\psi t}}{2}\right)\right)$$

$$= u_\omega^T \Sigma_B^{-1} u_\omega + v_\omega^T \Sigma_B^{-1} v_\omega$$

$$+ \left(\sum_{\omega=0}^{\Omega} \frac{(u_\omega^T \Sigma_B^{-1} u_\omega + v_\omega^T \Sigma_B^{-1} v_\omega + iv_\omega^T \Sigma_B^{-1} u_\omega + iu_\omega^T \Sigma_B^{-1} v_\omega)e^{2i\omega t} + (u_\omega^T \Sigma_B^{-1} u_\omega + v_\omega^T \Sigma_B^{-1} v_\omega - iv_\omega^T \Sigma_B^{-1} u_\omega - iu_\omega^T \Sigma_B^{-1} v_\omega)e^{-2i\omega t}}{2}\right.$$

$$+ 2\sum_{\psi=\omega+1}^{\Omega} \left(\frac{\left(u_\omega^T \Sigma_B^{-1} u_\psi - v_\omega^T \Sigma_B^{-1} v_\psi + iv_\omega^T \Sigma_B^{-1} u_\psi + iu_\omega^T \Sigma_B^{-1} v_\psi\right)e^{i(\omega+\psi)t} + \left(u_\omega^T \Sigma_B^{-1} u_\psi - v_\omega^T \Sigma_B^{-1} v_\psi - iv_\omega^T \Sigma_B^{-1} u_\psi - iu_\omega^T \Sigma_B^{-1} v_\psi\right)e^{-i(\omega+\psi)t}}{2}\right.$$

$$\left.\left. + \frac{\left(u_\omega^T \Sigma_B^{-1} u_\psi + v_\omega^T \Sigma_B^{-1} v_\psi - iv_\omega^T \Sigma_B^{-1} u_\psi + iu_\omega^T \Sigma_B^{-1} v_\psi\right)e^{i(\psi-\omega)t} + \left(u_\omega^T \Sigma_B^{-1} u_\psi + v_\omega^T \Sigma_B^{-1} v_\psi + iv_\omega^T \Sigma_B^{-1} u_\psi - iu_\omega^T \Sigma_B^{-1} v_\psi\right)e^{-i(\psi-\omega)t}}{2}\right)\right)$$

$$= c_B + \sum_{\omega=0}^{\Omega} r_{B,\omega} \cos\left(2\omega t + \xi_{B,\omega}\right) + h(t)$$

That is, a consant term c_B plus a sinusoidal component at double the frequency of each sinusoidal component of the evoked response, plus additional harmonic components $h(t)$ at the sum and difference of each pair of frequencies in the evoked response:

$$h(t) = 2\sum_{\psi=\omega+1}^{\Omega} s_{\omega,\psi} \cos\left((\psi + \omega)t + \zeta_{\omega,\psi}\right) + t_{\omega,\psi} \cos\left((\psi - \omega)t + \theta_{\omega,\psi}\right)$$

The variables take the following values:

$$
\begin{aligned}
c_B &= \sum_{\omega=0}^{\Omega} u_\omega^T \Sigma_{\mathrm{B}}^{-1} u_\omega + v_\omega^T \Sigma_{\mathrm{B}}^{-1} v_\omega \\
r_{B,\omega}^2 &= \left(u_\omega^T \Sigma_{\mathrm{B}}^{-1} u_\omega - v_\omega^T \Sigma_{\mathrm{B}}^{-1} v_\omega\right)^2 + \left(2 u_\omega^T \Sigma_{\mathrm{B}}^{-1} v_\omega\right)^2 \\
\tan \xi_{B,\omega} &= \frac{2 u_\omega^T \Sigma_{\mathrm{B}}^{-1} v_\omega}{u_\omega^T \Sigma_{\mathrm{B}}^{-1} u_\omega - v_\omega^T \Sigma_{\mathrm{B}}^{-1} v_\omega} \\
s_{\omega,\psi}^2 &= \left(u_\omega^T \Sigma_B^{-1} u_\psi - v_\omega^T \Sigma_B^{-1} v_\psi\right)^2 + \left(u_\omega^T \Sigma_B^{-1} v_\psi + v_\omega^T \Sigma_B^{-1} u_\psi\right)^2 \\
\tan \zeta_{\omega,\psi} &= \frac{u_\omega^T \Sigma_B^{-1} v_\psi + v_\omega^T \Sigma_B^{-1} u_\psi}{u_\omega^T \Sigma_B^{-1} u_\psi - v_\omega^T \Sigma_B^{-1} v_\psi} \\
t_{\omega,\psi}^2 &= \left(u_\omega^T \Sigma_B^{-1} u_\psi + v_\omega^T \Sigma_B^{-1} v_\psi\right)^2 + \left(u_\omega^T \Sigma_B^{-1} v_\psi - v_\omega^T \Sigma_B^{-1} u_\psi\right)^2 \\
\tan \theta_{\omega,\psi} &= \frac{u_\omega^T \Sigma_B^{-1} v_\psi + v_\omega^T \Sigma_B^{-1} u_\psi}{u_\omega^T \Sigma_B^{-1} u_\psi - v_\omega^T \Sigma_B^{-1} v_\psi}
\end{aligned}
$$

Converting back to the polar coordinates used throughout the paper, we have:

$$
\begin{aligned}
c_B &= \sum_{\omega=0}^{\Omega} Tr\left(A_\omega \Sigma_B^{-1} A_\omega \cos\left[\phi_\omega - \phi_\omega^T\right]\right) \\
r_{B,\omega}^2 &= \left(Tr\left(A_\omega \Sigma_B^{-1} A_\omega \cos\left[\phi_\omega + \phi_\omega^T\right]\right)\right)^2 + \left(Tr\left(A_\omega \Sigma_B^{-1} A_\omega \sin\left[\phi_\omega + \phi_\omega^T\right]\right)\right)^2 \\
\tan \xi_{B,\omega} &= \frac{Tr\left(A_\omega \Sigma_B^{-1} A_\omega \sin\left[\phi_\omega + \phi_\omega^T\right]\right)}{Tr\left(A_\omega \Sigma_B^{-1} A_\omega \cos\left[\phi_\omega + \phi_\omega^T\right]\right)} \\
s_{\omega,\psi}^2 &= \left(Tr\left(A_\omega \Sigma_B^{-1} A_\psi \cos\left[\phi_\omega + \phi_\psi^T\right]\right)\right)^2 + \left(Tr\left(A_\omega \Sigma_B^{-1} A_\psi \sin\left[\phi_\omega + \phi_\psi^T\right]\right)\right)^2 \\
\tan \zeta_{\omega,\psi} &= \frac{Tr\left(A_\omega \Sigma_B^{-1} A_\psi \sin\left[\phi_\omega + \phi_\psi^T\right]\right)}{Tr\left(A_\omega \Sigma_B^{-1} A_\psi \cos\left[\phi_\omega + \phi_\psi^T\right]\right)} \\
t_{\omega,\psi}^2 &= \left(Tr\left(A_\omega \Sigma_B^{-1} A_\psi \cos\left[\phi_\omega - \phi_\psi^T\right]\right)\right)^2 + \left(Tr\left(A_\omega \Sigma_B^{-1} A_\psi \sin\left[\phi_\psi - \phi_\omega^T\right]\right)\right)^2 \\
\tan \theta_{\omega,\psi} &= \frac{Tr\left(A_\omega \Sigma_B^{-1} A_\psi \sin\left[\phi_\psi - \phi_\omega^T\right]\right)}{Tr\left(A_\omega \Sigma_B^{-1} A_\psi \cos\left[\phi_\omega - \phi_\psi^T\right]\right)}
\end{aligned}
$$

We conclude that the broadband information content is given by

$$
I\left(X_t, Y\right) = f\left(c_B + \sum_{\omega}^{\Omega} r_{B,\omega} \cos\left(2\omega t + \xi_{B,\omega}\right) + \mathrm{h(t)}\right)
$$

F. Mutual information of the complex-valued Fourier signal

Consider the complex signal $W_{\omega,t}$ as specified in Eq. (8). It can be seen that this is a special case of the model specified in Appendix B by substituting the following:

$$
m = \left[u_\omega; v_\omega\right]
$$

$$
S = \begin{bmatrix} \Sigma_\omega & 0 \\ 0 & \Sigma_\omega \end{bmatrix}
$$

It therefore follows that the information content is given by:

$$
I\left(W_{t,\omega}, Y\right) = f\left(2 m^T S^{-1} m\right)
$$

Substituting the above values we observe that:

$$
\begin{aligned}
2 m^T S^{-1} m &= 2\left(u_\omega^T \Sigma_\omega^{-1} u_\omega + v_\omega^T \Sigma_\omega^{-1} v_\omega\right) \\
&= 2 c_\omega
\end{aligned}
$$

We conclude that the information content in the complex signal $W_{\omega,t}$ is given by:

$$
I\left(W_{t,\omega}, Y\right) = f\left(2 c_\omega\right)
$$

G. Mutual information for a Gaussian mixture model with different covariances

The modelling above assumes that the induced response (i.e. the changes in power that have no phase alignment to the stimulus) has the same distribution over both stimulus conditions. This corresponds to an assumption that the narrowband covariance matrix is invariant over stimulus conditions. To explore how our results generalise to the case of stimulus-specific induced effects, let us return to the result of Appendix B and now define a new random variable $\tilde{B}$ of dimension $P \times 1$ as follows:

$$\tilde{B} = B + \tilde{\varepsilon}$$

Where the new residual terms have distribution $P(\tilde{\varepsilon}|Y) = N(0, \mathrm{U}_Y)$ will be used to model induced effects. We assume that $P(\tilde{\varepsilon}|Y)$ is independent of $P(\tilde{B}|Y)$. We previously defined B as a Gaussian mixture model with two components (corresponding to the cases $Y = 1$ and $Y = -1$) and equal covariances; thus the new random variable has a distribution given by:

$$P\left(\tilde{B}|Y\right) = N\left(\mu + Y\mathrm{m}, \mathrm{S} + \mathrm{U}_Y\right)$$

Which corresponds to a Gaussian mixture model with some common covariance given by S as well as some stimulus-specific covariance given by U_Y. The mutual information for this variable is not tractable, however we instead obtain an upper bound by observing that $\tilde{B}$ is a linear function of B and $\tilde{\varepsilon}$, therefore the data processing inequality tells us that:

$$I\left(\mathrm{Y}, \tilde{B}\right) \leq I(Y, [B; \tilde{\varepsilon}])$$

Where $[B; \tilde{\varepsilon}]$ is the vector obtained by concatenating the random variables B and $\tilde{\varepsilon}$. We shall now obtain an expression for the mutual information between this new term and the stimulus labels Y. Note that the marginal distribution is given by:

$$\begin{aligned}
P([B;\tilde{\varepsilon}]) &= P(Y=1)P([B;\tilde{\varepsilon}]|Y=1) + P(Y=-1)P([B;\tilde{\varepsilon}]|Y=-1) \\
&= \frac{1}{2\sqrt{2\pi|S||\mathrm{U}_1|}} e^{-\frac{1}{2}(\mathrm{B}-\mu-m)^T S^{-1}(\mathrm{B}-\mu-m) - \frac{1}{2}\tilde{\varepsilon}^T \mathrm{U}_1^{-1}\tilde{\varepsilon}} + \frac{1}{2\sqrt{2\pi|S||\mathrm{U}_{-1}|}} e^{-\frac{1}{2}(\mathrm{B}+\mu+m)^T S^{-1}(\mathrm{B}+\mu+m) - \frac{1}{2}\tilde{\varepsilon}^T \mathrm{U}_1^{-1}\tilde{\varepsilon}} \\
&= \frac{1}{\sqrt{2\pi\begin{vmatrix} S & 0 \\ 0 & U_1 \end{vmatrix}}} e^{-\frac{1}{2}\left(\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix} - \begin{bmatrix}\mu+m\\ 0\end{bmatrix}\right)^T \begin{pmatrix} S & 0 \\ 0 & U_1 \end{pmatrix}^{-1} \left(\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix} - \begin{bmatrix}\mu+m\\ 0\end{bmatrix}\right)} \left(\frac{1 + \frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|} e^{-2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mu+m) - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_{-1}^{-1} - \mathrm{U}_1^{-1}\right)\tilde{\varepsilon}}}{2} \right)
\end{aligned}$$

Following the same approach of Appendix B, we obtain the following for the entropy:

$$\begin{aligned}
H([B;\tilde{\varepsilon}]) &= -\int P([B;\tilde{\varepsilon}]) \log P([B;\tilde{\varepsilon}]) d[B;\tilde{\varepsilon}] \\
&= -\frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu+m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_1 \end{bmatrix}\right]} \left(\log\left(N\left(\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix} \middle| \begin{bmatrix}\mu+m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_1 \end{bmatrix} \right) \right) \left(\frac{1 + \frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|} e^{-2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_{-1}^{-1} - \mathrm{U}_1^{-1}\right)\tilde{\varepsilon}}}{2} \right) \right) \\
&\quad -\frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu-m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_{-1} \end{bmatrix}\right]} \left(\log\left(N\left(\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix} \middle| \begin{bmatrix}\mu-m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_{-1} \end{bmatrix} \right) \right) \left(\frac{1 + \frac{|\mathrm{U}_{-1}|}{|\mathrm{U}_1|} e^{2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_1^{-1} - \mathrm{U}_{-1}^{-1}\right)\tilde{\varepsilon}}}{2} \right) \right) \\
&= \log 2 - H([B;\tilde{\varepsilon}]|Y) - \frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu+m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_1 \end{bmatrix}\right]} \log\left(1 + \frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|} e^{-2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_{-1}^{-1} - \mathrm{U}_1^{-1}\right)\tilde{\varepsilon}} \right) \\
&\quad -\frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu-m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_{-1} \end{bmatrix}\right]} \log\left(1 + \frac{|\mathrm{U}_{-1}|}{|\mathrm{U}_1|} e^{2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_1^{-1} - \mathrm{U}_{-1}^{-1}\right)\tilde{\varepsilon}} \right) \\
&= \log 2 - H([B;\tilde{\varepsilon}]|Y) - \frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu+m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_1 \end{bmatrix}\right]} \log\left(1 + \frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|} e^{-2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_{-1}^{-1} - \mathrm{U}_1^{-1}\right)\tilde{\varepsilon}} \right)
\end{aligned}$$

Applying the chain rule we have:

$$\begin{aligned}
\mathrm{I}(Y, [B;\tilde{\varepsilon}]) &= \log 2 - \frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu+m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_1 \end{bmatrix}\right]} \log\left(1 + \frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|} e^{-2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_{-1}^{-1} - \mathrm{U}_1^{-1}\right)\tilde{\varepsilon}} \right) \\
&\quad -\frac{1}{2}\mathbb{E}_{N\left[\begin{bmatrix}\mathrm{B}\\ \tilde{\varepsilon}\end{bmatrix}\middle|\begin{bmatrix}\mu-m\\ 0\end{bmatrix}, \begin{bmatrix} S & 0 \\ 0 & U_{-1} \end{bmatrix}\right]} \log\left(1 + \frac{|\mathrm{U}_{-1}|}{|\mathrm{U}_1|} e^{2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m - \frac{1}{2}\tilde{\varepsilon}^T\left(\mathrm{U}_1^{-1} - \mathrm{U}_{-1}^{-1}\right)\tilde{\varepsilon}} \right)
\end{aligned}$$

Now noting that $\log(1+e^{-X})$ is a convex function, and that the expectations over B and $\tilde{\epsilon}$ can be separated as they are independent, we can apply Jensen's inequality to the expectation over $\tilde{\epsilon}$ terms to obtain:

$$\mathrm{I}(Y,[B;\tilde{\varepsilon}]) \leq \log 2 - \frac{1}{2}\mathbb{E}_{N(\mathrm{B}|\mu+m,S)} \log\left(1+\frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|}e^{-2(\mathrm{B}-\mu)^{\mathrm{T}}\Sigma^{-1}m-\frac{1}{2}Tr\left(\mathrm{U}_{-1}^{-1}\mathrm{U}_1-I_P\right)}\right) - \frac{1}{2}\mathbb{E}_{N(\mathrm{B}|\mu-m,S)} \log\left(1+\frac{|\mathrm{U}_{-1}|}{|\mathrm{U}_1|}e^{2(B-\mu)^{\mathrm{T}}\Sigma^{-1}m-\frac{1}{2}Tr\left(\mathrm{U}_1^{-1}\mathrm{U}_{-1}-I_P\right)}\right)$$

Let us now define a generalisation of the previously defined function $f(\alpha)$ to the following:

$$f_\rho(\alpha) = \log 2 - \int \frac{1}{2\sqrt{\pi\alpha}} e^{-\frac{1}{4\alpha}(u-\alpha)^2} \log(1+\rho e^{-u}) du$$

This allows us to write the information content upper bound as:

$$I\left(Y,\tilde{B}\right) \leq \frac{f_{\rho_1}(\alpha)+f_{\rho_2}(\alpha)}{2}$$

where α is the same term specified in Appendix B, and the new terms $\rho_1 = \frac{|\mathrm{U}_1|}{|\mathrm{U}_{-1}|}e^{-\frac{1}{2}Tr(\mathrm{U}_{-1}^{-1}\mathrm{U}_1-I_P)}$ and $\rho_2 = \frac{|\mathrm{U}_{-1}|}{|\mathrm{U}_1|}e^{-\frac{1}{2}Tr(\mathrm{U}_1^{-1}\mathrm{U}_{-1}-I_P)}$.

H. Upper bound for narrowband information content with induced effects

Let us now model the narrowband signal with stimulus dependent induced effects as follows:

$$\tilde{Z}_{n,t,\omega} = Z_{n,t,\omega} + \tilde{\varepsilon}_{\omega,t}$$

Where the new residual terms have distribution $P(\tilde{\varepsilon}_{\omega,t}|Y) = N(0,\Lambda_{Y,\,\omega})$. Note this is the same form as the model given in Appendix G, by substituting:

$$m = A_\omega \cos\left(\omega t + \phi_\omega\right)$$

$$S = \Sigma_\omega$$

$$U_Y = \Lambda_{Y,\,\omega}$$

From Appendix G, we deduce the following:

$$I\left(Y,\tilde{Z}_{n,t,\omega}\right) \leq \frac{f_{\rho_1}\left(c_\omega + \mathrm{r}_\omega\cos\left(2\omega t+\xi_\omega\right)\right) + f_{\rho_2}\left(c_\omega+\mathrm{r}_\omega\cos\left(2\omega t+\xi_\omega\right)\right)}{2}$$

Where $\rho_1 = \frac{|\Lambda_{1,\omega}|}{|\Lambda_{-1,\omega}|}e^{-\frac{1}{2}Tr(\Lambda_{-1,\omega}^{-1}\Lambda_{1,\omega}-I)}$ and $\rho_2 = \frac{|\Lambda_{-1,\omega}|}{|\Lambda_{1,\omega}|}e^{-\frac{1}{2}Tr(\Lambda_{1,\omega}^{-1}\Lambda_{-1,\omega}-I)}$ are both constant with respect to time, and c_ω, r_ω and ξ_ω are the same terms specified in Appendix D. Thus, the narrowband information content associated with condition-dependent evoked *and* induced effects has an upper bound which is a sinusoidal function translated to double the original narrowband signal frequency (i.e. a slightly modified function of the same dynamics previously characterised for the case where only evoked effects are stimulus-dependent).

I. Upper bound for broadband information content with induced effects

Let us now model the broadband signal with stimulus dependent induced effects as follows:

$$\tilde{X}_t = \mu_t + \sum_{\omega=0}^{\Omega}\tilde{Z}_{\omega,t}$$

This is equivalent to the model of Appendix G by substituting:

$$m = \sum_{\omega=0}^{\Omega} A_\omega \cos\left(\omega t+\phi_\omega\right)$$

$$S = \sum_{\omega=0}^{\Omega}\Sigma_\omega$$

$$U_Y = \sum_{\omega=0}^{\Omega}\Lambda_{Y,\,\omega}$$

We therefore deduce that:

$$I\left(Y,\tilde{X}_{n,t}\right) \leq \frac{f_{\rho_1}(\alpha)+f_{\rho_2}(\alpha)}{2}$$

Where $\alpha = c_B + \sum_{\omega}^{\Omega} r_{B,\omega}\cos(2\omega t+\xi_{B,\omega}) + \mathrm{h(t)}$, (i.e. the same as in the case of Appendix E where only evoked effects were modelled), c_B, $r_{B,\omega}$, $\xi_{B,\omega}$ and h(t) are all as specified in Appendix E, and the new terms $\rho_1 = \frac{|\sum_{\omega=0}^{\Omega}\Lambda_{1,\,\omega}|}{|\sum_{\omega=0}^{\Omega}\Lambda_{-1,\,\omega}|}e^{-\frac{1}{2}Tr((\sum_{\omega=0}^{\Omega}\Lambda_{-1,\,\omega})^{-1}(\sum_{\omega=0}^{\Omega}\Lambda_{1,\,\omega})-I)}$ and $\rho_2 = \frac{|\sum_{\omega=0}^{\Omega}\Lambda_{-1,\,\omega}|}{|\sum_{\omega=0}^{\Omega}\Lambda_{1,\,\omega}|}e^{-\frac{1}{2}Tr((\sum_{\omega=0}^{\Omega}\Lambda_{1,\,\omega})^{-1}(\sum_{\omega=0}^{\Omega}\Lambda_{-1,\,\omega})-I)}$ are both constant with respect to time.

References

Angjelichinoski, M., Banerjee, T., Choi, J., Pesaran, B., & Tarokh, V. (2019). Minimax-optimal decoding of movement goals from local field potentials using complex spectral features. *ArXiv*.

Brookshire, G., 2022. Putative rhythms in attentional switching can be explained by aperiodic temporal structure. Nature Human Behaviour doi:10.1038/s41562-022-01364-0.

Carlson, T.A., Hogendoorn, H., Kanai, R., Mesik, J., Turret, J., 2011. High temporal resolution decoding of object position and category. Journal of Vision 11 (10), 1–17. doi:10.1167/11.10.9.Introduction.

Carlson, T., Tovar, D.A., Alink, A., Kriegeskorte, N., 2013. Representational dynamics of object vision: The first 1000 ms. Journal of Vision 13 (10), 1. doi:10.1167/13.10.1.

Cichy, R.M., Pantazis, D., 2017. Multivariate pattern analysis of MEG and EEG: a comparison of representational structure in time and space. NeuroImage 158 (December 2016), 441–454. doi:10.1016/j.neuroimage.2017.07.023.

Cichy, R.M., Pantazis, D., Oliva, A., 2014. Resolving human object recognition in space and time. Nature Neuroscience 17 (3), 455–462. doi:10.1038/nn.3635.

Cichy, R.M., Pantazis, D., Oliva, A., 2016. Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. Cerebral Cortex 26 (8), 3563–3579. doi:10.1093/cercor/bhw135.

Dijkstra, N., Ambrogioni, L., Vidaurre, D., van Gerven, M., 2020. Neural dynamics of perceptual inference and its reversal during imagery. ELife 9, 1–19. doi:10.7554/eLife.53588.

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., Düzel, E., 2010. Theta-Coupled Periodic Replay in Working Memory. Current Biology 20 (7), 606–612. doi:10.1016/j.cub.2010.01.057.

Gennari, G., Marti, S., Palu, M., Fló, A., Dehaene-Lambertz, G., 2021. Orthogonal neural codes for speech in the infant brain. Proceedings of the National Academy of Sciences of the United States of America 118 (31), 1–11. doi:10.1073/pnas.2020410118.

Goddard, E., Carlson, T.A., Dermody, N., Woolgar, A., 2016. Representational dynamics of object recognition: Feedforward and feedback information flows. NeuroImage 128, 385–397. doi:10.1016/j.neuroimage.2016.01.006.

Grootswagers, T., Wardle, S.G., Carlson, T.A., 2017. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. Journal of Cognitive Neuroscience 29 (4), 677–697. doi:10.1162/jocn.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. PLoS Biology (12) 11. doi:10.1371/journal.pbio.1001752.

Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., Woolrich, M., 2021. Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. Neuron 109 (5), 882–893. doi:10.1016/j.neuron.2020.12.007, e7.

Higgins, C., Vidaurre, D., Kolling, N., Liu, Y., Behrens, T., Woolrich, M., 2021. Spatiotemporally Resolved Multivariate Pattern Analysis for M/EEG. BioRxiv doi:10.1101/2021.08.17.456594, 2021.08.17.456594.

Hogendoorn, H., Burkitt, A.N., 2018. Predictive coding of visual object position ahead of moving objects revealed by time-resolved EEG decoding. NeuroImage 171 (December 2017), 55–61. doi:10.1016/j.neuroimage.2017.12.063.

Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., Schyns, P.G., 2017. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. Human Brain Mapping 38 (3), 1541–1573. doi:10.1002/hbm.23471.

Ince, R.A.A., Jaworska, K., Gross, J., Panzeri, S., van Rijsbergen, N.J., Rousselet, G.A., Schyns, P.G., 2016. The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. Cerebral Cortex 26 (11), 4123–4135. doi:10.1093/cercor/bhw196.

Jafarpoura, A., Horner, A.J., Fuentemilla, L., Penny, W.D., Duzel, E., 2013. Decoding oscillatory representations and mechanisms in memory. Neuropsychologia 51, 772–780. doi:10.1016/j.neuropsychologia.2012.04.002.

Kalafatovich, J., Lee, M., Lee, S.W., 2020. Decoding Visual Recognition of Objects from EEG Signals based on Attention-Driven Convolutional Neural Network. In: *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 2020-Octob*(Mi), pp. 2985–2990. doi:10.1109/SMC42975.2020.9283434.

Kerrén, C., Linde-Domingo, J., Hanslmayr, S., Wimber, M., 2018. An Optimal Oscillatory Phase for Pattern Reactivation during Memory Retrieval. Current Biology 28, 3383–3392. doi:10.1016/j.cub.2018.08.065.

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., Kriegeskorte, N., 2019. Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences of the United States of America 116 (43), 21854–21863. doi:10.1073/pnas.1905544116.

Kikumoto, A., & Mayr, U. (2018). Decoding Hierarchical Control of Sequential Behavior in Oscillatory EEG Activity. *BioRxiv*, 1–36. https://doi.org/10.1101/344135

King, J.R., Dehaene, S., 2014. Characterizing the dynamics of mental representations: The temporal generalization method. Trends in Cognitive Sciences 18 (4), 203–210. doi:10.1016/j.tics.2014.01.002.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: Integrating cognition, computation, and the brain. Trends in Cognitive Sciences 17 (8), 401–412. doi:10.1016/j.tics.2013.06.007.

Kunz, L., Wang, L., Lachner-Piza, D., Zhang, H., Brandt, A., Dümpelmann, M., Reinacher, P.C., Coenen, V.A., Chen, D., Wang, W.X., Zhou, W., Liang, S., Grewe, P., Bien, C.G., Bierbrauer, A., Schröder, T.N., Schulze-Bonhage, A., Axmacher, N., 2019. Hippocampal theta phases organize the reactivation of large-scale electrophysiological representations during goal-directed navigation. Science Advances 5 (7), 1–18. doi:10.1126/sciadv.aav8192.

LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R., 2013. Decoding Attended Information in Short-term Memory: An EEG Study. Journal of Cognitive Neuroscience 25 (1), 127–142. doi:10.1162/jocn_a_00305.

Linde-Domingo, J., Treder, M.S., Kerrén, C., Wimber, M., 2019. Evidence that neural information flow is reversed between object perception and object reconstruction from memory. Nature Communications (1) 10. doi:10.1038/s41467-018-08080-2.

Mohsenzadeh, Y., Qin, S., Cichy, R.M., Pantazis, D., 2018. Ultra-Rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. ELife (7) e36329. doi:10.7554/eLife.36329.

Robinson, A.K., Grootswagers, T., Shatek, S.M., Gerboni, J., Holcombe, A., & Carlson, T.A. (2020). Overlapping neural representations for the position of visible and imagined objects. *ArXiv Preprint ArXiv:2010.09932*.

Samaha, J., Sprague, T.C., Postle, B.R., 2016. Decoding and Reconstructing the Focus of Spatial Attention from the Topography of Alpha-band Oscillations. Journal of Cognitive Neuroscience 28 (8), 1090–1097. doi:10.1162/jocn_a_00955.

Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping 38 (11), 5391–5420. doi:10.1002/hbm.23730.

Schyns, P.G., Thut, G., Gross, J., 2011. Cracking the code of oscillatory activity. PLoS Biology 9 (5). doi:10.1371/journal.pbio.1001064.

Valentin, S., Harkotte, M., Popov, T., 2020. Interpreting neural decoding models using grouped model reliance. PLoS Computational Biology 16 (1), 1–17. doi:10.1371/journal.pcbi.1007148.

van de Nieuwenhuijzen, M.E., Backus, A.R., Bahramisharif, A., Doeller, C.F., Jensen, O., van Gerven, M.A.J., 2013. MEG-based decoding of the spatiotemporal dynamics of visual category perception. NeuroImage 83, 1063–1073. doi:10.1016/j.neuroimage.2013.07.075.

Van Es, M.W.J., Higgins, C., Quinn, A.J., Vidaurre, D., Gould Van Praag, C.D., Fabus, M.S., Woolrich, M.W. (2022). Representational Dynamics Simulator. Zenodo. doi: 10.5281/zenodo.6579997. Available at representational-dynamics.herokuapp.com (June 10th, 2022).

van Es, M.W.J., Marshall, T.R., Spaak, E., Jensen, O., Schoffelen, J.M., 2020. Phasic modulation of visual representations during sustained attention. European Journal of Neuroscience 1–18. doi:10.1111/ejn.15084, *December 2020*.

Wolff, M.J., Ding, J., Myers, N.E., Stokes, M.G., 2015. Revealing hidden states in visual working memory using electroencephalography. Frontiers in Systems Neuroscience 9 (september), 1–12. doi:10.3389/fnsys.2015.00123.

Xie, S., Kaiser, D., Cichy, R.M., 2020. Visual Imagery and Perception Share Neural Representations in the Alpha Frequency Band. Current Biology 30 (13), 2621–2627. doi:10.1016/j.cub.2020.04.074, e5.

Zhan, J., Ince, R.A.A., van Rijsbergen, N., Schyns, P.G., 2019. Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior. Current Biology 29 (2), 319–326. doi:10.1016/j.cub.2018.11.049, e4.

Zubarev, I., Zetter, R., Halme, H.L., Parkkonen, L., 2019. Adaptive neural network classifier for decoding MEG signals. NeuroImage 197 (March), 425–434. doi:10.1016/j.neuroimage.2019.04.068.