Uncovering Temporal Structure in Neural Data with Statistical Machine Learning Models



Cameron Higgins

Wadham College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

In loving memory of my little bro,
Elliot

Contents

List of Figures **vi**

List of Abbreviations **viii**

1 Introduction **1**

1.1 Motivation 1

1.2 Experimental Task Design and Analysis in Neuroscience 4

1.3 Statistical models of neural dynamics 7

1.4 Towards Dynamic Modelling of Task Data 12

1.5 Structure of this thesis 14

2 Inferring Behavioural States from LFP Recordings in the Prefrontal Cortex **17**

2.1 Introduction 18

2.2 Experimental Outline and Methods 22

2.3 Multivariate Results: HMM States reflect both Spontaneous and Stimulus-Induced Effects 30

2.4 Univariate Results: Time-Frequency Analysis is Less Sensitive to these Effects 40

2.5 Discussion 43

2.6 Conclusion 47

3 Sequential Encoding and Decoding Frameworks **48**

3.1 Introduction 48

3.2 A Dynamic Framework for Regression Models 57

3.3 Synthetic Data Results 66

3.4 Discussion 72

3.5 Conclusion 73

4 Mapping Sequential Visual Stream Activation in MEG 74
4.1 Introduction . . . 74
4.2 Experimental Setup . . . 77
4.3 Results: Overall Classification Accuracy . . . 79
4.4 Decoding Model Results . . . 81
4.5 Encoding Model Results . . . 85
4.6 Conclusion . . . 88

5 Exploring Replay Dynamics with Sequential Decoding Models 90
5.1 Introduction . . . 91
5.2 On the Identification of Replay in Human MEG Studies . . . 99
5.3 A Survey of Classification Methods and Sequenceness Tests . . . 106
5.4 Temporal Patterns in Replay Activation over Different Areas . . . 114
5.5 On the Relationship between Sparsity and Replay Detection . . . 120
5.6 Conclusion . . . 126

6 Understanding Replay Correlations with Resting State Networks 128
6.1 Introduction . . . 129
6.2 Methods . . . 136
6.3 On the Temporal Structure of MEG Resting State Networks . . . 138
6.4 Characterising the Resting State Networks . . . 143
6.5 Resting State Network Activity Predicts the Onset of Replay . . . 145
6.6 Longer Range Temporal Dependencies . . . 150
6.7 Discussion . . . 154
6.8 Conclusion . . . 156

7 Conclusion 158
7.1 Contributions of this Thesis . . . 158
7.2 Future Research Directions . . . 160
7.3 Closing Remarks . . . 162

Appendices

A Variational Bayesian Inference 165
A.1 Factorising the posterior . . . 165
A.2 Variational Free Energy . . . 165

B Variational Inference for Latent State Dynamics 167

C Variational Inference for a Dynamic Encoding Model 169

D Variational Inference for Dynamic Categorical Decoding Model 171

E Rebalancing Adjustment **173**

F LFP Analysis Supplementary Information **174**
F.1 Average LFP response over cues 174
F.2 Channel-wise univariate analysis 176

G Sequential Visual Stream Activation Supplementary Information **180**
G.1 Performance of other methods 180
G.2 Frequency Domain Analysis 180

H Replay Dynamics: Supplementary Information **183**

I Replay and RSN Correlation: Supplementary Information **185**

Bibliography **193**

List of Figures

1.1 High level branches of neuroscience . . . 3
1.2 Orthogonal task design in fMRI . . . 6
1.3 Graphical model outline of a Hidden Markov Model . . . 10

2.1 Experimental task outline . . . 23
2.2 Time Delay Embedded model outline . . . 26
2.3 Principal Component Analysis isolates two main trial trajectories . 31
2.4 Main mode of state timecourse variation predicts error rate . . . 33
2.5 State distribution prior to trial start predicts error rates and reward effects . . . 34
2.6 State distribution predicts cue value . . . 36
2.7 Power spectral densities per state . . . 37
2.8 Distributed network parameters that differentiate states . . . 38
2.9 LFP time-frequency analysis is sensitive to different whole-trial effects 41
2.10 LFP time frequency analysis is less sensitive to the stimulus-induced effects . . . 42
2.11 Example of univariate and multivariate analysis . . . 44

3.1 Temporally Unconstrained Decoding Analysis . . . 50
3.2 Trivial example . . . 54
3.3 Sequential Hidden Markov Models . . . 56
3.4 Fitting dynamic encoding and decoding models . . . 58
3.5 Overfitting issues of an unconstrained dynamic decoding model . . 64
3.6 Results of analysis on synthetic dataset . . . 70

4.1 Feedforward model of the visual stream . . . 76
4.2 Accuracy plots on MEG data . . . 80
4.3 Decoding model generalisation plots . . . 82
4.4 Decoding accuracy as function of frequency content . . . 83
4.5 Inferred encoding model parameters . . . 86
4.6 Subject level activity map comparison . . . 87

5.1 Replay outline, adapted from Ji and Wilson [147] 92
5.2 Liu et al. [31] Task Outline 100
5.3 A conceptual outline of sequenceness 102
5.4 Survey of classification methods: analysis pipeline 107
5.5 Regularisation and its effect on classifier coefficients 110
5.6 Survey of classification methods: results 112
5.7 Examining replay over the visual hierarchy 116
5.8 Replay is activated in top-down sweeps 117
5.9 A cascade model of the visual stream response 121
5.10 Main ERF features 123
5.11 Simulated results show temporal specificity of sparse decoders . . . 125

6.1 Anatomy of the Default Mode Network 131
6.2 Alpha temporal coding 134
6.3 States are hierarchically organised in time 139
6.4 State temporal statistics 142
6.5 State spectral features 144
6.6 State 7 represents the anterior DMN 146
6.7 State 8 represents the posterior DMN 147
6.8 Replay onset is predicted by resting state network activity 149
6.9 State distribution relative to functional localiser images (FLIs) . . . 150
6.10 Replay temporal statistics 151
6.11 Replay burst evoked response 153

F.1 Mean LFP response to each cue 175
F.2 Channel-by-channel time frequency analysis: channels 1-7 177
F.3 Channel-by-channel time frequency analysis: channels 1-7 178
F.4 State timecourse Principal Component Analysis 179

G.1 Decoding accuracy of alternative classifiers 181
G.2 Decoding accuracy as function of frequency content 182

H.1 Sequenceness results computed between each pair of HMM classifier states 184

I.1 State 1 Profile 186
I.2 State 2 Profile 187
I.3 State 3 Profile 188
I.4 State 4 Profile 189
I.5 State 5 Profile 190
I.6 State 6 Profile 191
I.7 Replay Evoked Response for each HMM classifier state 192

List of Abbreviations

ACC . . . . . Anterior cingulate cortex

ARD . . . . . Automatic relevance determination

BOLD . . . . . Blood oxygenation level dependent

DLPFC . . . . Dorsolateral prefrontal cortex

DAN . . . . . Dorsal Attention Network

DMN . . . . . Default Mode Network

EEG . . . . . Electroencephalography

ERF . . . . . Evoked Response Field

FFT . . . . . Fast Fourier Transform

FIR . . . . . Finite impulse response

FMRIB . . . . Oxford Centre for Functional Magnetic Resonance Imaging of the Brain

FSL . . . . . FMRIB Software Library

FLI . . . . . Functional localisation imaging

HMM . . . . . Hidden Markov Model

ICA . . . . . Independent Component Analysis

IIR . . . . . Infinite impulse response

LDA . . . . . Linear Discriminant Analysis

LFP . . . . . Local field potential

MEG . . . . . Magnetoencephalography

MRI . . . . . Magnetic Resonance Imaging

NNMF . . . . Non-negative matrix factorisation

OHBA . . . . . Oxford Centre for Human Brain Activity

OFC . . . . . Orbitofrontal cortex

OSL . . . . . OHBA Software Library

PCA . . . . . Principal component analysis

PCC . . . . . Posterior cingulate cortex

PFC . . . . . Prefrontal cortex

RSN . . . . . Resting state network

SNR . . . . . Signal to Noise Ratio

SPM . . . . . Statistical Parametric Mapping

TIDE . . . . . Time Delay Embedded

TUDA . . . . . Temporally unconstrained decoding analysis

VMPFC . . . Ventromedial prefrontal cortex

Chapter 1

Introduction

Contents

1.1 Motivation 1
1.2 Experimental Task Design and Analysis in Neuroscience 4
1.3 Statistical models of neural dynamics 7
1.4 Towards Dynamic Modelling of Task Data 12
1.5 Structure of this thesis 14

1.1 Motivation

The brain possesses a remarkable capacity to generate sophisticated patterns of activity in response to the diverse scenarios that one may encounter. To successfully navigate our way through life requires constantly integrating immediate sensory input with long term knowledge, recent items stored in working memory, inferred aspects of social context and a multitude of other factors accumulated from long before the immediate moment. The sensory input to the brain is practically identical when an image of a venomous snake is projected onto a cinema screen as when a potentially lethal threat is slithering towards you; the appropriate response in these two scenarios is entirely different. In fact, the neural architectures that enable this diversity of responses also ensure through stochasticity that no two responses by the same person to the same stimulus will ever be identical.

Rather than being a reactive system that can be characterised by simple input to output mappings, modern unifying theories of brain function outline a highly

dynamic system that actively generates coherent activity to form predictions [1, 2], complete patterns from incomplete information [3, 4], filter attention to the most relevant and salient features [5–7], and modulate activity based on contextual cues [8, 9]. Anyone who has ever tried to voluntarily cease all thought will appreciate the spontaneous nature of activity patterns in the mind. The inherent subjectivity of such thought experiments mostly prohibits their use in science, but nonetheless give a qualitative nature to a phenomena that can be observed at multiple levels of the brain. Spontaneous patterns of activity have been documented at all scales of neural data recordings, from widespread activations in human resting state studies [10, 11], mesoscale spontaneous activity in animal field potential recordings, [12] down to the level of individual cells [13].

These models support an understanding of neural systems as highly dynamic environments, where spontaneously generated activity seems the governing state of play that is then modulated by the incoming signals induced by some stimulus. Herein lies a central challenge to neuroscience research; while experiments universally seek to disentangle the driving factors behind a neural response to a given stimulus, inherent variations require sampling of copious quantities (from the perspective of any experimental subject) of trials in the hope that these variables will average out to reveal the specific effects the researcher is questioning. Anyone who has participated in a human research scan and experienced the variations in attention, motivation and arousal between the first trial and the two hundredth trial will appreciate the irony that this very repetition is necessary to average over these and numerous other unknown variables. While this has been a productive approach that has enabled enormous progress in the understanding of the brain, more may be understood by an integrated approach that identifies such variations and separates them from the direct effects of the variables under investigation.

It is worth at this point separating some very broad 'branches' of neuroscience to provide the context and motivation for this thesis. One could separate the field into behavioural, observational and modelling studies. This grouping is by no means intended to be comprehensive, nor suggest any of these fields are mutually exclusive,

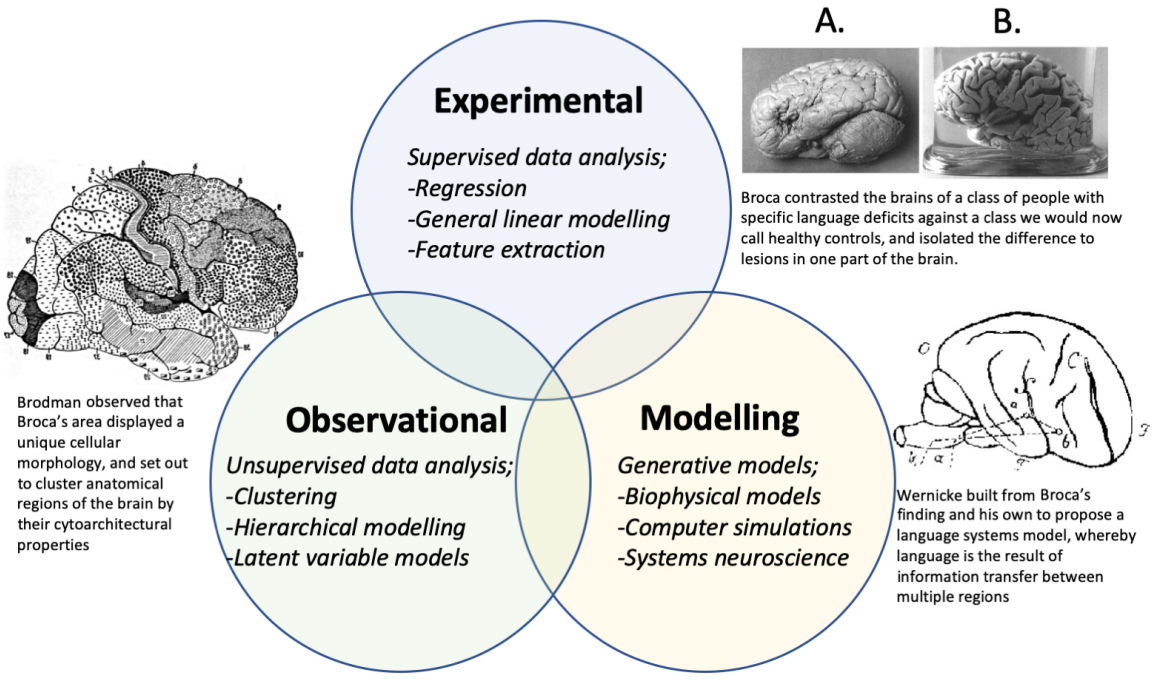


Figure 1.1: High level overview of the branches of neuroscience; three broad approaches to neuroscientific investigation and some of their associated methods. While each approach has particular methodologies, neuroscience benefits from the dialogue between levels, as demonstrated by the evolution from Broca's experimental finding to Wernicke's integrative model of language production, to Brodmann's classification of brain areas.

but rather to emphasise traditional differences in the approach to neuroscience that are relevant to our discussion here. Experimental or behavioural studies ask specific questions that can be experimentally controlled, seeking validation by observing how data varies with the manipulated variables. On the other hand, observational or passive studies identify how data varies intrinsically, looking for correlations between different observations and identifying meaningful patterns. We separate this again from the field of theoretical modelling, where specific mathematical relationships are proposed between different quantities in a manner that can be simulated or tested against real observations. While there is constant dialogue between these levels, we feel it is relevant that the tools and methodologies used for investigation are mostly different, with only limited degrees of overlap.

To loosely base this in a classic textbook example, consider the pioneering work of Broca, Wernicke and Brodmann outlined in Figure 1.1. Each of these is an early

example of one or other of these approaches, using slightly different techniques as were available at the time. Their methods were different, but in each case complementary and informed by the others' work; Broca's specific analysis was incorporated into Wernicke's model of area specialisation, which in turn motivated Brodmann to look for evidence of other regional differentiation across the cortex. To frame this more generally, the dialogue between these three branches is essential to advancing our overall understanding of the brain.

This thesis represents an attempt to reconcile our understanding of brain dynamics into our approach to analysing experimental task data. In this regard, it sits somewhere at the intersection of the observational and experimental branches discussed above. We wish to address the question of how spontaneous variations in processing affect our responses in experiments, for which we use many tools developed for unsupervised, observational analysis of data. From a methodological standpoint, we develop novel analysis techniques that build upon successful implementation in resting state analysis and apply them to supervised experimental designs. From a scientific standpoint, this allows us to link certain physiologically understood phenomena with experimentally induced conditions, providing novel insights into how brain dynamics support and explain behaviour.

1.2 Experimental Task Design and Analysis in Neuroscience

Neuroscience historically emerged as a discipline when the study of the brain and behaviour moved from philosophical enquiry to scientific methodology. More specifically, this refers to the controlled comparison of an observed quantity (the *dependent variable(s)*) under the experimental manipulation of at least one other measurable quantity (the *independent variable(s)*). Experimentalists look for statistically significant correlations between the dependent and independent variables to confirm or reject their hypotheses. The main challenges to good experimental design in this regard amounts to controlling variables that may correlate with the processes under investigation (referred to as confounds) while ensuring sufficient

subjects and trials are sampled for any relationship to be significant over the effects of other uncorrelated variables [14]. These latter variables are often referred to as noise, however are sometimes also called extraneous variables to emphasise that they frequently contain systematic variations that could potentially be characterised and understood.

To illustrate this with an example that serves to highlight both the steps involved in experimental design and subsequent analysis procedure, consider the design of a classic fMRI experiment to analyse working memory [15]. In this task, the haemodynamic response in particular regions of the brain was recorded while participants viewed a series of three faces, each presented for two seconds in a different location on a screen followed by a nine second memory delay, at which point the participants were asked to recall either the spatial location or identity of one of the stimuli. To separate the effect of working memory from the effects of motor preparation (as participants made their selection by pushing a button after the delay period), a control sequence was added where scrambled faces were shown and participants were required to push a button. As in figure 1.2, this resulted in a set of six independent variables modelled as influencing the BOLD response recorded. The timings of the task were designed so that all regressors were orthogonal; that is, they have zero correlation, allowing estimation of independent effects of each variable [16]. Finally, to reach statistical significance over extraneous variables such as physiological and measurement noise and individual anatomical deviations, the authors sampled a large number of trials from eleven different participants. This allowed the authors to identify a region localised to the superior frontal sulcus that is selectively responsible for spatial working memory [15].

This study serves as a classic example of experimental task design and data analysis. However there are assumptions inherent in this modelling approach that would not be suitable to many other forms of neural data. Task responses to the same stimulus will always vary over trials, perhaps in the precise timing of the response or maybe as a function of some extraneous variables such as concentration or attention. In fMRI, the slow temporal resolution and relatively course nature of the BOLD

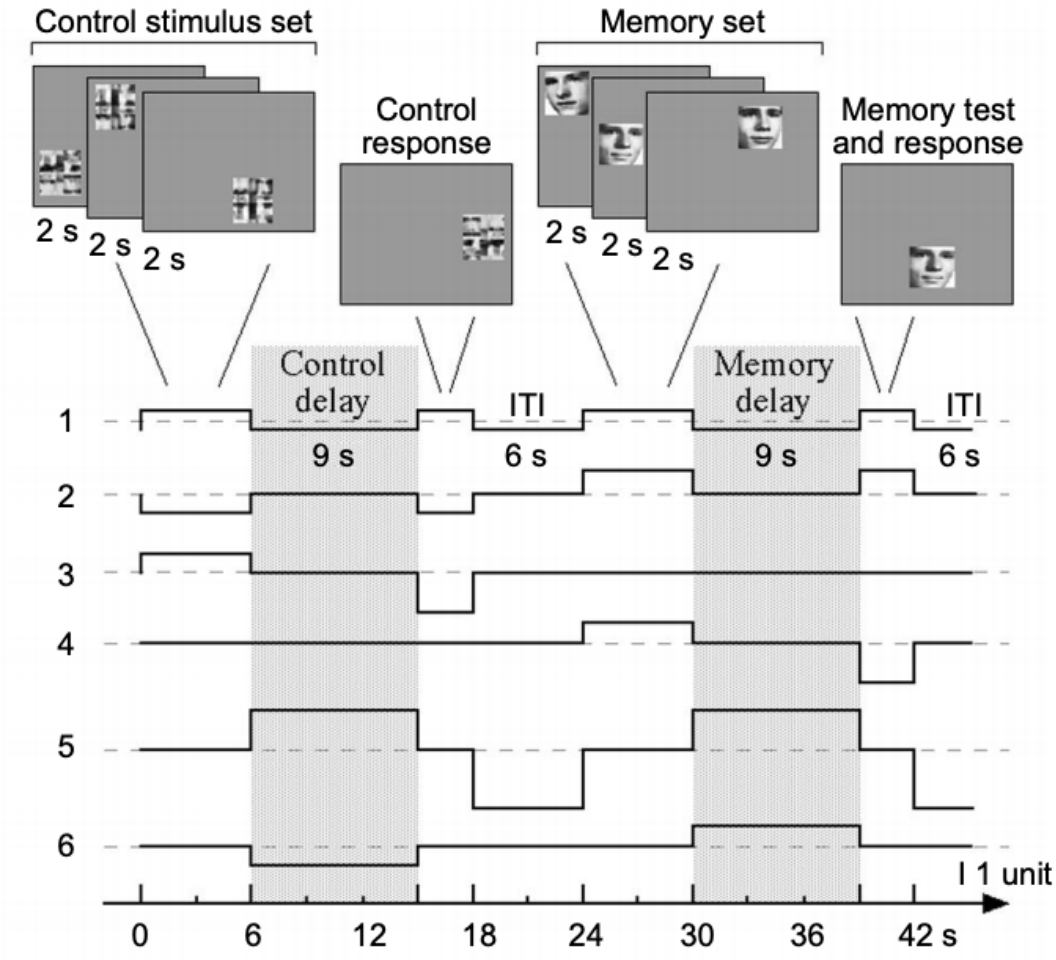


Figure 1.2: Experimental block task design in fMRI; reproduced from Courtney et al. [15]. Task designs arrange each trial into blocks where experimental conditions are varied. To adequately separate the conditions under investigation from the effects of confounds, a total of six variables are modelled: (1) visual stimulation versus no visual stimulation; (2) memory stimuli versus control stimuli; (3) control stimulus set versus control response; (4) memory stimulus set versus test stimulus and response; (5) delays during anticipation of response versus inter-trial intervals; (6) memory delays versus control delays. These formed the independent variables of a multiple regression model, with the dependent variable being the measured BOLD response in each voxel.

response (which measures blood oxygenation, an indirect measure of mass neural activity in a given brain region) make it more robust to these temporal deviations over trials than other forms of data - for example, the recorded action potentials fired by a specific cell following stimulus presentation. Furthermore, the ability to characterise the temporal response of fMRI reasonably well by the haemodynamic

response function (HRF) allows experimenters to prescribe temporal patterns in stimulus block trains and then look for linear correlations with the recorded signals. No equivalent impulse response function exists for most other forms of neural data, meaning experimenters cannot impose an expected temporal response characteristic for a stimulus and instead typically need to infer model coefficients independently at each timestep following some careful temporal alignment.

Trial to trial variations in the exact timing of induced effects can be enough to reduce statistical power and obscure relationships that are consistent over trials but not perfectly temporally aligned. In the worst case, where either through insufficient sampling or unevenly distributed noise, variations over trials may in fact correlate with independent variables and potentially support spurious conclusions. But in the best case however, such variations over trials can themselves be informative. They may reflect important changes in underlying neural processes and support a fuller understanding of the physiology underlying the response. Such an approach would require integrating the experimental analysis techniques discussed above with modelling of system dynamics.

1.3 Statistical models of neural dynamics

This thesis develops statistical machine learning models as a means of inferring dynamic structure in neural task data. More specifically, when using the term dynamics we are referring to systems whose response to a given stimulus is state-dependent; where the trajectory of each trial unfolds in time as a function of both the stimulus and the present state of the system. Where most approaches to modelling dynamics differ is how the concept of a state is defined and the complexity of the model mapping from states and stimuli to responses.

Our approach is largely based on Hidden Markov Models (HMMs), a dynamic statistical framework for which we develop customised observation models. This choice lies somewhat towards the simpler end of a scale of model complexity compared to alternatives reviewed below; we justify this choice on the basis of

model interpretability, availability of tractable methods for model inference, and suitability for the research questions we wished to ask.

1.3.1 Nonlinear Dynamic Systems Modelling

The traditional approach to modelling dynamic systems is grounded in deterministic mathematics; models of differential equations that propose specific, causal relations between parameters. Given an observed variable $x(t)$ the system trajectory through time can be computed by some nonlinear function f:

$$\frac{dx(t)}{dt} = f(x(t)) \tag{1.1}$$

In neuroscience, such modelling approaches have most famously advanced our understanding of cellular level dynamics of action potentials [17], and under generalising mean field assumptions contributed to a mechanistic understanding of how neural activity is propagated at scale [18]. However the scope of these models are necessarily very limited; they tend to be derived from a bottom-up approach where, rather than being inferred from observed data, the nonlinear function f is explicitly hypothesised based on strong assumptions about real biophysical parameters and simulated in a manner that can be compared to real data recordings and refined [19]. At scale this can be both computationally and technically challenging, and in any case seeks to address a fundamentally different question to that we are interested in; one of causation, or effective connectivity, rather than functional connectivity [20, 21].

In recent years significant advances in artificial neural networks has enabled functional connectivity between observed variables to be modelled through arbitrarily complex nonlinear functions [22]. In this approach, f is modelled as a recurrent neural network whose parameters are learned through stochastic inference, capable of modelling highly complex dynamics [23]. This approach has been shown to predict trial data trajectories to within an unprecedented level of accuracy [22]. However this model accuracy comes at a cost to model interpretability, with little information that can be read-out from a black-box neural network to actually explain how dynamics are predicted. Furthermore it remains unclear whether

the success of this method can generalise to other datasets with fewer samples or other data types. Hence while these models present unprecedented detail in capturing nonlinear dynamics, their challenges to interpretability necessarily limit the scope of research questions they can pursue.

1.3.2 Linear Dynamical Systems

One approach to equation 1.1 to make it tractable to smaller datasets is to constrain f to be a linear function of x. Under this condition, the equation can be solved; indeed this encompasses a number of popular autoregressive and multivariate autoregressive models that are frequently fit to dynamic data. Reformulating this linearised form of 1.1 into the more commonly expressed discrete time, probabilistic equivalent gives us:

$$X_t = \sum_{m=1}^{M} A_m X_{t-1} + \epsilon_t \tag{1.2}$$

where parameters A_m can be determined so as to minimise the variance of the error term, which is normally modelled as $\epsilon_t \sim N(0, \sigma^2)$. These methods are more closely aligned to the approach we have taken in terms of research objectives; they amount to finding a number of parameters that characterise the main dynamic properties of the data.

In the univariate case, where the data X_t is of a single dimension, a simple autoregressive model creates a relatively low dimensional representation of the basic frequency spectra of a signal [24]. Extending this to a multivariate case where each sample X_t is a vector of arbitrary dimension allows some interesting questions to be pursued regarding the relationships between different channels as a function of time, enabling a low dimensional representation of the main frequency coherence patterns [25]. Furthermore, compared to the nonlinear dynamic models typically associated with biophysical causation, these allow the quantification of more flexible measures of directional information flow such as Granger causality that have proved very popular in neuroscience [26].

However, linear modelling of systems that are fundamentally nonlinear has obvious limitations. In particular, data that is nonstationary in nature may be

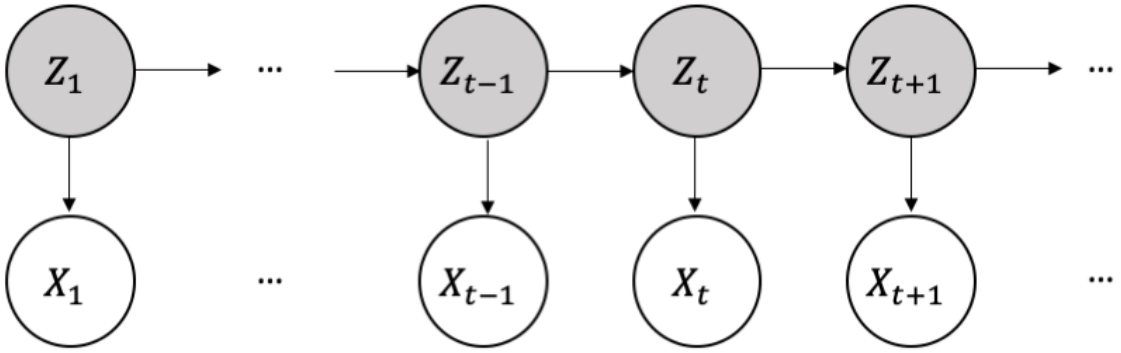


Figure 1.3: Graphical model outline of an HMM. Where previous models effectively considered the distribution of dynamic data X_t given its history $X_{1:t-1}$, the HMM introduces a latent state Z_t which is conditioned only on its value one time step before. This allows us to model X_t as conditionally independent of its history given Z_t, while the Markov assumption allows the efficient inference of the full state path $Z_1, Z_2 \dots Z_T$.

poorly approximated with linear dynamics, even if such a model is a good local approximation over some subset of the data [27]. For example, resting state data will typically show periods of heightened alpha activity, with a dominant 10Hz oscillation that activates transiently and irregularly. These individual segments may be well characterised by a linear dynamic system, however that same model may generalise poorly to the entire data recording, mixing the characteristics of the transient alpha periods with those of the remaining data. It is partly this limitation that motivates our extensive use of HMMs.

1.3.3 Hidden Markov Modelling

The methods above predict the future distribution of observed data based on its present value. These are referred to as statespace models, where the system state at any point in time t is defined by the entire vector X_t. Having a statespace of the same dimension as the original data does not aid interpretability, and the potentially quite high parameter space raises issues regarding computational tractability and model generalisation. For example, a model satisfying equation 1.2 of order M on data with N dimensions requires the fitting of MN^2 parameters; consequently most practical applications require some form of dimensionality reduction and/or regularisation.

An alternative approach is to discretise the statespace. This brings us to the HMM which will be used repeatedly throughout this thesis in a variety of contexts and so is reviewed here in greater detail. An HMM introduces a latent state variable that can only take discrete values $Z_t = [1, 2, \dots K]$, with some *observation model* explaining how the state variable at any point in time describes some data observations X_t. Critically, the observations are modelled as conditional only upon the latent state variable Z_t and the observation model parameters which are static with respect to time, making the observations conditionally independent of all previous and all future datapoints given the latent state at that point in time. All temporal variation between points is now described by state variable transitions [28].

The state variable itself is then modelled as varying over time via the first order Markov property; that is, it is conditioned only upon the value of the latent state variable in the immediately preceding timepoint. This makes it conditionally independent of all values of the latent state more than one timepoint prior, a modelling assumption that significantly reduces the computational complexity of the model by allowing us to write the joint distribution of all data observations, latent state variables and observation model parameters $\boldsymbol{\theta}$ of our graphical model of figure 1.3 as the following:

$$P(X_1, X_2, \dots X_T, Z_1, Z_2, \dots Z_T, \boldsymbol{\theta}) = P(X_1|S_1, \boldsymbol{\theta})P(Z_1)P(\boldsymbol{\theta}) \prod_{t=2}^{T} P(X_t|Z_t, \boldsymbol{\theta})P(Z_t|Z_{t-1}) \tag{1.3}$$

This then conveniently factorises into three iterative components:

•$P(X_t|Z_t, \boldsymbol{\theta})$, which is the probability distribution specified by the observation model;

•$P(Z_t|Z_{t-1})$, which describes the transition probabilities between states, modelled by a transition matrix $\boldsymbol{\Theta}$ which has entries $\Theta_{i,j} = P(Z_t = j|Z_{t-1} = i)$;

•$P(Z_1)$ and $P(\boldsymbol{\theta})$, the prior probabilities over the initial state value and the model parameters.

Model Inference

We can now turn to inference of the posterior distribution over model parameters for a sequence of data observations. This factorised posterior enables computationally efficient approximate inference - it is particularly amenable to variational methods, as efficient message parsing algorithms [28] allow the iterative computation of latent state dynamics followed by observation model parameters, converging to an approximation of the true posterior. As we utilise variational inference extensively in the ensuing work, full technical details are given in appendices A and B.

It is worth emphasising however that while these methods are computationally efficient, they do place certain bounds on the form of the observation model, dependent upon the existence of specific conjugate priors to ensure tractability of the variational approximation. The ensuing work explores several forms of observation model, some of which are directly tractable and others which require further mathematical approximations to ensure tractability. These practical limitations constrain the potential of variational methods, especially as these limitations can be overcome using recently developed methods for model inference with artificial neural networks [29]. These methods have transformed the field by enabling much greater freedom in model specification. However as a relatively recent field of development, these still face problems in implementation, with no generally agreed guidelines on network parameters required (such as what node numbers, transfer functions and depth are necessary for convergence). In contrast, variational methods are understood in some detail - both in terms of what they can and what they can't do - and their limitations have not proved restrictive over the range of HMM observation models explored in the ensuing chapters.

1.4 Towards Dynamic Modelling of Task Data

We have now outlined two basic methodologies for analysing data. The fMRI block design setup of figure 1.2 is a classic example of *supervised* analysis, whereby a direct relationship is tested between specific experimental variables and the observed data. In contrast, methods for analysing intrinsic neural dynamics or time varying

patterns of correlations are examples of *unsupervised* analysis, whereby data is clustered based on intrinsic patterns expressed over time. The focus of this thesis is on how to bridge these two fields, and develop models for supervised data analysis that explicitly model the time-varying nature of neural data in task designs.

1.4.1 Potential Advantages

Our approach primarily brings the two following benefits:

• **Improved Model Accuracy:** traditional approaches to supervised analysis impose strong assumptions of temporal synchrony over trials. When this assumption is poor, the inferred model parameters will be affected by a temporal smearing that reflects this misalignment of common activity patterns over trials. The models that we develop will explicitly infer the temporal structure of each trial, resulting in model parameters that better reflect the data and improve overall model accuracy.

• **Parameterisation of trial-specific event timings:** additionally, these models explicitly parameterise the timing of specific events or changes in activity patterns on an individual trial basis. This allows us to further investigate and quantify the relationship between the timing of these events and specific behavioural variables, such as reaction times and error rates, at the level of individual trials.

1.4.2 New Directions of Neuroscientific Enquiry

The models we develop are broadly generalisable to multiple datasets and different types of tasks, however we focus on two particular studies for which these methods allow us to establish novel lines of neuroscientific enquiry.

By using unsupervised models in chapters 2 and 6, we infer widespread patterns of distributed activity that provide an index of changing cognitive states. We can then ask specifically how these endogenous variations in brain state predict behaviour (as in chapter 2) or the onset of specific spontaneous neural events called replay (as in chapter 6).

This work then motivates the development of supervised time-varying models, introduced in chapter 3, which infer on an individual-trial basis the more subtle stages of processing of specific trial variables. This allows us to better characterise

the feedforward stages of visual ventral stream processing in MEG (chapter 4), then to explore for the first time the dynamics of spontaneous replay when these items are reactivated from memory (chapter 5).

These analyses represent a significant number of novel directions of neuroscientific enquiry, with findings that advance our understanding of specific aspects of neural dynamics. Nonetheless these research directions present just a few examples of the broad potential of the methods here developed, which we hope will find more widespread application in years to come.

1.5 Structure of this thesis

The brief review in sections 1.2 and 1.3 does not do justice to the vast field of data modelling in neuroscience, but it was not intended as a literature review so much as a high level overview of the field to contextualise and motivate the body of work that follows. Given the different types of data and lines of interrogation pursued over subsequent chapters, relevant literature will be reviewed individually at the start of each chapter. This pertains to both methodological work of relevance as well as scientific literature that informs each chapter.

1.5.1 Data sets utilised

The tools developed in this thesis are deliberately intended to be generalisable across multiple types of neural data. In line with this approach, two distinct data sets are introduced to motivate and validate the methods developed, and subsequently these datasets are analysed in detail to yield novel insights regarding neural function. We did not acquire the data ourselves in either study.

- **Dataset 1:** Extracellular electrophysiology recordings from macaque prefrontal cortex; this dataset comprises recordings from electrodes implanted intracranially in two rhesus macaque monkeys, and include isolated single cell recordings along with local field potential data from simultaneously recorded sites distributed across prefrontal cortex. Recordings were made throughout an information gathering and value comparison task, outlined fully in section 2.2 and published in Hunt

et al. [30]; all data is freely available online at the CRCNS data repository (http://crcns.org/) under dataset pfc-7.

• **Dataset 2:** human MEG recordings in a replay task; this dataset comprises MEG recordings from two distinct but related studies of spontaneous neural replay, where participants complete a task that requires them to learn and repeat sequences, outlined fully in section 5.2 and published in [31]; study one collected MEG data from 21 participants, study two from 22 participants.

Further details of each dataset will be outlined in the relevant chapters as they are more fully explored.

1.5.2 Chapter outline

Chapter 2 sets out a more comprehensive motivation for the ensuing tools that will be developed. We formally compare the most widespread method for evaluating time varying parameters in tasks (namely the use of sliding windows to estimate dynamic parameters over time and how these vary over experimental conditions) against a multivariate modelling approach previously introduced for use in MEG data [32]. This analysis is conducted on dataset 1 referred to above and demonstrates the overall motivation for HMM models in general, while suffering from certain drawbacks that motivate the development of supervised HMM observation models.

Chapter 3 introduces the novel methodology that is the primary technical contribution of this thesis to the machine learning literature. We extend the common generalised linear model (GLM[1]) framework to include a latent Markovian variable, allowing individual trials to exhibit variations in precise response timings. This extends work by Vidaurre et al. [33] to encompass encoding and decoding model variants, with a particular focus on classification tasks for which a novel logistic regression framework is implemented.

[1]In an unfortunate anomaly of method naming conventions, two ubiquitous, closely related yet distinct statistical frameworks were named the Generalised Linear Model and the General Linear Model. The widespread adoption of the acronym GLM to refer to one, the other, or both methods interchangeably is a source of persistent confusion in the field - we use GLM to refer to the former, and General LM to refer to the latter.

Chapter 4 then tests this model against an MEG dataset to investigate the activation of the ventral visual stream. We demonstrate that this model leads to improvements in decoding accuracy, with more interpretable activation maps and state timecourses that reliably correlate with behaviour on individual trials.

Chapter 5 then focusses on the study of neural replay. We show that the classifiers trained in chapter 4 can be fit to resting state data to identify spontaneous reactivations of learned items while participants are at rest. We then review the challenges of identifying sequential neural replay, and show that some of these challenges are overcome by the model introduced in chapter 4.

Finally in chapter 6 we return to the unsupervised model of neural dynamics introduced in chapter 2, using this to characterise the main resting state networks that vary alongside spontaneous replay, and demonstrate strong correlations between these networks, providing novel insights into the physiological basis of replay.

Methodological shortcomings and future work directions are then briefly recapped in chapter 7.

Chapter 2

Inferring Behavioural States from LFP Recordings in the Prefrontal Cortex

Contents

2.1 Introduction 18
2.2 Experimental Outline and Methods 22
2.3 Multivariate Results: HMM States reflect both Spontaneous and Stimulus-Induced Effects 30
2.4 Univariate Results: Time-Frequency Analysis is Less Sensitive to these Effects 40
2.5 Discussion 43
2.6 Conclusion 47

Methods for analysing LFP data are predominantly univariate and reliant upon windowed extraction of features, typically frequency band power and coherence, for specific hypothesis testing. Multivariate models have the potential to identify more subtle changes in patterns of correlations that might be distributed over multiple channels, however to date their application to LFP data has been somewhat limited. In order to explore the relative benefits and drawbacks of multivariate models over traditional analysis methods and to justify the models developed in subsequent chapters, this chapter applies both methodologies to a dataset of macaque LFP recordings and contrasts what information is reflected by each. Whereas univariate regression analyses allow specific questions to be asked about

each channel individually, multivariate models characterise the overall patterns of distributed network activity. We show that this results in a more robust encoding of specific experimental variables such as value, while also providing a measure of the endogenous behavioural state that is predictive of error rate on individual trials and likely reflects variations in cognitive attention to the task.

2.1 Introduction

Neural activity can be recorded at a range of scales, from micro level recordings of the action potentials fired by individual cells, to the macro level of whole brain recordings collected in MEG, EEG or fMRI. Local Field Potential (LFP) recordings are referred to as a mesoscopic scale measurement, as they reflect the electric potential at a localised point in the extracellular space within the cortex. Consequently this data type captures certain features reflective of micro level processes of local cells, certain features that reflect macro level processes of coordinated distributed activity, and certain features that are unique to the local potential recording at that particular site [34, 35]. This makes LFP recordings a particularly useful data type to probe questions about how these different levels interact, and a large volume of literature seeks to make sense of how cellular activity both contributes to and is influenced by larger scale oscillatory activity that is propagated over widespread regions of cortex [36–38]. The literature that links LFP traits to behaviour, however, is considerably smaller and tends to focus on specific identified phenomena such as cross frequency coupling [39–41]. Particularly in the context of emerging technologies that allow simultaneous LFP recording across multiple sites of cortex, we can also start to ask how changes in local activity may be distributed across multiple sites and, ultimately, how these variations guide behaviour.

2.1.1 LFP Physiology

One of the biggest appeals of the LFP signal is that it reflects both local and larger scale processes, and a central goal of researchers has been linking observed LFP fluctuations with known physiological processes. It is interesting then, and

speaks to the complexity of the LFP signal, that some of the biggest questions in this regard remain controversial. The apparently simple question "how local is the source of the LFP signal?" [42] is a good case in point; conflicting evidence and modelling approaches suggest a very nuanced answer that remains elusive [43–49]. This is at least partly explained by the degree of local synchronisation which can vary by cortical site and time [50, 51]. Furthermore, cell types are not uniformly distributed in space; they vary by cortical location and layer depth, in a manner that appears functionally relevant for the organisation and propagation of oscillatory activity [52] but which is not fully characterised.

These points are relevant to our discussion as they suggest caution in interpreting observed changes in LFP signals. They also explain one of the primary challenges of the data we are working with; the data has been recorded from multiple sites in monkey prefrontal cortex, but there can be no consistent interpretation from one recording session to the next. The practical limitations of this apparatus required daily reinsertion of electrodes and, on occasions, their movement to entirely new locations. As LFP recordings only a few millimetres apart can be strongly correlated or anticorrelated [50], and as certain key features such as high gamma activity may be completely absent as a function of neocortical column depth [53], care must be taken with interpreting observed results. Each recording offers a window onto multiple LFP sites, but each new session offers a view from a new and different window, with no means of referencing these to any common framework. In particular, any multivariate analysis that infers effects over multiple channels simultaneously will face challenges in scaling to other sessions, as those other sessions present an entirely new set of sampled sites. Nonetheless, a number of key LFP features have been quantified over time that can be analysed over each session individually, and are worth describing here in greater detail.

2.1.2 LFP features of interest

The transmission of information between different parts of cortex appears to be mediated by frequency band multiplexing, whereby oscillations at different

frequencies support distinct anatomical and functional forms of information transfer [37, 54, 55]. Of particular relevance for this study, which recorded from sites in prefrontal cortex (PFC) during a value processing task, is activity in the low frequency bands of delta (1-4Hz) and theta (4-8Hz) bands.

Coherence in the delta band between disparate regions of PFC has been shown to correlate with active decision making [56], visual attention [57, 58], and entrainment to relevant over irrelevant stimuli [59]. This may reflect an underlying functional significance of delta band activity whereby this frequency band inhibits information that is not relevant to the task at hand [60].

Theta band activity is most traditionally associated with the medial temporal lobe and subcortical areas; these areas have numerous projections to prefrontal cortex, and theta band coherence between prefrontal and subcortical areas has been implicated in successful memory encoding and retrieval [61, 62], temporal context awareness [63] and effective goal switching [64]. More broadly, an integrated view is emerging whereby frontal theta activity arises in situations requiring heightened cognitive control, a term used to collectively refer to executive functions such as attention, working memory, error monitoring or flexible planning that together achieve fluid goal-directed behaviour [65, 66]. On a physiological level this frequency band activity is believed to reflect widespread activation and coordination of prefrontal, parietal, medial temporal and subcortical structures, interpreted as an executive control mechanism for flexible maintenance of widespread networks [66].

One final frequency band signal of interest to our study here is quite distinct; activity in the high gamma band (greater than 50Hz) has been shown to correlate with local neuronal firing rates [67, 68]. Although the exact origins of high gamma signals are still controversial [69, 70], the prevailing view is that activity in this frequency band does not reflect regular oscillatory patterns but rather high frequency transient noise reflecting the scale of localised neuronal firing [68]. In this regard activity in this frequency band is quite distinct; it is not expected to exhibit any strong coherence with disparate areas, but is a useful indicator of aggregate heterogenous cell activation in the cortical columns adjacent to the electrode

[71]. Furthermore, whereas lower frequency oscillations appear to be indicative of information transmission between areas, high frequency power has been shown to encode specific features of stimuli presented [71–74], presumably reflecting the localised nature of encoding specific parameters of the environment.

2.1.3 LFP Analysis Methods

In order to detect changes in the above frequency bands, the vast majority of LFP analysis methods apply one form or other of univariate statistics, where desired features are tested one by one for statistical correlations with each of the variables under consideration. These mostly rely upon some formulation of the GLM framework introduced in section 1.2, with the key differences being how the variables are defined. Where the variable is just the time domain LFP signal, this is called an evoked response analysis; however as this approach is only sensitive to the LFP components that are phase-locked to the stimulus onset, they tend to be reserved for quite specific research questions [75–77].

The most common way to analyse LFP data is to fit the same GLM model to spectral features. This method will be outlined in detail below and employed to analyse the available data, but allows one to regress experimental variables against band limited power computed over sliding windows to characterise how power in specific frequency bands vary over experimental conditions [59]. While bandlimited power is the most commonly described feature, the same analysis can be equally applied to other measures such as band limited coherence [56] or phase locking index [62], however questions remain about the robust computation of such features on a trial-by-trial basis [78, 79].

The alternative approach that we wish to explore is to fit an unsupervised model of latent dynamics and then test whether the model parameters covary with experimental variables. Application of this approach to LFP data has to date been very limited; the most relevant studies applied multivariate autoregressive models over defined conditions and compared the inferred parameters [80, 81]. As discussed

in section 1.3 the inherent nonstationarity of neural data creates challenges to interpretation and robust estimation of parameters over trials.

HMM applications in particular to LFP data have been quite rare, and mostly limited to inference of quite specific physiological states like up-down transitions in slow wave sleep [82]. We propose this methodology could allow us to infer spontaneous changes in underlying brain activity over trials that may represent critical fluctuations in cognitive engagement, attention and numerous other endogenous variables of cognition and behaviour. This could allow us to explore and characterise more fully how such spontaneous fluctuations of latent brain state interact with specific responses to experimentally controlled variables.

2.2 Experimental Outline and Methods

2.2.1 Experimental Outline

The experiment has been outlined in detail in Hunt et al. [30], but is repeated here for clarity.

Task structure

Two adult male rhesus monkeys (Maccaca Mulata) were trained to complete a task evaluating two competing options for a potential reward. Each option, presented on the left and right hand sides of the screen respectively, comprised two prelearned picture cues that denoted either the probability of receiving reward or the magnitude of reward that would be received, should the monkey select that option. At the start of each trial, the cues were hidden; one of the cues at random would then be highlighted, and revealed after a saccade to that position on the screen. After 300msec of fixation, that cue would be covered and a second cue would be highlighted, to then be revealed following a saccade to that location by the monkey. The second cue was always either horizontally or vertically opposing the first cue, but not diagonally opposite, ensuring that at this point the monkey had complete information about *either* one option (i.e. all information pertaining to the left or right option) or one attribute (i.e. information from both sides pertaining to one

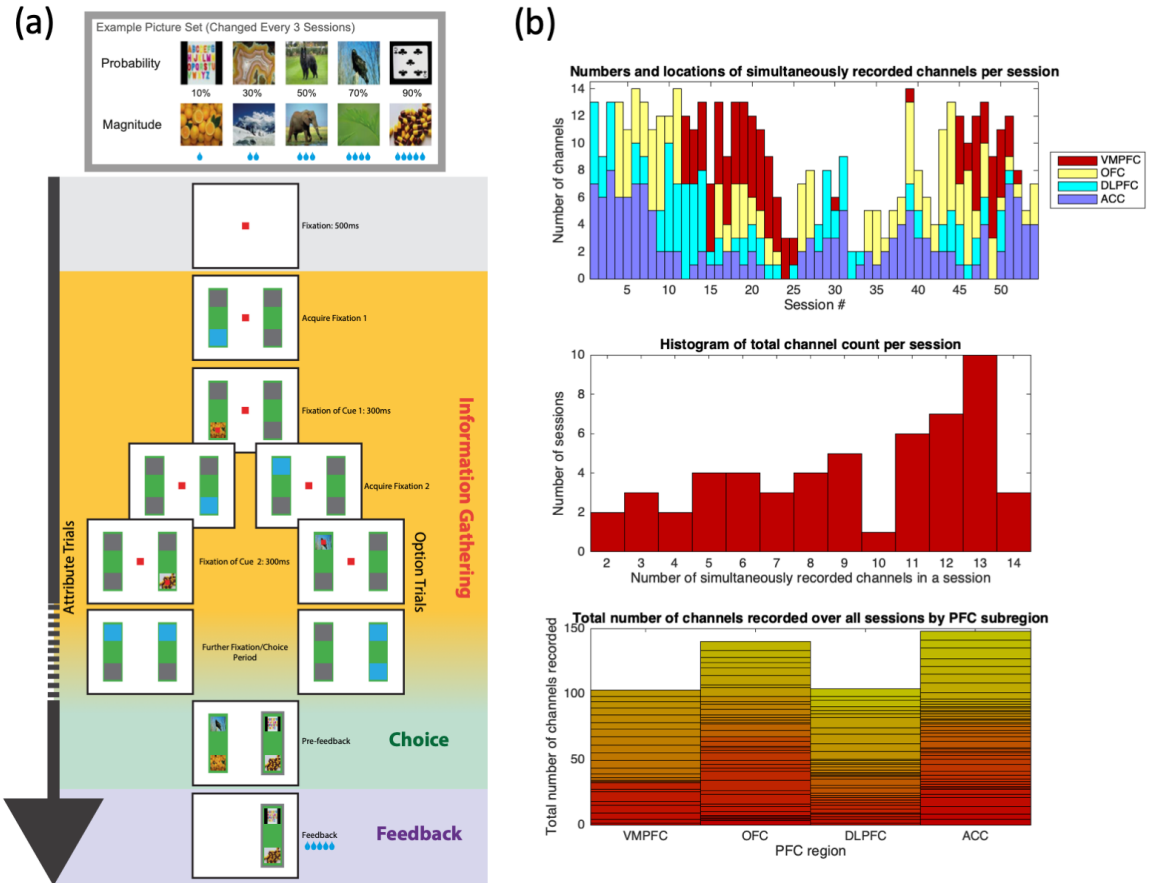


Figure 2.1: Experimental task outline; reproduced with permission from Hunt et al. [30]. (a) Top panel: the ten stimulus images and their associated task parameters; below, the structure of each trial, from initial fixation through successive cues to choice and feedback periods. (b) Heterogeneity of data recorded over sessions; the distribution of the quantity and location of recorded LFP channels within prefrontal cortex over different sessions is highly diverse, limiting the generalisability of multivariate methods.

of either the magnitude or probability of reward). From this point in each trial, the monkey could decide whether to continue exploring the remaining cues (which would each be revealed following a saccade to the location), or could select the left or right option by moving a joystick to the desired side. Following joystick movement, all cues were revealed, and a juice reward was administered according to the selected probability and magnitude. Note that the picture cues for each option and the cues assigned to be presented first and second were pseudorandomly assigned on each trial from a uniform distribution.

Data recordings

During this task, recordings were made of extracellular potentials in prefrontal cortex using implanted electrodes that were lowered into position each day. Activity was recorded from anterior cingulate cortex (ACC), dorsolateral prefrontal cortex (DLPFC), orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (VMPFC). ACC recordings were primarily from the dorsal bank of the cingulate sulcus; OFC recordings primarily from the medial orbital gyrus; DLPFC recordings from the dorsal and ventral banks of the sulcus principalis. Data was recorded at 40kHz using a Plexon Omnibus system; crucially as the original authors were primarily interested in single unit recordings, these were isolated and any channels where single units were not isolated were discarded. Furthermore the recording electrodes were removed at the end of each session's recording and lowered into position again at the start of the following session.

Analysis of spontaneous dynamics requires simultaneous recordings. In this case, recordings were made over multiple days, with each individual session recording from up to a maximum of 14 channels (see figure 2.1), with no consistent interpretation possible between the electrodes. In order to fully explore the role of simultaneous dynamics and the potential benefits of multivariate methods, in the ensuing chapter we focus solely on a single session that had the equal highest number of utilisable channels. There are obvious challenges to replicability and interpretation of the remaining data which will be discussed in the conclusion to this chapter.

Data Preprocessing

50Hz line noise was removed with an IIR notch filter of order 20 and bandwidth 1Hz, with notch centre frequencies at 50Hz and harmonic frequencies at 100Hz and 150Hz. LFP data was then downsampled to 250 samples/second for computational expediency.

2.2.2 HMM Time Delay Embedded (TIDE) Model

We wish to fit an unsupervised model that can identify the main patterns of spectral power and coherence between LFP recording sites, as we believe these properties may reflect key changes in cognitive state. For this purpose we now introduce a multivariate model previously used for analysis of MEG data [32]. It is worth emphasising that this model will be used again in chapter 6 for analysis of resting state MEG data.

Temporal Embedding

This model was designed to capture both spectral power features of the data and cross channel spectral coherence, and characterise how these change over different trials. These features have an equivalent formulation in the time domain, with spectral power being described by a signal's autocorrelation while the cross channel coherence is captured by cross correlation between channels. For this reason the model employs a temporal embedding, so each state Z_t defines the distribution of data across a temporal embedding of length W around time t, as in figure 2.2.

Observation Model

We can then completely specify the HMM observation model in probabilistic terms as follows:

$$P(X_{t-\tau:t+\tau}|Z_t = z) \sim N(0, \Sigma_z) \tag{2.1}$$

Under this model, the data is described by a zero-mean multivariate Gaussian observation model of dimension PW, with P denoting the dimensionality of the data X_t and $W = 2\tau + 1$ denoting the length of the embedding. Each latent state Z_t is then associated with a distinct covariance matrix Σ_z, effectively capturing different patterns in both the autocorrelation of data on each channel (reflected by the diagonal blocks of the covariance matrix) and the cross correlation between different channels (reflected by the off-diagonal blocks of the covariance matrix).

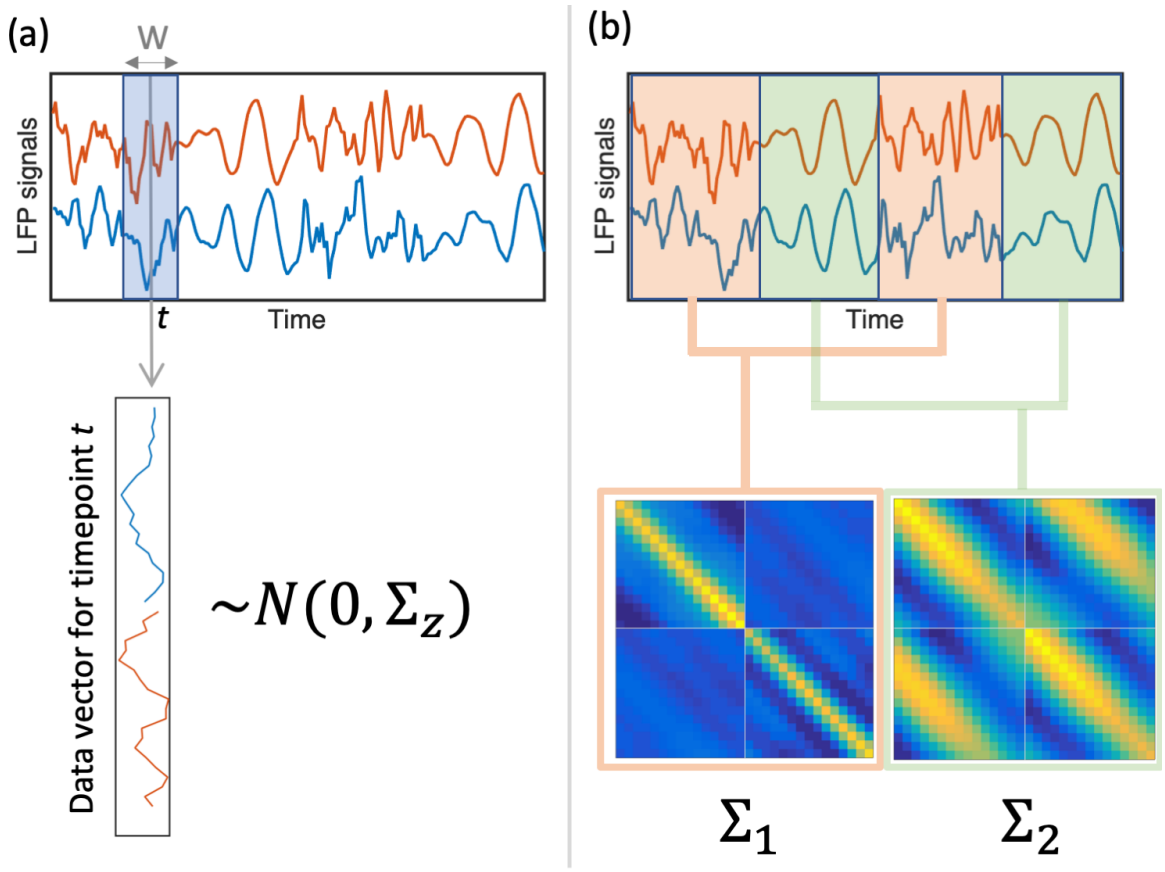


Figure 2.2: Time Delay Embedded model outline. (a) Temporal Embedding: for a given data point at time t, we extract a temporal embedding of total width W around the timepoint and stack the data from all channels into a vector. Our observation model is then specified over this vector as a zero mean, full covariance Gaussian distribution. (b) We then infer the latent state timecourse, with shading in red and green reflecting the active states vs time; and their respective covariance matrices Σ_1 and Σ_2. Due to the temporal embedding, the blocks on the diagonal of this covariance matrix capture patterns in autocorrelative structure on each channel, while the off-diagonal blocks capture cross-correlation patterns between pairs of channels. This data is summarised by two states; state 1 captures periods where the two channels are uncorrelated, whereas in state 2 the two channels display much stronger time-lagged correlation.

Applied Example

The basic features of this model can be explored with a simple simulated data example, shown in figure 2.2. We simulate structured data in two dimensions using the linear dynamic system model of equation 1.2, but switching the generative model parameters through the trial to simulate two distinct states. Data generated in the

first state exhibits some autocorrelation but no cross correlation between channels, while the second state exhibits closer coupling between channels at a specific frequency band. Fitting the time delay embedded model effectively captures the two states and the timings of their switch. The distinct spatiotemporal correlation patterns that define each state are evident in the covariance matrices, with the diagonal blocks reflecting the autocorrelation patterns as a function of timelag, whilst the off-diagonal blocks reflect the cross correlation between channels at distinct timelags. Note that these time-domain parameters can be equivalently converted to frequency domain representations, for example by the use of multitaper methods [83], which we make extensive use of due to the interpretability of activity in canonical frequency bands.

Selected Model Parameters

We fit a $K = 8$ state model with a temporal embedding length of $W = 10$ samples to the $P = 14$ channel data to identify the state sequence through time on each individual trial, the associated state parameters that describe the spatiotemporal features of each state, and how these correlate with deviations in trial variables.

2.2.3 Time-Frequency Analysis

In order to contrast univariate and multivariate analysis methods, we then conduct a sliding window time-frequency analysis to characterise the results this method obtains. Using a sliding window of length 200msec, the band-limited power was computed in 1Hz width bins from 1Hz up to 125Hz (the Nyquist frequency) by Fast Fourier Transform (FFT). Data from each trial were z-scored to focus on relative changes in the power spectrum. Channels within the same area were then averaged to create a single time-frequency reading per PFC subregion. Epochs were then extracted from 200msec prior to each cue presentation to 600msec after. The mean of each epoch is displayed in appendix F.

2.2.4 Regression Framework

We will analyse how the parameters of the TIDE model covary with experimental conditions, and then contrast this with how the time-frequency spectral features vary over experimental conditions. In each case this is analysed using a regression framework, where we iteratively compute regression coefficients for each timepoint within trial t as follows:

$$X_t = Y_t\beta_t + \epsilon_t \tag{2.2}$$

where β_t are the coefficients describing the overall effect of each regressor and $\epsilon_t \sim N(0, \sigma^2)$. Rows of X_t and Y_t correspond to trials. Y_t is the design matrix of regressors, with dimension $[N \times Q]$ where N is the number of trials and Q the number of experimental variables. X_t is the dependent variable, of dimension $[N \times P]$, which we vary over the two methods being contrasted: in the multivariate case X_t will be the active state variable vs time, whereas in the time-frequency analysis X_t will be the windowed power spectral density per frequency band per region per timepoint.

More specifically, we specify $Q = 9$ independent variables as the columns of our design matrix Y_t:

Column	Variable	Description
Y_1	Left vs Right	1 if Left, 0 if Right
Y_2	Probability vs Magnitude	1 if probability cue, 0 if magnitude
Y_3	Value	1 - 5 denoting the value of the cue
Y_4	Interaction Left / Prob	Y_1Y_2
Y_5	Interaction Left / Value	Y_1Y_3
Y_6	Interaction Prob / Value	Y_2Y_3
Y_7	Error rate	1 if trial resulted in an error, 0 otherwise
Y_8	Trial number	1 - T denoting trial number in the session
Y_9	Last trial rewarded	1 if previous trial was rewarded, 0 otherwise

The first three of these pertain to the main characteristics of the cue: Left vs Right, Probability vs magnitude, and value on a scale from 1-5; a further three pertaining to the interactions between each combination of these variables; and a further three that relate to the longer term trial dynamics. These last variables comprise one binary variable reflecting whether the ultimate decision on that trial

is a rational choice (defined as a selection which maximises the expected reward), to see if the inferred features are predictive of error; the trial number within the session, to regress out any progressive variations over the course of the session; and one final variable to indicate whether the last trial was rewarded, to separate any residual effect of a recent reward on the LFP signal. In the interests of brevity, we only present results pertaining to the first cue presented on each trial; subsequent cues were analysed and revealed largely similar information.

Regression Framework for Whole Trial Dynamics

One variation on this analysis is employed to characterise whole-trial effects. In order to investigate the macro-level deviations in multivariate state dynamics on the level of whole trials, we use Principal Component Analysis (PCA) to identify the main modes of variation in state dynamics over trials, then investigate how these predict the whole-trial variables of Y. Specifically, we take the probability of each state's activation within each trial and stack them into a matrix Γ of dimension $[N \times KT]$ (where N is the number of trials, K the number of states and T the trial length), and infer the PCA decomposition

$$\Gamma = AB \tag{2.3}$$

where B of dimension $[N \times KT]$ captures the modes that maximally explain the variance in the data and A of dimension $[N \times N]$ denotes their weight of expression on each trial.

We then compute how the first column of A, the coefficient weighting of the main mode of variance, predicts whole trial variables; specifically the three last columns of Y:

$$Y_i = g(A_1\beta_i) \tag{2.4}$$

for $i = 7, 8, 9$. Given the categorical distribution of variables Y_7 and Y_9, we use a logistic regression link function for g in these cases; for the discrete trial number Y_8 we use a Gaussian link function.

2.3 Multivariate Results: HMM States reflect both Spontaneous and Stimulus-Induced Effects

Applying the TIDE model to the raw LFP data gives us a set of inferred states over time, which we hypothesise reflect underlying changes in cortical network activity that may ultimately correlate with key features of the trials. We test this at two levels; firstly we investigate whether the macro patterns of state transitions over whole trials are predictive of behaviour within trials; we then investigate whether state changes reflect the parameters of specific cues.

2.3.1 Main Mode of Variation in State Timecourses Predicts Error Rates

Multivariate models can infer sophisticated patterns in neural data, but for these patterns to be relevant to a task they should ideally correlate with behaviour. In this section we set out to characterise the dominant state dynamics over trials and show that their main mode of variation is predictive of error rates, providing strong evidence that they reliably index some underlying cognitive state.

Trial Dynamics

Figure 2.3 plots the maximum likelihood state over trials as a raster plot; the first three principal components over these state trajectories (i.e. the first three rows of B from equation 2.3) and the corresponding coefficients (first three columns of A from equation 2.3). Note that the raster plot of figure 2.3 has sorted trials along the y axis based on the PCA scores, to enable easier visualisation of different trajectories.

The first principal component reflects the overall mean state distribution as a function of time. This highlights that different states are clearly associated with distinct periods within a trial; the mean distribution is not uniform with respect to time, but displays clear alignment of particular states to the main trial epochs. Notably the average onset of the first cue is 0.5 seconds into the recording, and the average time of decision is 3 seconds into the trial; these times clearly align with

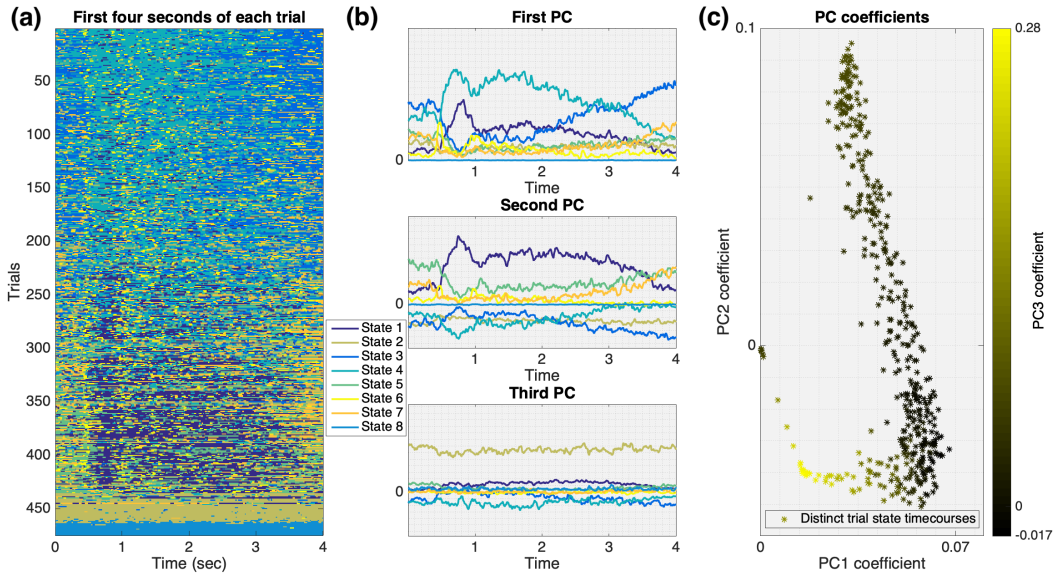


Figure 2.3: Principal Component Analysis isolates two main trial trajectories.
(a) Plotting the inferred state timecourses over all trials in a rasterplot, where the y axis is different trials and the x axis is time within a trial, the colours denote the active state at any point in time. At least two macro patterns are evident, with one cluster towards the top of the plot and the other emerging lower down. Note that to enable visualisation, trials on the y axis are not ordered consecutively but sorted by their PCA coefficients.
(b) The first three principal components describing variations in inferred state timecourses over trials. The first and second PCs capture two distinct trajectories over trials, while the third PC explains another subset that turns out to be outlying trials with high noise.
(c) Plotting the coefficients of the first three PCs (with the colour denoting the third PC coefficient) demonstrates they are highly anticorrelated. The strength of anticorrelation between PC1 and PC2 coefficients motivate our use of the first coefficient as a feature to predict whole trial experimental variables.

transitions in the mean state distribution, confirming that inferred state transitions reflect changes in activity induced by distinct trial epochs.

It is upon analysing the second principal component however that two distinct, whole-of-trial trajectories emerge. One trajectory involves transitions between states 1, 5, and 7 (respectively in blue, green and brown in the principal component plot; trials expressing these dynamics occupy rows 250-450 of the rasterplot), while an anticorrelated trajectory follows similar timings but substitutes these states with 3 and 4 (orange and purple respectively, dominating trials on rows 1-200 of the rasterplot). Figure 2.3c plots the expression strength of each of these

components on each trial, demonstrating that they are not independent but very strongly anticorrelated - hence representing a single axis of variation between two opposing overall trial trajectories. Note that, as state timecourses are mutually exclusive, a degree of anticorrelation should be expected, but the strength of this pattern is nonetheless striking.

Rejection of Outlying Trials and Artefacts

One particular state is activated immediately prior to each cue; later spectral analysis confirms that this state reliably separates saccade motion artefacts when they occur, both aligned to cue onset and free saccades during the trial. The third PC explains a small subset of outlier trials described by state 2; similarly the fourth PC (see appendix F) explains a smaller subset of trials described by state 8. These trials occupy rows 450 to 480 in the rasterplot, and comprise the string of points nearest the origin in figure 2.3c, supporting the notion that they are a subspace of trajectories mostly distinct from those described by PCs 1 and 2. Later analysis of frequency domain features of these states suggests a much higher broadband power that is likely to be some kind of artefact; we will focus our subsequent discussion on the remaining states and the two trajectories identified above.

State Trajectories predict behaviour

Figure 2.4 plots the results of the regression analysis outlined in equation 2.4. Both the error rate ($p = 1.3 \times 10^{-4}$) and the trial number ($p = 7.2 \times 10^{-3}$) were significantly predicted by a model based solely on the coefficient of PC1; that is, the main mode of variation between state time courses on individual trials. This confirms that these distinct trial trajectories are reflective of meaningful neural processes, with the dominant state trajectory associated with lower error rates while being more concentrated on trials towards the start of the session. No significant relationship was observed between trial number and error rate; nor was any significant relationship observed between PC1 and the last trial rewarded.

The relationship with the error rate suggests a strong behavioural link for the inferred state trajectories, which we argue reflect spontaneous variations in higher

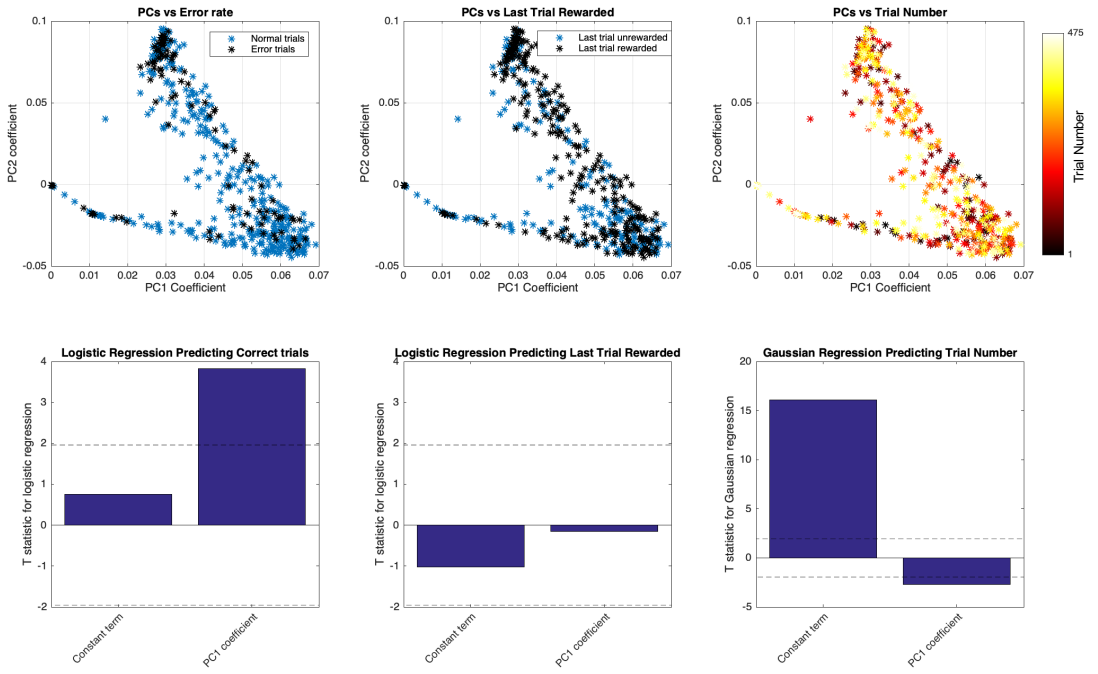


Figure 2.4: Main mode of state timecourse variation predicts error rate. Error trials can be strongly predicted by logistic regression using only the first principal component coefficient (left). This is taken to as preliminary evidence that this mode of variation reflects endogenous changes in cognitive state of relevance to the task. The same coefficient does not have any relation to the recent presence of reward (centre); a weaker relationship exists between this coefficient and the trial number within session (right), which may reflect slower scale fluctuations in cognitive engagement over the course of the session.

order cognitive processes associated with successful task completion such as attention, cognitive control, or motivation. The fact that this relationship emerges immediately from the first principal component reinforces the view that these features are robustly encoded within the LFP when analysed simultaneously across multiple sites.

2.3.2 HMM States reflect Stimulus Parameters

The original aim of the experiment was to understand how specific stimulus variables, outlined in the design matrix Y, were reflected in the LFP signal. Having identified meaningful correlations between state variables and trial parameters at the whole trial level, we now turn our attention to individual epochs within trials, to show firstly that the correlations identified above generalise to individual trial epochs; and then to show that state variations further predict the parameters of individual cues.

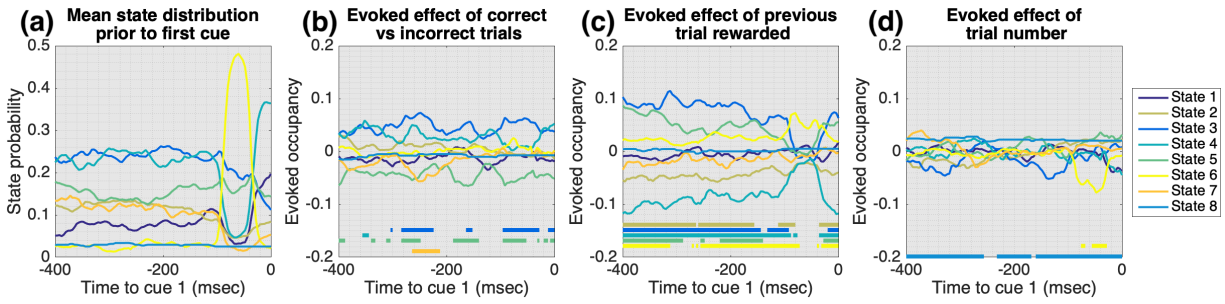


Figure 2.5: State distribution prior to trial start predicts error rates and reward effects. Plots show the evoked effect of each variable on each state in the 400msec prior to cue 1 presentation; significance bars at base of plots mark points significant to $p < 0.005$. (a) The mean state distribution prior to the first cue onset has several states that may be active vs time, reflecting the PCA-based analysis conducted previously. (b) Significant correlations were found between the activation of states 3 and 4; an anticorrelation was found with activation of states 5 and 7. This shows the state distribution before the trial has begun is predictive of the ultimate outcome of the trial as correct or error. (c) Very strong correlations were found between the state distribution prior to cue onset and the receipt of a reward at the end of the previous trial, confirming that this leaves a strong residual effect in the network activity. (d) Effects relative to trial number did not predict activity in any of the states of interest (state 8 explains a cluster of particularly noisy trials that occured towards the end of the session).

State Distribution Prior to First Cue Predicts Error Rate

One possible confound in the above analysis is that error rates may derive from the actual parameters of successive cues (i.e. errors may be more likely on trials where all cues are low in value, or are similarly distributed); and that the main mode of variation over trials may not reflect the endogenous cognitive state we propose but just an aggregate value correlate over successive cues. To eliminate this possibility, we isolate data from the 400msec period prior to the presentation of the first cue and analyse its relationship with the error rate; this is after the trial has been initiated but prior to any relevant information being revealed.

Fitting the regression model of equation 2.2 to the HMM state distribution over this this epoch identifies significant correlations between the active state and the likelihood that trial will result in an error; in line with the above analysis of overall trial trajectories, increased activation of states 3 or 4 (linked to the main trial trajectory) prior to the first cue are associated with correct trials, whereas

activation of states 1, 5, or 7 (linked to the second state trajectory) are associated with error trials. Points are marked as significant at p=0.005 levels.

Additionally, a strong relationship is also expressed between the state distribution and the recent receipt of reward. Activation of states 3 or 5 reflects residual reward effects from a previous trial; states 4 and 7 with the absence of recent reward. Alongside the effect of error rate, we can observe a clear dissociation of the trials prior to the very first cue presentation as follows:

• State 3 is associated with trials in trajectory one (that reflect a behavioural state linked to correct trial performance) and the residual effect of recent reward;

• State 4 is associated with trials in trajectory one (that reflect a behavioural state linked to correct trial performance) where there is no residual effect of recent reward;

• State 5 is associated with trials in trajectory two (that reflect a behavioural state linked to poorer trial performance) and the residual effect of recent reward;

• State 7 is associated with trials in trajectory two (that reflect a behavioural state linked to poorer trial performance) where there is no residual effect of recent reward.

State Distribution Following Cue Presentation Reflects Value Processing

The regression model of equation 2.2 allows us to furthermore quantify the effect of different cue parameters on the state distribution immediately after cue onset, as in figure 2.6. Little effect is observed as a result of the presented cue being of probability or magnitude in nature, nor by its positioning on the left or right hand side, however high value cues are strongly correlated with a transition into state 4 ($p = 10^{-11}$) and to a lesser extent (at later timepoints) state 3 ($p = 10^{-4}$); low value cues strongly correlate with a switch into state 1 ($p = 10^{-16}$).

2.3.3 Characterising the features that define behavioural states

Having established a strong link between the active state and the experimental variables, we now turn to analysing the parameters that define each state, in order to understand what specific patterns in the LFP signal they reflect. Interpretation

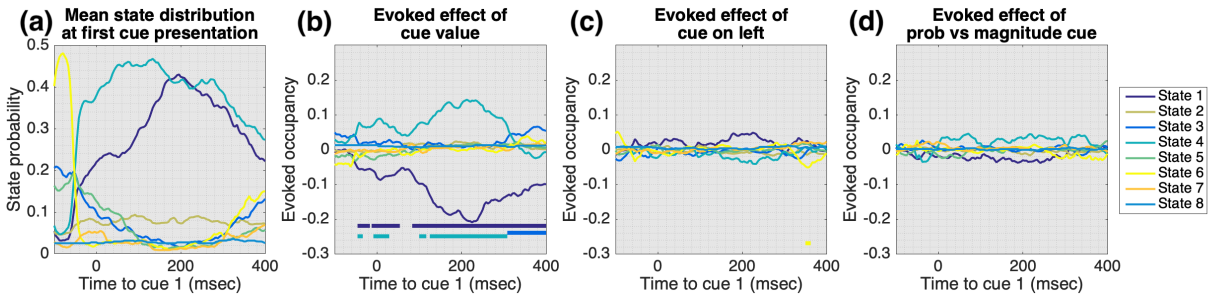


Figure 2.6: State distribution predicts cue value. Plots show the evoked effect of each variable on the state distribution following cue 1 presentation; significance bars at base of plots mark points significant to $p < 0.005$. (a) Mean distribution shows states 1 and 4 dominate processing following cue onset. (b) These states are very significantly associated with the cue value ($p = 10^{-16}$ for state 1, $p = 10^{-11}$ for state 4). (c) No significant correlation is observed between state distribution and left / right presentation; nor (d) whether the cue denotes probability or magnitude information.

of multivariate parameters however can be challenging, as robust state descriptions often are derived from many subtle, distributed variations in network parameters. We convert those parameters to the frequency domain through the use of the multitaper [83]; while this allows us to plot the frequency features of each state, there is no clearly defined statistical tests for significance of these values on a single session of data. The following results are therefore analysed without reference to significance, a shortcoming that will be discussed in section 2.5.

Power Spectral Analysis

Plotting the average power spectrum of each state as in figure 2.7 allows us to characterise some key differences between states. States 2, 6 and 8 stand out immediately as outliers with respect to the remaining states, an interpretation consistent with their irregular time-domain activation. We conclude that they reflect some form of artefact - in the case of state 6 we can infer from the temporal alignment to saccades that these are saccade artefacts; states 2 and 8 appear to also be artefactual in origin although no direct interpretation can be specified. These states are excluded from the remaining analysis.

The remaining states occupy a relatively tight band of the spectrum; some slight differences can be read out, especially in the high gamma band power of DLPFC

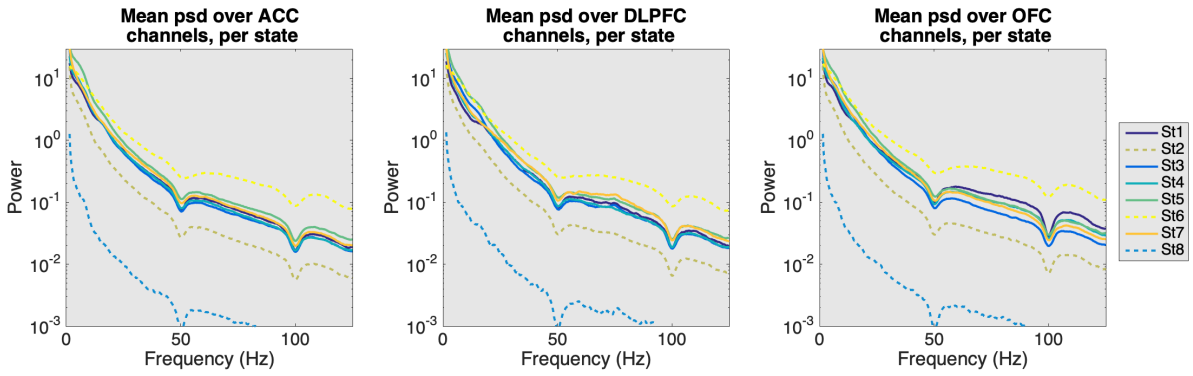


Figure 2.7: Power spectral densities per state, averaged over channels within each brain area. This immediately identifies outlier states which are marked in dashed lines and not considered in further analysis. The remaining states demonstrate some minor differences, for example in OFC high gamma, but overall these appear subtle and suggest the raw power spectra is not sufficient to fully explain the differences between states.

channels in state 5 and 7 or OFC channels in state 1, however the differences are subtle. A fuller understanding of the variations that characterise different states requires analysing power alongside coherence as function of each channel.

Analysis of Power and Coherence

For ease of interpretability we apply a non-negative matrix factorisation to the high-dimensional coherence matrices to decompose the frequency domain into three modes; this is a data-driven approach to find the best decomposition of spectral content, that normally nonetheless converges to canonically identified oscillatory patterns [84]. The inferred modes for this dataset are specified at the top of figure 2.8, and delineate one very low frequency band (<4Hz); a second that contains two distinct spectral peaks (one at 5Hz and another at 15Hz); and finally a high gamma band mode expressing frequency components above 30Hz.

To characterise some sense of baseline activity averaged over all trials, we determine first the mean power spectral density and coherence over active states (i.e. excluding states 2, 6, and 8). This results in the distribution of power and coherence in each of the specified frequency bands shown in figure 2.8(a), which provides a sense of the main spatial distribution of power in each frequency band. This shows, for example, that high gamma activity is on average concentrated in two channels

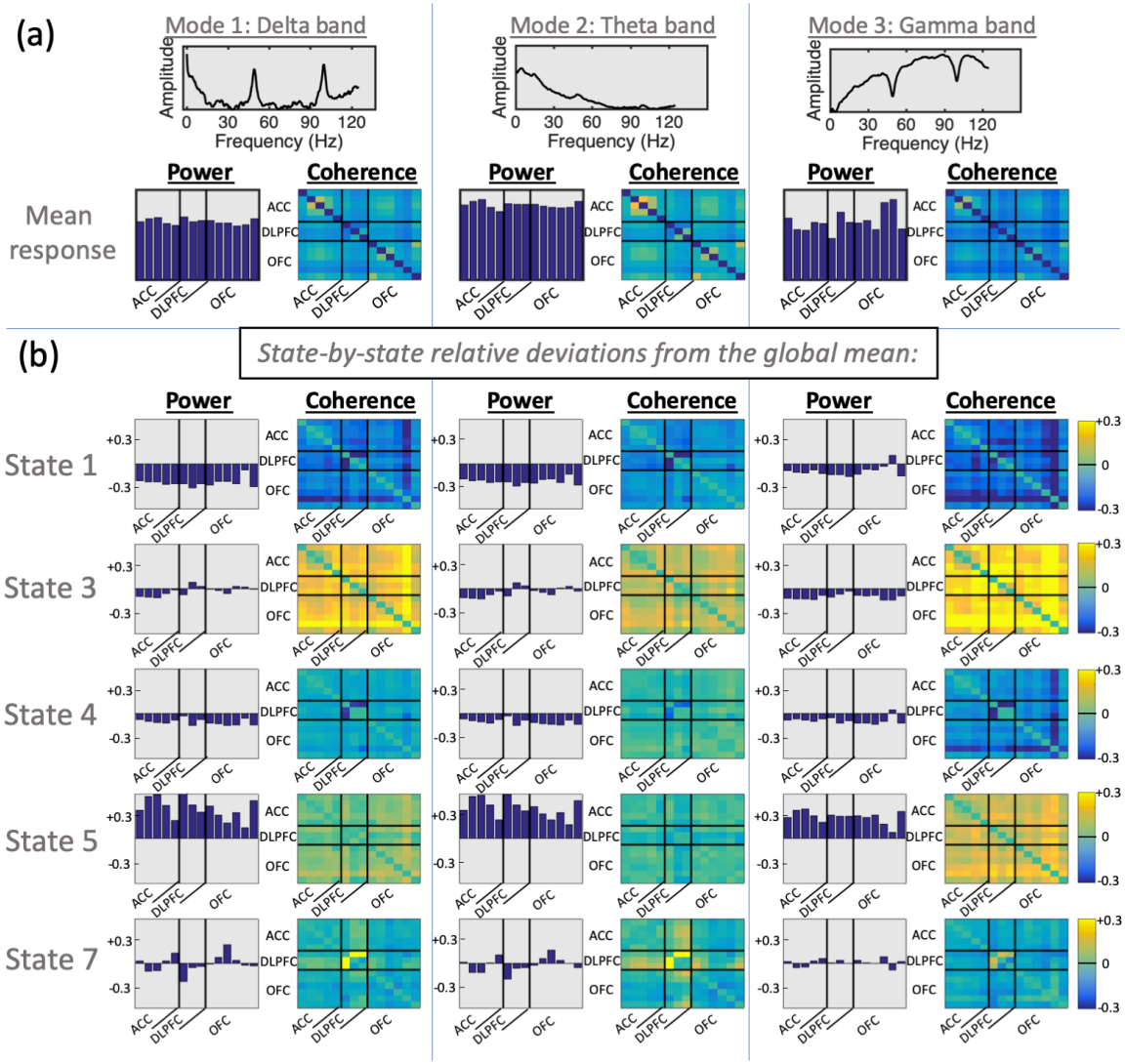


Figure 2.8: Distributed network parameters that differentiate states: each state is associated with particular spectrally resolved features, however these features are high dimensional and not straightforward to interpret directly. (a) We first compute the mean power per channel and coherence per pair of channels within each of three inferred frequency modes that explain the most variation in the data. This highlights the heterogeneity of responses, even over channels within the same PFC subregion. (b) We then plot the relative deviation from this mean in the statistics expressed by each state. Significant changes in scaling of power and coherence are evident between states; some of these are explored in more detail in the text, but in the absence of objective statistical significance tests these are difficult to quantify.

located in OFC. Turning to coherence, the spatial patterns expressed are surprisingly similar over different frequency modes, with the exception of the high gamma band that shows lower overall coherence while preserving the same spatial patterns. This supports the notion of high gamma being locally generated, hence less coherent with respect to other areas; the consistency of the spatial patterns does suggest that correlations between recording sites are preserved over different frequency bands.

In order to interpret individual state parameters, we scale each of these with respect to the global mean so that plotted values in figure 2.8(b) reflect how each state deviates from a common baseline. Merging this with the behavioural associations identified in section 2.3.1 allows us to interpret how certain trial parameters are expressed in the distributed LFP signal.

Correct trials linked to coherence of low frequency oscillations

In section 2.3.1, trials were separated into meta trajectories that were predictive of the error rates on individual trials. We can now identify the main differences between states associated with each of these trajectories. For clarity we focus here on the states prior to cue 1 onset, to disambiguate these from potential effects of presented cues.

State 3, predictive of correct trial outcomes, demonstrates elevated low frequency coherence compared with states 5 and 7 (associated with error trials), along with distributed changes in the sites with most low frequency power. State 3 also expresses significantly lower power in the gamma band than either of these two states. State 7 in contrast expresses some unique low frequency patterns not seen in any other states, including a prominent lateral synchronisation and increase in low frequency power on OFC channels that do not express a strong relationship with encoded parameters. This could reflect disengagement from the task, as state 7 is associated with poor performance on subsequent tasks and the absence of any residual reward signals from the previous trial.

2.3.4 Value processing robustly reflected in spatially distinct coherence patterns

Turning to cue induced effects, state 1 was strongly linked to low value cues, and state 4 to high value cues.

In low frequencies, state 4 exhibits higher power and coherence over nearly all channels compared to state 1. These changes appear widespread over all channels and brain areas. In the high gamma band, several channels in OFC appear to display greater variance in their distributions between state 4 and state 1. The established link between high gamma and local neuronal firing suggests this may reflect the presence of local cells encoding high or low value cues - which is consistent with the original studies finding that OFC cells encoded the cue value [30]. Nonetheless, the evidence here suggests cue value is not only encoded in the high gamma range but also in distributed changes in lower frequency bands.

We now turn to the more traditional time frequency analysis to see if this similarly identifies such changes and the specific channels on which they occur.

2.4 Univariate Results: Time-Frequency Analysis is Less Sensitive to these Effects

We can compare the results obtained through this multivariate analysis to the univariate time-frequency approach outlined in section 2.2.3.

2.4.1 No Univariate Features Correlate with Error Rate

Having established that the main mode of variation in state timecourses was predictive of error rates, we expected to find univariate correlates of this in the LFP power spectra. No correlation was found when analysing the epoch around presentation of the first cue, as shown in figure 2.9. Instead, site specific LFP power was found to modulate with the other two whole-trial regressors. This identified a strong residual effect of any reward that was administered on the previous trial, which explained a large portion of the variance in high gamma activity in ACC and DLPFC, as well as the power in beta band (approximately

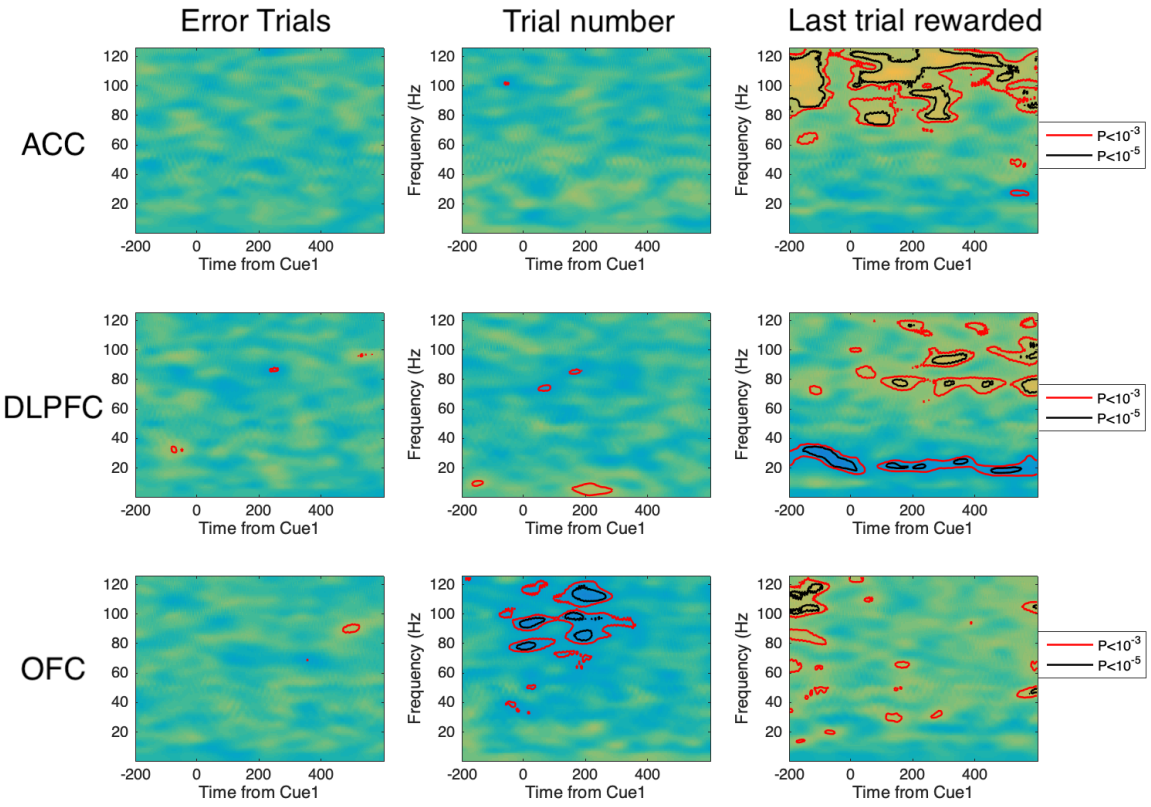


Figure 2.9: LFP time-frequency analysis is sensitive to different whole-trial effects: the regression coefficients β for each of the last three columns of the design matrix, pertaining to variables that deviate over whole-trials. Red and black lines denote statistical significance to $p < 0.001$ and $p < 10^{-5}$ respectively. No relationships were found between LFP features and error trials. OFC high gamma (which specifically encodes cue value) also showed an interaction with trial number within the session. Strong interactions were found with residual reward from the previous trial, especially in ACC high gamma and DLPFC beta bands.

20Hz) in DLPFC. Furthermore, a significant portion of the variance in OFC high gamma immediately following cue presentation is explained by the trial number within the session, demonstrating progressive decreases in OFC high gamma power over the course of each recording session.

2.4.2 Univariate LFP features show weaker correlations with stimulus value

Figure 2.10 highlights the effects of individual cue parameters on the observed LFP spectra. Out of all the regressors and frequency bands, the only reliable effect is

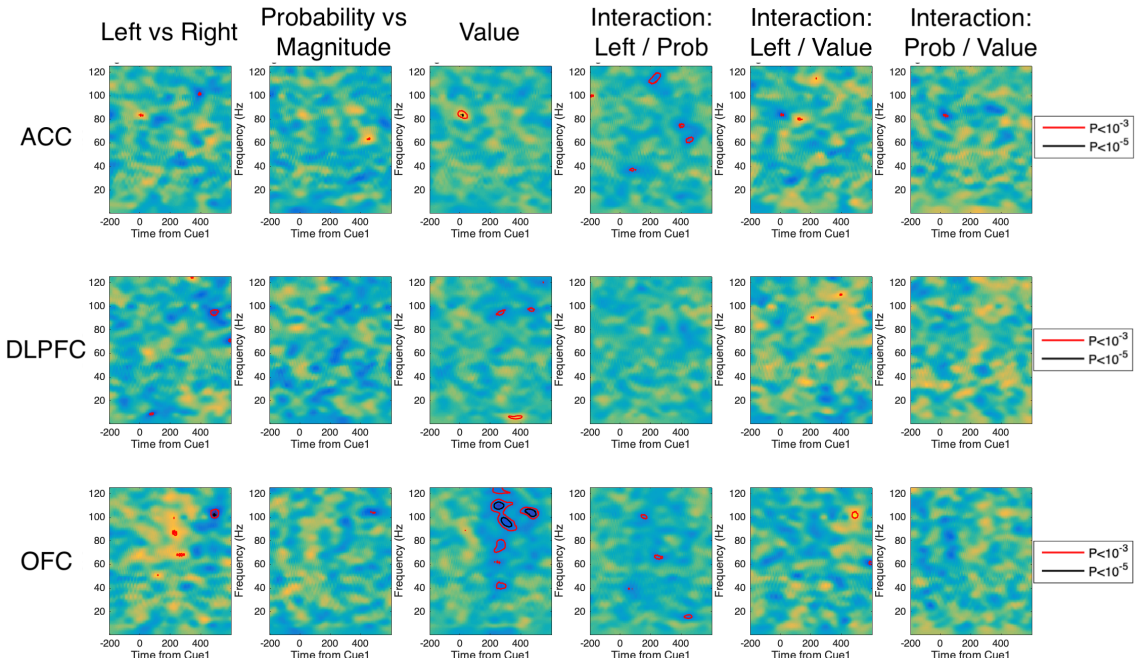


Figure 2.10: LFP time frequency analysis is less sensitive to stimulus-induced effects: the independent variables Y were regressed iteratively against the LFP power in each frequency band, PFC subregion and timepoint. The first six inferred regression coefficients β are plotted here, with red and black lines denoting points that are significant to $p < 10^-3$ and $p < 10^-5$ respectively. Overall, few features were identified that reliably correlated with stimulus variables, other than a relationship between cue value and OFC high gamma power.

that of cue value which is encoded in high gamma power in OFC sites. Analysing individual channels reveals this effect is only significant on a single channel (see appendix F). This is consistent with the finding that OFC cells encode cue value in their firing rates, peaking approximately 300msec post cue presentation [30]; as high gamma reflects local firing, presumably the OFC cell population adjacent to this electrode encoded low value cues. It is notable that no robust relationship is expressed betweeen the LFP power and any of the interaction terms, suggesting the data is sufficiently explained by the three first order regressors. Compared to the robust differentation of value between states 1 and 4, which showed distributed changes in low frequency power and coherence, no correlation is here observed betweeen low frequency components and cue value.

2.5 Discussion

These findings highlight a number of differences in the types of information that can be obtained using univariate modelling as opposed to multivariate models. Overall, the time delay embedded model was able to infer robust network states that reliably correlated with behaviour, whereas the windowed feature analysis only identified a few hallmark features confined to certain channels that correlated with experimental variables. We here explore in some detail the reasons for this, and the relative drawbacks of sliding window based analyses as opposed to HMM based analyses that provides the central justification for much of the methodology developed in subsequent chapters.

2.5.1 Multivariate Sensitivity to Distributed Variation

Time-frequency analysis involves a simple regression model being fit recursively over each timepoint, channel and frequency band. This means that parameters are independent (with the exception of adjacent timepoints where windows used to compute features overlap), and so each regression parameter is directly attributable to that channel, timepoint and frequency band. In contrast, the TIDE model infers a simultaneous distribution over timepoints within a window, computing the likelihood of each different state based on the distribution of all of these points.

This is a critical distinction that motivates multivariate modelling in general, and much of the methods developed in subsequent chapters, so we explore it more fully with the trivial example of figure 2.11. Consider the case of a two dimensional variable that varies over two conditions as in the figure shown; in the context of the analysis above, this could be the windowed power spectral density over two distinct channels at the same timepoint and frequency bin. The procedure conducted above amounts to projecting these points onto each axis consecutively, represented by the histograms of points on the X_1 and X_2 axes respectively, and testing each axis' distribution for statistically significant differences over conditions. In this case neither of the axes in isolation contains a significant degree of variation as to infer a statistically significant relationship in the data. Suppose instead however we

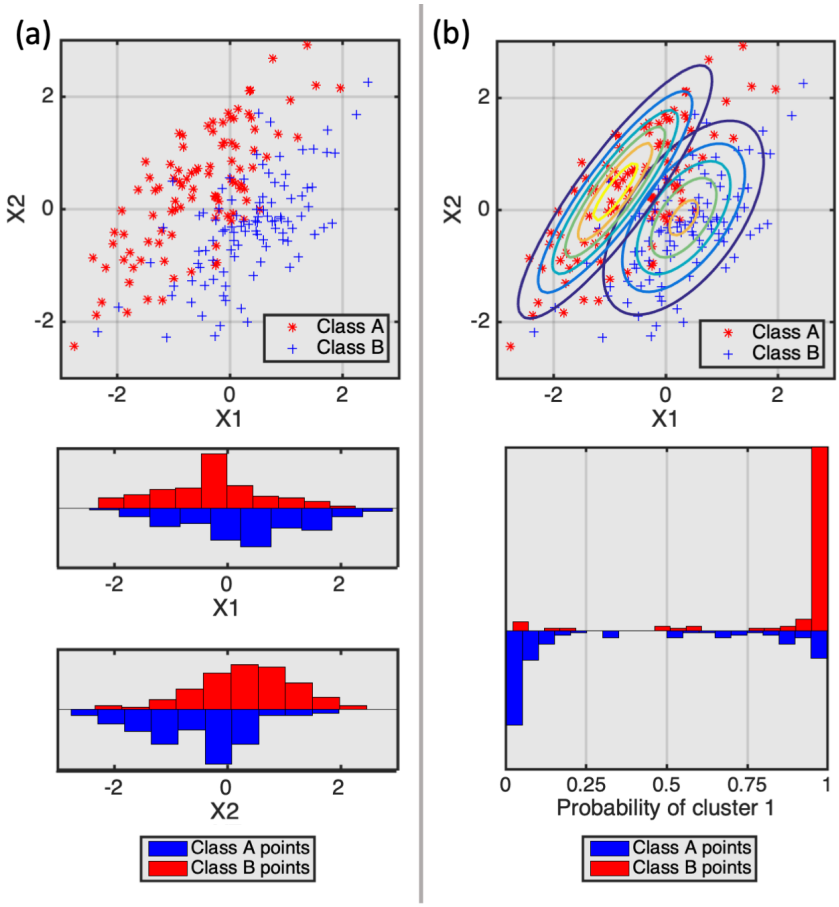


Figure 2.11: Example of univariate and multivariate analysis: (a) Suppose we have the two dimensional data shown and wish to investigate how it varies between the two classes A and B. A univariate analysis separates each dimension under observation and compares the distributions of $X1$ and $X2$ in turn - each of which have highly overlapping distributions that are not significantly different between classes. (b) If we instead employ a multivariate model and cluster the data in its original space, we extract features that better represent the correlative structure of the data. We then find that the resulting cluster membership probability very effectively separates the two classes.

set out to cluster the multidimensional data first - for simplicity in this example we employ a Gaussian mixture model with 2 components, but the same principle applies to clustering in time with HMMs with suitable observation models. This clusters the points by assigning a likelihood of membership to each cluster based on the inferred distributions, taking into account the simultaneous distribution of each point across all data dimensions. If we then ask whether a predictive relationship exists between the experimental variable and the membership of either cluster, a robust relationship emerges.

This appears to be the driving factor behind the result obtained in the above analysis. Figure 2.8 demonstrates that a number of states are differentiated not by variations that are uniformly expressed over channels, but which are distributed across the network. Multivariate models will remain sensitive to large numbers of small variations across and between channels, whereas univariate analyses will only be sensitive to variations on each channel in isolation.

2.5.2 Shortcomings of Sliding Window Analyses

One of the interesting features of this analysis was the relatively poor performance of time frequency models in identifying changes in power induced by experimental variables. We performed this analysis using a sliding window width of 200msec - this represents a commonly used value for this parameter, that trades off the more stable feature estimates produced by longer window sizes with the reduced sensitivity to short duration, bursty effects. It is possible that there may be a more optimal window size for this analysis, however methods to objectively determine a fixed window length (for example by cross-validation) can be computationally demanding, and in any case will remain suboptimal to HMM based analyses whenever variable duration or nonstationary effects are present.

A second challenge is the types of features that can be reasonably assessed using sliding window based analyses. The HMM analysis found frequency-limited coherence to be an important factor in determining state distributions, yet in the sliding window analysis we focussed only on power. This is because of established challenges in computing band limited coherence over sliding windows that often result in very noisy and unreliable estimates of this parameter [85, 86]. By comparison, the HMM model groups together points from all trials that share a similar distribution; spectral coherence is then computed through a multitaper that analyses all of this data, producing more robust estimates for each state.

2.5.3 Shortcomings of Multivariate Modelling

While the TIDE model has inferred important dynamics that explain quite significant variations between trials, a number of critical shortcomings remain.

• **Scalability:** This model is dependent upon simultaneous recording from all channels. All the analysis presented has been performed on only one of 54 different recording sessions, and consequently represents only a small portion of the available data. While univariate analysis techniques could easily be extended to each channel recorded over all 54 sessions, the TIDE model cannot scale up as it is dependent on both the types of channel present per session, and their relative proportions. As overarching principles for how LFP features should vary from one channel to another are presently not well understood even within the same brain region, developing a framework for extending multivariate models to this kind of data is a nontrivial task and ultimately beyond the scope of this thesis.

However it is important to note that separate TIDE models were fit to the two other sessions that contained 14 independently recorded channels. Emphasising that each of these sessions were recorded from different sites in different proportions to the session analysed and presented above, in one of these two sessions the fitted model reproduced the key findings of section 2.3.1 - albeit with some qualitative differences in state definition and expression vs time. In the other session the result was not significant but trended the same direction. This validates the present analysis and reinforces the view that states can be inferred from LFP data that robustly predict error rates and behaviour; the exact relationship between states and behaviour will however depend on the specifics of the data recorded in each session.

• **Specificity to research questions:** Another major advantage of univariate models is that they allow very specific questions to be asked regarding the relationship between observations and variables. In contrast, the TIDE model infers the states that best describe the intrinsic modes of variation in the data, regardless of whether these vary with experimental variables. In the study above, the researchers were primarily interested in dissociating how different cue parameters were encoded in the brain. All of the analysis conducted only identified relationships with one of these variables, the cue value; based on our HMM modelling we could state nothing about the relationships between coherence and the other parameters, such as whether

the encoded information related to the left or right option, aspects of which may be represented in the data but in a manner too subtle to be captured in distinct states.

2.6 Conclusion

We have presented a novel analysis of LFP data, and demonstrated - through direct comparison with more common sliding window analyses - the benefits of multivariate dynamic models. Fitting a model previously developed for MEG data [32], we show that trial variables such as value are robustly encoded in distributed network parameters in ways that may not be detectable to univariate methods. Furthermore, we show that the main mode of variation of whole trial state trajectories is a meaningful reflection of underlying cognitive processes not controlled by the experiment, such as attention and cognitive engagement, as it is predictive of error rates on individual trials. These findings demonstrate the methodological potential of multivariate modelling for LFP data, which could support a broader understanding of how experimental variables are encoded in the brain within the context of ongoing network activity and spontaneous variations over trials. Nonetheless, this was not pursued further due to fundamental shortcomings of the multivariate model as it currently stands. The data structure lacks a common reference frame over multiple sessions, which prohibits the model in its current form from scaling to all sessions in a manner that is consistently interpretable. Furthermore, the unsupervised modelling approach does not necessarily support direct questions about network wide encoding of specific experimental variables; the model is fit independently of the variables under investigation, then post-hoc tested whether a relationship exists between inferred states and experimental manipulations. For these reasons, we now turn our attention to formulating explicitly supervised multivariate models.

Chapter 3

Sequential Encoding and Decoding Frameworks

Contents

3.1 Introduction 48
3.2 A Dynamic Framework for Regression Models 57
3.3 Synthetic Data Results 66
3.4 Discussion 72
3.5 Conclusion 73

3.1 Introduction

As established in the previous chapter, multivariate models can provide robust inference of large scale dynamics within trials. We have shown that these can be very useful to infer the background patterns of activity within the brain that provide a meaningful readout of an abstract brain 'state'; while abstract in definition and not guaranteed to describe differences that are easily interpretable, they nonetheless find patterns in distributed information that can much more reliably correlate with behaviour than variations on any single channel.

However, these models in their current form find the modes of data that explain the most variance, a definition which is not guaranteed to find the variations relevant to known experimental variables. In the example of the previous chapter, researchers interested in how data presented pertaining to the left vs right option

are encoded could draw no conclusions from the analyses conducted, as the inferred states do not correlate at all with this variable. This problem is compounded when we go from a number of focussed, intracranially inserted recording probes to non-invasive, whole brain recordings such as those obtained by MEG or EEG. In such scenarios, where there is typically much poorer signal to noise ratios, fitting an unsupervised model to task data may result in a set of states that describe the macro patterns of variation in the data without capturing any meaningful differences between experimental conditions.

For these reasons, we now turn to the development of supervised time-varying models, extending the work of Vidaurre et al. [33] to allow us to ask more specific and targeted questions about the relationships between data observations and experimental variables.

3.1.1 Background

Information is represented and processed in the brain through specific patterns of dynamic activity in large cell assemblies. Simple everyday tasks, such as recognising an object, involve a volley of neural activity that propagates through specific feedforward and recurrent architectures, starting at lower sensory areas and spreading towards higher levels of the visual hierarchy [87, 88]. Our ability to consistently associate a particular object with its name, properties and appropriate response, is a result of these pathways propagating a consistent and repeatable sequence of activity upon each presentation of the same stimulus. This activity is stochastic in nature and never identical over any two stimulus presentations, but each time distributed by common statistical patterns that reliably elicit a suitable behavioural response.

Such sequential patterns of activity are ubiquitous in neural data. At the cellular level, there is evidence that tasks elicit regular dynamic sequences in ensemble firing rates, with specific variables encoded in either a dynamic or persistent fashion on some subspace of an overall task sequence [89–92]. Where such cellular encoding regimes are well characterised, such as in the spatial navigation system, specific

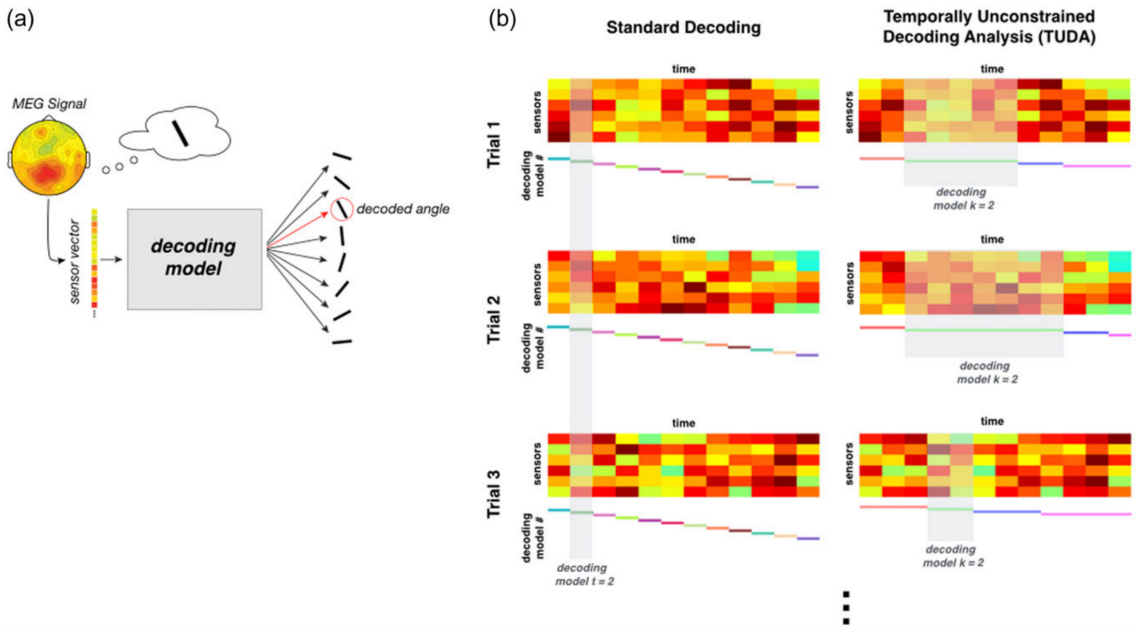


Figure 3.1: Temporally Unconstrained Decoding Analysis: the concept proposed by Vidaurre et al. [33] that this chapter builds on and extends. (a) Data is recorded over multiple trials of different types, in this case MEG recordings of a subject viewing an angle of orientation. (b) The standard approach to decoding involves aligning all trials in time and training successive models over successive timepoints. Conceptually, this paper proposes inferring a Markov state sequence on each trial, with each state associated with parameters that maximise the overall model fit. Figure reproduced from Vidaurre et al. [33].

relationships have been observed between sequence activation at the cellular level and mesoscopic sequences observable in the LFP [93–96]. The speed at which the LFP follows a regular sequence in individual trials of a task also explains the degree to which specific firing rates encode key task variables [77]. Studies have also found that gamma and beta band power changes in specific spatiotemporal sequences are a key mechanism underlying mechanisms of memory [97, 98]. Such sequences are also ubiquitous in whole brain data, with evidence that learned sequences are reactivated in human MEG recordings [99, 100]. Sequences continue to be important in neural data recoded at slower timescales; studies in fMRI have shown that even within task blocks, the metric of functional connectivity is not static but evolves through a reliable pattern on each block [101].

We argue that the precise timing at which these sequences unfold may be

an important source of variation in the brain. For example, while there may be a reasonably consistent sequence of patterns on each trial, the precise timing and structure of these patterns will differ on a trial-to-trial basis. This between trial variability may reflect important aspects of cognitive attention, strategy or arousal levels.

3.1.2 Regression models of encoding and decoding

Linear regression models have been widely used to identify relationships between experimental variables and MEG sensor data [102–104], source space data [105], EEG sensor data [106–109], cellular firing rates [110, 111] and local field potentials [41, 112] to name but a few influential examples. Furthermore, the simple inclusion of a nonlinear link function extends this to the generalised linear modelling (GLM) framework; where this link function represents a haemodynamic response function (HRF), this forms the General LM framework used for most univariate fMRI analysis [113–115]; other link functions such as the logistic sigmoid function allow this framework to generalise to classification problems [116, 117].

The vast majority of these GLM applications could be classed as either encoding or decoding models. These two frameworks are fundamental to neuroscience research, both for interpreting neuroimaging data, but also more fundamentally as models of essential functions performed by the nervous system. Encoding models explain how information can be encoded within the brain; decoding models explain how that code be read out by downstream brain areas.

The distinction between encoding and decoding models is a trivial swap of the dependent and independent variables in the GLM framework. Supposing we have a data matrix X of dimension $[N \times P]$, and design matrix of regressors Y of dimension $[N \times Q]$, then an encoding model would be any that satisfy:

$$P(X|Y, \alpha) = g(Y\alpha) \tag{3.1}$$

where α is a $[Q \times P]$ matrix of regression coefficients and g the appropriate link function for the model (e.g. in the case of the standard gaussian model,

$g(Y\alpha) = Y\alpha + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$). On the other hand, a decoding model is of the form:

$$P(Y|X, \beta) = g(X\beta) \tag{3.2}$$

where β is a $[P \times Q]$ matrix of regression coefficients and g the appropriate link function.

The trivial swap of dependent and independent variables has a few important consequences. Firstly, one can select distributions or link functions which may not be analytically invertible, but which best characterise the data. For example, where Y is categorical and X continuous, a logistic regression decoding model can be fit, where X can vary uniformly in any plane parallel to the logistic regression parameters. There is no equivalent encoding model that can be fit without imposing additional assumptions over the distribution of X (e.g. assuming Gaussian noise distribution as in [118]). Secondly, we are able to set priors over our model parameters that may not have an equivalent form when the dependent and independent variables are swapped. Lasso regularisation [119], for example, has the often desirable quality of feature selection, enforcing sparsity over model parameters; especially in cases where P and Q are significantly different in scale, imposing such sparsity on the coefficients of an encoding model is not equivalent to imposing sparsity over a decoding model parameter set, and no explicit form exists for an equivalent constraint in the inverse model. Regularising priors can also be placed over the temporal relationships between variables, a topic we explore more fully below.

The standard application of GLM encoding and decoding models is to align all points to some temporal marker (e.g. stimulus onset) and treat each timepoint as if sampled from a distribution independent of adjacent samples but consistent over trials. Notwithstanding commonly used smoothing and windowing functions that may be applied to the data, the model itself implies the following inherent assumptions:

• Data is *synchronous*; the data at any aligned time $t = \tau$ for a given trial is sampled from the same underlying neural process as that from $t = \tau$ at any other

trial.

• Data from adjacent timepoints are independent: the data at time $t = \tau$ is completely independent of the data at time $t = \tau + 1$.

• There is a sufficient number of trials to obtain a reasonable estimate of the underlying model parameters independently at each timestep.

These are strong assumptions which do not generally hold for real neural data.

We illustrate this point with a trivial example in figure 3.2, where we have generated 2-dimensional data X that follows a noisy trajectory on each trial as a function of one of two class labels Y. Generating this data from sinusoidal terms allows us to simulate, through a noisy phase increment at each timestep, the temporal misalignment over multiple trials that we believe better reflects neural activity:

$$X = \left[\begin{smallmatrix} sin(\theta_t) \\ Y sin(4\theta_t) \end{smallmatrix} \right] + \epsilon_X \tag{3.3}$$

$$\theta_t = \theta_{t-1} + \epsilon_\theta \tag{3.4}$$

where $\epsilon_X \sim N(0, \sigma_X^2 I)$ is measurement noise and $\epsilon_\theta \sim N(c, \sigma_\theta^2)$ represent a noisy phase increment at each timestep, and Y is a binary variable denoting the trial type, taking values $Y = [-1, 1]$ reflected in the two different colours in figure 3.2. Thick coloured lines plot the mean trajectory vs time for each trial type, originating at the origin and unfolding through the plane. Critically, while our generated data X is two dimensional, only one of these dimensions, $X2$, varies with respect to the trial type Y; the other, $X1$, is independent of Y but nonetheless following a dynamic trajectory on each trial.

Fitting a standard, synchronous Gaussian encoding model amounts to determining a mean and covariance matrix over X for each condition, at each timestep. Figure 3.2 plots the temporally aligned data from all trials, sampled at two example timesteps, along with the fitted mean and covariance for each condition shown in contour lines. This has resulted in inferred means for each condition that vary slightly in the dimension $X1$ that should be independent of Y. More clearly both distributions have inferred correlations in the data covariance that are asymmetric

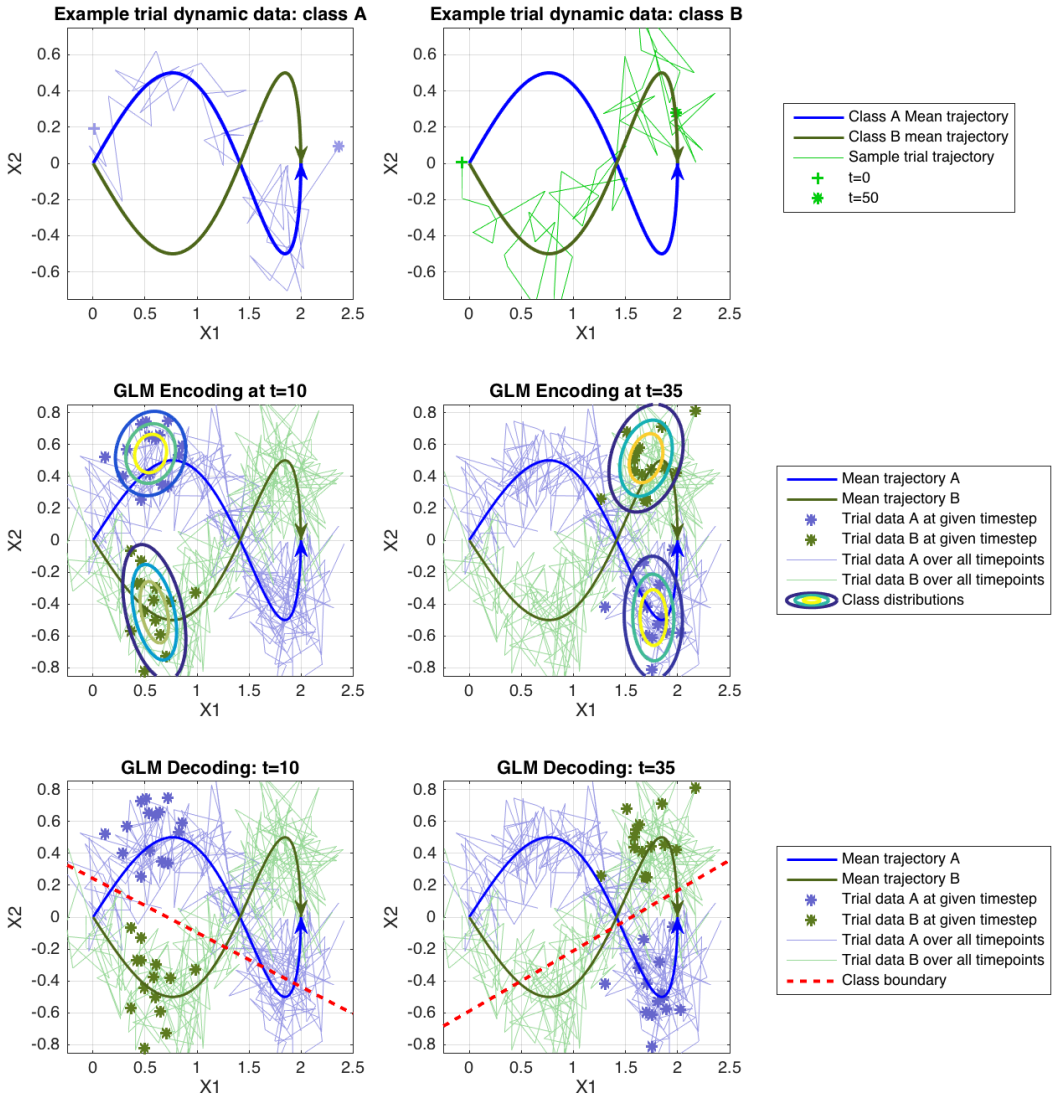


Figure 3.2: Outlining the approach with a trivial example. Top panel: we simulate dynamics for multiple trials, where each trial corresponds to one of two different experimental conditions, A and B (denoted by blue and green respectively), and plot the trajectories over time on one sample trial of each. Each trial begins at the origin and noisily follows a dynamic trajectory in two dimensions. Middle panel: standard, synchronous encoding model for two example time steps, which isolates a single point in time and fits a Gaussian distribution to the observed data. Whereas the optimal encoding model Gaussian distributions should have means symmetric across the $X2$ axis and have isotropic covariance, inferred parameters are skewed by noise and poor phase alignment. Bottom panel: decoding model for two example timesteps, which fits a classification boundary between points. In both cases the optimal decode plane should correspond to the line $X2 = 0$, but due to noise and phase shifts the inferred decode plane is significantly tilted.

across the $X2 = 0$ line and have been induced by measurement noise and inter trial phase shifts.

Fitting a standard, synchronous logistic regression decoding model to this data in a similar manner, we should expect to observe classification boundaries that are roughly horizontal, following the line $X2 = 0$ as closely as possible. Again, we instead observe that the model parameters have been significantly shifted from the ideal distribution, having been offset by both measurement noise and inter trial differences. Note that in this example both measurement noise and phase noise were controlled to very low levels for illustrative purposes, but nonetheless induced significant deviations in model accuracy - these issues are compounded in high dimensional, noisy neural data.

3.1.3 Models of Sequential Dynamics

A number of models exist that could better reflect the dynamic nature of neural data. As discussed in chapter one, HMMs are an attractive candidate for our purposes, having been utilised extensively for inference of unsupervised dynamics in electrophysiological data [32, 83, 84, 120], and more recently supervised decoding as outlined in [33]. HMMs present a number of advantages in neuroimaging analysis; by constraining the dynamic state space to a discrete set of mutually exclusive states, the output is a typically low dimensional sequence that is often more interpretable than continuous state space models - in particular for resting state analyses, state distributions tend to correlate strongly with known functional modes such as visual or motor networks. HMMs also impose a temporal smoothness constraint that offers increased robustness over outlying data points. In the context of supervised regression models, HMMs offer a means by which to infer dynamic sequences over a low dimensional, discrete state space which is suitably constrained for data that may have a high number of dimensions and low signal to noise ratios.

HMMs also present an opportunity to impose reasonable assumptions about state dynamics, an avenue that has not been significantly explored in HMM neuroimaging applications. The prior on the transition matrix can prohibit undesired state

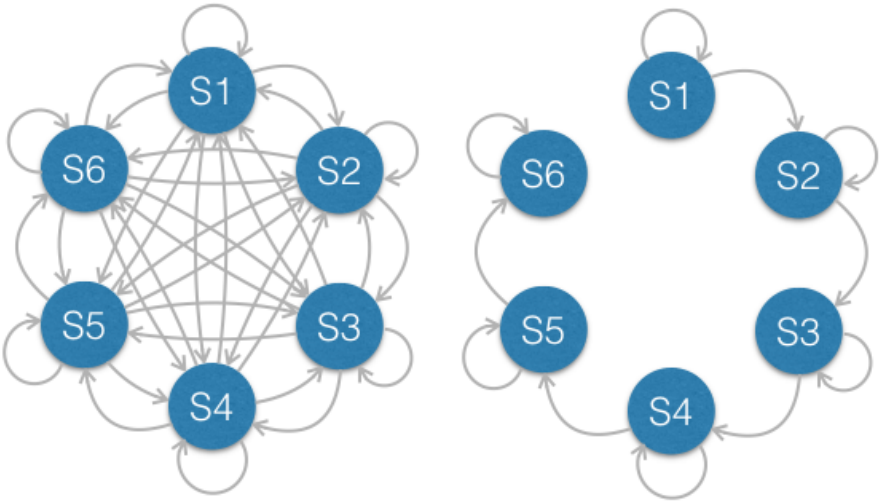


Figure 3.3: Sequential Hidden Markov Models. There is a finite number of discrete states, and any active state is free to transition to any other state in the system (left panel). Prohibiting state jumps and only allowing the active state at each timepoint to remain active or transition to the next consecutive state (right panel), we enforce a sequential state structure that acts as a strong temporal regulariser and ensures our model captures sequential dynamics.

transitions, acting as a powerful means of temporal regularisation by enforcing the temporal structure that best reflects underlying assumptions about the data. Where data is believed to have strongly sequential characteristics, constraining state transitions to a single step while prohibiting backwards transitions as illustrated in figure 3.3 allows us to extract the most salient features of *sequential* representations, separating these from transient sources of noise that could otherwise drive state dynamics.

Applying this conceptually to the same trivial simulation example already discussed illustrates the potential of this approach (note that the observation model details we be outlined subsequently in some detail - we here outline these results only at a conceptual level to motivate the work that follows). Fitting a simple two state sequential encoding model would pool all of the trial data points together into two states that best explain each condition by a Gaussian mean and covariance (as shown in figure 3.4). Compared to the synchronous case discussed above, the means and covariance are now symmetrical across the $X2 = 0$ line as desired.

Similarly fitting a logistic regression model to this data results in state assignment that maximises the linear separability of the data by fitting horizontal classification planes. In both cases the exact timing of the inferred transition from one state to the other varies on each trial, reflecting the dynamic nature of the data. This is manifest in the slight differences in timing over trials in the transitions from state 1 to state 2 in the right hand panel of fig 3.3, although for illustrative purposes in this simulation the between trial variability in timing is quite low, and is expected to be much higher in real data. Also for illustrative purposes, this model uses just 2 states, however this number could be increased which would result in states with more refined spatial distributions that should, due to their alignment in space rather than time, all lead to more accurate parameter inference than in the synchronous model.

3.2 A Dynamic Framework for Regression Models

Given the dynamic nature of neural data, we argue that low dimensional dynamic models are better suited to data analysis than the synchronous GLM approach outlined in section 3.1.2. Vidaurre et al. [33] introduced an HMM framework for supervised data analysis with a dynamic decoding model for a single Gaussian stimulus variable, and demonstrated the improvements in decoding accuracy that this achieved. This innovation [33] was the central motivation for the work that follows in this chapter, where we extend this approach in a number of ways:

• Generalisation to the multinomial case, whereby the regression is fit to an arbitrary number of stimulus variables simultaneously, while the Markovian states maximise the model fit over all of these variables;

• This allows us to generalise the model further to fit both encoding and decoding frameworks of arbitrary dimensionality;

• We implement nonlinear link functions, specifically logistic regression, allowing use of categorical dependent variables;

• We introduce temporal regularisation by constraining state HMM sequences to

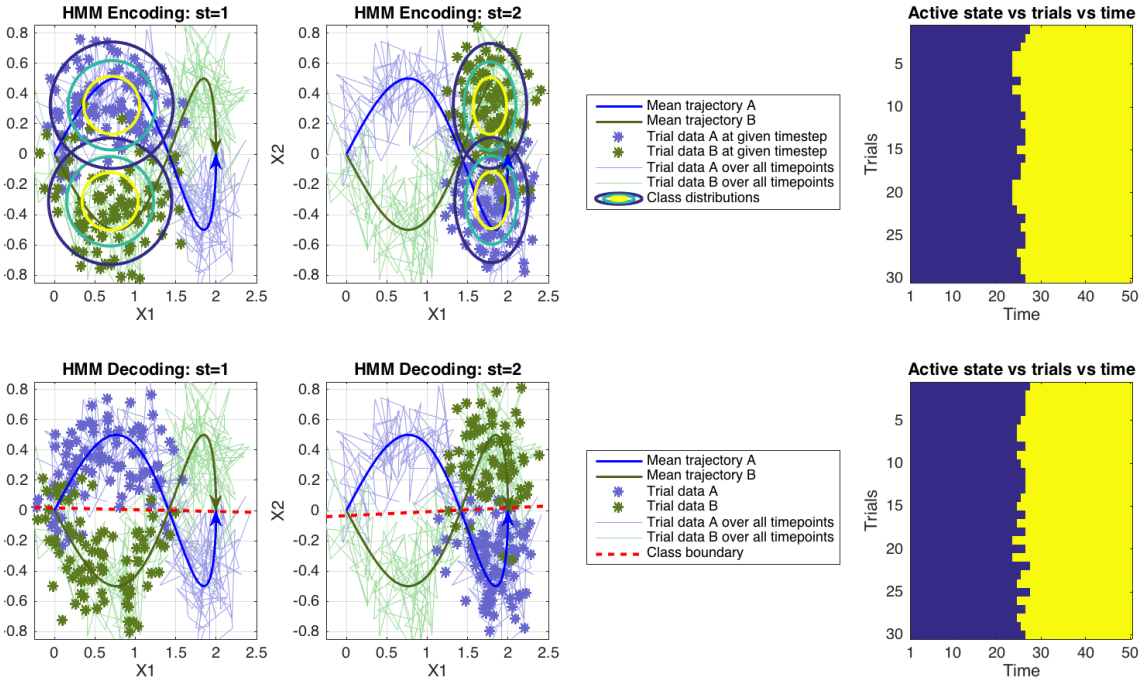


Figure 3.4: Fitting dynamic encoding and decoding models. Conceptually outlining the approach, we can infer two condition dependent states in either an encoding (top panel) or decoding (bottom panel) framework. The left side plots the assignment of data points into states and the observation model fit to the data; right side plots the active state vs time on each trial, ordered so trials 1-15 are of class A and trials 16-30 are of class B and with the colours dark blue and yellow denoting states 1 and 2 respectively. Top panel: an HMM encoding model determines the state active at each timepoint and fits a Gaussian distribution over all points assigned to that state. This results in parameter estimates that better reflect the characteristics of the data (i.e. symmetry and isomorphic noise covariance). Bottom panel: similarly for the HMM decoding model, an active state is determined for each data point and the corresponding logistic regression parameters inferred that best separate the data. The result is classification planes that are much more closely aligned to the optimal parameter set.

follow a deterministic order, i.e. a sequence, on each trial.

3.2.1 Sequential HMM

Taking the HMM framework outlined in section 1.3.3, we have a latent state variable denoted by $Z_{n,t} \in \{1, 2, ...K\}$ which describes the system state at time t on trial n. The Markov assumption states that this variable is conditionally dependent only on its value at the previous timepoint; in section 1.3.3 we parameterised this by the

transition matrix $\mathbf{\Theta}$. We now further impose the sequential structure illustrated in figure 3.3 by setting a limited Dirichlet prior on this transition matrix as follows:

$$P(Z_{n,t} = j|Z_{n,t-1} = i) = \begin{cases} \Theta_{i,j} \text{ for } j \in i, i+1 \\ 0 \text{ for } j \notin i, i+1 \end{cases}, P(Z_{n,1} = k) = \begin{cases} 1 \text{ for } k = 1 \\ 0 \text{ for } k \neq 1 \end{cases} \quad (3.5)$$

where $\Theta_{i,j}$ denotes the $\{i, j\}$th entry of the transition matrix $\mathbf{\Theta}$ between Markov states:

$$\mathbf{\Theta} = \begin{Bmatrix} \Theta_{1,1} & \Theta_{1,2} & 0 & 0 & ... & 0 \\ 0 & \Theta_{2,2} & \Theta_{2,3} & 0 & ... & 0 \\ 0 & 0 & \Theta_{3,3} & \Theta_{3,4} & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \end{Bmatrix} \quad (3.6)$$

Each non zero value in this constrained transition matrix is a random variable to be inferred; to each of these we fit a second order Dirichlet prior with concentration parameter of 1. With this model, we can now condition regression coefficients upon the latent state variable $Z_{n,t} \in \{1, 2, ...K\}$.

3.2.2 A Dynamic Encoding Model

We now incorporate this latent variable $Z_{n,t}$ with defined state sequence into the standard encoding framework of equation 3.1. Choosing a simple Gaussian distribution for the link function enables us to model our data likelihood as:

$$P(X_{n,t}|Y_{n,t}, Z_{n,t}, \alpha^{(1)}, \Lambda^{(1)}, \alpha^{(2)}, \Lambda^{(2)}, ...\alpha^{(K)}, \Lambda^{(K)}) \sim N(Y_t\alpha^{(Z_{n,t})}, \Lambda^{(Z_{n,t})}) \quad (3.7)$$

where $X_{n,t}$ is the $[1 \times P]$ vector of data recorded at time t on trial n; $Y_{n,t}$ is the $[1 \times Q]$ vector of stimulus variables (i.e. the design matrix row at time t on trial n); and $Z_{n,t} \in \{1, 2, ..., K\}$ denotes the active state. Note that our discrete state variable $Z_{n,t}$ now acts to select the appropriate set of regression coefficients $\alpha^{(k)}$, each of which are of dimension $[Q \times P]$. Importantly, the columns of each matrix $\alpha^{(k)}$ are mutually independent, whereas the state timecourse variable at each timestep is dependent upon the encoding across all dimensions of the data. This ensures a sensible overall state distribution while permitting an easily computable factorised representation over each column of the matrix $\alpha^{(k)}$.

To complete the Bayesian hierarchy we assign priors over each of our parameters. We model the noise precision in each state $\Lambda^{(k)}$ as a strictly diagonal distribution, with each entry on the diagonal (denoting the ith diagonal entry by $\Lambda^{(k)}_{i,i}$) being gamma distributed. Furthermore we set an automatic relevance determination prior [121] over each (i,j)th element $\alpha^{(k)}_{i,j}$ of the regression coefficient matrix in each state. This introduces a unique hyperparameter $\rho^{(k)}_{i,j}$ for each of these entries that acts to prune irrelevant features in a data-driven manner, assigning each entry an uncorrelated normal distribution. Completing the Bayesian hierarchy, we have:

$$P(\Lambda^{(k)}_{i,j}) \sim Gamma(a_0, b_0) \tag{3.8}$$

$$P(\alpha^{(k)}_{i,j}|\rho^{(k)}_{(i,j)}) \sim N(0, \frac{1}{\rho^{(k)}_{i,j}}) \tag{3.9}$$

$$P(\rho^{(k)}_{i,j}) \sim Gamma(a_0, b_0) \tag{3.10}$$

$$P(Z_t = j|Z_{t-1} = i) = \begin{cases} \Theta_{i,j} \text{ for } j \in [i, i+1] \text{ and } j < K \\ 0 \text{ for } j \notin [i, i+1] \end{cases} \tag{3.11}$$

$$P(\Theta_{i,j}) \sim Dirichlet(c_0) \tag{3.12}$$

where we use uninformative values for each of the hyperparameters a_0, b_0, c_0. We can then compute our full model posterior distribution as follows:

$$P(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\rho}|\mathbf{X}, \mathbf{Y}) = \frac{\left[\prod_{n=1}^{N}\prod_{t=1}^{T} P(X_{n,t}|Y_{n,t}, Z_{n,t}, \alpha^{(1)}, \Lambda^{(1)}, ..., \alpha_K, \Lambda^{(K)})\right] P(\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\rho}, \mathbf{Z})}{P(\mathbf{X})} \tag{3.13}$$

where we have used $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ as shorthand for the full set of data over N trials of length T, and $\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\rho}$ as shorthand for the full set of parameters over all unique states. This posterior contains conjugate priors, and so is tractable using traditional variational Bayesian methods, as outlined in more detail in appendix C.

3.2.3 A Dynamic Decoding Model for Categorical Variables

We could outline an equivalent Gaussian decoding model to the above by trivially swapping the dependent and independent variables. This would be similar to the approach used in Vidaurre et al. [33]. However, if the labels in $\mathbf{Y}$ are distributed categorically, as is the case for much of the data we will consider, a more suitable

decoding model is that of logistic regression. A brief note on notation: as we are now changing our observation model but not the underlying dynamic state model, we maintain the notation of $Z_{n,t}$ for the latent state variable; for the observation model parameters, the regression coefficients - labelled as α and their hyperparameters ρ for the encoding model - will for the decoding model be denoted as β and τ respectively.

If our binary data $\mathbf{Y}$ takes values $[1, -1]$, we can write:

$$P(Y_{n,t} = 1|X_{n,t}, Z_{n,t}, \beta^{(1)}, \beta^{(2)}, ...\beta^{(K)}) = \sigma(X_{n,t}\beta^{(Z_{n,t})}) \tag{3.14}$$

$$P(Y_{n,t} = -1|X_{n,t}, Z_{n,t}, \beta^{(1)}, \beta^{(2)}, ...\beta^{(K)}) = 1 - \sigma(X_{n,t}\beta^{(Z_{n,t})}) = \sigma(-X_{n,t}\beta^{(Z_{n,t})}) \tag{3.15}$$

where $\sigma(x) = \frac{e^x}{1+e^x}$ refers to the logistic sigmoid function, and our discrete latent variable $Z_{n,t} \in \{1, 2, ...K\}$ selects the appropriate set of coefficients $\beta^{(k)}$ which are each of dimension $[P \times 1]$. Note that by arbitrarily setting $\mathbf{Y}$ to take values of $[1, -1]$, we can generalise equations 3.14 and 3.15 to the following:

$$P(Y_{n,t}|X_{n,t}, Z_{n,t}, \beta^{(1)}, \beta^{(2)}, ...\beta^{(K)}) = \sigma(Y_{n,t}X_{n,t}\beta^{(Z_{n,t})}) \tag{3.16}$$

Extending to the case of multinomial logistic regression with N mutually exclusive classes, we extend the dimension of Y to $Q = N - 1$, allowing for one arbitrary reference class over which the log odds of each other class are defined. Our full data likelihood function for multinomial, dynamic logistic regression can now be written as:

$$P(Y_{n,t}|X_{n,t}, Z_{n,t}, \beta^{(1)}, \beta^{(2)}, ...\beta^{(K)}) = \prod \sigma\Big((Y_{n,t}) \cdot (X_{n,t}\beta^{(Z_{n,t})})\Big) \tag{3.17}$$

where $\cdot$ denotes the element wise product, $\sigma(x)$ applied to a vector quantity is taken to imply element-wise application of the logistic link function above; and $\prod$ denotes the product of vector elements. As in the case of the encoding model, each Markov state now has an associated matrix of coefficients $\beta^{(k)}$ of dimension $[P \times Q]$. Again the columns of this matrix are statistically independent of each other conditioned on the state variable, allowing a natural factorisation over columns, whereas the state variables $Z_{n,t}$ at each timestep are conditionally dependent on all classification dimensions to infer the best overall state sequence.

Specifying priors to complete the Bayesian hierarchy, we again use an ARD prior [121] over the regression coefficients to prune out features with limited evidence:

$$P(\beta_{i,j}^{(k)}|\tau_{i,j}^{(k)}) \sim N(0, \frac{1}{\tau_{i,j}^{(k)}}) \tag{3.18}$$

$$P(\tau_{i,j}^{(k)}) \sim Gamma(a_0, b_0) \tag{3.19}$$

$$P(Z_t = j|Z_{t-1} = i) = \begin{cases} \Theta_{i,j} \text{ for } j \in [i, i+1] \text{ and } j < K \\ 0 \text{ for } j \notin [i, i+1] \end{cases} \tag{3.20}$$

$$P(\Theta_{i,j}) \sim Dirichlet(c_0) \tag{3.21}$$

Where each $\tau_{i,j}^{(k)}$ models the precision of the (i,j)th entry of the matrix $\beta^{(k)}$. We set uninformative values for hyperparameters a_0, b_0 and c_0. We can then compute our full model posterior distribution as follows:

$$P(\mathbf{Z}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \frac{\left[\prod_{n=1}^{N}\prod_{t=1}^{T} P(Y_{n,t}|X_{n,t}, Z_{n,t}, \beta^{(1)}, \beta^{(2)}, ...\beta^{(K)})\right] P(\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z})}{P(\mathbf{X})} \tag{3.22}$$

where $\boldsymbol{\beta}, \boldsymbol{\tau}$ denote the full set of parameters over all unique states. Note that the logistic likelihood function does not have a corresponding conjugate prior, making it intractable to conventional variational Bayesian methods. It can however be well approximated by a convex lower bound to the logit function that is widely used [122, 123], and admits a conjugate prior enabling variational inference. Further details of this are given in appendix D.

3.2.4 Correcting for Unbalanced Data

One general issue with logistic regression emerges when data is not balanced; that is, the samples from one class are more numerous than those from another class. This produces a bias in the inferred regression parameters, as the overall likelihood is maximised by assuming one class is more likely than another.

This problem is exacerbated when we introduce a free state variable in the form of the HMM states, as a bias will be introduced anytime a particular state is active over more samples of one class than it is for another. Furthermore, as variational iterations proceed, this encourages more samples to be assigned to the state that biases them most, leading to states that fit exclusively to one class. In

other words, the inclusion of a freely inferred state variable creates a situation where data samples assigned to each particular state may be unbalanced, even if the data over all trials is balanced. This problem is not addressed by the sequential constraint, as this does not impose any equal assignment of states over different classes, it only constrains the global transition structure.

Returning to the trivial example from figure 3.2, fitting our logistic regression decoding model to this data, results within a few iterations in two states that have a mutually exclusive relationship with the two conditions, as in figure 3.5, top panel. The classification boundaries of each state can no longer be seen on the visible plane as their parameters have grown very large - effectively assigning all data points with very high confidence into one of the two classes. This achieves perfect classification accuracy on the test data but cannot generalise to unseen data at all, as the states no longer describe any temporal feature of the data.

Fortunately, this problem can be corrected by rebalancing the weights when computing the regression parameters, as outlined in appendix E. This widely used procedure [124–126] effectively removes the bias by re-weighting the sampled data to reflect class distribution equally.

While the issue of unbalanced data is a widely acknowledged problem for logistic regression, it is nonetheless present for Gaussian regression models also. A similar outcome, where states are mutually exclusively assigned to the given conditions, can be established as a result of very poor initialisation. But whereas the logistic regression likelihood function would drive the accumulation of samples in those states that biased a certain condition, the Gaussian cost function is much more locally stable. In our experience, experiments with real data have yet to show this outcome under any sensible initialisation. Nonetheless, should this problem be encountered with novel data sets it can be trivially dealt with by an equivalent weight rebalancing step.

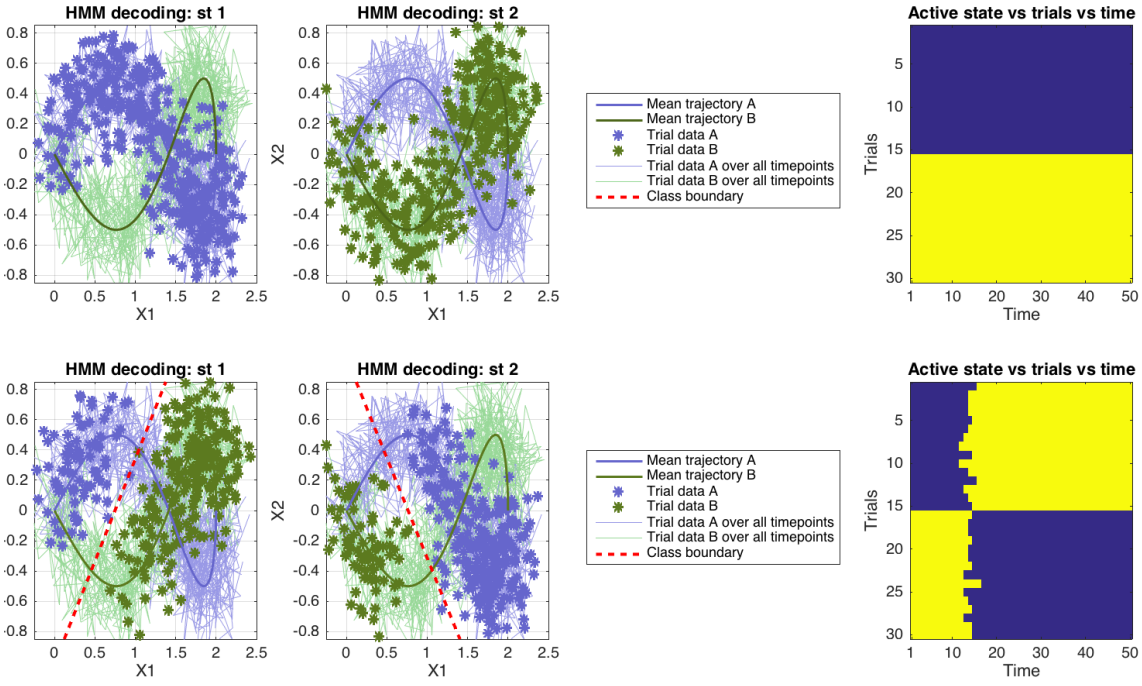


Figure 3.5: Overfitting issues of an unconstrained dynamic decoding model. The upper and lower panels here show the model output converging to two different kinds of overfitting problems; the left side plots the assignment of data points into states and the classification boundaries fit to the data; right side plots the active state vs time on each trial, ordered so trials 1-15 are of class A and trials 16-30 are of class B. Top panel: where there is no data rebalancing adjustment and no constraints on state dynamics, states can be fit exclusively to experimental conditions. This result has achieved a global maximum likelihood solution by setting the classification boundary to be very far from the data in each state, and assigning all points within each state a near infinite likelihood of class assignment. This model cannot generalise to unseen data. Bottom panel: in some cases a logistic regression model may satisfy the likelihood of maximal linear separability of data points by maximal misalignment of state timecourses on trials of different conditions. While this solution classifies the training set well, it generalises poorly to unseen data.

3.2.5 Motivation for Sequence Constraint

In section 3.1.3, we outlined that the primary motivation for constraining the permissible state transitions to sequential patterns was that this reflected our prior belief regarding the nature of much trial-based neural data. This could therefore help extract the most salient features of sequential activity while allowing for trial-by-trial variations in progression through these sequences. Here, we will additionally demonstrate that without the sequence constraint, supervised models can overfit in

ways that will not generalise to unseen test data - even with the above outlined rebalancing adjustment. Imposing sequential transitions therefore goes beyond just reflecting our prior assumptions about the data, and additionally imposes temporal regularisation that ensures the model extracts the most desirable and generalisable characteristics of the data.

Returning to the same trivial example and fitting an HMM decoding model with data rebalancing but with an uninformative prior on the transition matrix (i.e. allowing any state transitions as in the left panel of figure 3.3) results in the inferred parameters shown in figure 3.5, lower panel. The inferred class boundaries, which would ideally match the $X2 = 0$ line, have instead converged to a very different solution. The logistic regression model maximises the linear separability of data in different classes assigned to each state - in this case, maximal linear separability has been achieved by maximal misalignment of states over different classes, which can be seen most clearly in the right panel showing the active state as a function of trial type. The expected state sequence for trials of one class is the exact opposite of that for the other class. While this solution has perfectly classified the data points in the test set, it has not captured useful information that could be fit to new data where the class label is not already known.

A technical note must be made here: in this particular example where the two conditions are exactly opposite in sign, taking the mean of the two state decode parameters would in fact result in a decoding plane that approximates the horizontal axis, as desired. One might also conceive of other ways to fit states to this particular example that generalise to similar test data, however, this particular recourse extends only to this trivial illustrative example; in higher dimensional data simulations where this problem was replicated, both such solutions failed to achieve classification accuracy any higher than chance levels.

This further motivates our sequentiality constraint; it would effectively reduce the probability of state misalignment problems outlined here, while also reflecting our reasonable assumptions about the nature of the data. Returning to our trivial

example, by incorporating both the sequentiality constraint and the unbalanced data correction we recover the sensible state timecourses displayed in figure 3.4.

3.3 Synthetic Data Results

3.3.1 Data Generation

The simulated data used thus far was designed for low dimensional visualisation of the distinct features of the model. We now wish to test the model more thoroughly, and will simulate data of greater complexity for this purpose. We wish to replicate the following characteristics of neural data; that the data follows a certain general dynamic trajectory through state space on each trial; that deviations from this mean trajectory define the difference between different stimuli; and that there is noise present in both the instantaneous measurement but also the phase along these dynamic trajectories. We do this by extending the simple sinusoidally generated data in figure 3.2 to an extended Fourier series with random coefficients across multiple dimensions:

$$X_{n,t} = A(\theta_{n,t}) + Y_i B(\theta_{n,t}) + \epsilon_{n,t} \tag{3.23}$$

$$\epsilon_{n,t} \sim N(0, \sigma^2) \tag{3.24}$$

$$\theta_{n,t} \sim N(\theta_{n,t-1} + c, c\delta) \tag{3.25}$$

where n denotes the trial and t the timing in the trial; $X_{n,t}$ is of dimension $[1 \times P]$, $Y_{n,t}$ is of dimension $[1 \times Q]$, A is $[1 \times P]$ and B is $[Q \times P]$. Note that Y here could be allowed to vary over time within a trial and could be continuously valued in each dimension; however to maintain consistency comparing a Gaussian encoding model and logistic regression decoding model we will constrain $Y_{n,t}$ to be mutually exclusive labels that are static over the course of a trial, with one column containing a 1 for all timepoints of a given trial and all other columns containing 0 for all timepoints in that trial. Critically, on each trial, X now follows a continuous, stimulus invariant trajectory $A(\theta)$, with an additional additive stimulus-specific component $B(\theta)$, with noise present in both the phase increment at each step (in order to simulate between trial variability in the timing of the activity) and the Gaussian measurement noise

across all sensors at each timepoint. We generate both the mean trajectory $A(\theta)$ and each column of the condition-specific trajectories $B(\theta)$ as a Fourier sum:

$$A(\theta_{n,t}) = \sum_{a=1}^{100} \eta_a sin(a\theta_{n,t}) \tag{3.26}$$

$$B_y(\theta_{n,t}) = \sum_{b=1}^{100} \zeta_{b,y} sin(b\theta_{n,t}) \tag{3.27}$$

where each vector $\eta_a \sim N(0, \frac{1}{a}I)$ and each vector $\zeta_{b,y} \sim (0, \frac{R}{b}I)$, where I is the identity matrix of dimension $P \times P$. This replicates the $\frac{1}{f}$ spectrum typical of electrophysiological data whilst allowing us to control how well separated the conditions are through the parameter R. Note also we use a Fourier sum with zero phase offset; this ensures each trial trajectory begins from a common reference point at the origin, modelling a consistent neural initial state with differences between experimental conditions only manifested as time increases. Furthermore, the constant c, that defines the mean phase increment, is defined in the range $0 < c < \frac{\pi}{T}$, where T is the period of the trial. This ensures that the slowest component of the Fourier series has a period more than double the mean trial length, creating data that does not revisit statespace manifolds (i.e. it is not oscillating over the course of the trial). Figure 3.6 plots a sample of two dimensions of this data, demonstrating the mean trajectory generated on the first two dimensions of X for two given stimuli conditions Y.

For the data tested, we used a data dimensionality $P = 10$, stimulus dimensionality $Q = 10$, trial length $T = 30$, condition separation $R = 0.5$, and with variable test parameters measurement noise $0.01 < \sigma < 2.5$, phase increment noise $0.01 < \delta < 10$ (where each range was chosen to capture values over which accuracy goes from near perfect to approximately chance), and number of states $K \in \{4, 8, 16\}$.

Importantly for the results that follow - this entire simulation is based on a continuous statespace generative model. This is in contrast to the discretised statespace model imposed by the HMM. In section 1.3.3 we argued that one of the advantages of HMMs over continuous dynamic models was greater robustness when dealing with poor signal to noise ratios and high dimensionality - even if these

modelling assumptions are a poorer reflection of the true underlying origins of the recorded signals. This simulation allows us to test this assertion directly.

3.3.2 Cross validation testing

We test all data through tenfold cross validation, over twenty different random instantiations using the model outlined above. In each case we separate a test fold from the remaining training data, and infer the posterior distribution over our model parameters using the training data – using the same approach for both model frameworks, for the encoding model we infer the full posterior $P(\boldsymbol{\alpha}, \mathbf{Z}_{train}|\mathbf{X}_{train}, \mathbf{Y}_{train})$ and for the decoding model the posterior $P(\boldsymbol{\beta}, \mathbf{Z}_{train}|\mathbf{X}_{train}, \mathbf{Y}_{train})$.

Unfortunately, as outlined in Vidaurre et al. [33], the fact that our model fits state timecourses $\mathbf{Z}$ conditioned on the trial labels $\mathbf{Y}$, leaves us with no direct means to infer the corresponding test set state timecourses without violating the principle of separation of training and test datasets. We however take a different approach to Vidaurre et al. [33], and propose a means to estimate trial-specific state timecourses in an unbiased manner using an equivalent unsupervised learning HMM independent of the test labels. Once again we benefit here from the sequence constraint, as the temporally regularised state trajectories are easier to estimate than unconstrained state timecourses used elsewhere.

3.3.3 State Timecourse Estimation by Equivalent Unsupervised Observation Model

Specifically, having inferred a set of state timecourses conditioned on the neural data and condition labels, $P(\mathbf{Z}_{train}|\mathbf{X}_{train}, \mathbf{Y}_{train})$, we now try to estimate an *equivalent unsupervised model*. Qualitatively, we pose the question: if the state timecourses already inferred had been derived by an unsupervised observation model, how would the parameters defining each state be distributed? This effectively provides a reasonable means of estimating the state timecourses independently of the labels, given some hyperparameters to be learned from the data:

$$Q(\mathbf{Z}|\mathbf{X}, \mu_1, \Sigma_1 ... \mu_K, \Sigma_K) \approxeq P(\mathbf{Z}|\mathbf{X}, \mathbf{Y}) \tag{3.28}$$

Setting these hyperparameters to describe a Gaussian distribution over $\mathbf{X}$ with parameters unique to each state, we can then infer their marginal values given the training data and state timecourses fit with the supervised model using variational Bayesian inference:

$$P(\mu_1, \Sigma_1 ... \mu_K, \Sigma_K | \mathbf{X}_{train}, \mathbf{Z}_{train}) \propto P(\mathbf{X}_{train}, \mathbf{Z}_{train} | \mu_1, \Sigma_1 ... \mu_K, \Sigma_K) P(\mu_1, \Sigma_1 ... \mu_K, \Sigma_K) \tag{3.29}$$

$$\propto \Big[\prod_{n=1}^{N} \prod_{t=1}^{T} N_{X_{n,t}}(\mu_{Z_{n,t}}, \Sigma_{Z_{n,t}}) \Big] P(\mu_1, \Sigma_1 ... \mu_K, \Sigma_K) \tag{3.30}$$

where the hyperparameters have ARD priors of the same form specified in equation 3.10. Importantly this inference is independent of the regression parameters so can be applied interchangeably to encoding and decoding frameworks; furthermore it is entirely compatible with our dynamic priors over permissible state sequences, allowing us to estimate sequential state timecourses without recourse to the labels:

$$Q(\mathbf{Z}_{test} | \mathbf{X}_{test}) \propto \mathbb{E}_{P(\mu_1, \Sigma_1 ... \mu_K, \Sigma_K | \mathbf{X}_{train}, \mathbf{Z}_{train})} \Big[P(\mathbf{X}_{test} | \mathbf{Z}_{test}, \mu_1, \Sigma_1 ... \mu_K, \Sigma_K) P(\mathbf{Z}_{test}) \Big] \tag{3.31}$$

where $\mathbb{E}_{P(X)}$ is used to denote the expectation with respect to the probability distribution $P(X)$. From here, a classification output over labels can be computed either explicitly (for the decoding model) or through maximum likelihood (for the encoding model). This procedure is unbiased as the test set labels are never used in making the prediction. Compared to the previous approach of Vidaurre et al. [33], we now have trial specific estimates of state timecourses without recourse to bias removal methods.

3.3.4 Encoding Model Results

The results of the encoding model fit to the data are summarised in figure 3.6. Our model allows us to compute the likelihood of each trial label as a function of time over the test data; we then compute accuracy as the proportion of timepoints over which the maximum likelihood label matched the true value. When averaging over the noise parameters and plotting against time, our dynamic encoding model outperforms

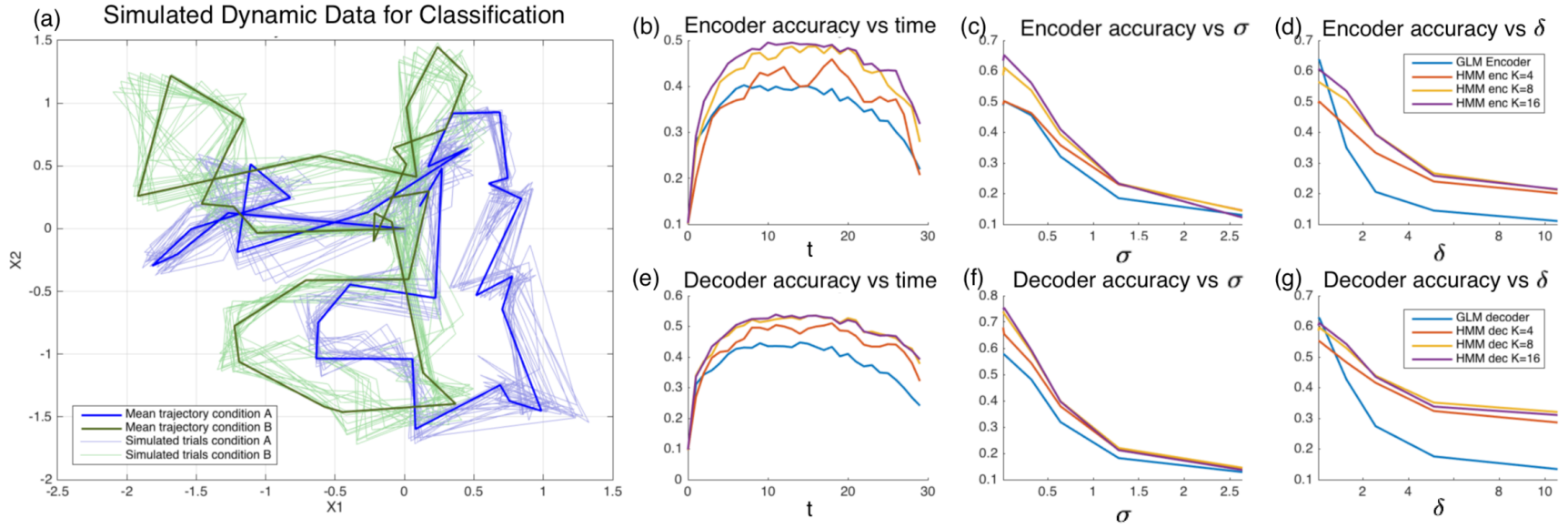


Figure 3.6: Results of analysis on synthetic dataset. (a) Two exemplar dimensions of the data, demonstrating the randomly generated trajectories of two conditions, and the noisy trial data generated belonging to each of those conditions. For these trials, $\sigma = 0.02$, $\delta = 0.08$ (b) Encoder accuracy as a function of time, averaged over σ and δ values; (c) Encoder accuracy as a function of σ, averaged over time and δ values; (d) Encoder accuracy as a function of δ, averaged over time and σ values; (e) Decoder accuracy as a function of time, averaged over σ and δ values; (f) Decoder accuracy as a function of σ, averaged over time and δ values; (g) Decoder accuracy as a function of δ, averaged over time and σ values.

the synchronous GLM model fitted to individual timepoints. Increasing the state resolution increases the performance, but the model nonetheless outperforms the synchronous GLM model even for the lowest number of states tested, $K = 4$. As anticipated, the synchronous GLM model's performance degrades as time increases, due to the poor alignment of data as phase noise accumulates further from the origin.

As a function of measurement noise, all models have an exponentially decreasing performance curve. There appears to be a slight relationship between the number of states K and the rate of performance decay; the model with the lowest number of states has the poorest performance for the lowest noise condition, but the best performance under the highest levels of noise simulated. This can be understood through the degree of temporal regularisation, with fewer state models having fewer parameters to fit complex trajectories where noise is low, albeit being more heavily regularised to still extract the main features in high noise environments.

On the other hand, increasing the amount of between-trial variability in the timing of the activity, i.e. the phase noise δ, dramatically degrades model performance, with a synchronous GLM exhibiting much more rapid decay than the HMM models. Notably, where phase noise is minimal, the synchronous GLM outperforms all the HMM models; however as soon as significant amounts of phase noise are introduced this relationship is reversed. The HMM models continue to perform quite a significant amount above chance (10% being chance level) after the synchronous GLM performance has dropped to chance levels.

3.3.5 Decoding model results

As the data labels $\mathbf{Y}$ are mutually exclusive, we can arbitrarily use the same data in a logistic regression decoding model to dynamically classify the data, results of which are also shown in figure 3.6. Our model predicts the log odds of each trial label as a function of time over the test data; we then compute decoding accuracy as the proportion of timepoints for which the highest log odds trial label matched the true value. We see an improvement in classification performance by

applying logistic regression, which is unsurprising given the model is explicitly formulated for classification tasks.

Similar to the encoding model, we observe performance that decreases exponentially with either the measurement noise σ or the phase noise δ, and a similarly very modest trend between number of states and the rate of decay. Comparatively however, the decoding model seems much more robust as phase noise increases, with performance deterioration only marginal over the highest levels of phase noise simulated.

3.4 Discussion

3.4.1 Sequential Models are more Robust to Temporal Misalignment

The clearest conclusion from this analysis is that, as expected, sequential models are more robust to the temporal misalignment introduced by phase noise. This is the source of noise that these models explicitly characterise, hence it is unsurprising that they perform better than the standard synchronous models that ignore this source of variation. Nonetheless, the speed of degradation of the standard synchronous models in figure 3.6 (d) and (g) is somewhat striking, and highlights the potential of these models to address this often overlooked source of variation in recorded data.

3.4.2 Classification using Encoding and Decoding Models

We have used classification models based on both encoding and decoding frameworks. The decoding models explicitly classify data based on a logistic regression distribution; in the case of the encoding models, classification is performed based on the likelihood of the inferred Gaussian distribution of each class. This latter form of classification is equivalent to Linear Discriminant Analysis (LDA); notably while this chapter has gone by the more neuroscience oriented title of *Sequential Encoding and Decoding Models*, we could equivalently have framed it in machine learning terminology as *Sequential LDA and Logistic Regression Classifiers*.

Although both LDA and logistic regression classifiers fit a linear plane that separates classes, the results obtained with simulated data are significantly different; encoding models did not perform as well under increasing phase noise or for low numbers of states, e.g. $K = 4$. The reasons for this lie in the assumptions imposed upon the data; while logistic regression models assume that data is linearly separable, LDA models assume the data in each class has a Gaussian distribution, which is generally considered to be a much stronger assumption. If the data is well described by a Gaussian distribution, then LDA can perform better than logistic regression models; however where this assumption is poor - for example where the continuous data is forced to be explained by a low number of discrete states - LDA is known to perform more poorly than logistic regression [127].

3.4.3 Discretised Statespace Models can Explain Continuous Data

HMMs that discretise the statespace here perform well despite using data generated from continuous trajectories. This supports our original argument in favour of HMMs, that they enforce an interpretable low dimensional statespace that is robust over poor SNRs, and emphasises that the underlying data does not itself need to be discretised for these models to be useful.

3.5 Conclusion

We have proposed a novel framework for encoding and decoding models that broadens and extends the original concept of Vidaurre et al. [33]. In applying this to simulated data, we have demonstrated that the standard approach of aligning trials in time results in significant performance degradation where there is temporal misalignment over trials. These sequential models are much more robust over this source of variation, and we now turn to applying them to real data to fully explore their benefits.

Chapter 4

Mapping Sequential Visual Stream Activation in MEG

Contents

4.1 Introduction 74
4.2 Experimental Setup 77
4.3 Results: Overall Classification Accuracy 79
4.4 Decoding Model Results 81
4.5 Encoding Model Results 85
4.6 Conclusion 88

Having established a methodological framework and demonstrated its advantages on simulated data, we now turn to real MEG data to explore its potential further.

4.1 Introduction

4.1.1 Visual Stream Processing

For a number of reasons, the human visual system is an ideal candidate for testing these models. MEG has reasonably high sensitivity over visual cortex areas, enabling relatively high levels of decoding accuracy that has made visual stimuli one of the dominant paradigms of MEG decoding studies [88, 128–131]. Furthermore, a number of localisation studies and computational models have contributed to a relatively advanced understanding of visual system pathways and timings of activation, providing us with a solid theoretical baseline to which we can compare our results [88, 129, 132, 133].

The very first visual information is understood to arrive in V1 approximately 60msec following stimulus onset [88]. This information corresponds to relatively low level visual features, akin to the edges of lines in an image [134]. From V1, the visual system can be very loosely modelled as a feedforward architecture, with the low level features being fed forward through a hierarchy of stages designed to identify successively higher-order visual features. From a first level that detects edges, successive stages of the visual hierarchy may for example recognise basic shapes; then subsequently group a set of shapes into some object (as in figure 4.1 (a), where successive stages piece together a whole letter from component lines). This model has most famously been used as a basis for convolutional neural networks that have found widespread application in automated image classification [135, 136].

While this model helps us to broadly understand the visual hierarchy conceptually, it is necessarily a gross abstraction from the underlying biology. Anatomical studies long ago confirmed multiple feedforward and feedback pathways in the visual stream[138], as well as numerous top-down forms of modulation such as attention and other biases [4]. While the timing and function of specific feedback pathways are a very active area of research [139], certain broad gradients are nonetheless discernible. Namely, from lower visual areas information is projected along a ventral stream through lateral visual cortex and towards inferior temporal cortices [140], arriving approximately 250msec post stimulus onset [132]; this is generally associated with the semantic content of the image. A second dorsal pathway projects information to parietal cortical areas; this is more generally associated with the spatial content of an image [140]. Together the visual ventral and dorsal streams are often referred to as processing respectively the 'what' and 'where' aspects of visual sensory input.

Relating this to MEG decoding, numerous studies now have established visual system peaks in decoding around 150msec post stimulus onset [88, 130, 132], at which point visual information is understood to be moving through lateral and dorsal visual areas, but not yet in temporal or parietal cortices. This provides us with a solid basis to which to compare our results.

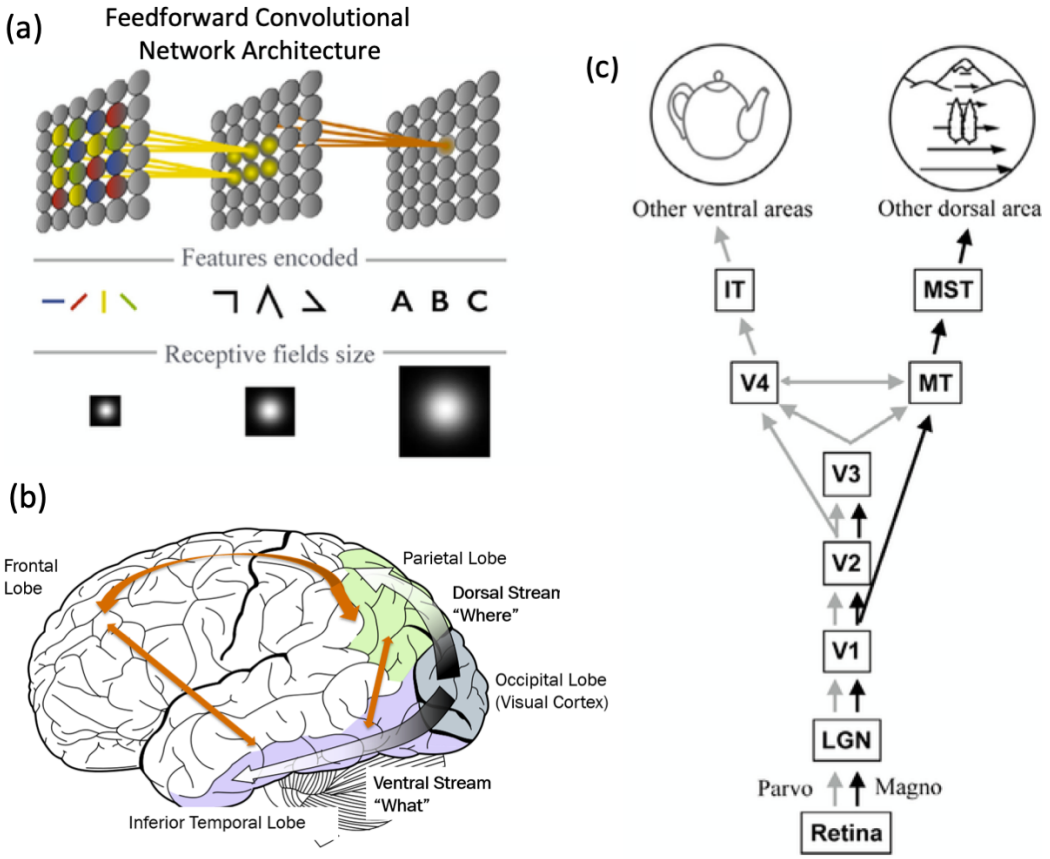


Figure 4.1: Feedforward model of the visual stream: (a) Feedforward neural networks are loosely modelled on visual hierarchical processing, where successive stages of the hierarchy detect increasingly complex features. (b) This is only an abstract model of the human visual system, which combines top down modulation and multiple feedback processes. Importantly however, the human visual hierarchy has two major pathways, the ventral and dorsal streams. (c) To a first approximation, these architectures can be unified in the model shown, where lower level features propagate forward through visual areas before separating into the ventral (semantic, or 'what' content) and dorsal (spatial, or 'where' content). Image adapted from [137].

4.1.2 Motivation: Replay in the Visual System

Both the encoding and decoding models are in this chapter applied to MEG data collected by Liu et al. [31]. This study will be reviewed much more fully in chapter 5, but was focussed on the detection of replay, the spontaneous sequential reactivation of the neural representations of previous experience.

While this phenomenon has been extremely well studied in rodent models,

where direct electrophysiological recordings of multiple cells have enabled robust identification and extensive investigation of when it occurs, replay has so far only been observed in a handful of human MEG studies - namely Kurth-Nelson et al. [99] and Liu et al. [31][1]. The relatively low accuracy of MEG classification methods has yielded results that are statistically significant albeit not sufficiently validated to promote rapid progress in the scientific study of this phenomenon, which will form the central focus of chapter 5. Here, we use the functional localiser data from this study as a testbed for the new methods developed, to see whether classification improvements by enhanced sensitivity to temporal variation might improve the detection and characterisation of replay.

4.2 Experimental Setup

Task: For brevity we repeat here only the details pertaining to the portion of the experiment (i.e. the functional localiser session) that these methods have been applied to. A total of $N = 21$ participants were trained and scanned in the task. Subjects were trained to recognise eight different visual stimuli selected to elicit diverse neural responses. Subjects fixated on a cross on screen; when stimuli were presented they were asked to identify if the stimulus was correctly oriented by pushing a button. Stimuli were presented in a random order with a total of 20 presentations of each stimulus.

Data processing: Data was processed using OHBA Software Library (OSL) and Statistical Parametric Mapping (SPM). Data was sampled at 600Hz on a CTF scanner, bandpass filtered between 0.5Hz and 50Hz and downsampled to 100Hz. Artefact rejection was done using ICA. Data from trials where stimuli were inverted were discarded; data from correctly oriented trials were aligned to stimulus presentation. Data were epoched to a window from 4 seconds prior to stimulus presentation to one second post stimulus. We then applied unbiased trial

[1]To disambiguate terminology, some studies have used the term 'replay' to describe a distinct (but possibly related) paradigm where participants are cued to visualise or recall a past learned stimulus. In this context however, we are referring only to the uncued, spontaneous reactivation of stimulus representations in specific task relevant sequences, which we posit is a very distinct phenomena - and for which the cited work is the only known evidence.

normalisation, whereby data are normalised with respect to a mean and standard deviation computed only on trial data preceding stimulus presentation. This avoids introducing a bias as the mean and standard deviation from the pre stimulus period contain no information about the channel signals. Dimensionality reduction was performed by PCA, taking the first 55 components which represented on average 90% of the variance in the data; this was similarly done in unbiased manner, with principal component axes determined by the covariance only of data up to stimulus presentation, and not beyond. The trial data was then epoched again to a period of analysis from 0-500msec following stimulus presentation.

Classification: This data was analysed by both encoding and decoding frameworks. Models were fit to each subject separately using tenfold cross validation. Encoding was then applied using a conventional temporally aligned, synchronous Gaussian GLM model; and subsequently using the Gaussian HMM encoding model outlined above, for which we estimate the state timecourses for the test set using either 1) the state timecourses averaged over the training trials, or 2) state timecourses estimated by applying an unsupervised HMM with fixed equivalent unsupervised observation models on the test data (as introduced in section 3.3.3). Trial timepoints were marked as correct if the maximum likelihood stimulus given the model parameters matched the true stimulus, otherwise were marked as incorrect. Mean accuracies were computed per subject, then a mean accuracy computed over all subjects.

Decoding was implemented using a conventional temporally aligned, synchronous decoder using logistic GLM with lasso regularisation (regularisation penalty $\lambda = 0.01$); then with the dynamic HMM logistic decoder outlined above. Lasso regularisation was chosen specifically for comparison - despite being different to the ARD prior used in the HMM observation model - as it has been the primary classification method used in comparable replay studies [99]. The effect of this choice is explored more fully in Appendix G and Chapter 5. The regularisation penalty was chosen as the value that achieved highest classification accuracy over the range of values tested. Comparisons with L2 regularised logistic regression and with linear support

vector machines (see appendix G) demonstrated that these produce similar results - slightly higher decoding accuracy in the case of L2 regularised logistic regression, but still below the levels achieved by the dynamic HMM logistic decoder.

Again, this model was fit with two different ways of estimating the state timecourses for the test set, i.e. using either 1) the state timecourses averaged over the training trials, or 2) state timecourses estimated by the equivalent unsupervised modelling method of section 3.3.3 on the test data. Test trial points were marked as correct if the logistic classifier assigned the highest probability to the true stimulus, and marked as incorrect otherwise.

4.3 Results: Overall Classification Accuracy

The accuracy of each model vs time is plotted in figure 4.2. Accuracy was computed through tenfold cross validation; as discussed earlier in section 3.3.2, there is no direct way in which to fit state timecourses to unseen test data, as these timecourses are conditioned on the labels. We can however estimate the state timecourses using the equivalent unsupervised model; or just take the mean state timecourse computed over all training trials and use that on each test set trial. We assess two measures of accuracy, one being the time specific mean classification accuracy over subjects; the second being how well the classifiers trained around one timepoint generalise to adjacent timepoints [141].

In each case a higher classification accuracy is achieved by the best performing dynamic statespace methods than by the equivalent synchronous method, although this difference is much greater in the case of decoding models than encoding models. Somewhat surprisingly, where test data state timecourses were estimated on a trial specific basis using the equivalent unsupervised observation models, the classification accuracy was still not as good as when these test data were fit using just the state timecourse averaged over training trials. There are two potential reasons for this; firstly, it may be that the method of estimating state timecourses by the equivalent unsupervised observation models is a poor estimate of the actual state timecourses in the full model. Alternatively, it could be that the trial-averaged state timecourses

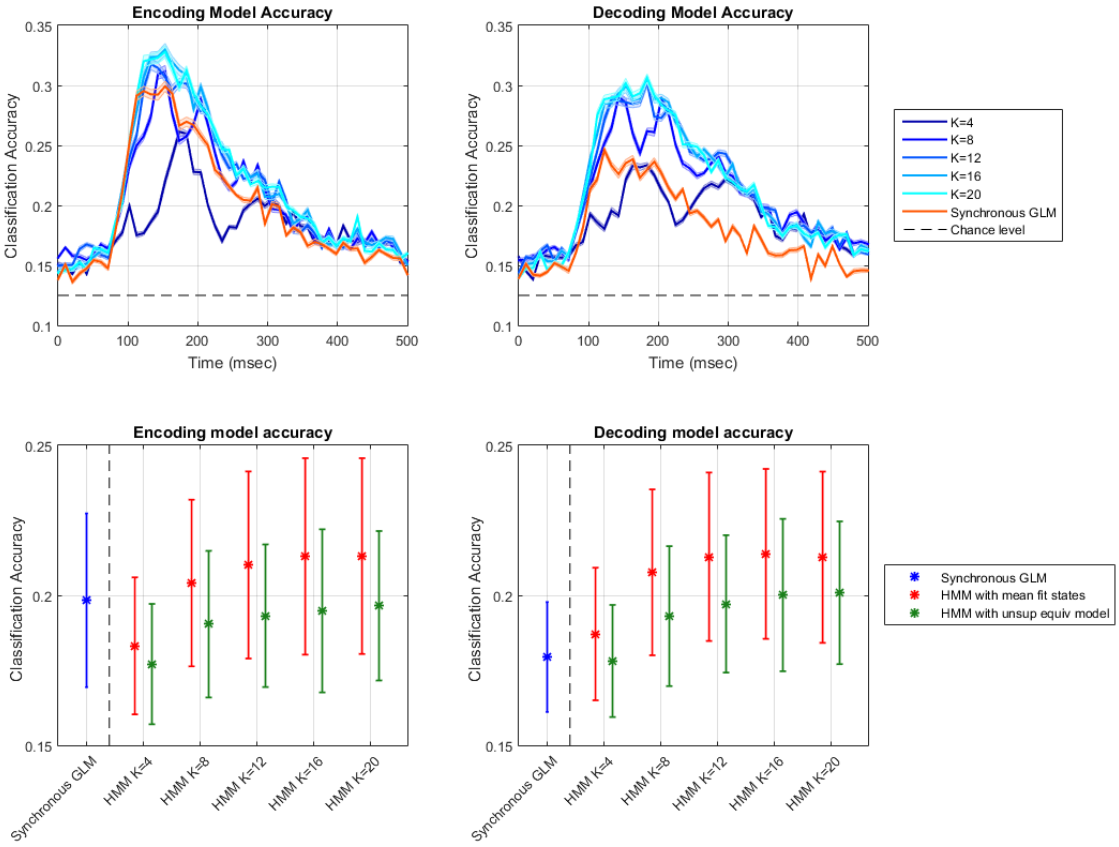


Figure 4.2: Accuracy plots on MEG data. Left: Encoding Model classification accuracy computed as a function of time, where $t = 0$ denotes stimulus onset (top panel; mean +/- s.e. over subjects) for different values of K, plotted using mean fitted state timecourses; and as function of classifier parameters and state fitting method (bottom panel; mean +/- s.e. over subjects and timepoints). Right: Logistic decoding model accuracy as a function of time (top panel); and as function of classifier parameters and state fitting method (bottom panel).

apply a smoothed linear combination of decoding parameters over time - this temporal smoothing may in fact better represent the data than the hard, discrete state changes fit in the estimated state time courses.

As a function of K, performance increases with more states, plateauing for approximately $K = 12$. This is notable given such a model has significantly fewer parameters than the equivalent synchronous GLM model. It is also significant that models with few states nonetheless achieved nearly equivalent classification accuracy to higher state models for $t > 300$msec; this could reflect the emergence of a more stable neural representation for these timepoints that does not require

rapidly changing states to optimally decode.

The encoding model achieves a comparable accuracy to the decoder. Over the period from 100-200msec post stimulus, the encoding models perform slightly better, whereas over the period after 200msec post stimulus the decoding models perform slightly better. This may reflect characteristics of the data itself; noting that our synthetic data tests in section 3.3 demonstrated greater resilience of the decoder to between-trial variability in the dynamics, this may suggest there is poorer alignment of data further from the stimulus onset, with accumulated phase delays being better decoded by the logistic regression based methods. The high performance of the encoding models however remains slightly surprising for a classification task. The advantage of this result is that encoding models allow a direct interpretation of encoding parameters over sensor space, something that is obscured with decoding models, where regression parameters over space reflect both signal and covarying noise components [118]. This is an advantage we explore more fully in section 4.5.1.

4.4 Decoding Model Results

4.4.1 Generalisation of decoders

The time-generalisation curves of figure 4.3 outline how the model parameters, trained on a particular timepoint, cross generalise to test data from a separate timepoint [141]. For adequate comparison with the conventional synchronous decoding models, the x axis denotes the time of training (for which, in the case of our HMM state based model, we use the state timecourses averaged over training trials to identify the appropriate state-specific decoding model) and the y axis denotes the time of test data. In comparing the two models, it is apparent that the HMM state-based decoder achieves greater accuracy on the diagonal, but also that it achieves this with a weaker gradient, demonstrating greater generalisability over adjacent timepoints. This is unsurprising given that the model is temporally regularised to find state parameters that by necessity must fit multiple timepoints. Furthermore, a weak rebound effect on the off-diagonal, where decoding is significantly less than chance, is more evident in the dynamic

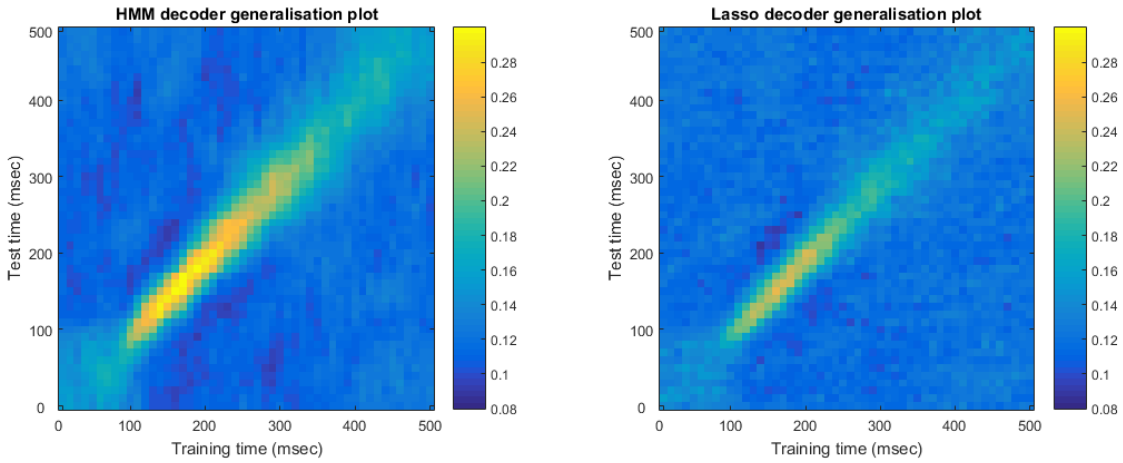


Figure 4.3: Decoding model generalisation plots; one common analysis technique in the decoding literature is to plot how each classifier generalises over time, by plotting the accuracy obtained using classifiers trained on one timepoint (x axis) and tested on the test data from all other timepoints (y axis). The HMM decoding model (left) demonstrates a stronger and broader pattern of decoding accuracy along the diagonal compared to the synchronous GLM model (right), where $t = 0$ denotes the stimulus onset time.

model generalisation plot than the GLM plot. This is consistent with previous studies in visual cortex dynamics that observe a repolarisation effect with about 50msec delay on channels associated with stimuli [88].

4.4.2 Decoding Accuracy as a Function of Frequency Content

We can compare the frequency domain sensitivity of our new HMM decoder/encoder to the conventional synchronous GLM approach. The results discussed in this section refer exclusively to the decoder model fit, although similar results can be obtained using the encoding approach. We wished to test the hypothesis that by explicitly aligning points in time and training only on samples from exactly that time step, that the conventional synchronous GLM approach was restricting the frequency content of information accessible to it. The results shown here are derived using the same cross validation analysis as above, the only difference being in the final classification of test data, where we substituted bandpass filtered data X. All model parameters (state timecourses and regression coefficients) are trained on unfiltered data, it is only the test set which is bandpass filtered. For computational speed,

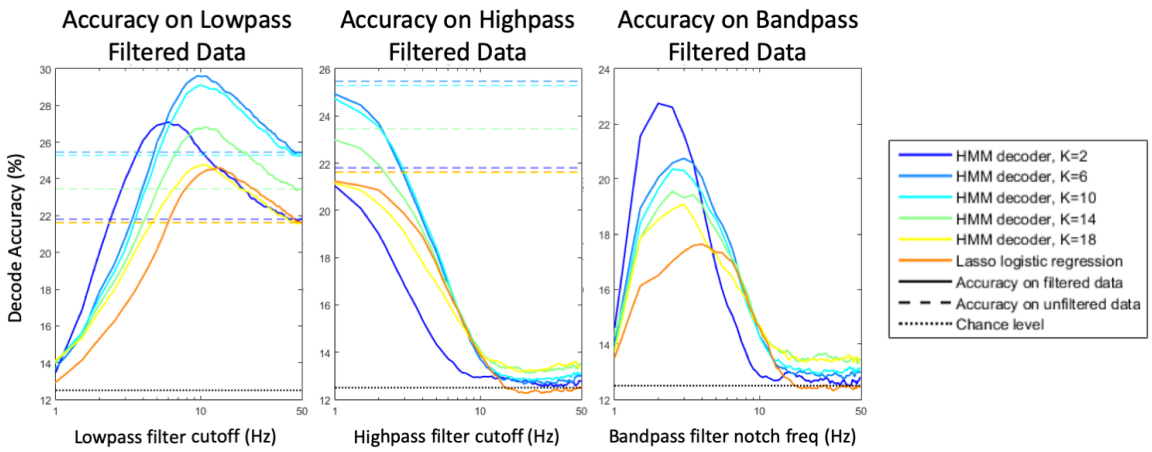


Figure 4.4: Decoding accuracy as function of frequency content. Left panel: mean accuracy of a range of different models when test data is lowpass filtered, demonstrating dominance of low frequency information content and increased sensitivity of low state number models to this content. Middle panel: mean accuracy when test data is highpass filtered further emphasises dominance of low frequency information content; also demonstrates existence of high frequency information content that can only be decoded by dynamic models with higher numbers of states. Right panel: mean accuracy on bandpass filtered test data (with a notch width of 2Hz) similarly demonstrates selectivity of different choices of K for frequency of information content.

the results shown were computed over the limited time window of 100-300msec post stimulus, with accuracy averaged over this window.

We believed that by varying the number of states K we may alter the frequency sensitivity of the different decoders, as - by virtue of imposing a sequence of HMM states - each state has an average lifetime that is roughly inversely proportional to K. Results demonstrate that this is indeed the case, with low pass filtered data below $10Hz$ demonstrating quite a clear delineation in accuracy as a function of the number of states K. Low state models pool over more data points, but are consequently poor at detecting transient, higher frequency content in the signal. Models with higher numbers of states are better able to detect transient shifts, but less able to detect low amplitude, low frequency fluctuations spread over many adjacent samples. In this case this trade-off leads to an optimal value overall for K of approximately 6, where the balance results in the best overall accuracy.

The bandpass filtered data results reflect a spectrum of information content

that very roughly follows the $1/f$ spectrum typical of electrophysiological data. Notably in high frequencies, despite the much poorer quality of decoding there is nonetheless high frequency gamma band information content that demonstrates a similar delineation of accuracy by the number of states, where models with more states do a better job of decoding these (chance level is shown as 12.5%; only the $K = 14$ and $K = 18$ state models had significantly greater than chance accuracy at $p < 0.05$). Notably, any improvement in decoding accuracy in this frequency band was entirely dependent upon trial specific alignment of state timecourses, with synchronous methods performing at chance levels for all content $> 20Hz$. We can further demonstrate that the performance of the HMM decoder models on this frequency band is entirely dependent on having trial specific state timecourses, by running the same analysis using only the mean state timecourses over all trials (see appendix G.2).

This proves that the traditional approach of assuming synchronicity over trials and aligning dynamic data by time, biases methods towards specific frequency bands, omitting information at lower and higher frequencies. We can demonstrate that there is decodable information at low frequencies - and emphasise the fact that a model with only $K = 2$ states nonetheless outperforms the traditional approach, by detecting this lower frequency information content. This should challenge certain claims often made by temporal generalisation studies that link poor generalisation of decode coefficients over time as evidence of highly transient neural representations - this could merely be an artefact of the frequency bands biased by the method. Further, high frequency content, despite being low in power in the overall signal, is here shown to be significantly decodable only after trial specific alignment of activity, without which it cannot contribute in the traditional decoding frameworks.

4.5 Encoding Model Results

4.5.1 Encoding Model Extracts Interpretable, Behaviourally Salient Parameters

We now turn our attention to the explicit interpretation of the inferred model parameters, focussing now on the encoding model. Figure 4.6 plots the trial specific state timecourses inferred over all subjects for an encoding model with $K = 8$ (this parameter was chosen for ease of visualisation compared to higher state models; all results discussed are generalisable to higher state models). Variations in trial specific state sequences can be understood with reference to the mean state timecourse, where each state has a clear time at which it is most likely to be active, but where each state also exhibits significant trial-specific variations in exact state activation timings. States fit to time periods where $t > 200$msec seem to show a broader distribution, possibly reflecting the result of accumulated between-trial variability in the timing of activity as you move further from the timing of stimulus presentation.

As further evidence that the extracted states have captured the evolution of true underlying physiological changes, we investigated their relationship with behavioural measures in the tasks. The reaction time of participants (the length of time between stimulus onset and button push) was tracked over trials and correlated with the probability of each state being active at time t. As shown in the bottom left panel of figure 4.5, there is a strong diagonal pattern visible. Note that blue areas denote negative correlation, implying that if the corresponding state is active at that time then it is consistent with a reduced reaction time, whereas yellow denotes delayed responses. Trials where state timecourses peaked slightly ahead of the global mean for that state were correlated with reduced reaction times; whereas those trials where the same states were active later than their mean activation times resulted in slower participant responses. This is strong evidence that the state dynamics are a meaningful description of underlying neural processes and how they vary on a trial by trial basis. It is particularly notable given that reaction times were not a relevant or salient task variable - there was no instruction or motivation given

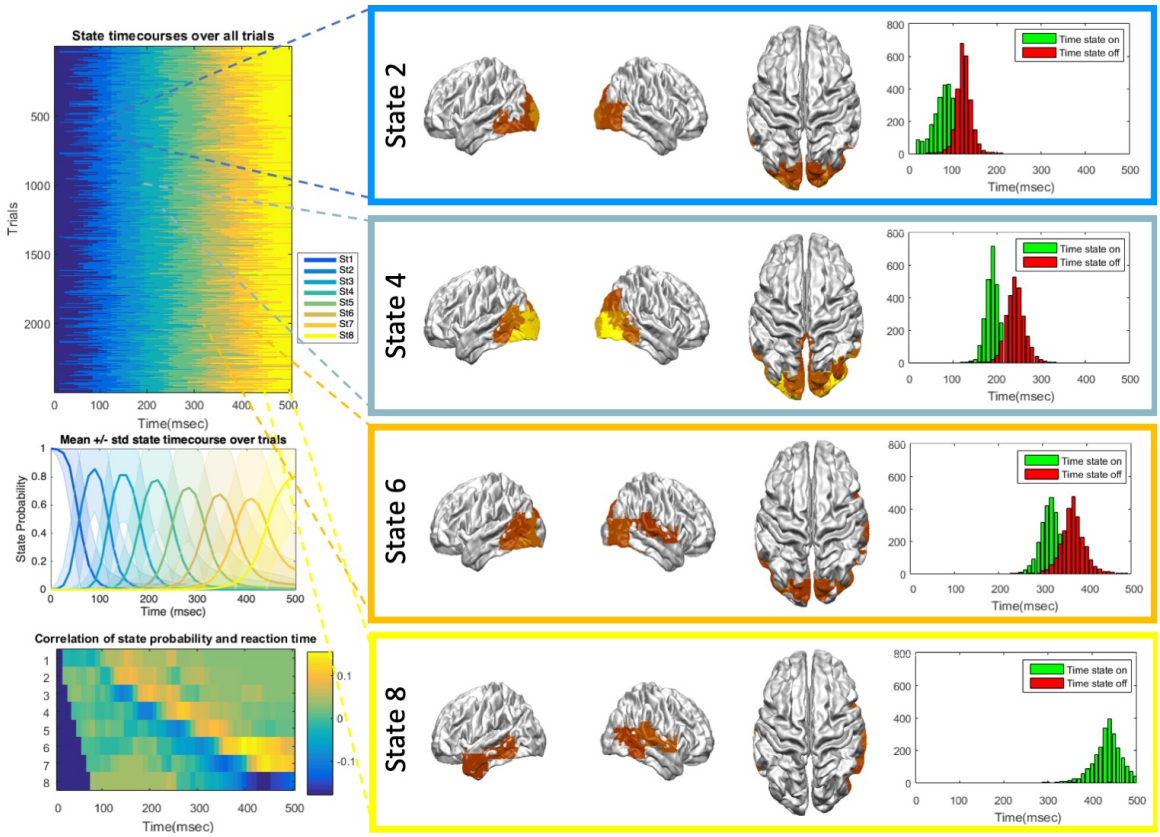


Figure 4.5: Inferred encoding model parameters, $K = 8$. Left: state timings plotted on a trial specific basis demonstrate variations in activity over different trials (top); these variations correlate with the participant reaction time (bottom). Right: spatiotemporal distributions of each state - after projecting the model into source space, the expected variance over different conditions was computed over subjects and is plotted here; histograms plot the timings when each of these states are activated and deactivated over the set of trials.

to the participant to respond quickly, variation in responses simply arose due to passive variations in neural processing over trials, demonstrating the importance of these variations and further validating the approach.

An advantage of having encoding models that equivalently classify the data is that encoding model parameters are more directly interpretable as a function over space than their equivalent decoding model parameters [118]. As discussed in section 4.1 the dynamics of visual stream object recognition have been well characterised [88, 129, 132, 133] which gives us a strong prior expectation over where information should be present in the signal. As a function of time, we expect initial information to be focussed around V1, then in subsequent states

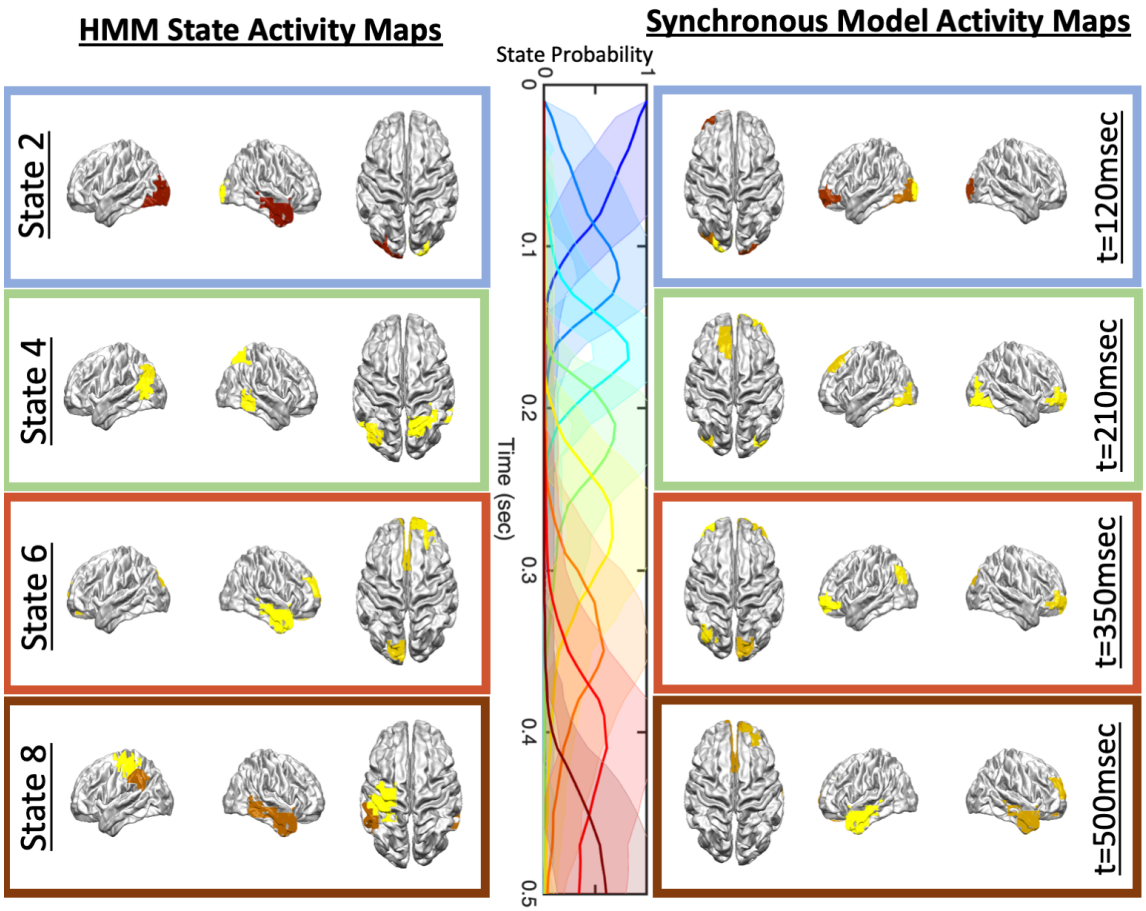


Figure 4.6: Subject level activity map comparison: spatial map of the variance over inferred stimulus encoding patterns for an example subject using an HMM dynamic encoder (left), compared to the equivalent synchronous model maps (right) locked to each state's peak activation time. Agreement between the maps is relatively poor. Previous literature suggests these timings should respectively correspond roughly to lower visual activation; higher visual areas; inferior temporal areas; and fronto-temporal areas. The HMM maps seem to match this progression more closely, albeit still with significant noise as may be expected at subject level.

progress more laterally. Information content in lower sensory areas may nonetheless be expected to be re-engaged at later timepoints by recurrent projections. As such information takes approximately 300msec to reach prefrontal cortex [133], we would not expect states focussed on processing prior to this time to show significant activation in frontal regions.

Figure 4.5 also plots the expected variance over the different classes of the encoding model distribution, per state, projected into source space. At the group

level we observe a clear progression of information along the ventral visual stream (figure 4.5). Reconciling this with the times of each state activation, we see that information content is concentrated in visual cortex areas up until approximately 300msec post stimulus onset, when it starts to progress towards temporal areas, arriving in lower temporal areas around 450msec post stimulus onset. In particular, these maps suggest quite localised representations right up to the later stages of the visual hierarchy, which previous authors have argued is more distributed and includes lower level visual areas [131, 132]; it is possible that these conclusions resulted from the poorer temporal alignment in these later representations.

We can also compare how this looks at the level of an individual subject, as in figure 4.6, and compare it to the equivalent synchronous model at each state's peak activation timing. Correspondence between the two methods is fairly poor; notably the maps for individual timepoints have power concentrations in areas such as frontal sensors that are more likely to represent artefacts than true signal, whereas the maps for the HMM encoding model roughly follows our expectation of visual stream activation.

4.6 Conclusion

The highly dynamic nature of neural data presents a challenge to the most commonly used linear regression based methods of data analysis, which have no explicit way of recognising dynamic patterns and rely on assumptions of data alignment over trials. By introducing an HMM framework with explicitly sequential structure, we have demonstrated how these methods can be augmented to align data dynamically in time by identifying common sequential characteristics, allowing for much greater accuracy in encoding and decoding models even while significantly reducing the number of parameters used. Both on simulated data and with real data from a MEG experiment, we have shown that both encoding and decoding models can benefit from this approach, in either a gaussian regression or logistic regression model for classification. Furthermore, we have demonstrated that this increase in accuracy is due to a heightened sensitivity to information content at different frequencies and

greater robustness to noise by pooling over multiple timepoints. We propose that this framework could be more widely used for the analysis of neural data.

Chapter 5

Exploring Replay Dynamics with Sequential Decoding Models

Contents

5.1 Introduction 91
5.2 On the Identification of Replay in Human MEG Studies 99
5.3 A Survey of Classification Methods and Sequenceness Tests 106
5.4 Temporal Patterns in Replay Activation over Different Areas 114
5.5 On the Relationship between Sparsity and Replay Detection 120
5.6 Conclusion 126

Replay is a particularly well studied phenomenon in cellular neuroscience, where individual cells known to encode higher level features of the environment are reactivated in sequential sweeps that rapidly replay recent learned experience. This mechanism is believed to be important for memory consolidation and building internal models of the world. Recently it has been shown that replay of visual information in humans can be decoded in MEG [31, 99], with rapid sequential reactivation of the representations of learned stimuli occurring spontaneously in rest periods. These sequences were identified using the traditional synchronous decoding approach discussed in the previous chapter; while finding a statistically significant result, a number of unresolved questions remained. Specifically, replay was identified using classifiers trained at one specific time point (200ms post image

presentation) of the full visual stream response; this specific component was found to reactivate in the sequential pattern that defines replay, but the finding generalised poorly to classifiers trained on other timepoints. It was also reported that sparse classifiers were able to identify replay where other classifiers were not, a result that had no theoretical justification in light of sparse classifiers' poorer decoding accuracy as shown in the previous chapter.

These findings warranted a thorough investigation of the methodological approaches to replay detection. We review a number of different classifiers to show that HMM-based methods that utilise sparse L1 regularisation infer broader patterns of replay across multiple stages of the visual hierarchy. We then analyse the temporal relationships between the same sequences activating in different visual areas, and show that these reverse the original feed-forward direction of information transfer, now activating in top-down sweeps through the visual hierarchy. Finally, we return to the question of why sparsity improves replay inference, and show through simulations that this is a serendipitous effect of the reversal of replay activation; sparse classifiers are more robust to the effects of such temporal order reversals.

5.1 Introduction

In 1994, Wilson and McNaughton [142] made the seminal discovery that hippocampal place cells in sleeping rats replay the specific patterns of activation reflecting that animal's recent wakeful experience. Having initially postulated that this represented a mechanism of memory consolidation, decades of research since have shown that the authors had in fact stumbled upon a phenomena significantly more widespread and fundamental to neural function than their original research had anticipated. From the initial discovery of individual cell firing rates [143] and the formal linking of their ensemble activity to past experience [142], the ability to measure replay spawned a vast field of productive research that used this phenomenon to probe wider neural function. This included quantification of numerous LFP phenomena and their association with cellular replay [144], discovery of replay during wakeful rest[145] and the discovery that replay instantiated numerous cortical representations

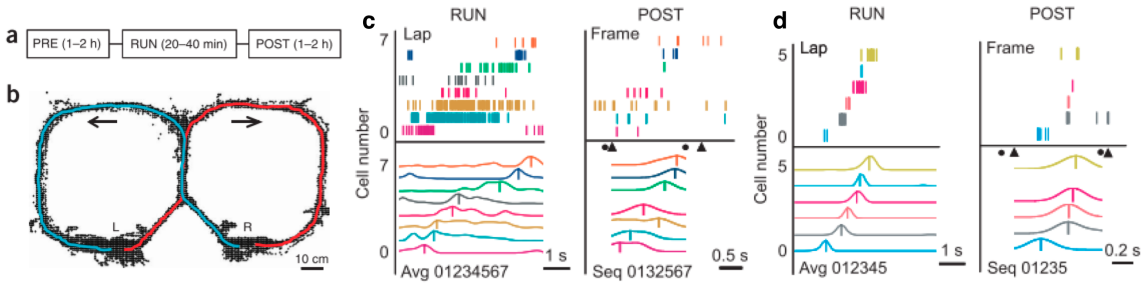


Figure 5.1: Replay outline, adapted from Ji and Wilson [147]. (a-b) When a rat runs along a track, hippocampal place cells encode the current location. (c) During active exploration (the RUN phase) these cells exhibit sequential sweeps as the rat traverses successive place fields. Later in wakeful rest (the POST period), the same cells are reactivated in the same sweeping fashion, but at temporally compressed rates. (d) Cells in visual cortex that encode the visual features of the environment are also seen to fire in sweeps during active exploration; furthermore, they too replay this experience later in rest.

[146, 147], generally showing that a more widespread and distributed phenomenon than the original recordings of small cell ensembles in the hippocampus suggested.

Indeed, the activation of widespread cortical ensembles during wakeful rest led Kurth-Nelson et al. [99] to hypothesize that representations of experienced stimuli that are decodeable in human MEG studies may also replay in manner that could be decoded. Their investigations were rewarded with the discovery of a non-invasive signature of replay in human MEG recordings, opening the door to numerous avenues of investigation as to how this phenomenon is expressed in sophisticated cognitive tasks. However, the recency of this discovery also left a number of questions unanswered, from both a technical and neuroscientific standpoint, prompting a more robust analysis of replay data with novel decoding techniques.

This work is divided into two chapters. In this chapter, we review the literature regarding what representations are replayed and under what contexts; this motivates the ensuing work which is focussed on understanding and enhancing replay detection in MEG using HMM classification models. Chapter five then reviews what is known about network mechanisms that support replay, motivating an ensuing analysis of the MEG resting state networks and their relationship with replay events.

5.1.1 Replay in Hippocampal Place Cells

The hippocampus is one of the most studied areas of the brain, perhaps as it is one of the few higher brain areas for which we have developed a sophisticated understanding of both the region's functional role in behaviour and the principles of local information encoding [148]. Famously implicated in episodic memory formation [149], the hippocampus has been modelled as an index that records specific spatiotemporal signatures associated with experiences, supporting their coherent retrieval at a later time [150, 151]. This model is based on empirical evidence of specific cellular encoding regimes, with the discovery of cells that fire selectively in response to specific locations, called place cells [152]. This was followed by successive discoveries of head direction cells [153], grid cells [154] and time cells [155], each of which records in ensemble firing rates some specific component of a general spatiotemporal frame of reference. Taken together, this prompted a question to researchers whether this spatiotemporal reference system would always reflect the immediate time and position, or whether it may at times coherently simulate some other frame of reference to that presently experienced [156].

Pavlides and Wilson [143] made the early discovery that individual place cells were more likely to fire during sleep if their respective environments had been recently traversed. Later mapping of ensemble firing rate activity demonstrated that the specific temporal correlation structure recorded during wakeful experience was also preserved in reactivations during sleep [142, 157, 158]. Specifically, when rats were taught to traverse a track, the cells that encoded successive positions on that track later reactivated during sleep in the same sequence. However these sequences were reactivated at much faster rates, displaying a temporal compression factor on the order of twenty times what was actually experienced [159]. These studies together established the view that the neural signatures of wakeful experience were systematically reactivated during sleep, a process hypothesised to support the consolidation and maintenance of memory traces [145, 160].

5.1.2 Replay Activates Widespread Networks Beyond the Hippocampus

Models of episodic memory differentiate recent and long term memory, where information of recent events is initially encoded in the hippocampus but gradually transferred to the neocortex [161]. These models also postulate that hippocampal activity orchestrates reinstatement of experience-specific activity patterns in the neocortex, with repeated excitation responsible for consolidation into long term memory traces in the neocortex [162]. Qin et al. [146] demonstrated that the same patterns of temporal sequence reinstatement observed in hippocampal place cells were observed in posterior parietal cells, confirming that replay generalised beyond the hippocampus. Subsequent studies showed similar reactivation of sequences in visual cortex [147], medial prefrontal cortex [163, 164], the ventral striatum [165–167] and the motor cortex [168].

Importantly these distributed activations did not occur in isolation, but consistently aligned to high frequency events called sharp wave ripples - which themselves are one of the largest scale synchronous events observed in the mammalian brain [169, 170]. The whole-brain patterns of activity linked to replay will be discussed more fully in chapter 6, but this suggests that replay as a mechanism may be a significantly more widespread event than originally anticipated, instantiating the synchronous reinstatement of distributed representations of recently learned experience in numerous regions of neocortex [164]. In particular, the discovery of robust replay in visual cortex [147] was to eventually arouse the interest of MEG researchers, given the ability to non-invasively decode visual imagery at fast timescales on the order of that which cellular replay is observed.

5.1.3 Replay of Cortical Ensembles in Wakeful Rest

Initial studies assumed that memory consolidation occurred during sleep, as then the brain is unburdened by stimuli and sensory inputs that could otherwise interfere with these processes [171]. However the strong association observed between replay of cellular sequences and sharp wave ripple events [142, 169], which are

known to occur both during sleep and awake states, raised the possibility of replay being observed during wakeful rest. Kudrimoti et al. [145] demonstrated this relationship, first showing replay in the awake state and finding that population discharge characteristics of replay events were indistinguishable from sleep - but nonetheless dependent on ripples.

Later studies of wakeful replay, however, began to unearth some significant differences. Where previous studies of replay relied upon extensive repetition of a running track, awake replay was shown to occur immediately after a single run on a novel track [172]. Even more surprisingly, it was found to occur in both the forward and reverse directions, rapidly simulating both the forward and backward trajectories along the track [172–176]. This finding significantly challenged the interpretation of replay as fundamentally a recapitulation of past experience, as by flipping the ordering of items it was now reactivating an abstraction from the original experience [156, 175]. Further work showed that the length of replayed sequences corresponded to the topology of behavioural trajectories through an environment [177, 178], and then that replay could piece together previous experiences into whole trajectories that had not been experienced together previously [175, 179]. Rather than a function solely of memory consolidation, replay began to look more like a mechanism for forming an internal model of the world [156] and possibly supporting functions like cognitive planning [171].

5.1.4 Controversial and Poorly Understood Aspects of Replay

The model of replay as solely a memory consolidation mechanism was further confounded by observations of the same sequences that would be referred to as replay occurring spontaneously in hippocampal place cells prior to the actual experience itself [180]. The authors coined the term ‘preplay’ to distinguish this from replay, however noting that it had largely identical features other than the temporal precedence. This and studies since link preplay to a more mechanistic role of a-priori establishing correlation structures between cells that could subsequently be mapped onto variables such as place [181–183]. An analogy is the creation of a

blank movie reel, where frames are setup in a specific order and only subsequently have content deposited onto them [156]. This picture is far from clear, with strong theoretical frameworks linking preplay and replay in terms of functionality yet to meet widespread acceptance [184]. The very notion of preplay remains controversial, with significant questions regarding how it can be robustly measured [185].

Further confusion derives from conflicting evidence of links between replay and behaviour, with some studies finding direct links between volume of exposure and subsequent replay [186, 187] while others found no link at all [175]. Some have further speculated that, in addition to consolidating memory traces, replay may at times function to erase recently learned information from the highly plastic parts of the hippocampus [188].

On the one hand, replay is a reasonably well-defined phenomenon that has been well characterised in both hippocampal and neocortical regions alongside sharp wave ripple events; on the other hand, its functional role and significance are the source of significant debate. The picture now seems more nuanced than initial memory consolidation models suggested, and a full understanding of how replay supports memory, planning, cognition and behaviour is likely to be an intense focus of neuroscience research for some time into the future.

5.1.5 Detection of Replay in Human MEG studies

It is only very recently that replay has been studied in humans, with the realisation that it reinstated widespread neural representations across the neocortex that could potentially be measured noninvasively. This opens up an extremely valuable means of investigating the functional significance of replay, given the implication that its role was more broader than mere memory consolidation. There are also practical limitations to any study of memory and planning using rodents that can be overcome with humans, where sophisticated controlled experiments can be established and explained to participants. We will focus in particular on studies in MEG by Kurth-Nelson et al. [99] and Liu et al. [31], reviewed in detail in section

5.2, which established MEG as a means for investigating the functional relevance of replay in planning and cognition.

However alongside the emergence of these studies in humans, the terminology of 'replay' has proliferated through a number of subtly distinct contexts, which are all likely related but create some ambiguity. While convergent evidence from animal models and now a number of human MEG studies would suggest that replay may be a general neural mechanism supporting a number of cognitive functions, there is not sufficient evidence to conclude that any of these studies are necessarily looking at precisely the same phenomena. It seems likely that each of these are related but distinct in important ways that are yet to be clarified.

Spontaneous vs Cued Replay

The first important distinction is between study designs that involve cued and spontaneous replay. Michelmann et al. [100] for example design a MEG task where participants are presented with a sequentially structured visual or auditory stimulus, and then subsequently instruct participants to mentally replay the stimulus upon a cue. This gets at an interesting and important, yet unanswered question; how similar are the neural processes underlying cued, volitional visual or auditory imagery from the spontaneous replay of learned patterns associated with learned images? While one may suspect significant overlap, it would be unwise to assume equivalence. By framing this in the terminology of replay, this stands in contrast to early hypotheses that replay occurred during offline periods such as sleep and wakeful rest, in order to optimally allocate neural resources and eliminate interference from sensory inputs [142, 171].

In the interests of clarity, our use of the term replay is restricted to spontaneous patterns of reinstatement - that is, the detection of patterns associated with a stimulus in the absence of an external, experimenter-provided instruction to deliberately evoke the content associated with a stimulus. We also differentiate this from the working memory style task of Huang et al. [189], which identified fast reverse reactivation of sequences during a 5 second memory retention interval, similar in content to a cued memory instruction. However we furthermore do

not suggest that the spontaneous reinstatement of these patterns is necessarily absent of a deliberate and conscious effort to do so, just that we do not require it (and importantly, participants did not report such conscious efforts to reinstate sequentially the visual images of the task).

Sequential Replay vs Static Reinstatement

Another important distinction is between replay of a specific sequence of neural representations, and reinstatement of a single static representation. A series of recent fMRI studies have shown an intriguing correlation between the spontaneous reinstatement of stimuli representations in the hippocampus and subsequent behavioural performance [190, 191]. The authors report this correlation over both wakeful rest and sleep periods, and conclude that this reflects replay of the learned stimuli to support memory [190] or active planning [191]. In another recent study, Zhang et al. [192] demonstrated an equivalent correlation between memory performance and the reinstatement of gamma power distribution over intracranial EEG sensors during sleep in humans.

However the static nature of these reinstatements is difficult to immediately reconcile with certain central notions of replay established through rodent physiology. Detailed models explain how temporal compression of sequences within individual theta cycles supports memory consolidation by inducing long term potentiation effects specifically for the learning of sequential structure [93, 193, 194]. The strength of these models have led some to argue that sequential patterns are the defining feature of replay [195]. The very slow timescale of the BOLD response would appear to be incompatible with these timescales; it remains entirely plausible however that the static representations measured by Deuker et al. [190] and Momennejad et al. [191] are a rapid temporal sequence of hippocampal activity that nonetheless manifest as static patterns in the BOLD signal due to its slow temporal resolution. There is also the controversial argument that BOLD recordings are indeed sensitive to activations that are consistently 100-200 milliseconds apart, allowing for rapid sequential reactivation to be measured this way [195]. For the sake of our argument, we simply acknowledge that the sequential nature of activations is a crucial ingredient

of replay that is well understood in rodent physiology, and that the link between sequential patterns and static reinstatement is for the moment unclear and an important line of enquiry.

5.2 On the Identification of Replay in Human MEG Studies

We here review the methods implemented by Kurth-Nelson et al. [99] and Liu et al. [31] in some detail, motivating the specific questions that we then address using the sequential decoding model.

5.2.1 Task Requirements

The cognitive task components necessary to induce replay in participants are likely to be a source of research for some years to come, as they are presently quite poorly understood. Based on the limited evidence no strong statements can yet be made, other than listing what has so far been observed. Each task to demonstrate replay required participants to learn and recall some sequential information; in the work of Kurth-Nelson et al. [99] this was a cognitive map linking different stimuli that participants had to navigate, in the work of Liu et al. [31] this was a learning a specific sequential arrangement of stimuli. From a methodological standpoint however, the detection of replay requires cognitive tasks to have distinct items that can be decoded independently. In the papers discussed, visual stimuli are utilised for which robust decoding has been well established, although in theory any form of stimulus could be used.

The following analysis has been conducted on the second study contained in Liu et al. [31]; for completeness we review in detail here the task specifics. Each participant was required to attend two days, the first for learning the structure of the task and the second for the actual MEG scan. On Day one, participants learned that they would be presented with a set of stimuli in a specific ordering (the 'visual sequence'), and they were given a rule by which to unshuffle this into the target ordering (referred to as the 'structure sequence'), as in figure 5.2(a). For example,

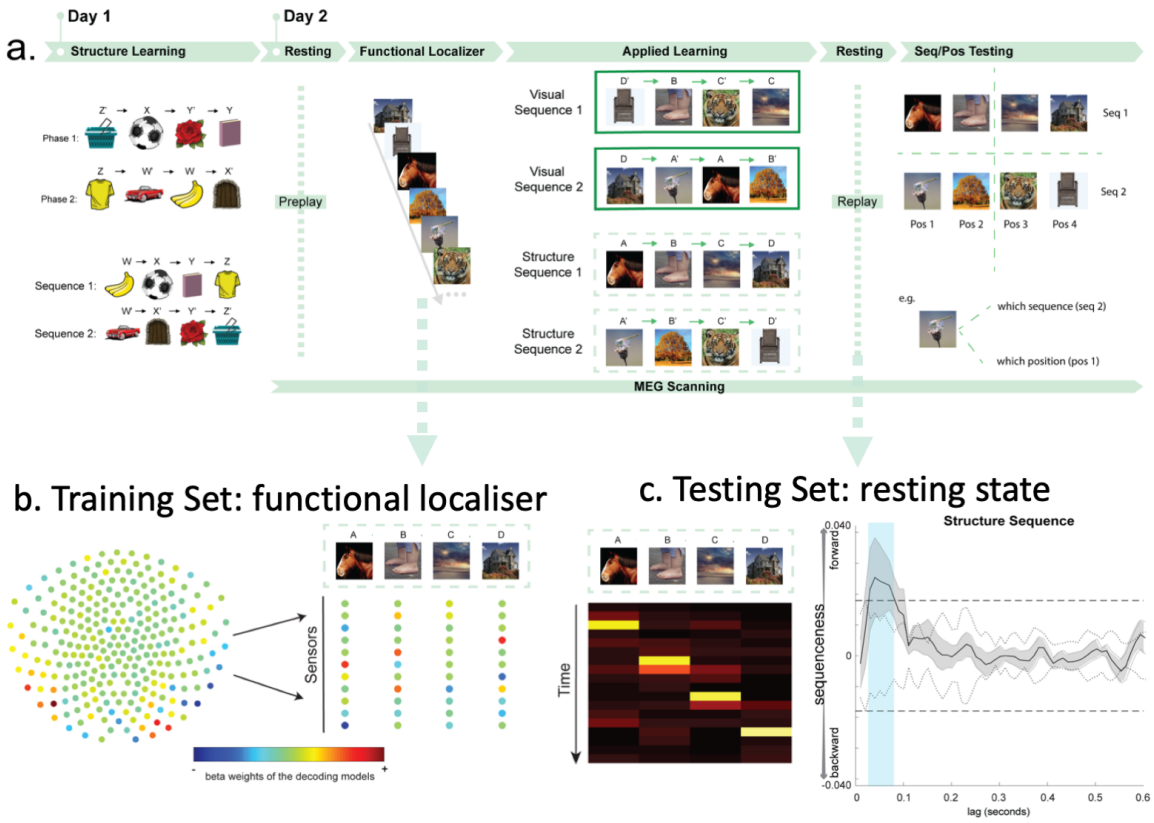


Figure 5.2: Liu et al. [31] Task Outline: (a) On Day One, participants learned the structure of the task, specifically learning a rule by which to unshuffle stimuli that are presented from a shuffled sequence (the 'visual sequence') into the desired ordering (the 'structure sequence'). They were also told to memorise this rule as it would be required for them to complete the task the following day. On Day 2, while undergoing a MEG scan, participants underwent an initial resting state scan; then a passive viewing of the novel stimuli (the 'functional localiser' session for classifier training); they were then presented with the stimuli in the deliberately shuffled order, before undergoing a second resting state scan prior to being tested for correct recall of the structure sequence. (b) The analysis pipeline involved training a set of classifiers on the functional localiser data; (c) Then classifying the entire resting state scan and testing for significant sequential reactivation of the stimuli in the target sequence. The authors found significant evidence for replay of the structure sequence at periods of 0.04-0.08 seconds, when the classifiers were trained specifically at 200msec post image presentation in the functional localiser data. Figure adapted from Liu et al. [31].

they learned that the first item presented to them visually (visual sequence item 1) belonged to the last position of structure sequence 2, and so on until all swaps were learned. They were also advised that this rule would be the same in the scanner the following day, except that a novel set of visual images would be used. On day 2, while in the scanner, after an initial resting state scan and the passive presentation of functional localisers of the full stimulus set, participants were presented with the items in the visual sequence; they then had one more resting state session before being tested for correct rearrangement of the stimuli into the structure sequence, as in figure 5.2(b). The authors then looked for evidence of replay of the structure sequence during the second resting state scan, after participants had learned the task structure but had not yet been tested on it.

5.2.2 Analysis methods

Classifier training

For each participant, classifiers must be trained independently on each stimulus. Both Kurth-Nelson et al. [99] and Liu et al. [31] utilise sparse Lasso-regularised logistic regression classifiers, training models with a range of λ regularisation penalty values, with a single value to be selected through cross validation. They select a specific training point of 200msec post stimulus onset, citing that this temporal value is associated with the highest accuracy for categorical decoding [128]. The trained classifiers are then completely specified by their decode coefficients, denoted by β of dimension $[P \times Q]$, and a set of intercept terms of dimension $[1 \times Q]$, where P denotes the number of MEG sensors and Q denotes the number of visual stimuli to be decoded.

Classification of resting state data

Each subject's resting state data, denoted as X of dimension $[N \times P]$ with N denoting the number of timepoints, could then be classified by a linear transformation into a set of predictions $R = X\beta$, where R is a $[N \times Q]$ matrix indicating the extent to which each visual item is represented in the data at each point in time. The true classification probability is then obtained by a logistic sigmoid

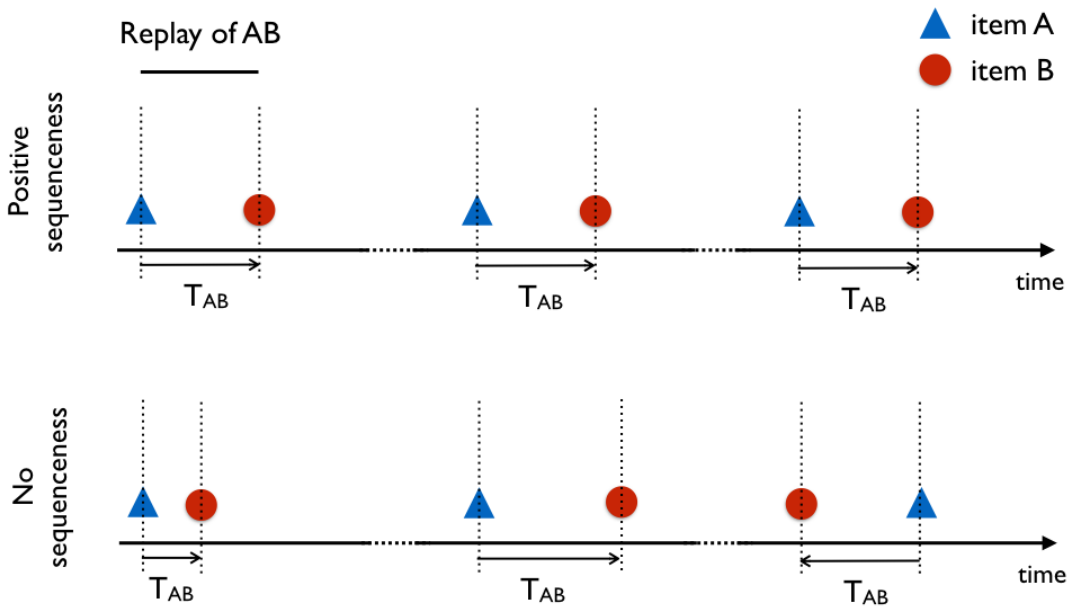


Figure 5.3: A conceptual outline of what sequenceness measures: the overarching goal of the methodology introduced by Kurth-Nelson et al. [99] is to test for evidence of a regular interval between activations of an item A and a second item B. This is loosely based on a test of the asymmetry of the cross correlation between valid pairs; where an observed sequence $A \rightarrow B$ repeats at a regular interval T_{AB}, this should result in positive sequenceness (upper panel); where this interval is irregular, no sequenceness should be observed (lower panel).

transform on $Y = \sigma(R)$; however as this results in Beta distributed values on the interval $[0, 1]$ whereas the data pre-transform is approximately Gaussian distributed on the interval $[-\infty, \infty]$, the subsequent analysis continues on the data in the pre sigmoid-transformed space.

Sequenceness test

Kurth-Nelson et al. [99] introduced a novel test for replay referred to as *sequenceness*, loosely inspired by the original work of Hoffman and McNaughton [196], which identifies the average asymmetry of the cross-correlation between two encoded items as a measure of their sequential reactivation. In its original form, for a sliding time lag τ the authors fit a vector autoregression model to the set of scores over all stimuli R_t of dimension $[1 \times Q]$ as follows:

$$R_t = R_{t-\tau}\alpha_\tau + \epsilon_t \tag{5.1}$$

where ϵ_t is a vector of error terms of dimension $[1 \times Q]$. This results in a set of matrices α_τ, each of dimension $[Q \times Q]$ which describe the degree to which one stimulus score predicts another at a given timelag. The authors then define some template of *valid* forward and backward sequential transitions (i.e. the stimulus sequences that are considered valid for the cognitive task requirements), defined as α_f and α_b for the forward and backward templates respectively, and regress these templates against each matrix α_τ, obtaining a set of coefficients w for forward and backwards transitions as a function of timelag:

$$\alpha_\tau = \alpha_f w_{f,\tau} + \alpha_b w_{b,\tau} + \epsilon_t au \tag{5.2}$$

In order to overcome the high autocorrelation of the observed sequences, the specific test statistic of 'sequenceness' is then defined as the difference between weights $S_\tau^{valid} = w_{f,\tau} - w_{b,\tau}$, effectively modelling the asymmetry of the cross-correlation of valid sequences.

To test for significance of this test statistic, a null distribution is generated by randomly permuting the templates α_f and α_b to obtain a set of *invalid* sequences (i.e. stimulus sequences that do not form part of the cognitive task) and for each of these permutations determining the relevant weights and sequenceness score $S_\tau^{invalid}$ by the same procedure. This allows a significance test of the null hypothesis that valid sequence transitions are no more likely to be observed in the data than invalid sequences.

Amendment for alpha interference

In the work of Liu et al. [31], a small amendment to this procedure was necessary as the authors were investigating resting state replay where there was high alpha band activity. This derived from the focus of the work on replay during the resting state session, whereas Kurth-Nelson et al. [99] had analysed replay in periods within the trials, that consequently didn't have a strong alpha component. They amended equation 5.1 to regress out correlated affects at alpha periods, as follows:

$$R_t = R_{t-\tau}\alpha_\tau + \Big(\sum_{k=1}^{K} R_{t-\tau-kT_{10Hz}} \alpha_{\tau,k} \Big) + \epsilon_t \tag{5.3}$$

where T_{10Hz} refers to the period of a 10Hz alpha wave at the sampling frequency; K was set so as to cover the maximum timelag investigated (600msec). The analysis was then continued as above on each α_τ.

Cross validation for L1 constraint

This complete pipeline was performed for each subject for multiple values of the L1 classifier penalty λ. A single value of λ was then selected through hold-one out cross validation over subjects, with a value selected for each held-out subject that maximised the test statistic for sequenceness S_τ^{valid} over all other subjects.

5.2.3 Conclusions and Unresolved Questions

These studies in combination established the ability to measure fast replay of temporally compressed sequences using MEG. They identified that the neural representations of distinct visual stimuli are reactivated in rapid succession, with a period of approx. 40msec between successive items. They also showed that the replayed sequences relate specifically to cognitive task structure, supporting the theory that replay supports a planning function, potentially by strengthening links in an internalised generative model of external structure.

Nonetheless, a number of methodological questions remained. In each case the results reported were statistically significant but only by a relatively small margin (p=0.036 in Kurth-Nelson et al. [99] and p=0.045 in Liu et al. [31]). As highlighted in recent discussions about scientific reproducibility, statistical significance is not an absolute quantity, but a gauge of confidence to be interpreted carefully with reference to prior work, modelling assumptions and methods [197]. Confidence in the reported p-values of these studies was challenged by the following empirical findings:

Temporal specificity

Both authors trained classifiers using just a single timepoint of the functional localiser data, isolated to 200msec post stimulus onset. This was justified through previous work by Kurth-Nelson et al. [128] establishing a peak in the accuracy for decoding object categories approximately 200msec post stimulus presentation.

However, the timing of peak decoding accuracy is a function of several factors including the type and category of stimuli used (with faces having a different time of peak decoding accuracy for example than houses [132]), subject differences and numerous other variables that vary between experiments; consequently the time of 200msec is by no means universal and is in fact divergent with a number of other studies in the field [88, 132]. This would not be a problem if the replay result generalised to classifiers trained at these other timepoints, however it was found that it was highly temporally specific to a band of data between 180 and 220 msec post stimulus (returned to in more detail in figure 5.6). There is currently no clear reason why an effect would be temporally localised to this specific timepoint of training.

Dependence on sparsity

Both authors trained sparse classifiers using L1 regularised logistic regression, with regularisation penalty chosen by a cross validation procedure that maximised sequenceness. L1 regularisation imposes sparsity on the classifier coefficients, meaning that a significant portion of those coefficients are exactly zero. Experimentation with other forms of classifiers struggled to reproduce the same result, or at least resulted in patterns of sequenceness that were less clean than the original (if retaining the same trend). There are a number of empirical cases where sparsity has improved performance for reasons that are not well understood theoretically [198, 199]; this could justify the observed dependence, however the absence of a theoretical understanding why this would be the case remains concerning. In particular, given the superior decoding performance of L2 regularised logistic regression in functional localiser decoding tests, the failure of the same method to reliably reproduce the result was of significant concern.

Replication over multiple studies

On the other hand, the exact data processing pipeline replicated the same significant result of sequential replay in two published studies [31, 99] and in a third, as yet unpublished, study [200]. This significantly increases confidence in the results, even as the above concerns emerged. The chances of obtaining a significant result

on three independent studies is remote, lead to the idea that maybe there were methodological reasons for the above dependencies. These could either help us understand the underlying nature of replay, or alternatively cast further doubt on the findings. This motivated a more detailed survey of the classification methods that could be used and the results they obtained.

5.3 A Survey of Classification Methods and Sequenceness Tests

In the previous chapter we developed methods to identify the main modes of sequential activity that discriminate different visual stimuli, and showed that these propagate forward through the anatomical areas associated with the ventral visual stream. The temporal specificity of replay identified in the above studies, which use a decoder trained only at $t = 200$msec post-stimulus in the functional localiser data, suggests either that replay only instantiates one very specific anatomical component of the ventral visual stream; or alternatively that replay instantiates broader patterns of visual stream reinstatement but these are not detectable using existing methods.

We hypothesise that the latter statement is true; that multiple components of the visual stream should demonstrate sequential patterns in their reinstatement, and that the temporal specificity of inferred replay is a result of the limited sensitivity of decoding methods. We set out to test this hypothesis by reviewing a number of alternative decoding models. The pipeline of analysis to review different classifiers and gauge their ability to detect replay is summarised in figure 5.4; individual steps are explained in detail in the following sections.

5.3.1 Classifiers Tested

Following from the above analysis, we set out to quantify what methodological constraints were necessary to identify replay in the dataset of Liu et al. [31]. Specifically, we wanted to analyse a series of variants of logistic regression, which vary in how they regularise the classifier coefficients and how they identify datapoints to train over. For the following discussion, we refer to a classifier trained on an

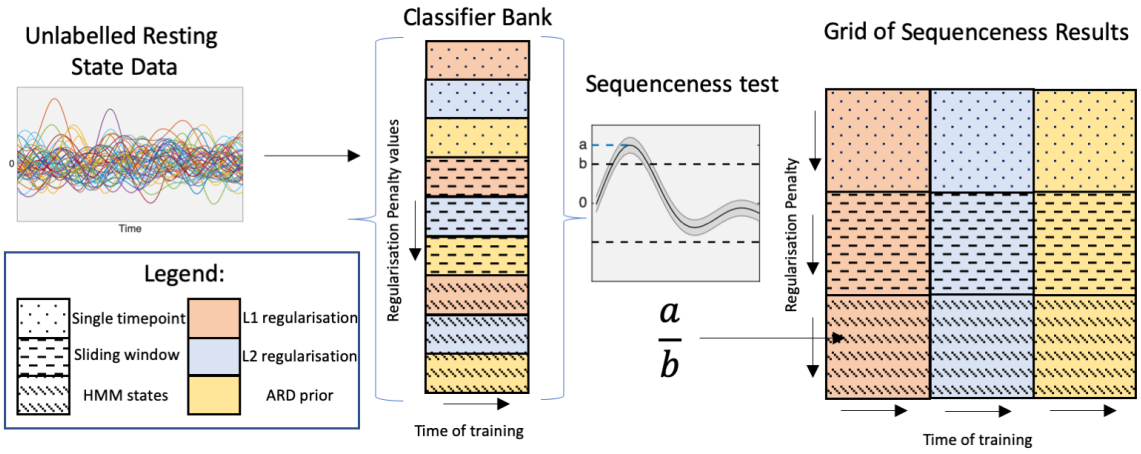


Figure 5.4: Survey of classification methods: analysis pipeline. All subjects' resting state data was fed through a bank of classifiers containing all permutations of the variables under investigation; this included (1) the type of regularisation, (2) the treatment of time, (3) the regularisation penalty applied, and (4) the training data timepoint (or HMM state) the classifier was associated with. Sequenceness results were computed for each classifier. For ease of visualisation, we determined the ratio of the peak sequenceness score to the peak permutation threshold score, and plotted this statistic over a grid representing all possible classification methods.

$[N \times P]$ matrix of data points X with a corresponding $[N \times Q]$ vector of binary labels Y, fit by a logistic regression model:

$$P(Y) = \sigma\Big(X\beta\Big) \tag{5.4}$$

where $\sigma(x)$ denotes the logistic sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. Values of β can then be fit so as to maximise the likelihood of the training data.

Regularisation of coefficients

During training, we can impose L1, L2 or ARD constraints over the inference of classifier coefficients. Regularisation is often desirable to reduce overfitting; in logistic regression specifically it is frequently necessary as when data is perfectly linearly separable (defined as the existence of any hyperplane that can perfectly separate the data by class assignment), the logistic regression likelihood term is unconstrained and has no unique solution. Perfect separability is guaranteed when $P > N$ (i.e. there are more data dimensions than samples) as is the case for the data we use, so regularisation is a necessity.

In section 3.2.3 we introduced logistic regression with an ARD prior. We repeat this here for clarity in its Bayesian formulation, amending to the logarithmic form of the posterior to emphasise consistency with the conventional L1 and L2 equations below:

$$log(P(\beta, \lambda_{ARD}|X, Y)) \propto log(P(Y|X, \beta)) + log(P(\beta|\lambda_{ARD})) + log(P(\lambda_{ARD})) \tag{5.5}$$

$$P(\beta|\lambda_{ARD}) \sim N(0, \lambda_{ARD}^{-1}) \tag{5.6}$$

$$P(\lambda_{ARD}) \sim Gamma(a_\lambda, b_\lambda) \tag{5.7}$$

This amounts to finding the parameters β that maximise the likelihood of the data observations, under the constraint imposed by the prior term. L1 regularisation can be phrased in very similar terms, maximising the same likelihood subject to a regularisation penalty on the L1 norm of the coefficient vector:

$$\beta = \underset{\beta}{\text{argmax}} \Big[log(P(Y|X, \beta)) + N\lambda_{L1}||\beta||_1 \Big] \tag{5.8}$$

Finally, L2 regression can be viewed in a similar manner, as a problem of maximising the likelihood subject to a regularisation penalty on the L2 norm of the coefficient vector:

$$\beta = \underset{\beta}{\text{argmax}} \Big[log(P(Y|X, \beta) + N\lambda_{L2}||\beta||_2 \Big] \tag{5.9}$$

In this formulation it is easy to see the similarities between these equations. Under certain conditions the penalty terms are in fact equivalent; namely if $\lambda_{ARD} = I$ where I is the identity matrix, then equation 5.5 simplifies to the form of equation 5.9; similarly if a_λ is set to the identity matrix, then the prior of 5.5 simplifies to a Laplace prior which is equivalent to 5.8.

There are two key differences that concern us. Firstly by convention, L1 and L2 regression scale the penalty term with the number of data observations N, whereas Bayesian convention dictates that the prior term is not scaled as it should become obsolete in the limit of many data points. Secondly, we solve equation 5.5 by finding a full posterior, while equations 5.8 and 5.9 are solved by finding

point estimates. Broadly speaking, full posterior estimation has a range of benefits over point estimates which has motivated our broad use of it throughout this thesis - however specifically in regards to inferring sparsity there are established methodological challenges that are better addressed through point estimation; this is discussed further below.

In the fitting of each of these classifiers, we have one parameter that controls the regularisation penalty. Ten values for this parameter were chosen for each classifier to cover a reasonable range; as they operate in different ways they are not directly comparable other than in the general directional sense that increasing this parameter makes the model more heavily regularised. An example is given in figure 5.5. This highlights similarities in the profiles of the ARD and L2 approaches, both of which constrain coefficients to shrink in size asymptotically towards zero as the penalty increases; in contrast, the sparsity constraint of the L1 classifier sets many coefficients to exactly zero, and does not display smoothness or asymptotic behaviour but rather a discontinuous threshold function of the regularisation penalty for each channel.

Critically, where previous authors used cross validation to select a single value of the penalty, using a criteria that maximised sequenceness statistic over held-out subjects, we have instead mapped out comprehensively how sequenceness varies over a range of values of the penalty. This is due to our different objectives; we are less interested in obtaining a single maximal value than understanding how this test statistic varies with different regularisation regimes and penalties.

Treatment of temporal dimension

The methods we test also differ in their treatment of time. As discussed in chapter 3, the conventional approach to decoding is to train independent classifier coefficients at each timestep. Although a widely used approach, this can result in poor parameter estimates due to insufficient sampling.

One alternative is to fit a sliding window to the data, training classifier coefficients at each timepoint using data pooled from a window around that time. This can increase accuracy as shown in appendix G, but at the potential cost of decreased

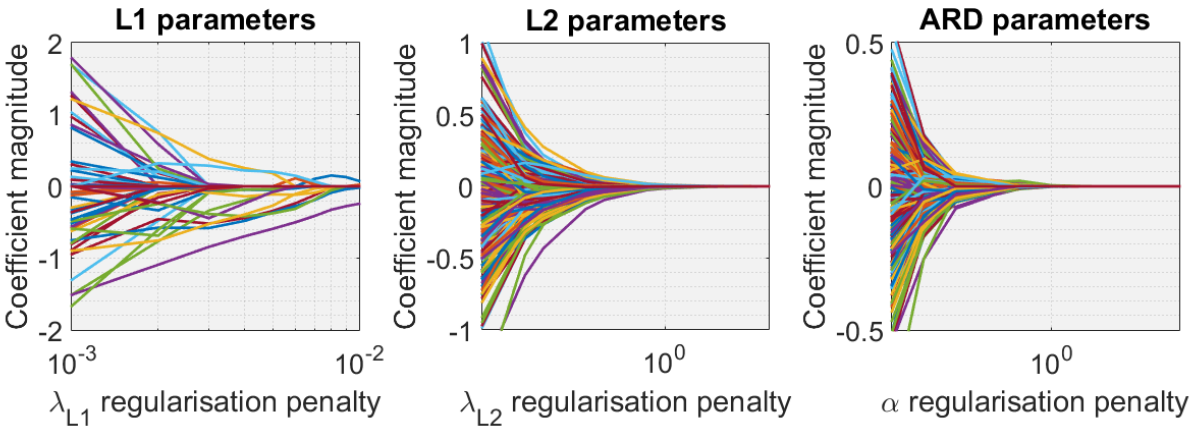


Figure 5.5: Regularisation effects on inferred classifier coefficients: each line represents a given sensor for a sample subject, specifically plotting how the inferred classifier coefficient varies as a function of the regularisation penalty parameter for each regularisation scheme. Note that in L1 regularisation, each coefficient becomes exactly zero beyond some particular threshold, whereas the L2 and ARD coefficients only approach zero asymptotically.

sensitivity to high frequency signal components. The high temporal specificity of classifiers to detect replay reported above bring into question whether this windowing would improve replay detection or undermine it; we therefore test this method and compare it to the standard approach.

Finally, we can fit the dynamic decoding model of section 3.2.3, allowing the allocation of data points to suitable state classifiers to be inferred.

This leaves us with a total of three different regularisation methods and three different approaches to the treatment of time. By quantifying the sensitivity of each of these methods to replay, we should be able to determine more specifically what features of a classifier support replay inference.

5.3.2 Sequenceness as a Metric

We will use the sequenceness metric introduced by Kurth-Nelson et al. [128] and outlined above as a gauge of each classifier's sensitivity to replay, however a few caveats must first be noted. Sequenceness was introduced as a test statistic to replace the lack of ground truth during the resting state scan by a test for timelagged correlation between scores. This makes it dependent upon permutation tests to

prove this time lagged correlation is significantly greater than a non-parametric null distribution. For our purposes it is important to note that this null distribution does vary systematically with different classifiers; changing the regularisation penalty or temporal scope of training results in variations to both the magnitude and frequency sensitivity of the resulting classifier. For this reason we normalise each sequenceness score to a common frame of reference; as in figure 5.4, denoting the peak sequenceness score as a and the peak null distribution score as b, we plot as a single metric for each classifier the quantity $\frac{a}{b}$, where a value greater than 1 denotes significant forward sequence detection; a value less than -1 denotes significant backward sequence detection; and values in between are not significant.

It should also be noted that the task structure permits only 23 possible invalid sequence permutations that together form the null distribution for each classifier. This admits a second source of uncertainty, as this is limited resolution from which to estimate the quantity b for each classifier, and limits the lowest possible p value obtainable for any classifier to $p = 0.044$. This was designed in the original study for practical reasons, but as we are now surveying a number of methods and parameters we should correct for multiple comparisons and require more stringent levels of significance. This is not possible given the structure of the task; we compensate by qualitatively assessing the temporal structure of results marked as significant; high variance and/or oscillatory patterns in this temporal structure undermine confidence in a result, even if it is significant to $p = 0.05$ levels. We also in section 5.4.1 then apply more stringent tests that the inferred patterns are not statistically independent in time, identifying common temporal structure around their activations.

5.3.3 Findings

The nine types of classifier were used to classify the resting state data, allowing the measure of sequenceness to be inferred as a function of the regularisation penalty and timepoint of training. In the case of HMM classification methods, time is replaced by the HMM state number fit to the data.

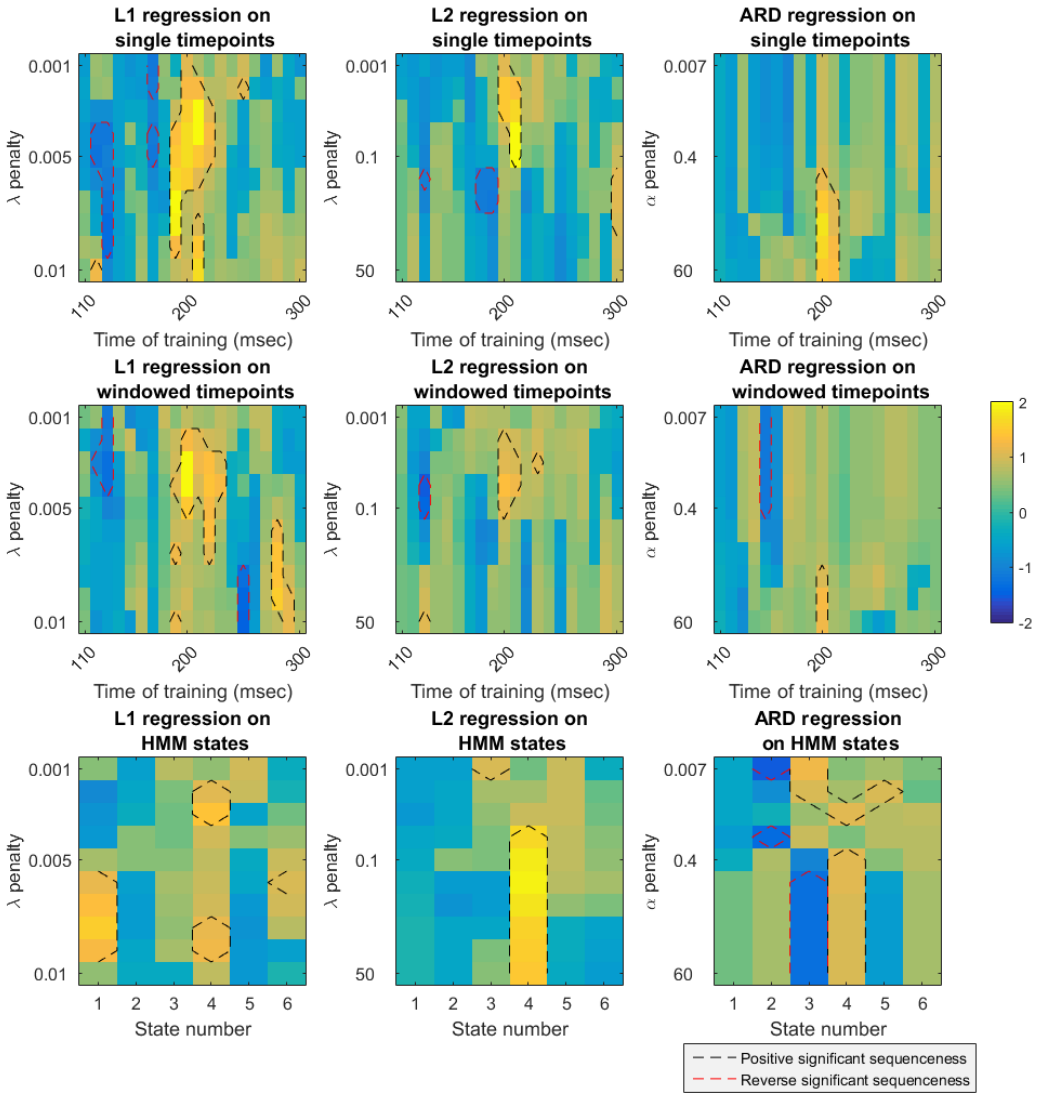


Figure 5.6: A survey of classification methods and inferred sequenceness: following from the analysis pipeline presented in figure 5.4, the full grid of results is plotted here. Note that scores greater than 1 are evidence for forward sequences to $p = 0.05$ levels (and marked out by black dotted lines); scores less than -1 are significant evidence for reverse sequences to $p = 0.05$ levels.

Sparsity Enhances Replay Sensitivity

Figure 5.6 demonstrates that, when trained on individual timepoints, L1 regularisation in training results in replay detection over a broader range of timepoints than any other classifier (top row). This is similarly true when data are windowed (middle row), and in the case of HMM states (bottom row); L1 regression consistently identifies more datapoints that result in a significant sequenceness result than L2 or ARD regression. This confirms the result more informally reported in section 5.2.3, that replay inference appeared dependent on the imposition of sparsity, and is consistent with results obtained using other datasets that for succinctness are not presented here.

This is a surprising finding however given that the use of L1 regression in functional localiser data resulted in slightly poorer decoding accuracy, and broader evidence that ridge regression is well suited to data that displays strong covarying noise (such as MEG data). Section 5.4.1 will analyse in detail the reasons for this, and establish through simulations that sparse classifiers are more robust to reinstatements that do not preserve the temporal sequence of activation over different areas.

5.3.4 Windowing data does not improve sequence detection

We also wished to test whether pooling over more timepoints, which may increase SNRs, would achieve better results without requiring HMM sensitivity to dynamic temporal alignment. The first and second rows of figure 5.6 compare the impact of windowing data as opposed to training on single timepoints. Note that in supplementary work not included in chapter 3, the same approach led to very modest performance increases compared to standard synchronous timepoint decoding. However, the results here are more ambiguous with multiple small clusters appearing to display sequential reactivation. A number of these clusters do not display qualitative features consistent with replay when the full temporal pattern of sequenceness is analysed, suggesting some of these may be spurious.

There is not sufficient evidence to conclude that windowing data systematically improves replay detection.

5.3.5 Inference of trial specific timings may enhance replay sensitivity

Using the temporal information obtained with the HMM state timecourses produced the results of the bottom row of figure 5.6. These results are not immediately striking, with a number of small clusters of significant sequential reactivation emerging around different states depending on the exact method used. However, between all of these methods, five out of the six states display significant sequenceness for some combination of regularisation methods and values of the penalties used. These also reflect temporal patterns consistent with replay found in previous studies. Even if some of these results are spurious, this represents a much broader scope of data for which sequential patterns are found *by some method* compared to the non-HMM alternatives in the top and middle row.

Returning to our hypothesis that replay is not exclusively related to the visual representation localised to approximately 200msec post stimulus, these results suggest that utilising trial-specific variations in timing inferred by HMM models may be necessary to identify representations in other visual areas that are replayed. Next, we shall conduct further tests of how the events inferred linked to different HMM classifier states are temporally related.

5.4 Temporal Patterns in Replay Activation over Different Areas

In the previous section we identified that sparse L1 classifiers seemed to improve replay sensitivity compared to other forms of regularisation, as did the HMM compared to single timepoint or sliding window classifiers. For this reason in the further analysis we restrict our focus to the classifiers obtained by using L1 regression with HMM state timecourses (i.e. the bottom left grid of figure 5.6). This is also one of the more interesting results to pursue, as the three states associated with

statistically significant sequential reactivation (states 1, 4 and 6) derive from data at quite distinct points temporally in the original training data (approximately 100, 200 and 300 msec post stimulus onset respectively). Hence establishing replay in each of these states represents stronger evidence to refute the localisation of replay to patterns of activation only around 200msec, compared to finding replay in states that are immediately temporally adjacent.

Importantly, we can now introduce two forms of sequenceness analysis, summarised in figure 5.7(a). We will refer to as *item* sequenceness the analysis of periods between successive items, within a given state - as denoted by the periods T_1, T_4 and T_6 in figure 5.7(a). These have already been established as significant for each of these states; nonetheless the full timescale of their sequenceness is displayed in figure 5.7(b). We can then also analyse the *state* sequenceness, by analysing the periods between the onset of a given sequence within any pair of states - as denoted by the periods T_{14}, T_{46} and T_{16} in figure 5.7(a).

As an aside, Kurth-Nelson et al. [99] and Liu et al. [31] both use a cross-validation procedure to select a single value of the regularisation parameter λ. This is because the sequential activation of images is their test of whether or not to reject the null hypothesis, and is directly dependent on this value. Our hypothesis however regards the temporal correlation of these sequences. This allows us to select the value of this parameter that maximises the sequenceness of images, and then determine whether the inferred sequences using this parameter are synchronised in time.

5.4.1 Replayed Sequences in Distinct Spatial States Activate in Top-Down Sweeps

In order to assess the correlation of replay sequences in different visual areas, we first define their activation through time. Given a probability of image A at time t in state k as $P(A_{t,k})$ and the probability of image B as $P(B_{t,k})$, we define $P(S_{t,k,A\rightarrow B}) = P(A_{t,k} \cup B_{t+\tau,k})$ to be the score for a replayed sequence $A \rightarrow B$ in state k at time t. We define $\tau = 40msec$ to be the period of maximum sequenceness obtained in section 5.3. This definition is identical to that employed by Liu et al. [31], we just extend it to multiple brain regions where we have observed sequential

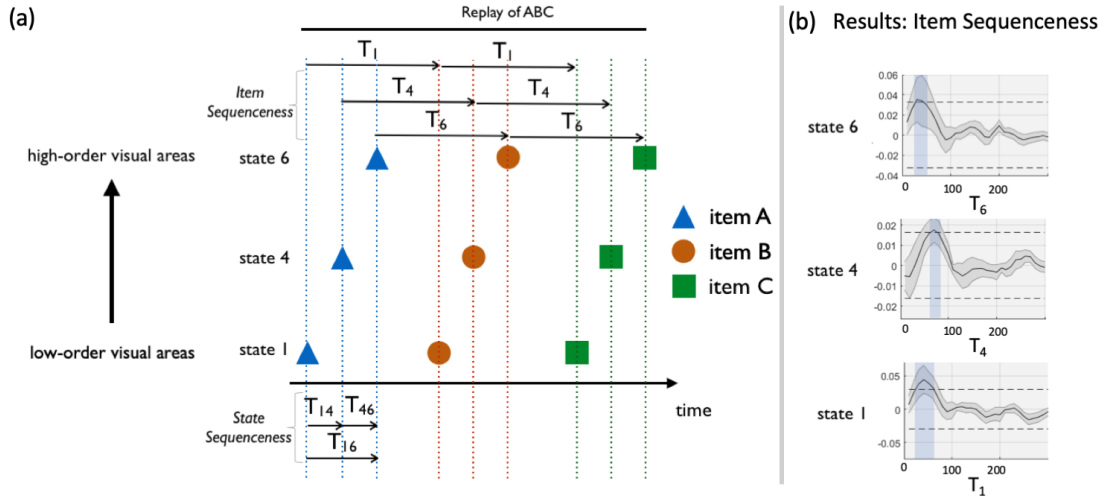


Figure 5.7: Examining replay over the visual hierarchy: (a) To outline conceptually the *item* and *state* sequenceness measures, consider the three states we reported to show significant evidence of replay in section 5.3. States 1, 4, and 6 showed significant evidence of *item* sequenceness; that is, a consistent positive trend in the periods T_1, T_4 and T_6 between activation of valid items within that state. We can additionally now investigate *state* sequenceness, by analysing the periods T_{14}, T_{46} and T_{16} between the activation of a valid sequence in each pair of states. If these are consistently positive, this is evidence for activations that start in lower visual areas and propagate forward to higher order visual areas, as in the original functional localiser data. (b) The results of the item sequenceness analysis for each state, showing the periods at which they are significant to $p = 0.042$.

reactivation, and ask whether these sequences are temporally synchronised through the measure of *state* sequenceness.

We can then fit equation 5.3 to the set of state scores in the same manner as already done for items; only now we set the timeseries R_t to be a $[nT \times K]$ matrix, columns of which denote the probability of replay in a given state $P(S_{t,k,:})$. We also stack over all valid sequences, hence our number of rows nT where T is the length of the time series and n is the number of valid sequences.

One minor deviation from the original methodology concerns the statistical tests applied; Liu et al. [31] and Kurth-Nelson et al. [99] applied a permutation test that determined whether valid transitions were more frequent than invalid transitions (where validity was defined by the task structure). As we are now dealing with scores in different brain regions, the concept of an invalid sequence no longer applies; $A \rightarrow B$ in any brain region can conceivably and indeed should have some temporal correlation with $A \rightarrow B$ in every other region. For this reason we instead determine

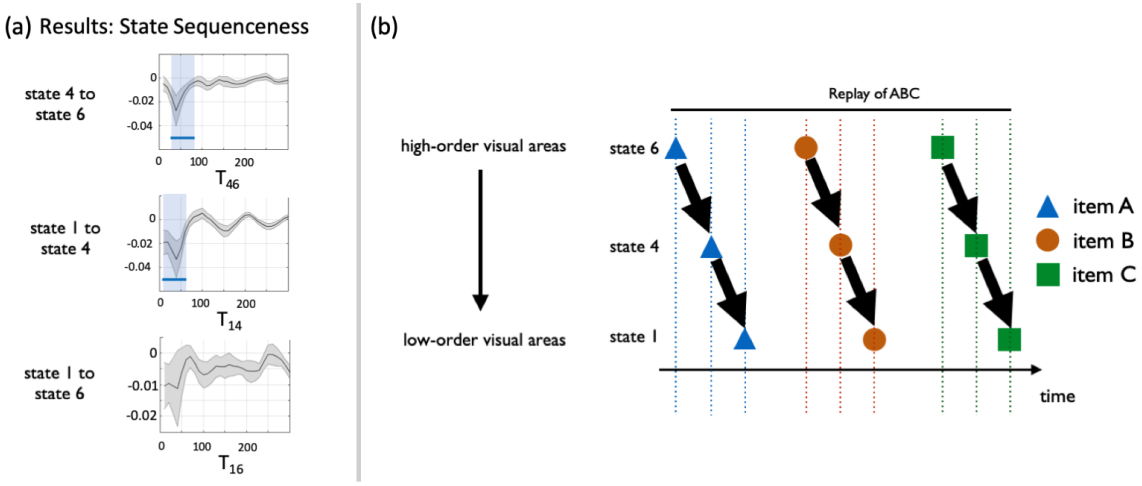


Figure 5.8: Replay is activated in top-down sweeps: (a) State sequenceness analysis shows significant evidence for *reverse* sequenceness between states 1 and 4, and again between states 4 and 6, suggesting that replayed sequences in state 6 lead state 4 which again lead the same sequences in state 1. No significant result is found however between state 1 and state 6. Significance tests reported to $p < 0.05$ using cluster sign flipping tests. (b) Together, these results show evidence that replay instantiates a top down sweep from higher levels of the visual hierarchy down to lower level visual areas, reversing the original feed-forward direction of information transfer.

significance using a paired-sample sign test, a non-parametric test to ascertain whether the forward sequenceness at a specified timelag is greater than the reverse sequenceness. Finally, after computing significance at each timepoint individually, to correct for multiple comparisons we apply a cluster threshold test.

As in figure 5.8, a significant relationship is inferred between state 1 and state 4 and again between state 4 and 6; but in both cases these detect a reverse direction of activation, (i.e. inferring that the periods T_{14} and T_{46} as introduced in figure 5.7(a) are significantly negative). This implies that activation of sequences in state 6 leads the activation of the same sequences in state 4, which in turn leads the same sequence activating in state 1. The relationship between state 6 and state 1 is not significant, however is nonetheless trending in the same direction and consistent with state 6 on average leading state 1.

This is the reverse direction to that which the information was encoded. Models of the visual hierarchy state that successive brain areas detect features at higher and higher levels of abstraction. Under this model, early visual areas detect features

such as edges and lines, successive areas then combine those lines to recognise shapes, with higher brain areas recognising the specific combination of shapes that corresponds to a certain person's face or some meaningful object. This finding suggests that when sequences are replayed, each item in the sequence is replayed in a top-down fashion, with representations in the upper levels of the visual hierarchy activating first and then instantiating the same representations in lower visual areas.

It is useful at this point to refer back to the physiology of replay, which is grounded in hippocampal sharp wave ripples. Observations of coupled hippocampal-cortical replay suggest that activity in the hippocampus leads reactivations elsewhere in the brain [167]. This supports the view that sharp wave ripples instantiate replay in the hippocampus through cell assemblies that index a particular spatiotemporal reference; these cells subsequently project out to cortical assemblies, instantiating replay of the cortical representations associated with that spatiotemporal reference. This could potentially support the top-down activation of the visual system, with hippocampal indexing of more abstract conceptual features that then project on to the lower level visual features that define them. In line with this, it should be noted that the hippocampus is much more densely connected to temporal areas associated with abstract semantic content than they are with lower level visual areas, although projections to those areas also exist [201].

5.4.2 Null Results Not Reported

To our knowledge, top-down hierarchical activations of the visual system that reverse the original encoding direction have not been detected before. The most similar work appears to be that of Linde-Domingo et al. [202], who recently demonstrated evidence of a reversal of direction between encoding and memory retrieval, specifically relating to semantic and perceptual content of stimuli. This would appear to support our own finding, albeit in very different context not specifically related to spontaneous replay eliciting reactivation of the visual stream. The novelty of our findings makes it particularly interesting, however also should necessitate a very conservative

evidence threshold. The evidence presented above is promising but not at this stage conclusive - in particular the following shortcomings must be acknowledged:

• Effect does not generalise well to other state combinations: after finding that states 1, 4 and 6 displayed significant sequential reactivation of images in section 5.3, we limited our analysis of inter-state sequential patterns to these states. Upon finding the above evidence of top-down sweeps, we expanded this to test for a more general effect over all states, for example whether state i always lagged state $i-1$. This was not the case in general, and results were quite mixed between different states (see appendix H for full results). This will be discussed further in the next section, where reasons are given why reverse sequences may not propagate uniformly back through the learned state sequence.

• Effect is unique to replayed sequences, not a property of stimulus reactivations in general; in the introduction we discussed the distinction between replayed sequences and the reinstantiation of a static representation. Our analysis of the resting state data begins by classifying the entire time series and obtaining a score for each stimulus and each brain area; the later replay score only represents a small subset of these events, hence we set out to also measure whether the scores for each stimulus displayed significant top-down directionality; no significant relationship was found, either for top-down or bottom-up directionality of processing. Again there are several possible reasons for this; it seems likely that the isolation of replay sequences removes a large number of spurious activations, as the chances of the spurious activation of two specific stimuli representations exactly 40msec apart should be much lower than that for a single stimulus at any timepoint. So it is possible that all reinstatements are being activated in a top down manner, just that the noise floor is too high for this to be detected when dealing with the raw stimulus scores. Alternatively, it is possible that static reinstatements are actually quite a distinct process that do not activate in the same top-down manner as those activated in sequence. The latter are believed to be instantiated by hippocampal projections, while there is really no unique prediction of what should precede a static reinstatement of a neural representation; hence there may be no

premise to expect that top-down patterns observed with replay should generalise to other stimulus reactivations.

While these results may undermine confidence in our reported findings, there are good theoretical reasons for each of these that we review in the next section.

5.5 On the Relationship between Sparsity and Replay Detection

As mentioned previously, sparse classifiers were found to perform slightly worse in cross validated decoding tests when compared to L2 or ARD models, yet they were much more likely to find evidence of replayed sequences in resting state data. This slightly paradoxical finding warranted further investigation. We hypothesised that there may be a methodological reason for this, which may be linked to the reversal of sequence directionality identified in section 5.4. Through simulations, we show that such temporal direction reversals significantly alter the observed spatiotemporal patterns reflected in the data. We further demonstrate that sparse classifiers remain robust over permutations and reversals of the temporal ordering of activity over channels, whereas L2 and ARD models see significant performance degradation. This provides an important theoretical justification for why L1 classifiers are necessary for decoding replay, and furthermore shows that this dependence derives from the top-down directionality of replay sweeps identified in section 5.4.

5.5.1 A Cascade Model of the Visual Stream Response

To see this effect, it is worth analysing how exactly a top-down projection of the visual hierarchy would be reflected in sensor-space data. Figure 5.9 plots a simple model of visual system activation in response to a stimulus; this convolutional model specifies a cascade of impulses through five visual areas, each of which has the same Evoked Response Field (ERF). Importantly, we model a top-down sweep as a reversal of the impulse timings while the ERF stays the same on each channel. The result of this simple model is that the full spatiotemporal response during a simulated backward sweep is markedly different from a simple reversal of the

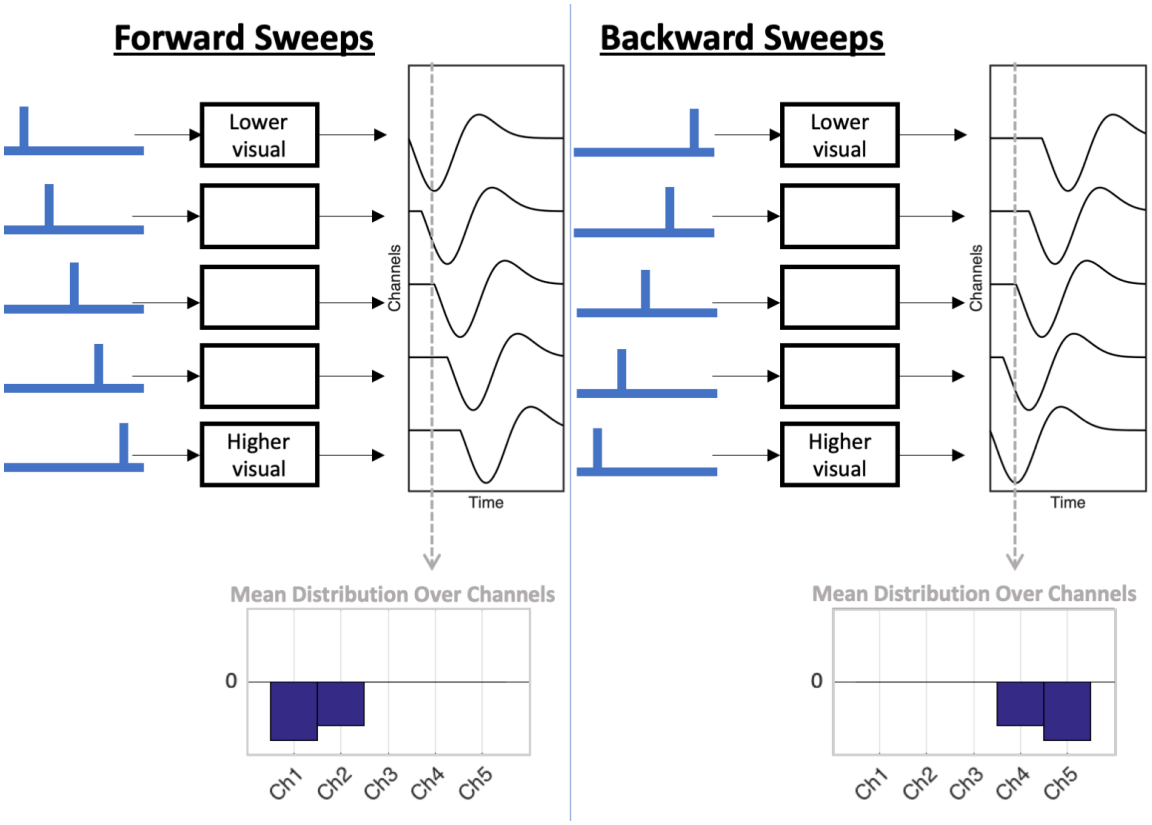


Figure 5.9: A cascade model of the visual stream response; part of the challenge in decoding reverse sweeps of the visual stream is that they are unlikely to reflect a simple reversal of the time domain of the original forward sweeps. If we model the forward sweep observed during visual stimulus presentation as a cascade of impulse responses moving progressively through the visual hierarchy, and assign an impulse response function to each of these areas, we can compare it to a backward sweep that reverses the timing of impulse functions but maintains the same impulse response. These two signals are not just a temporally reversed version of one another; comparing the mean distribution over channels timelocked to the middle of the impulse response train, for example, we observe a pattern that is completely different in forward sweeps as opposed to backward sweeps.

time domain of a forward sweep; this is emphasised by comparing the distribution over sensors timelocked to the middle of the impulse response train, where areas in the middle stages of the visual hierarchy are being activated. The observed signal over conditions is completely different; in fact, the multichannel signal recorded in the forward sweep at this point is never observed at any timepoint in the reverse sweep, as it reflects a combination of time-decaying signals in lower visual areas that derives from their specific temporal pattern of previous activation.

For simplicity this simulated ERF is uniform over all channels; it has been generated to replicate a key property of the observed data, namely that a strong

initial depolarisation is followed by a longer, lower amplitude deflection with opposite sign. Figure 5.10 plots the ERFs from several sample channels in the functional localiser data used in chapter 4, demonstrating some variations in shape whilst maintaining this property which has been widely characterised in the ERF literature. By implication, a top-down sweep that reverses the time of onset of these ERFs would result in a signal distribution over channels that does not directly match the distribution observed in the original, forward sweep training data.

5.5.2 L1 Regression Results in Sparsity over both Time and Space

Imposing an L1 penalty enforces an explicit spatial sparsity constraint over sensors, forcing a large number of classifier coefficients at any point in time to be zero and effectively selecting only a small subset of sensors from which to decode. In the standard approach where each timepoint is modelled as independent, this does not strictly enforce that sensor coefficients must be sparse in time; that is, there is no impediment to the classifier selecting exactly the same subset of sensors from which to decode, or from a given sensor contributing to the classification over all points in time.

However a quick analysis of the data we have shows this is not the case in practice, and that the sparsity constraint results in sensors contributing to the class assignment over a much narrower temporal window. Figure 5.10 plots, for four sample sensors, the expected information content derived from each as a function of time - that is, $\mathbb{E}[X_{i,t}\beta_{i,t}] = \mu_{i,t}\beta_{i,t}$, where i denotes the sensor and μ the ERF (average over trials). This quantifies how much a sensor contributes to the classification score on average. This identifies a stark difference between L1 and L2 inferred coefficients in terms of the temporal profile of activation. In particular, for increasing values of the L1 penalty, sensors contribute information content only in small temporal windows around their ERF peaks. In contrast, with L2 regression models, a given sensor contributes information to the classifier over a broader sample of timepoints, which reflect both timing of peak depolarisation and subsequent temporal decay components.

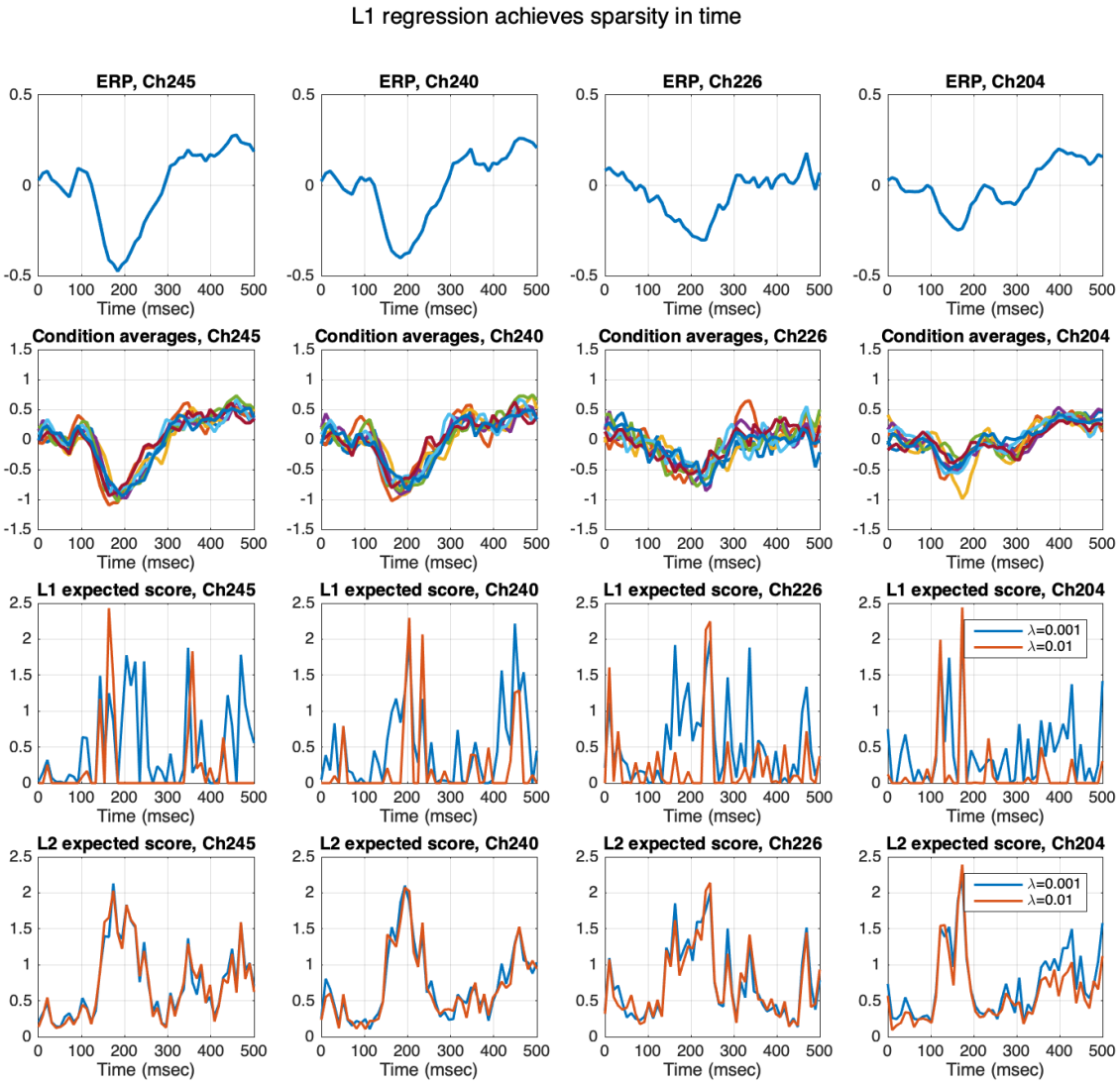


Figure 5.10: Comparing ERF features over different sensors: first row plots the ERF (over all stimuli presentations of the functional localiser data explored in chapter 4) of the four sensors, showing they all display a prominent depolarisation, in most cases followed by a repolarisation as broadly categorised in the ERF literature. Second row: condition averages, or ERF per stimulus on each of the four sensors. This highlights the consistency of the shape over different stimuli, and the subtlety of variations over any temporal stage of the waveform. Third row: the expected contribution of the sensor vs time to the classification score for lower (blue) and higher (red) sparsity parameters. As sparsity increases, each sensor has a narrower window of time in which it is used for the classification; this window mostly aligns to the ERF peaks on each sensor. Bottom row: by comparison, L2 regularisation results in the same sensors contributing to the classification score over a much wider window of time, incorporating information from the full waveform.

One may suppose that a sparse classifier that utilises only the information content contained in each channel's peak deflection might continue to perform robustly should the overall temporal order of peak deflections be varied, when compared to an L2 or ARD classifier that utilises both the information in peak deflections but also in residual temporal decay components from a previous deflection. This hypothesis can be tested more thoroughly with simulated data.

5.5.3 Simulating Hierarchical Reactivations

We generate data with $P = 30$ channels, with a cascade of activations propagated over subsequent channels in the hierarchy, each convolved with the same ERP waveform and with additive uncorrelated white noise. Denoting by E_t the single channel ERF time series extracted from Channel 245 (plotted in figure 5.10), we simulate data where one class models a cascade of this ERF over channels; and the other class is just white noise:

$$P(X_i|Y = 1) = \delta(\tau - i) \circledast E_\tau \Big|_{\tau=t} + \epsilon_{i,t} \tag{5.10}$$

$$P(X_i|Y = 0) = \epsilon_{i,t} \tag{5.11}$$

where X_i denotes the ith channel and $Y \in [0, 1]$ denotes the trial class. Results are shown in figure 5.11(a); qualitatively, the classifier coefficients fit to each of the sample channels shown replicate the above-mentioned behaviour; L2 classifier coefficients are distributed over all timepoints, while L1 classifier coefficients are sparse in time centred around the peaks of each channel's activation only.

We plot the accuracy using generalisation plots [141] in figure 5.11(b). We found that, for simulated data that had identical temporal structure to the training data, all classifiers performed well. The L2 classifier slightly outperformed both L1 classifiers, which replicates the findings of chapter 3 on functional localiser data. If, however, we generate data that reverses the temporal timing of cascades through channels, the performance of the L2 classifier degrades dramatically, while L1 classifiers maintain high performance. Furthermore, this property is not unique to reversals of the temporal structure; if the timecourse of activations are randomly

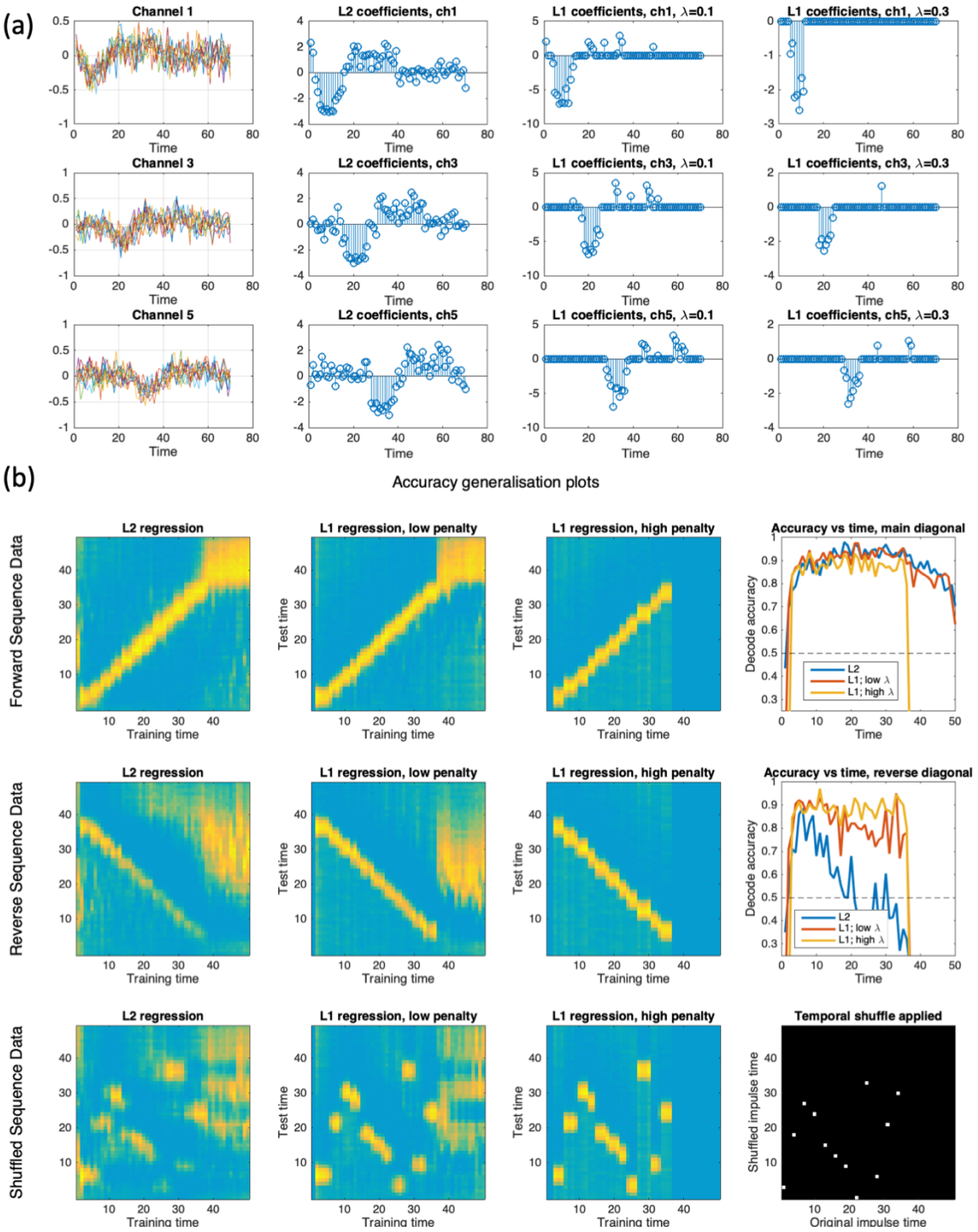


Figure 5.11: Simulated results show temporal specificity of sparse decoders: (a) Simulated data and inferred regression coefficients in time; three example channels from those simulated that show progressively higher areas of the visual hierarchy that depolarise later. Comparing their inferred classification coefficients shows that L1 classifiers are effectively sparse in time as well as space, with channels only used for decoding at times close to their peak depolarisation. (b) Testing how the classifiers generalise to data simulated in forward, reverse and random sweeps. L2 regression performs best for data that maintains the original temporal pattern; L2 performance degrades significantly when this ordering is reversed or randomised, whereas sparse L1 regression classifiers continue to perform well.

shuffled, as in the final row of figure 5.11, the L2 classifier fails to recover that structure whereas the sparse classifiers identify the reordering quite effectively.

5.5.4 Replay in Top-Down Sweeps Requires Sparse Decoders

In section 5.4.1 we showed that the sequences replayed are activated in top-down sweeps. The unexplained dependence of replay findings on sparsity can in this context be understood as these decoders are much more robust to permutations of the temporal sequences of activation. As we have shown that the sequences which are initially encoded in a feed-forward hierarchy are later replayed in reverse, the directionality of these top-down sweeps would necessitate the use of sparsity to accurately decode the true temporal patterns of reactivation.

5.6 Conclusion

We have investigated how several different classifiers vary in their sensitivity to replay detection. Initially motivated by the concerning specificity of reported methods to both a specific time of training and to the imposition of sparsity constraints, we have demonstrated the underlying cause of each of these dependencies, which in both cases has been illuminating as to the nature of replay itself. With regards to the temporal specificity of replay inference, we have shown that this may derive from the assumption of temporal synchronicity over trials, and that more broad patterns of replay may be observed when this assumption is relaxed with HMM classification models. With regards to sparsity, we have then shown through simulations that sparse classifiers are robust over permutations of the order of reactivation sequences, which would enable them to more accurately decode the temporal patterns of activation reinstated during rest. All of these results support the conclusion that replay can be observed at multiple areas of the visual hierarchy, and critically that these are activated in reverse sweeps that start at the upper levels of the visual hierarchy and propagate down to lower visual areas. Given the significance of

this claim however, we acknowledge that the statistical power limitations of this study probably require further confirmation.

Chapter 6

Understanding Replay Correlations with Resting State Networks

Contents

6.1 Introduction 129
6.2 Methods 136
6.3 On the Temporal Structure of MEG Resting State Networks 138
6.4 Characterising the Resting State Networks 143
6.5 Resting State Network Activity Predicts the Onset of Replay 145
6.6 Longer Range Temporal Dependencies 150
6.7 Discussion 154
6.8 Conclusion 156

In chapter 5, we applied the sequential decoding models developed in chapter 3 to the analysis of replay, and inferred sequential patterns of replayed sequences propagating in top-down sweeps through the visual hierarchy. We now wish to investigate how these sweeps relate to the more established resting state network activity [203, 204], by comparing the onset of replay sequences to the temporal activation of canonical resting state networks. In pursuit of this goal, we return to the time delay embedded HMM model introduced in chapter 2 to analyse the resting state MEG data from the replay experiment. The first part of this chapter is devoted to characterising the resting state network dynamics and showing that the

inferred states arrange into a temporal hierarchy between two distinct metastates, one of which comprises the default mode network and associated regions. The second part of this chapter then compares the activation of these states to the timecourse of replay, showing clear preferential activation during periods of default mode network activity on both short and long timescales.

6.1 Introduction

Resting state data refers to the type collected over periods absent of experimental stimulus; participants are typically instructed to fixate their attention to a point on a screen, but are otherwise free to let their minds wander with no objective or task for them to complete. Since early studies first identified robust patterns of correlation between different areas in fMRI studies [205], resting state fMRI has developed into a large field analysing which areas functionally correlate and what their relationships are in time [20, 206, 207]. These patterns of correlation have been shown to strongly correlate with networks mapped out in tasks [208] and certain aspects of structural connectivity [209], whilst also being predictive of individual traits and characteristics [210].

While the majority of resting state analyses have been conducted on fMRI data, similar networks can be identified in MEG that show good spatial alignment with many of the networks identified in fMRI [203, 204]. The higher temporal resolution of MEG reveals that states switch very rapidly, transiently activating for periods on the order of 100msec [120]. While this exceeds the temporal sensitivity of the BOLD signal, it seems possible that the BOLD measurements aggregate multiple state visits over each temporal sample, reflecting the same activity measured at different spatial scales [211]. Some caution however is advised in comparing these networks, as the physiological basis of each signal is quite distinct; MEG networks are defined with respect to specific frequency bands, while the BOLD signal derives from the more indirect haemodynamic response. Good correspondence has been shown between gamma band oscillations and the associated BOLD response [67];

however oscillations in lower frequency bands show mixed results that do not support a simple one-to-one interpretation [108, 212].

This is relevant to our discussion and the work that follows, as certain findings regarding fMRI resting state dynamics which have been more fully characterised in the literature cannot be assumed to apply equally to MEG resting state dynamics.

6.1.1 Structure of Resting State Activity

As we are interested in the temporal alignment of replay events to resting state networks, it is worth briefly reviewing what is known about the temporal structure and functional correlation between these networks.

Resting state analyses identify multiple networks that activate over time in any resting state scan, however their temporal patterns of activation are not random but significantly structured. Certain networks are more or less likely to activate following activity in certain other networks, with the data largely explained by relatively few high-level networks that further subdivide into distinct components [213]. These have variously been characterised as metastates in studies using HMM based methods [214], subgraphs [215] or temporal functional modes [213]. All of these distinct methods consistently highlight one particularly strong feature in the temporal structure of these networks; the anticorrelation between a subset of networks linked to the default mode network, and a second subset of networks associated with the dorsal attention network [213, 216, 217].

The specific relevance of the DMN will be discussed in greater detail below, but importantly the overall organisation of RSN temporal structure into these metastates could have a functional interpretation. Margulies et al. [218] outline a macroscale gradient distinguishing the different RSNs, where one end of this gradient contains networks associated with sensory/motor processing functions; and the other end comprises regions associated with the DMN that integrate information from different areas and support higher level cognition. They argue that this represents an overarching principle of neural function, with functional purposes linked to either sensorimotor input/output computations, or higher level cognition. Analysis of

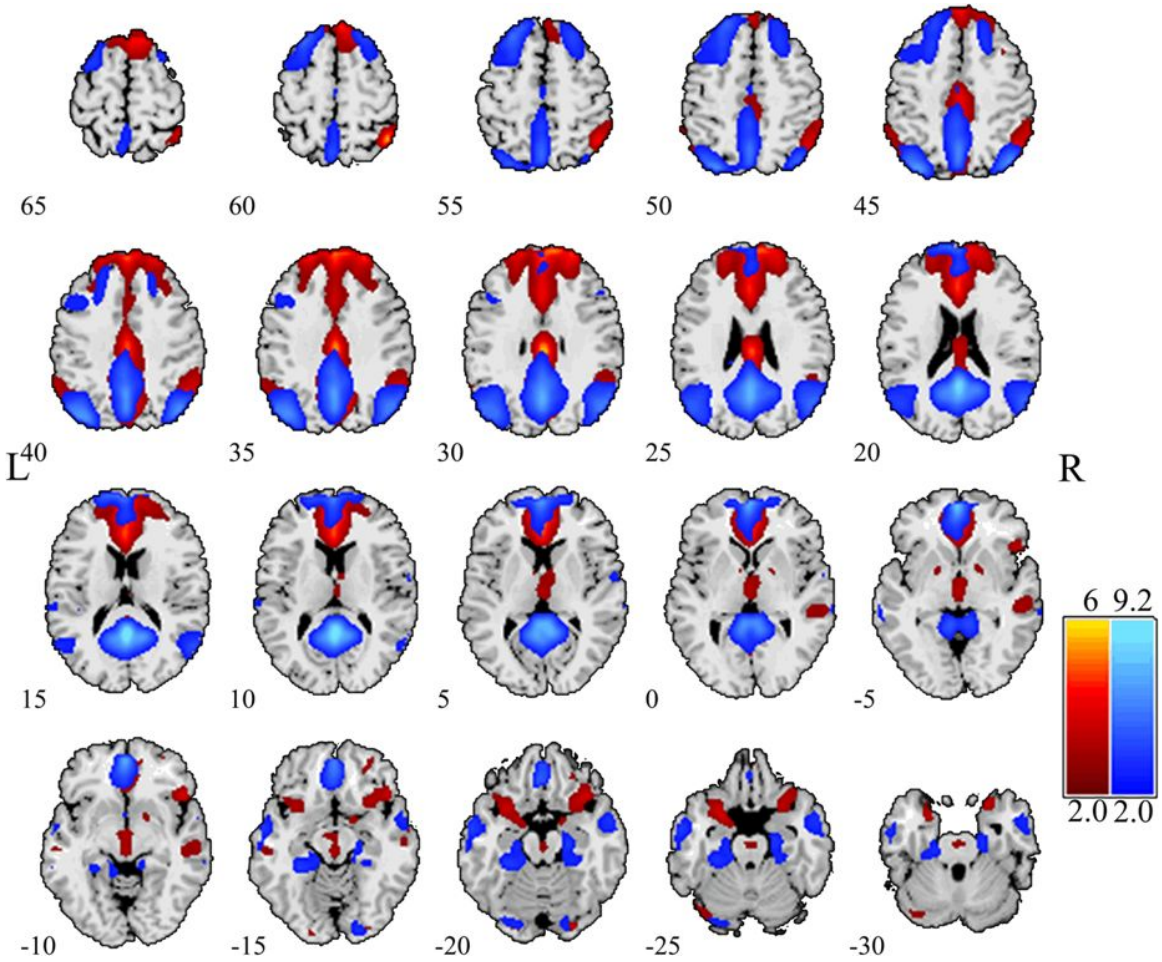


Figure 6.1: Anatomy of the default mode network, as defined by Xu et al. [219]. Regions of the anterior DMN are denoted in red, with regions of the posterior DMN demarcated in blue. Results were inferred by spatial ICA, with the two modes representing two of the inferred components.

the metastates identified by Vidaurre et al. [214] similarly identifies one metastate oriented to sensory area processing and the other associated with higher order cognition. Turning to MEG resting state studies, Baker et al. [120] showed a similar temporal delineation, with the DMN less likely to transition to sensory states. The data suggest that a large amount of the variance in resting state activity can be understood through these two functions, with resting activity on slower timescales moving through periods of focus on sensorimotor processing of external stimuli and periods of higher order cognition that integrates information from distinct networks.

6.1.2 The Brain's Default Mode Network

The collection of brain regions associated with this higher level cognitive function are collectively known as the default mode network. It is subdivided into two main regions (or subnetworks); the anterior part (aDMN) and posterior part (pDMN), as outlined in figure 6.1. This figure derives from a moderately low dimensional spatial ICA analysis of resting state fMRI data [219]; utilising more or less components can amalgamate these into a single network [220] or further subdivide them [20], however the anterior / posterior segmentation is one that can be detected in MEG using the same analysis methods we pursue here [32]. The aDMN is taken to include the dorsal medial prefrontal cortex, anterior cingulate cortex and anterior temporal lobes; the pDMN is taken to include the posterior cingulate cortex, the posterior inferior parietal lobule, hippocampal and temporal cortices [219].

The precise function of the default mode network, while extensively studied, remains difficult to conclusively pin down. It was originally referred to as a task-negative network, as it was discovered to anti-correlate with a wide variety of fMRI tasks [221] and became known as the default mode network as it was believed to represent the brain's resting baseline level of activity. Indeed, it has been shown that this network increases in activation during various stages of reduced consciousness such as sleep [222, 223] and anaesthesia [224].

However further analysis of its anatomy and the role of those regions may indicate this is to some degree a result of the types of tasks being presented, which generally involved orienting attention towards an external stimulus. This theory emphasises DMN function as broadly associated with various aspects of 'internally generated cognition'; and that these happen to typify certain aspects of mind-wandering activity that occur between trials in tasks. The components of the aDMN have been variously associated with processes that support self-referential mental activity and introspective thought [225], and has similarly been implicated in a number of disorders of emotional regulation, suggesting a role in processing the emotional content [226]. The pDMN is most closely associated with recollection of past experiences [227] and working memory retention [228], as well

as imagining the future [219, 229]. It seems plausible that spontaneous cognition between trials and during resting state scans tend towards these functions, for example through mental reconstruction of events from the past or future, that variously involve self-referential and emotional content; and that these processes are ceased upon presentation of a stimulus reorienting the participant to the task [230]. Nonetheless it is unclear whether such spontaneous, online thought processes alone would explain the persistence and degree of DMN activity in resting data - and indeed during sleep and anaesthesia - or whether the DMN may support other processes beyond deliberate cognition [227].

6.1.3 Resting State Alpha Oscillations

Much of the resting state literature, being dominated by fMRI studies, is blind to the role of oscillations in cognition. In a surprising number of ways the evolution of thinking on the role of alpha oscillations parallels that of the DMN; early frameworks that interpreted alpha activity as an idling or offline rhythm have now been overtaken by a view of alpha that quite explicitly supports and organises certain aspects of cognition.

The ubiquity of alpha oscillations is evident in the fact that they were the very first oscillatory patterns to be discovered, when Hans Berger famously measured the EEG potential over posterior scalp regions of awake subjects. Alpha activity, defined in the 8-13Hz range, remains by a significant margin the strongest electrophysiological signal measurable from the human waking brain [231]. Until relatively recently, the association of alpha desynchronisation with visual stimulation was viewed as evidence for alpha's role as an idling rhythm reflecting a lack of processing over the areas it was concentrated [232].

This view however has been refined in light of a number of studies associating spatial attention with alpha increases in areas contralateral to the attended stimuli [234–236]. Further work extended the understanding of alpha from a role in visual field attention to a role in working memory maintenance, with alpha power over parietal areas linked to memory retention [237–239]. These and other studies led to

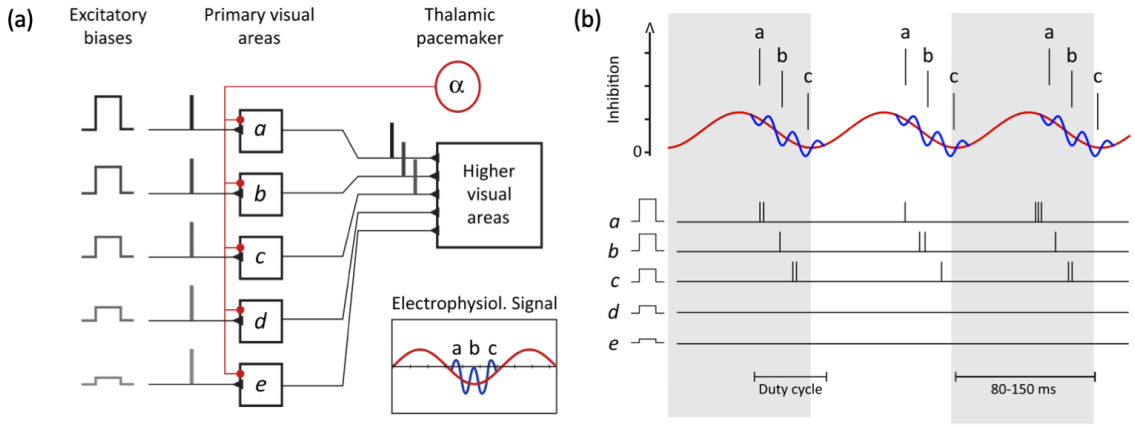


Figure 6.2: Alpha temporal coding: adapted from Jensen et al. [233], alpha oscillations both inhibit irrelevant stimuli and temporally order competing items into a coded sequence. (a) When multiple items are fed into lower visual areas, each have different levels of excitability. Centrally paced alpha oscillations provide inhibition over all of these areas, affecting the timing of their output to higher visual areas. (b) As inhibition is reduced on the decreasing edge of the alpha cycle, the most excitatory representations (in this case item a) fire first, followed consecutively by the remaining items in order of their excitability. Recurrent inhibition in an interneuronal network is proposed to keep representations apart in time.

the theory of alpha as an inhibitory signal that supports cognition by gating selected inputs; in the case of visual attention, capacity is enhanced by deliberately gating irrelevant stimuli [235]. In the case of working memory, alpha increases inhibit sensory information that might interfere with a maintained visual representation, thereby supporting effective maintenance of the previously acquired image [240].

Further to the association of alpha power with inhibition of sensory input, there is emerging evidence that alpha rhythms provide a central clock that times and organises individual events. There is significant evidence of phase dependent sensitivity to perception, for example detecting a very transient or masked stimulus [241–243]. It has also been shown that alpha phase modulates both the firing of individual cells [244] and the amplitude of gamma oscillations [245]. Putting these together supports a model where nested gamma oscillations clock the processing of successive items; this has been variously referred to as an 'alpha to-do list' [233] or a 'gamma slide-show' [246], and can be thought of as the visual-cortex equivalent of theta cycle phase coding in the hippocampus [247, 248]. It suggests

a much more central role for alpha in organising distinct neural events than was previously believed.

Given the parallel evolution of thinking around the DMN and alpha oscillations - from reflecting default or idling patterns respectively to far more central cognitive functions - it is perhaps fitting that the two have been shown to be linked. While caution must be exercised in comparing oscillatory power increases with the BOLD signal, combined EEG / fMRI studies show significant correlations between parietal alpha and activation of the default mode network [220, 249]. Furthermore MEG resting state analyses that replicate the same anatomical areas of the DMN find that pDMN power in particular is concentrated in the alpha band [32]. These analyses were the original basis of the TIDE-HMM model introduced in chapter 2, which we will again utilise in this chapter.

6.1.4 Network Mechanisms associated with hippocampal replay

Finally, we must ask what types of network activity might be expected to correlate with replay. While replay has never been studied within the context of whole brain RSN analysis, rodent electrophysiology has identified a number of specific localised phenomena that are associated with distinct forms of replay. These relate to replay during theta cycles and replay during sharp wave ripples.

Replay during Sharp Wave Ripples

The original discovery of replay of place cells was tied to events called sharp wave ripples, recorded during sleep in rodent hippocampus. Sharp wave ripples have been described as the most synchronous neural event in the mammalian brain [169]; they occur during 'offline' states, and are characterised by large amplitude negative polarity deflections in CA1 (the 'sharp wave') that are often, but not always, followed by high frequency oscillations in the 110-200Hz range also in CA1 (the 'ripple') [96]. It is these high frequency oscillations that replay the cellular representations of learned experience in rapid, temporally compressed sweeps.

In this context, 'offline' refers to either sleep or wakeful immobility. During slow wave sleep these events occur with particularly high density compared to wakeful rest, and are believed to be phase modulated by slow waves less than 2Hz in frequency [169], however wakeful state ripples occur in the absence of this slow wave activity. The propensity for sharp wave ripples to project from the hippocampus to widespread cortical regions would raise the possibility of association with brain networks that include the hippocampus; indeed, Kaplan et al. [250] found a correlation between EEG bursts in the 80-180Hz range, defined as ripples, and activation of the default mode network in monkeys.

Theta Cycle Replay

Replay during sharp waves is considered a different phenomenon to that observed during theta cycles. Compared to the 'offline' states above, this form of replay is entirely online during active exploration. Johnson and Redish [251] identified the replay of trajectories encoded in place cells occurring around critical decision points within a maze, that always swept ahead of the animal and simulated the possible options before it. Local field potential recordings showed these trajectories were explored within theta cycles [252, 253], and are proposed to represent choice evaluation processes packaged into theta cycles [254, 255]. While it is not known whether this would correlate with whole-brain networks, theta band activity is a hallmark of several MEG defined RSNs [32, 120] which could also correlate with the onset of replay in this experiment.

6.2 Methods

The model introduced in chapter 2 was a novel application to LFP data of a model previously introduced for MEG analysis [32]. We here return to use the same model in its original context, to analyse the resting state data collected in the replay study of [31]. This unsupervised model extracts the canonical resting state networks and their activations in time independently of the previous analysis of

replay events, allowing us to investigate whether these replay sequences correlate with specific whole brain network activity.

6.2.1 Source space processing

In addition to the preprocessing steps outlined in section 4.2, data from both resting state scans of all 22 subjects were projected onto an 8mm uniform grid of the MNI152 atlas using a linear constrained minimum variance beamformer [256, 257]. Beamformed data was then parcellated into 38 Regions of Interest (ROIs) covering the entire cortex excluding subcortical areas [258]. Leakage correction was then applied to each parcel using the method outlined in Pascual-Marqui et al. [259]; finally sign ambiguity corrections were made as in Vidaurre et al. [83].

6.2.2 Time delay Embedded Model

We return to the model introduced by Vidaurre et al. [32] and used in chapter 2, repeating here for clarity the observation model:

$$P(X_{t-\tau:t+\tau}|Z_t = z) \sim N(0, \Sigma_z) \tag{6.1}$$

where we set the number of temporal embeddings $W = 2\tau + 1 = 15$ samples, and set the number of states $K = 8$. Under this model, the data is described by a zero-mean Gaussian observation model. Each latent state Z_t is then associated with a distinct covariance matrix Σ_z, effectively capturing different patterns in both the autocorrelation of data on each channel and the cross correlation over different channels.

There is one difference in how this is applied to the MEG data compared to LFP data of chapter 2; given the moderate dimensionality of our source space MEG data and the poor signal to noise ratios, we operate on a reduced dimensionality space, taking the first 84 principal components of the source space data. In addition to the computational advantage this provides, the $1/f$ power spectrum of MEG means these components are typically dominated by lower frequency content that is believed to be a more reliable driver of state dynamics than for example gamma band activity

[32]. We also deviate from the standard variational inference algorithm and employ a stochastic learning variation that is more scaleable to larger group datasets [260].

6.2.3 State Spectral Analysis

The time-delay embedded HMM model parameterises frequency and coherence features in the time domain. As in chapter 2, we convert these to the frequency domain for easier interpretation. For this purpose we employ a multitaper due to its frequency estimation benefits over analytical methods [83]. Finally, to enable easier interpretation, we collapse the full frequency spectrum into a low dimensional number of modes that best explain the frequency domain features; this is done using non-negative matrix factorisation as in Vidaurre et al. [32] and chapter 2.

6.3 On the Temporal Structure of MEG Resting State Networks

The model fit to the resting state data allows us to interrogate the temporal structure of resting state network activation prior to considering the relationship with replay onset. We can analyse both the temporal statistics of state activation through time, in addition to the inferred transition matrix structure outlining the probability of transitions between each pair of states.

6.3.1 Hierarchical clustering of transition matrix structure

One of the inferred model parameters is the transition probability matrix Θ, whose (i,j)th entry denotes the probability $\Theta_{i,j} = P(Z_t = j|Z_{t-1} = i)$. As a parameter this transition matrix fully describes the Markovian system dynamics, and we can investigate whether any hierarchical structure exists such as that found with fMRI resting state networks in Vidaurre et al. [214] by analysing the structure of this matrix. This allows us to visualise the dynamics as a directed graph structure as in figure 6.3, where nodes' relative proximity is reflective of their transition probabilities [261]. Noting that state numbers are arbitrary, we label states consecutively from

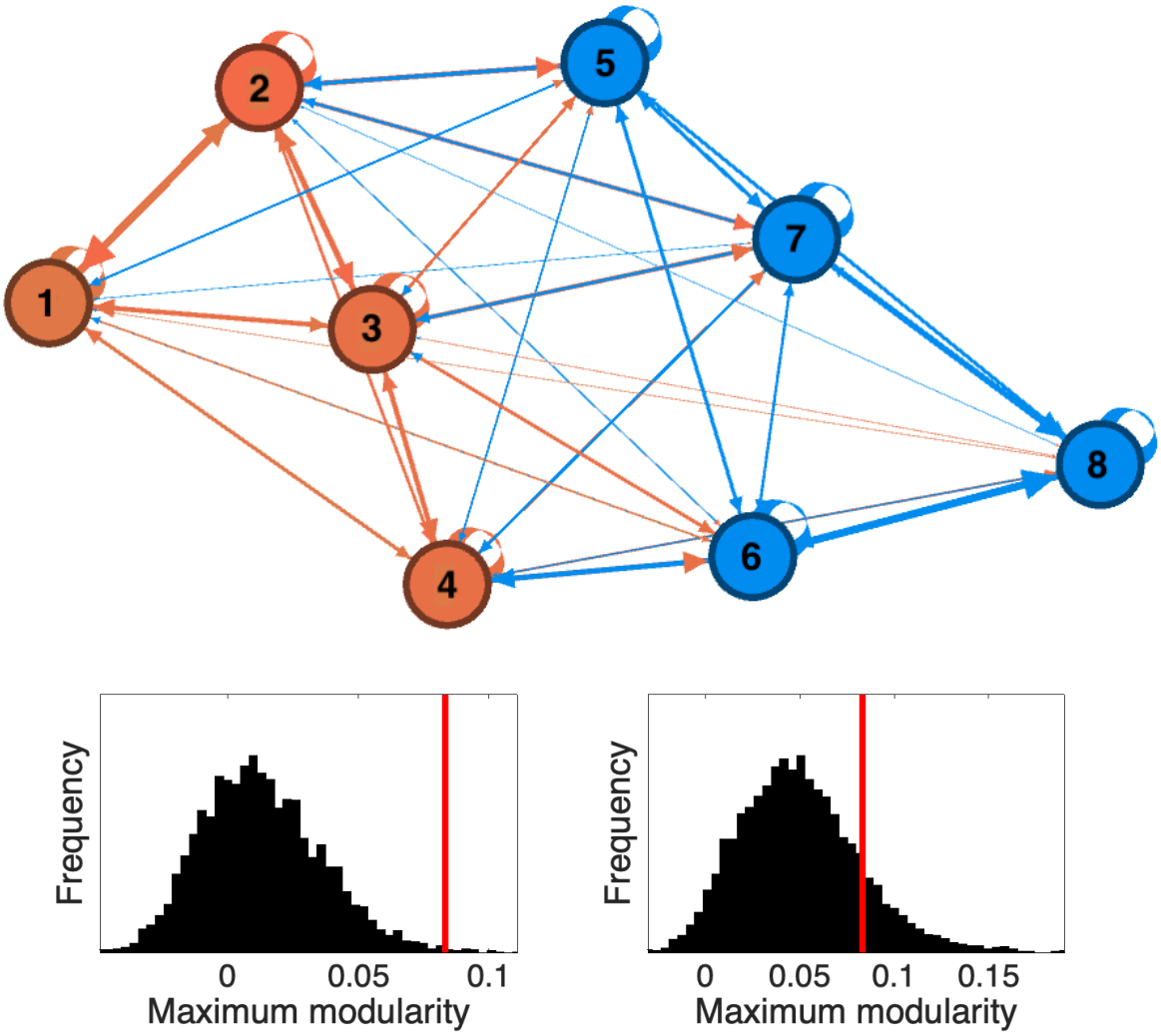


Figure 6.3: Visualising the HMM transition probability matrix from resting state MEG: [top] plot of the transition matrix in directed graph form [261] shows each state as a node, with distances between nodes reflecting their transition likelihoods. Applying a hierarchical clustering to this structure identifies two metastates, coloured in red and blue. This clustering achieves a modularity score of $M = 0.08$; statistical tests show that this hierarchical structure is significantly more modular than expected by chance when comparing to randomly generated directed graphs (bottom left, $p = 1.8 \times 10^{-3}$), however this result is no longer significant when comparing to modularity of randomly generated undirected graphs (bottom right, $p = 0.13$)

1 to 8 along the horizontal gradient that emerges in this visualisation, which will later be shown to have functional relevance.

As the transition matrix is not symmetrical, this must be analysed as a directed graph, and we therefore utilise the community discovery algorithm outlined by Blondel et al. [262] which can efficiently handle this bidirectional structure. Briefly, this algorithm finds hierarchical groupings of nodes that maximise the modularity of the system, where modularity is defined as follows:

$$Q = \frac{1}{K} \sum_{i,j} \Big[\Theta_{i,j} - \frac{k_i k_j}{K} \Big] \delta(c_i, c_j)$$

where K is the number of nodes, $\Theta_{i,j}$ is the i,jth entry of the transition matrix, $k_i = \sum_j \Theta_{i,j}$ is the sum of transition probabilities directed towards node i, c_i is the community to which node i is assigned and the delta function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise. Qualitatively, this is a metric that varies from -1 to 1, where higher values represent groupings of increasingly dense interconnections within groups and increasingly sparse connections between groups.

Fitting this algorithm to the transition matrix structure subdivides the eight states into two hierarchical metastates, denoted by the colours red and blue in figure 6.3. This tells us that the resting state data has a tendency to alternate between states 1-4 for extended periods of time, before switching into the second metastate described by states 5-8. This is similar to the finding of [214], which found a similar emergence of two hierarchical metastates within the much slower dynamics of fMRI data.

We then wish to ask whether this hierarchical structure is significant or could arise from random chance. We do this by comparing the modularity score obtained in the above clustering with a null distribution of modularity scores that might be obtained by random chance. This second step, however, is dependent upon strong modelling assumptions regarding what kinds of transition matrices may arise by chance - these assumptions are used to generate transition matrices, each of which allows us to derive a modularity score. We initially assume that each off-diagonal element of the transition matrix is generated from a uniform distribution

on the interval $[0, M]$, where M is defined as the maximum observed value in the original transition matrix inferred above. Diagonal elements were then computed as $A_{i,i} = 1 - \sum_{j \neq i} A_{i,j}$. We then generated 5000 such transition matrices and fit the same community discovery algorithm to find the modularity of inferred clustering. Under this assumption, we find that the modularity score obtained from the resting state network data is much more strongly organised into hierarchical clusters than could be expected by random chance ($p = 1.8 \times 10^{-3}$).

However, this analysis assumes that each element of the transition matrix is independent; for example that $A_{i,j}$ is completely independent of $A_{j,i}$. In practice, while transition matrices are not symmetrical there is a tendency for off-diagonal elements to covary strongly. For this reason we generated a second set of transition matrices that were perfectly symmetrical, with each $A_{i,j} = A_{j,i}$ generated from a uniform distribution on the interval $[0, M]$. Under this assumption, the modularity score was still greater than the mean of the null distribution but no longer statistically significant ($p = 0.13$). Consequently while we can state that the observed resting state dynamics are more hierarchically organised than a random generated undirected graph structure, we cannot make the same claim with regards to directed graph structures.

6.3.2 Temporal Statistics of Inferred States

We can also better understand the temporal dynamics of these metastates by investigating their temporal activation profiles.

Fractional Occupancies

The plots of figure 6.4 highlight the fractional occupancy of each state, or total proportion of time that each state is active over all subject's resting state data. Recall that states 1-8 were so labelled by their emergence on the horizontal axis of figure 6.3; it is therefore interesting that an almost monotonically decreasing pattern emerges in the fractional occupancies of these states. The first metastate comprising states 1-4 make up the four highest fractional occupancies; the remaining

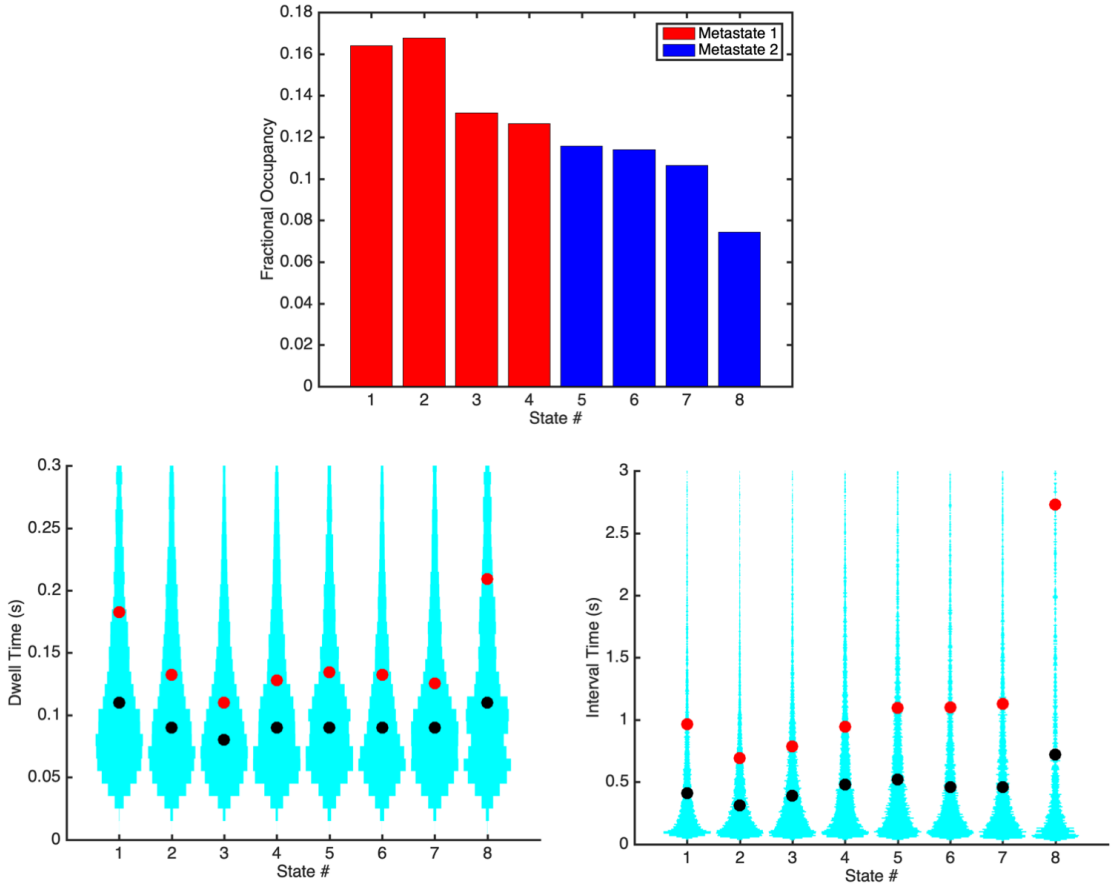


Figure 6.4: State temporal statistics. The fractional occupancies, dwell time and interval period statistics for each state. Clear patterns emerge that can be understood in terms of the metastates inferred previously, with metastate 1 explaining larger portions of the data and exhibiting more regularly timed intervals than the second metastate. States 1 and 8 that emerge at opposite ends of the network are both the most stable in terms of typical dwell times.

states together account for only 40% of the remaining data, with state 8 explaining markedly less data than any other state.

Mean lifetimes

Plotting the average lifetime of each state activation reveals that overall, states tend to be active for periods on the order of 100msec, displaying the same fast switching characteristic previously described by Baker et al. [120]. However we can further identify that states 1 and 8 each have much longer activation times than the remaining states. These states mark opposite ends of the horizontal gradient of

figure 6.3, and their longer lifetimes show that they are the two most stable states that tend to endure for longer than the intermediate states 2-7.

Mean Interval Periods

Plotting the distribution of intervals between state activations identifies a notably longer period between activations for state 8. Additionally, the difference between the mean and median interval is much more pronounced, suggesting the distribution of intervals for this state is longer-tailed than the remaining states.

Interpreting the horizontal axis of figure 6.3

Putting the information on these three plots together, we can see that resting state activity disproportionally concentrates in the first of the two metastates, with states further to the left of figure 6.3 active more often than states further to the right. Visits to states at the furthest ends of this axis are however also the most stable, with states 1 and 8 each having the longest dwell times of any states. Finally the intervals between successive state visits become both longer and more irregular towards the right hand side of this plot, with state 8 having markedly more irregular periods between state visits.

6.4 Characterising the Resting State Networks

To investigate further what types of activity these networks represent, we can inspect their spatio-spectral profiles. As discussed in chapter 2, the TIDE-HMM model separates states based on both power and coherence, and this can be computed for each state.

6.4.1 Global Frequency Domain Features

Global Power Spectral Density

Taking first the spectral characteristics averaged over all ROIs, figure 6.5 demonstrates significant differences between different states in two particular frequency bands. Significant variation between states is expressed in a low delta/theta band from 1-5Hz, with a second mode of variation in the alpha band of 8-12Hz.

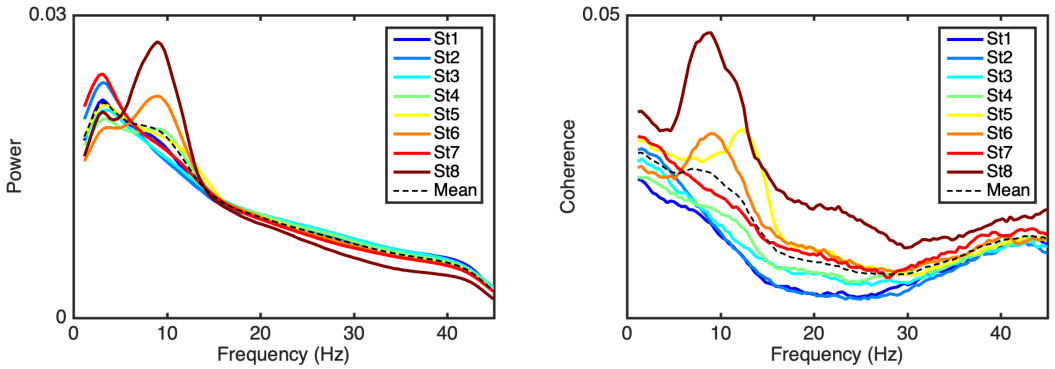


Figure 6.5: States by spectral features. The average power spectral density over all channels for each state shows states 7 and 8 are associated with the highest power in the alpha and delta/theta bands respectively. Averaging the coherence over all pairs of channels shows these states, and indeed all components of metastate 2, are associated with much higher overall coherence across all frequency bands.

In particular, states 7 and 8 represent maximal power in each of these delta/theta and alpha bands respectively.

Global coherence

Plotting the average coherence between all pairs of ROIs as a function of frequency reveals some striking differences between states. A broadband scaling of the overall coherence of each state is apparent, with states 1-4 that comprise the first metastate being associated with lower overall coherence than all states in the second metastate. State 8 in particular is associated with much higher overall network coherence across all frequency bands.

Turning our attention to specific frequency bands, it is apparent that states 7 and 8 reach their peak coherence in the delta/theta and alpha bands respectively, alongside this overall coherence increase.

6.4.2 Spatial Distribution of Power and Coherence

We can now turn our attention to how these changes in power and coherence are distributed spatially, with specific attention on states 7 and 8. Plotting the anatomical distribution of bandlimited power identifies these states as the anterior and posterior components of the canonical default mode network. State 7 has power concentrated in anterior cingulate and temporal areas in the delta/theta

range, with these areas also showing high network coherence within this band. State 8, in contrast, includes the precuneus along with parietal cortex, comprising the posterior regions of the DMN.

6.4.3 Combining the Temporal and Spatial Organisation of MEG Resting States

We have identified a series of rapidly switching brain networks that organise temporally into two hierarchical metastates; states at the furthest ends of this network are the most temporally stable in terms of dwell times, however they are not visited equally. States 7 and 8 comprise the regions of the default mode network, suggesting an overall functional association of that metastate with higher order cognition and network activity that is more coherently organised around certain frequency bands. These states are visited infrequently but also irregularly, as measured by the long tailed distribution of state intervals. Overall, the characteristics of states 7 and 8, and which correspond to anterior and poster DMN, replicate the findings in Vidaurre et al. [32].

States comprising the other metastate are characterised by more widespread desynchronisation as reflected by lower network coherence; their activity is concentrated around known anatomical networks that correlate with sensory areas (see appendix for full details); these states explain a larger proportion of the data, and intervals between state visits are more regularly spaced. Overall, the organisation of the states by their temporal structure in figure 6.3 parallels the metastate organisation found in fMRI, which positions the DMN at the far end of a scale from more direct sensory processing areas to higher order and more abstract cognition.

6.5 Resting State Network Activity Predicts the Onset of Replay

We now turn to whether the activation of these networks predicts the onset of replay. In particular, we are interested in whether individual replay events tend to occur during periods of default mode network activation, or during some other state.

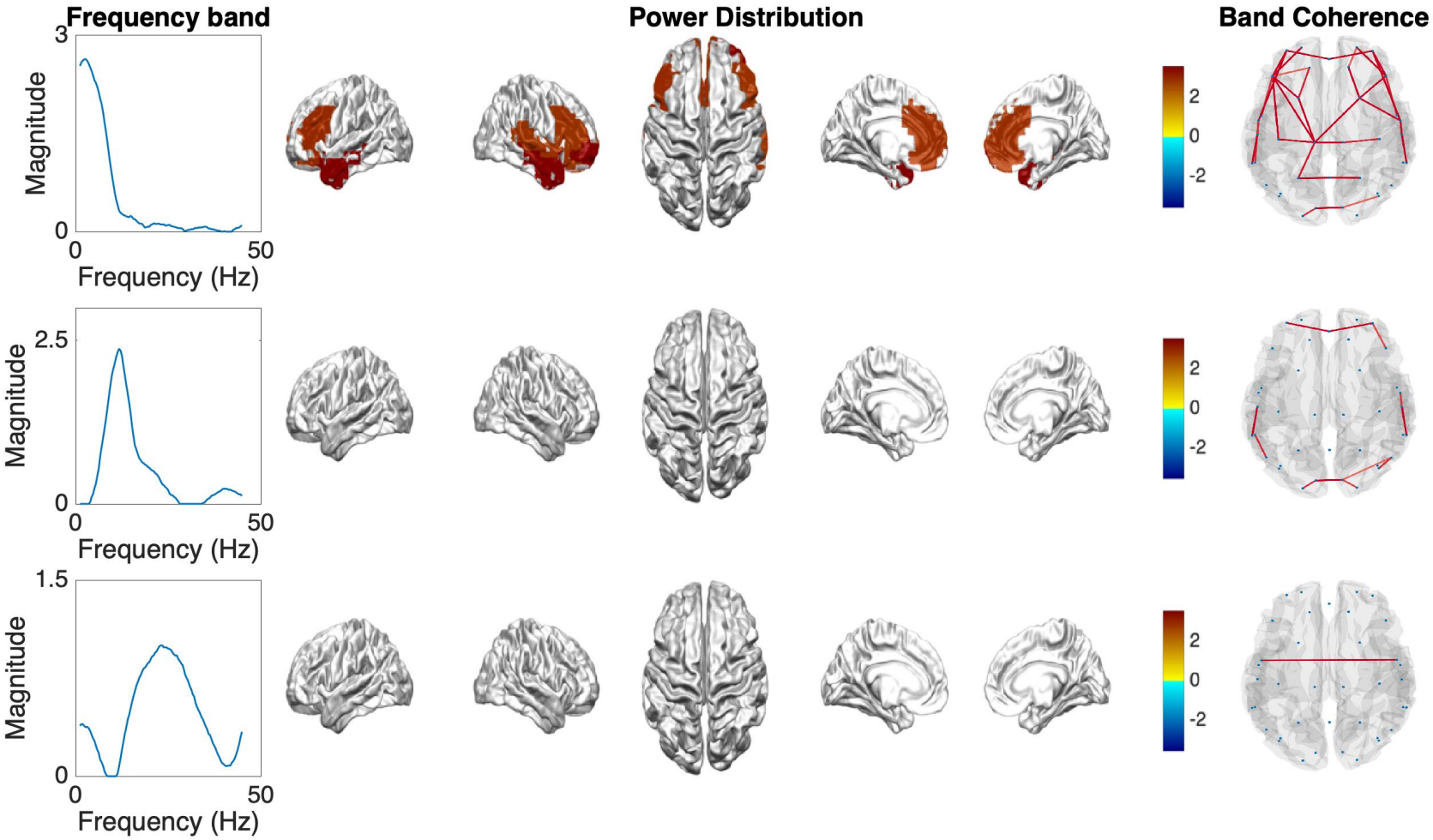


Figure 6.6: State 7 represents the anterior DMN: The three frequency bands inferred by NNMF that best explain the data, correspond to delta/theta, alpha and beta bands respectively. State 7 is associated with significantly increased power and coherence in the frontotemporal regions that characterise the anterior DMN; this is concentrated in the delta/theta band.

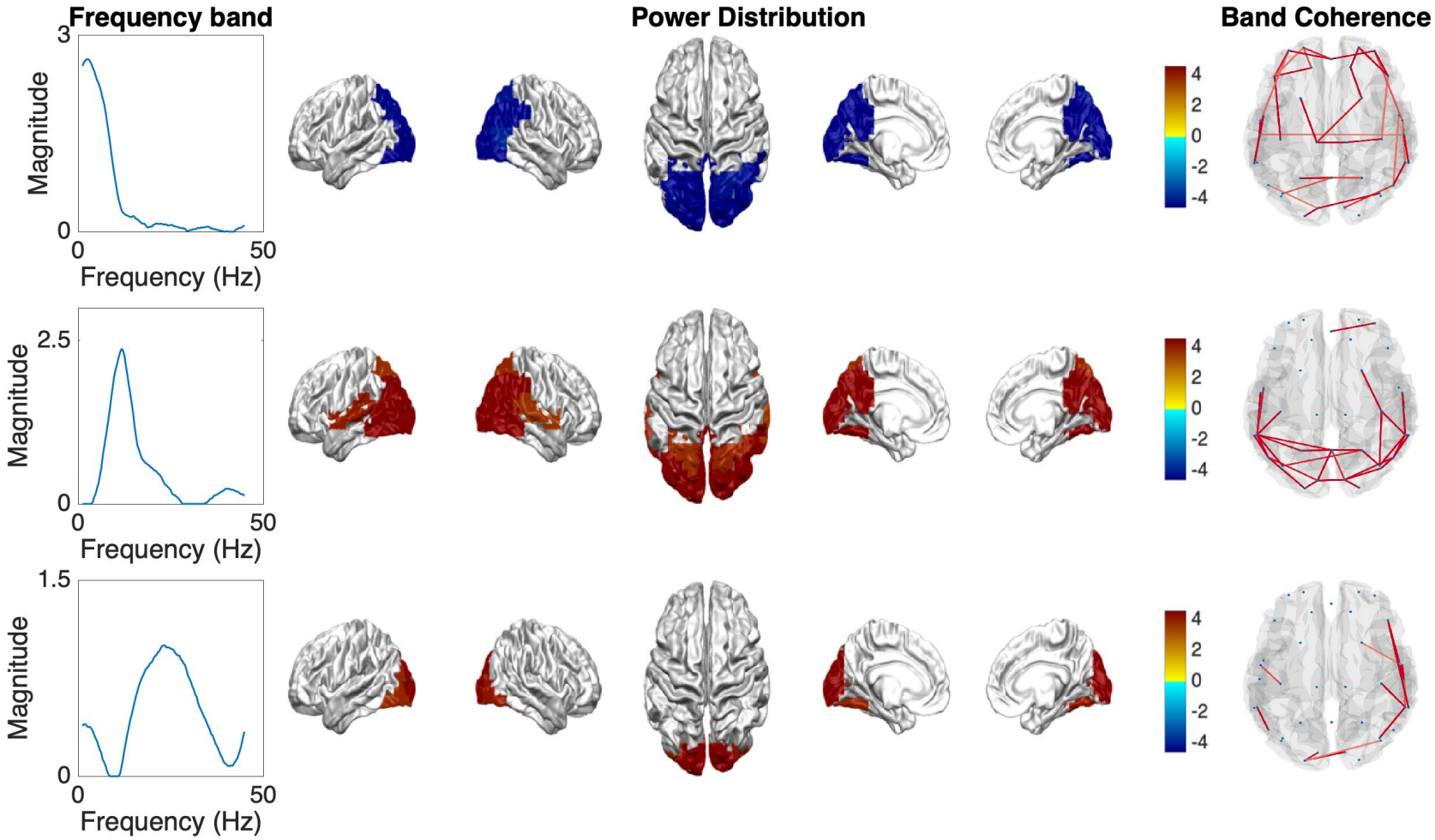


Figure 6.7: State 8 represents the posterior DMN: The three frequency bands inferred by NNMF that best explain the data, correspond to delta/theta, alpha and beta bands respectively. State 8 is associated with significantly increased power and coherence in the parietal regions that characterise the posterior DMN; this is concentrated in the alpha band.

First we must define a timecourse of replay. In section 5.3 we identified three HMM decoder states which showed significant evidence for sequential reactivation of items. For each of these three states, we identify the peak sequenceness period; that is, the time lag at which the sequenceness measure is maximised, which we define as τ_k for each state k. We then define a replay score in time as follows:

$$r_t^{(k)} = \sum_{a,b \in V} P(Y_t = a) P(Y_{t+\tau_k} = b) \tag{6.2}$$

$$= \sum_{a,b \in V} \sigma(X_t \beta_a^{(k)}) \sigma(X_{t+\tau_k} \beta_b^{(k)}) \tag{6.3}$$

where V is the set of valid sequences as defined by the experimental setup, σ is the logistic sigmoid function, β are the learned classifier coefficients for each stimulus and HMM state, X_t is the resting state data, and Y_t are the stimulus labels (which are unknown but being estimated for the resting state data). This whole equation can also be viewed in logic terms; the summation implements a logical *or* operation over different possible pairs of items a and b, while the inner product implements a logical *and* operation over the two representations with a timelag of τ_k.

6.5.1 DMN Correlates with Replay

We can now directly check for whether this replay score correlates with the activation of any particular resting state networks. Note that the resting network state timecourses are inferred using the TIDE-HMM in an unsupervised manner, i.e. with no knowledge of the timing of the replay events.

We identify candidate replay events by thresholding the replay score $r_t^{(k)}$ at the 99th percentile, and exclude superthreshold points that are immediately adjacent in time. Using these timestamps of replay onset and conducting an epoch analysis of the distribution of RSN states around each replay event, As shown in figure 6.8, identifies significant correlations of replay onset with state 8, the posterior components of the default mode network, and also with states 7 (the anterior DMN), and a temporally leading correlation with state 3 (a component of the same metastate). States 1-3, associated with the first metastate and anticorrelated with the default mode network, similarly exhibit strong anticorrelations with replay.

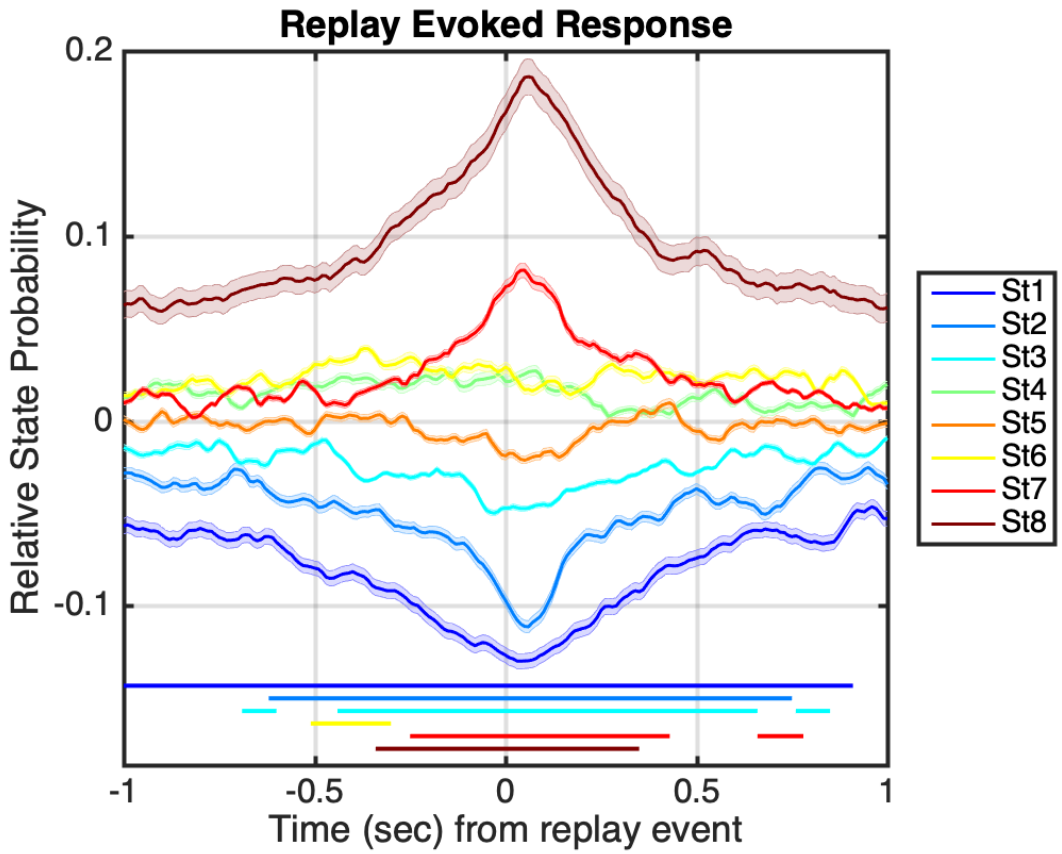


Figure 6.8: Replay onset is predicted by resting state network activity: Regressing the time from a replay event onto the inferred state timecourses shows a strong correlation between replay and resting network states that comprise the DMN. A strong anticorrelation is found with states 1 and 2 that are similarly known to be anticorrelated with the DMN. Significance bars show periods of significance to $p < 0.01$.

This finding is displayed for the HMM classifier state 4 from the previous chapter; it generalises to both other classifiers (i.e. classifiers derived from states 1 and 6 in figure 5.8 which were used to infer the top-down directionality of replay sweeps), which demonstrate very similar patterns (see appendix I).

6.5.2 DMN and Alpha Activity Anticorrelate with Original Stimulus Processing

One potential criticism of this analysis is that the neural activity defining states 7 and 8 may just be a closer approximation to the activity the classifiers are trained upon, leading to a correlation between these states and inferred replay. This would seem unlikely given the central notions of the DMN being a task-negative state

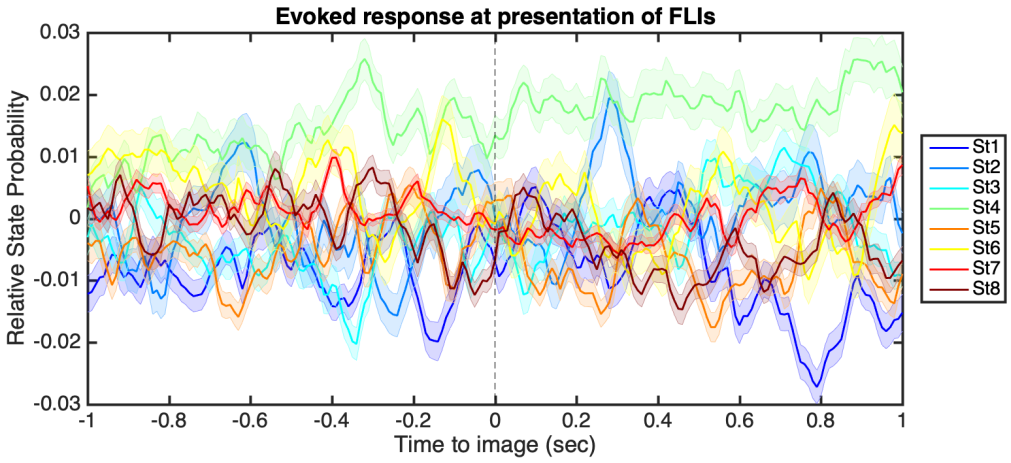


Figure 6.9: Distribution of states relative to functional localiser images (FLIs): the states that best predict replay onset do not merely match patterns of activity to that observed during functional localisers. States 7 and 8 best predict replay but have no relationship with the functional localiser data. While state 4 appears to have an elevated probability approaching and following stimulus presentation, and may be a state associated with visual processing, the result is not significant for any state.

that is anti-correlated with stimulus processing, but could not be ruled out. For this reason we refit the states to the functional localiser data, fixing the state parameters to those learned in the resting state data, to identify what states were most associated with stimulus encoding.

As shown in figure 6.9, no significant correlation between these states was observed; if anything the relationship with the DMN and states associated with it is trending towards a negative correlation, but none of these results were significant.

6.6 Longer Range Temporal Dependencies

Finally, we wished to investigate the existence of longer range temporal dependencies between distinct replay events and the activation of different metastates of RSN activity. We pursue this as a natural extension of the temporal structure between RSNs identified in section 6.3, and strong correlation between replay and the DMN states that displayed the most irregularly distributed, but also the most temporally stable temporal activation.

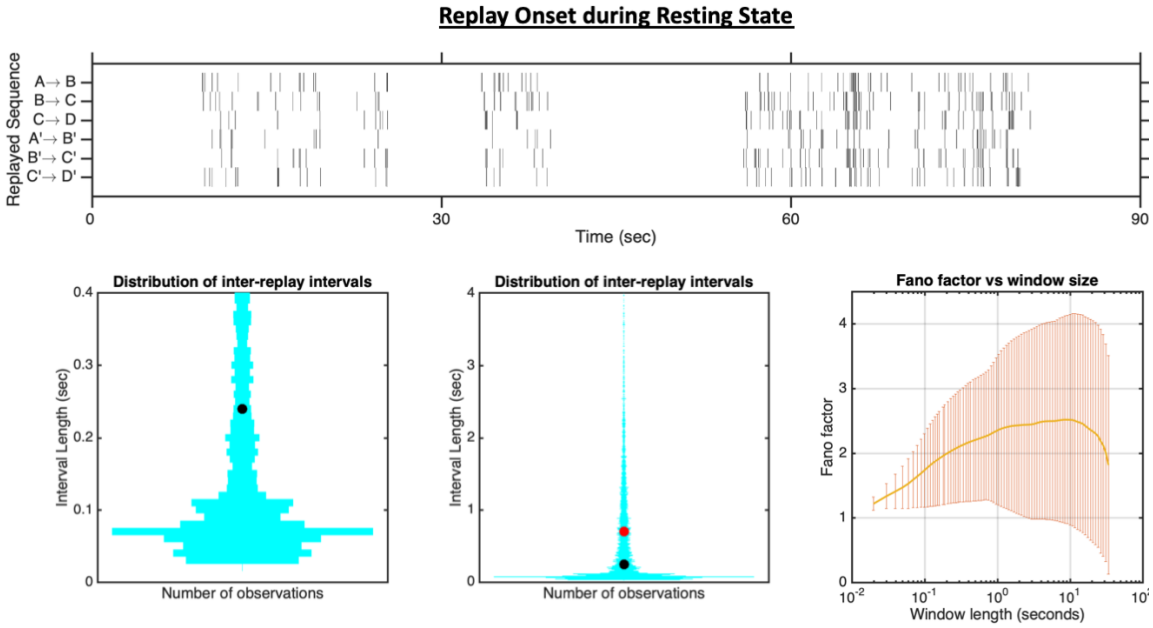


Figure 6.10: Replay temporal statistics: plotting a sample 90 second segment of one subject's replay onset timings shows a strong temporal organisation at long timescales. This can be quantified by analysing the inter-replay intervals, the distribution of which is particularly long-tailed; also by analysing the Fano factor as a function of window size, showing temporal organisation on the order of 10 seconds.

6.6.1 Replay Intervals are Irregularly Distributed

As replay is most strongly associated with the activation of state 8, and furthermore given that this state presented the most long-tailed, irregularly distributed reactivation periods, we then wished to investigate whether our estimated replay events displayed similar behaviour. This could be understood through analysing the periods between successive candidate replay events, mirroring the analysis of interval periods of the posterior DMN conducted previously, but also through analysing the relationship between the Fano factor for a variety of window sizes.

Figure 6.10 demonstrates that replay intervals indeed have a long tailed distribution, showing that these events are characterised by a number of events occurring in a burst, but also that these replay bursts are then interspersed by long periods of quiescence. This is evidenced by the degree to which the median exceeds the mean, and to which both of these statistics exceed the mode of the distribution.

6.6.2 Fano Factor Analysis

We can furthermore investigate the regularity of replay events through a model-based approach. Any event that has a constant probability of occurrence over a period of time can be characterised as a Poisson point process with rate λ. A truly Poisson process has constant emission probability regardless of recent history; that is, the rate λ is independent of the period of time since the last event. One useful corollary of this is that, for any given window of length W, the expected number of events μ emitted within that window is equal to the standard deviation σ of expected events. Defining the Fano factor as the ratio $\frac{\sigma}{\mu}$, we have a statistic that should approach unity for perfect Poisson processes, while Fano factors greater or lower than one reflect processes that are more or less regular respectively than a perfect Poisson process [263]. More simply, a high Fano Factor indicates that the events display bursting - i.e. there are periods of time where there are higher numbers of events interspersed with periods where there are lower numbers of events.

We then analysed this statistic for observed replay events as a function of window length, to estimate the periods over which this organisation is most variable. This showed replay events were most irregular over window lengths on the order of ten seconds, a surprisingly long period. This period should be interpreted loosely as it is partially dependent on the choice of threshold in defining replay events; nonetheless, it is very significant that this is on a timescale relevant to fMRI, and may partially explain some of the findings reported regarding replay in that modality [190, 191, 195]. Of relevance for our understanding of replay in the context of RSNs, note that the state most strongly correlated with replay was also the most irregular as measured by its interval distribution. This led us to investigate whether there was a more long term correlation between replay burstiness and resting state network switches.

6.6.3 Correlations with RSN Metastates

Having identified that replay events have a bursty nature, we finally wanted to investigate the distribution of RSNs around the onset and cessation of burst events.

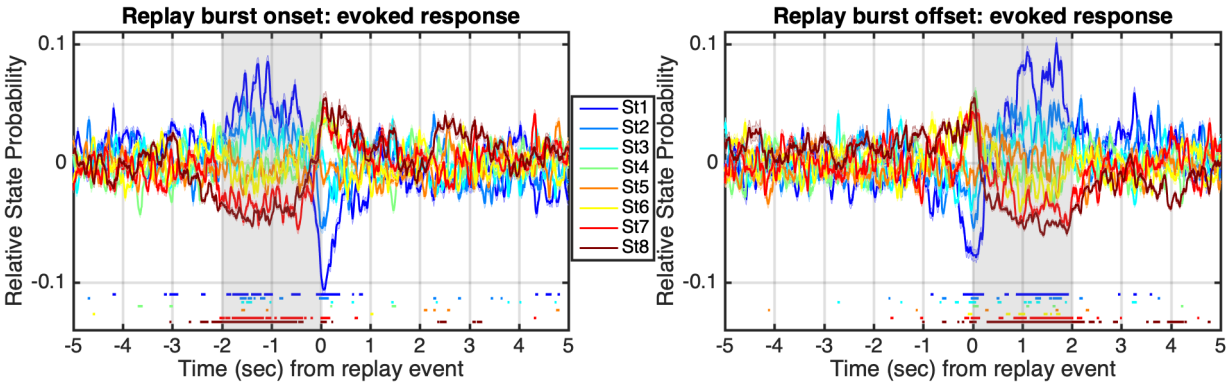


Figure 6.11: Replay burst evoked response: defining burst onset and offset as replay events for which the preceding or proceeding (respectively) interval is greater than 2 seconds, we conduct the same analysis of the evoked effects on state distributions. This confirms that the onset and offset of these burst events are strongly associated with a switch in the dominant metastate, with elevated probabilities of sensorimotor state activations and decreased probabilities of DMN-related state activations in the periods before or after a burst. Significance lines delineate points where effects are significantly different from zero to $p = 0.05$ levels.

Returning to the analysis of section 6.5, we ran a second epoch analysis, this time setting the epoch criteria as superthreshold points where either the preceding interval or proceeding interval was greater than 2 seconds (figure 6.11). This identified points that approximately represented the start or end of burst periods respectively. While the intention was to set an interval determined by the maximum Fano factor window size, this value of ten seconds was too large to permit sufficient samples; hence two seconds was selected as a compromise.

Figure 6.11 shows a strong relationship between the starts and ends of replay burst periods and switches of the dominant metastate. Immediately prior to burst onset, states that comprise the sensorimotor activity dominate the data; these rapidly switch to anticorrelate with approximately the first second of data following a burst. Similarly the end of a burst is associated with a rapid switch from DMN related states back into sensorimotor states. Note that this is partly explainable by the criteria of selection; the epochs marked in grey by definition contain no replay events, and we know from section 6.5 that this should mean an increase in states anticorrelated with replay. Nonetheless the speed of this switch is striking, with state 1 in both cases moving from significantly negatively correlated to significantly

positively correlated over a period of roughly 100msec in both cases. Furthermore, there is some evidence that this relationship is sustained beyond the two second window that defined the epochs, with burst offset still significantly correlated with decreases in DMN activity 4 seconds after replay cessation.

6.7 Discussion

6.7.1 Interpreting Replay's Correlation with the DMN

We have demonstrated a very strong relationship between the onset of replay and activation of the default mode network. Replay as a mechanism has been quite well characterised at the cellular level, where it is understood primarily as a mechanism for strengthening synapses between cells, for example to consolidate recently acquired memories. On the other hand, the default mode network has been broadly studied at the level of whole brain activation in humans, and is functionally understood in more abstract terms through associations with sleep and offline states, as well as various functions linked to higher order cognition. The reasons why these two phenomena are linked is not clear, and significant further evidence will be needed to make confident claims in this regard, however permitting some speculation allows several candidate theories to emerge.

The first is to interpret this link in light of the DMN's role in self-reflective thought and future planning. Liu et al. [31] argues that replay in humans does not reflect a mere recapitulation of prior experience, but rather supports the creation of a "cognitive map" structure into which these experiences can be stored. In this model, replay is not just a mechanism for recording memories as they are experienced in a linear fashion, but rather plays a role in structuring them by their relevance and salience for future needs. A number of studies have now established that not all experiences are replayed equally, but that replay rates appear linked to planning uncertainty [191]. In this light, it seems plausible that some higher level cognitive process would be required to prioritise memory engrams for future needs and infer the most useful structure (i.e. the "cognitive map") in which they should be stored. The DMN would appear to be the perfect candidate for this given its functional

associations with goal directed planning, episodic memory and internally generated cognition. This points to a model where the DMN engages brain areas that prioritise and structure past experiences, while spontaneous replay of these items then acts to consolidate this structure into a more stable long-term representation.

On the other hand, both replay and DMN activity were originally understood as offline processes. Replay was first studied in sleep and only later found to also occur during wakeful rest; the DMN was only found to associate with internally generated cognition some time after it was characterised as the brain's default, task-negative state. Indeed the two studies that most closely support the present finding were focussed on anesthetised monkeys, and observed that sharp wave ripples in EEG correlated with fMRI activation of the DMN [250] and silencing of subcortical centres linked to sensorimotor activities [264]. Replay occurs at much higher rates during sleep than in any awake state [170], and the correlation between sleep and memory consolidation is so strong it has led some to argue it is a central reason why we sleep [265]. This theory posits that sleep creates a brain state in which replay can be most effective, for example by eliminating potential interference from sensory inputs [266]. This prompts a second speculative theory for the coupling of the DMN with replay; this brain state engages oscillatory mechanisms such as increased alpha and heightened global coherence that may be necessary for replay to occur without excessive interference from sensory areas.

These are two plausible reasons for the observed association between the DMN and replay, however there are numerous other possibilities and the evidence at the moment is not nearly strong enough to favour any of these. What we can state with confidence is that the two phenomena are linked, and uncovering why is likely to shed light on fundamental aspects of brain organisation.

6.7.2 Replay and Alpha Oscillations

Putting aside the overarching reasons why the DMN and replay would be linked, more immediate insights might be made by investigating the role of alpha oscillations relative to replay. Alpha activity is understood as an inhibitory signal that acts

to gate irrelevant stimuli and potentially also to organise distinct visual features through a phase coding scheme [233]. This is the visual system equivalent of the more well characterised theta-gamma coding in the hippocampus [233, 247]. The state that replay is most strongly linked to is characterised by increased alpha power and coherence over posterior DMN regions; notably though, it is also linked to a weaker extent with the anterior DMN, which is characterised by power in frontotemporal regions in the delta/theta range. An open question is whether replay timings are functionally linked to either of these oscillatory signals.

The distribution of inter-replay intervals in figure 6.10 identifies a large number of intervals that are less than 100msec, followed by a long tail. This raises the possibility that these sequences are organised within cycles of an underlying oscillation. It is not clear from the literature whether one could expect these to be organised by alpha phase - the dominant frequency over visual areas and the frequency band we see most correlated with replay; or theta phase - the dominant frequency band of the hippocampus that may be the origin of replay projections. This forms a fascinating line of enquiry that will be pursued in further work, however is also considerably challenging due to the variability of alpha and theta phase over cortex and difficulty disambiguating a true neural relationship from a mere artefact of linear classification during strong alpha periods. Nonetheless, finding any mechanistic relationship would support our understanding of how network oscillations support replay, which would in turn contribute substantially to the question posed above of uncovering why replay is linked to the DMN.

6.8 Conclusion

In the previous chapter, we used the dynamic decoding model introduced in chapter 3 to identify replay of visual stimuli at multiple levels of the visual hierarchy, and showed that these propagate in the reverse direction to the original feedforward activation, instead activating in a top-down sweep. We here analysed the same data using the unsupervised model introduced in chapter 2 to characterise the resting state network activity and showed that it replicates the metastate organisation

previously characterised in fMRI studies. We then found a strong correlation between replay onset and the activation of networks states in resting MEG data that comprise components of the DMN, occupying one extreme of the network graph and functionally linked to aspects of higher order cognition. Finally, we showed that this association extended to longer timescales, showing that replay events are highly variable over longer timescales, and that the onset and offset of transient replay 'bursts' is predicted by transitions from one brain metastate into the other.

Chapter 7

Conclusion

Contents

7.1 Contributions of this Thesis 158
7.2 Future Research Directions 160
7.3 Closing Remarks 162

Neuroscientific research requires understanding both the spontaneous dynamic activity of the brain and the specific patterns induced by experimental stimuli. These parallel goals are often studied separately, with different tools employed to analyse one or other facet of recorded brain activity. This thesis has presented an attempt to merge some of these methods, creating novel frameworks for supervised data analysis that blend the traditional GLM approach to task data with the dynamic inference framework of the HMM, thus enabling a number of unique insights into neural function that were not apparent using standard analysis methods.

7.1 Contributions of this Thesis

Multivariate Modelling as a Tool for LFP Analysis

LFP data represents a unique insight into brain network activity at the mesoscale, typically recorded over multiple sites and normally alongside collection of single unit data. Despite the richness of these datasets, analysis pipelines overwhelmingly rely upon univariate feature extraction. In chapter 2 we proposed a multivariate model previously introduced for MEG analysis [32], and demonstrated the potential

advantages that it had over traditional univariate approaches, including heightened sensitivity to distributed variations in network activity, and the ability to infer behaviourally relevant latent states.

Dynamic Encoding and Decoding Models

Encoding and decoding models are the two fundamental approaches to supervised data analysis in neuroscience, but in their standard formulation they ignore variations in timings over different trials. We extended the work of Vidaurre et al. [33] to cover logistic regression and multivariate encoding models, providing a framework to infer meaningful variations over trials in classification tasks. This innovation resulted in improved classification accuracy; sensitivity to information content at broader frequencies; and more interpretable, robust activation maps.

Enhancing Replay Detection

Previous studies found that the detection of replay was limited to classifiers trained on a fairly narrow temporal window of data [31, 99]. We used the dynamic decoding models to demonstrate that this was a result of these classifiers' insensitivity to temporal variations over trials, as a dynamic model with inferred, trial-specific state timecourses was able to generalise replay detection to data at multiple timepoints. We furthermore clarified the relationship between sparsity and replay sensitivity, showing that this is likely explained by the top-down nature of replay projections that do not preserve the original spatiotemporal structure of the training data.

Replay Instantiates Top-Down Sweeps of the Visual Hierarchy

We could then investigate the temporal relationship between replayed sequences in these successive stages of the visual hierarchy. We showed that these activations are indeed temporally correlated, but the direction of this correlation is the reverse of that observed during the original feedforward processing. This reflects top-down sweeps through the visual hierarchy, where original stimulus processing progressed through a feedforward bottom-up manner.

Replay is Predicted by Activation of the Default Mode Network

Finally we showed that replay is tied to activation of the default mode network, in particular the posterior components that are strongly associated with alpha oscillations. This pattern was replicated at both short and long timescales, with replay occurring in bursts that are predicted by switches in the overall metastate of brain activity.

7.2 Future Research Directions

A number of avenues for future work exist, both from a methodological and from a neuroscientific standpoint; we here briefly explore just a few of these most likely to be pursued.

Joint Probability Modelling

There are two significant shortcomings with the supervised models introduced in chapter 3; the propensity for these models to overfit, and the inability of the model to infer state timecourses for held-out test data. In chapter two we introduced a sequential constraint to regularise the data, that was sufficient to overcome the overfitting problem. Experimentation since with other datasets has shown that this is by no means a universal solution, and that overfitting continues to be a problem with some datasets, depending on the dimensionality and signal to noise ratios. One potential solution that may solve both of these problems is to model the joint probability. Consider the logistic regression decoding model specified as:

$$P(Y|X, Z, \beta) \sim \sigma(X\beta_Z) \tag{7.1}$$

And the time delay embedded model specified as:

$$P(X_{t-\tau:t+\tau}|Z, \Sigma) \sim N(0, \Sigma_Z) \tag{7.2}$$

We can then specify a joint model as follows:

$$P(X, Y|Z, \beta, \Sigma) = P(Y|X, Z, \beta_Z)P(X|Z, \Sigma_Z) \tag{7.3}$$

This would effectively find the distribution of states over the raw data X that maximise prediction accuracy of Y. In theory, such an approach should reduce overfitting by requiring states to be fit according to a sensible parameterisation of X, rather than arbitrary mapping to the labels Y; furthermore the state parameters over X should support a reasonable estimation of state timecourses for unseen test data. Nonetheless, this is a significant piece of work that may or may not overcome the problems highlighted above; it is furthermore likely to require adaptations to one or both model specifications, hence it has not until now been developed.

Interpretation of Decode Coefficients

A widespread concern in multivariate analysis is the interpretation of decode coefficients. While some machine learning applications can tolerate uninterpretable black-box architectures that rely on the strength of their predictions, neuroscientists really need to understand how and why their model is computing a particular output. While significant theoretical work has been advanced in this regard [118], my own analyses (not included in this thesis for brevity) have highlighted certain limitations and pitfalls of some standard approaches to interpreting results. This work was not mature enough to include in this thesis, but nonetheless informed much of the data analysis conducted herein and stands to be developed further and formalised.

Correlates of LFP States in Single Unit Activity

One of the exciting aspects of the work outlined in chapter 2 was the inference of latent states that had strong behavioural correlations with error rates. This provides a novel gauge of an endogenous variable that is outside of normal experimental control, as a gauge of general cognitive engagement and focus within the task. This dataset includes single unit activity recorded across multiple spatial sites in the brain, for which the encoding regime of task variables has been well characterised [30]. An unanswered question is how endogenous variations in attention might affect encoding at the level of single units. With our inferred latent state we now have a variable that can be treated like any other experimental variable, and pursue whether changes in latent LFP states are reflected in cellular level encoding characteristics.

Investigating Alpha Phase Relationships with Replay

Finally, but perhaps more immediately, we have not yet been able to pursue in detail the relationship between replay onset and alpha oscillations, despite having characterised a very strong association between the two. As mentioned in chapter 6, there is a possibility that replay sequences are organised around alpha cycles; any evidence for such a mechanism would significantly advance our understanding of the relationship between the DMN and replay in much more physiological terms. However investigating this is highly nontrivial due to the heterogenous phase of alpha over cortex and methodological constraints to disambiguate a "true" phase alignment from an artefact of linear classification.

7.3 Closing Remarks

Neural systems are never quiescent, but rather characterised by spontaneous dynamic activity that is modulated by sensorimotor inputs. Many of these spontaneous patterns are highly structured, reflecting an organisation that can be studied and characterised. Nonetheless, most experimental analyses ignore this structure and instead impose assumptions of temporal synchrony in the hope that everything will average out over sufficient trials. In so doing, they overlook a valuable and interesting source of variation that can potentially provide greater insight into the underlying neural processes under observation.

We have taken methods for inference of latent structure in unsupervised settings and adapted them for use in supervised analysis. In so doing, we have shown that these can enhance methodological performance, yielding increases in accuracy and sensitivity to variations that may be consistent over trials but poorly temporally aligned. In applying these models to the study of replay in MEG and value encoding in LFP data we have characterised key relationships in the data that were otherwise overlooked. From both a methodological and a scientific standpoint, the show does not end here and there are numerous avenues for further work - but it is my hope that the methods outlined and the insights they have provided may

promote their further use as a meaningful and worthwhile means of understanding temporal variations in neural data.

Appendices

Appendix A

Variational Bayesian Inference

Variational methods involve approximating the true posterior $P(\boldsymbol{\theta}|\mathbf{Y})$ by an approximate posterior $Q(\boldsymbol{\theta})$ in which the parameters are statistically independent; crucially this approximate posterior is analytically tractable via an algorithm guaranteed to converge. It can be considered an extension of expectation maximisation from point estimation to full posterior estimation [267].

A.1 Factorising the posterior

The first step in these methods is to explicitly formulate the approximate posterior, in which the random variables are statistically independent of each other:

$$P(\boldsymbol{\theta}|\mathbf{Y}) \approx Q(\boldsymbol{\theta}) = Q_1(\theta_1)Q_2(\theta_2)...Q_n(\theta_n) \tag{A.1}$$

Where $\boldsymbol{\theta} = [\theta_1, \theta_2...\theta_n]$ is any arbitrary set of parameters and $\mathbf{Y}$ is the observed data.

A.2 Variational Free Energy

We can explicitly measure the degree by which Q differs from the true posterior P by measuring the relative entropy or Kullback-Liebler divergence of Q from P, defined as follows:

$$D_{KL}(Q||P) = \mathbb{E}_Q\Big[log\frac{Q(\boldsymbol{\theta})}{P(\boldsymbol{\theta})}\Big] \tag{A.2}$$

$$\propto \mathbb{E}_Q\Big[log\frac{Q(\boldsymbol{\theta})}{P(Y|\boldsymbol{\theta})P(\boldsymbol{\theta})}\Big] \tag{A.3}$$

$$\propto \mathbb{E}_Q\Big[\sum_{i=1}^{n} D_{KL}(Q_i(\theta_i)||P_i(\theta_i) - log(P(Y|\boldsymbol{\theta}))\Big] \tag{A.4}$$

Minimisation of this quantity, also called the free energy, is equivalent to finding the closest possible approximation $Q(\boldsymbol{\theta})$ to the true posterior $P(\boldsymbol{\theta}|\mathbf{Y})$. We can therefore proceed via a variational process in which we iteratively minimise the free energy over each component $Q_i(\theta_i)$ of the approximate posterior. This is equivalent to determining the distribution Q_i that meets the following:

$$log(Q_i(\theta_i)) \propto \left\langle log(P(\mathbf{Y}|\boldsymbol{\theta})) + log(P(\theta_i)) \right\rangle_{\not{\theta_i}} \tag{A.5}$$

Where $\left\langle \right\rangle_{\not{\theta_i}}$ denotes the expectation taken with respect to all other parameters θ_j for $j \neq i$. For further details refer to Beal [267], MacKay [268],

Appendix B

Variational Inference for Latent State Dynamics

This thesis makes broad use of the Hidden Markov Model framework, with a number of different observation models developed as suitable for each chapter. Inferring the distribution of this latent state variable can be done through the Baum-Welch forward backward algorithm; we summarise the key steps here in a general sense as it applies to any of the observation models used. As state dynamics are inferred independently on each trial given the observation model parameters, in the below notation t is taken to be the timing within any trial, with the forward-backward algorithm being computed for each trial independently.

We define the vector L_t of dimension $[1 \times K]$, whose kth entry is the likelihood of state k being active at a time step t given data $\mathcal{D}$ and observation model $\mathcal{M}$:

$$L_t = P(\mathcal{D}_t|Z_t, \mathcal{M}) \tag{B.1}$$

Utilising the forward-backward inference equations, we define the following recursive message functions:

$$m_f(Z_t) = P(Z_t, \mathcal{D}_{1:t}|\mathcal{M}) \tag{B.2}$$

$$m_b(Z_t) = P(Z_t, \mathcal{D}_{t+1:T}|\mathcal{M}) \tag{B.3}$$

which can then be computed as:

$$m_f(Z_t) = (m_f(Z_{t-1})\hat{\Theta}) \cdot L_t \qquad \forall t \in [2, 3, 4..., T] \tag{B.4}$$

$$m_f(Z_1) = \gamma_0 \cdot L_1 \tag{B.5}$$

$$m_b(Z_t) = \frac{\Big(m_b(Z_{t+1}) \cdot L_{t+1}\Big)\hat{\Theta}^T}{\sum m_f(Z_t)} \qquad \forall t \in [1, 2, 3, ..., T-1] \tag{B.6}$$

$$m_b(Z_T) = \frac{u_K}{\sum m_f(Z_t)} \tag{B.7}$$

where γ_0 denotes the prior belief in the starting state (for our sequential structure this specifies that the first state is active at $t = 1$ on every trial with probability 1); $\hat{\Theta}$ denotes the expected value of the transition matrix; $\sum m_f(Z_t)$ denotes element wise summation of the vector $m_f(Z_t)$ and u_K is a vector of length K whose values are all equal to one.

Having computed the forward and backward messages, we can compute the state probabilities as:

$$P(Z_t) = \gamma_t = \frac{m_f(Z_t) \cdot m_b(Z_t)}{\sum(m_f(Z_t) \cdot m_b(Z_t))} \tag{B.8}$$

where the summation in the denominator is over vector elements.

Appendix C

Variational Inference for a Dynamic Encoding Model

The model of section 3.2.2 can be inferred upon through Variational Bayesian methods. This involves approximating the model posterior by a factorised distribution as follows:

$$P(\mathbf{Z},\alpha^{(1)},\Lambda^{(1)},\alpha^{(2)},\Lambda^{(2)},...\alpha^{(K)},\Lambda^{(K)}|\mathbf{X},\mathbf{Y}) \approx \prod_{n=1}^{N}\prod_{t=1}^{T} Q(Z_{n,t}) \prod_{k=1}^{K} Q(\alpha^{(k)})Q(\Lambda^{(k)}) \tag{C.1}$$

This approximation allows us to compute analytic forms for each component function of the posterior, when a suitable conjugate prior is used. Utilising the priors specified in equation 3.8 results in a marginal distribution over the regression coefficients that is normally distributed, independently over each column (we use the notation below of $\alpha_{i,:}^{(k)}$ to denote the ith column of the kth state coefficient matrix). The noise precision in each state is distributed by a diagonal matrix whose elements are independently gamma distributed (we use below the notation of $\Lambda_{i,i}^{(k)}$ to denote the ith diagonal element of the kth state precision matrix). More explicitly, we have:

$$Q(\Lambda_{i,i}^{(k)}) \sim Gamma(a_i^{(k)}, b_i^{(k)}) \tag{C.2}$$

$$Q(\alpha_{i,:}^{(k)}) \sim Normal(\mu_{\alpha_i}^{(k)}, \Sigma_{\alpha_i}^{(k)}) \tag{C.3}$$

with update equations as follows:

$$a_i^{(k)} = a_0 + \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{n,t}^{(k)} \tag{C.4}$$

$$b_i^{(k)} = \left[b_0 + \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{n,t}^{(k)}(X_{n,t}(i) - Y_{n,t}\mu_{\alpha_i}^{(k)})^2 + Y_{n,t}Y_{n,t}^T\Sigma_{\alpha_i}^{(k)}\right]^{-1} \tag{C.5}$$

$$\Sigma_{\alpha_i}^{(k)} = \frac{1}{\mathbb{E}(\rho_i^{(k)})}\mathbf{I} + \sum_{n=1}^{N}\sum_{t=1}^{T}Y_{n,t}^T\gamma_{n,t}^{(k)}a_i^{(k)}b_i^{(k)}Y_{n,t} \tag{C.6}$$

$$\mu_{\alpha_i}^{(k)} = \Sigma_{\alpha_i}^{(k)}\left[\sum_{n=1}^{N}\sum_{t=1}^{T}Y_{n,t}\gamma_{n,t}^{(k)}a_i^{(k)}b_i^{(k)}X_{n,t}(i)\right] \tag{C.7}$$

where we have used shorthand $\rho_i^{(k)} = [\rho_{i,1}^{(k)}, \rho_{i,2}^{(k)}, ..., \rho_{i,P}^{(k)}]$; and $X_{n,t}(i)$ is taken to denote the ith element of the vector $X_{n,t}$. Finally our prior over the hyperparameter ρ results in its marginal distribution being gamma distributed, with update equations as follows:

$$Q(\rho_{i,j}^{(k)}) \sim Gamma(\psi_{i,j}^{(k)}, \chi_{i,j}^{(k)}) \tag{C.8}$$

$$\psi_{i,j}^{(k)} = a_0 + \frac{\mu_{\alpha_i}^2 a_i b_i^{(k)}}{2} \tag{C.9}$$

$$\chi_{i,j}^{(k)} = b_0 + \frac{1}{2} \tag{C.10}$$

Appendix D

Variational Inference for Dynamic Categorical Decoding Model

The model of 3.2.3 cannot be directly inferred using variational methods, as the non-linear logit link function introduces a nonlinearity that has no conjugate prior. Nonetheless we can utilise the following convex approximation to the logit link function:

$$\sigma(z) = \frac{1}{1+e^{-z}} \geq h(z) = \sigma(\xi) exp\left(\frac{z-\xi}{2} - \lambda(\xi)(z^2 - \xi^2)\right) \tag{D.1}$$

where $\lambda(\xi) = \frac{1}{2\xi}(\sigma(\xi) - \frac{1}{2})$. This introduces tight lower bound on the logit function, which can be optimised with respect to the additional parameter ξ.

Using this second order approximation to the likelihood, we can now use conjugate priors in the forms specified in equation 3.18, which allow the following factorised posterior:

$$P(\mathbf{Z},\beta^{(1)}, \beta^{(2)}, ...\beta^{(K)}|\mathbf{X}, \mathbf{Y}) \approx Q(\mathbf{Z}) \prod_{k=1}^{K} Q(\beta^{(k)}) \tag{D.2}$$

This approximate posterior is analytically tractable. Furthermore given the conditional independence of columns of each state's regression coefficient matrix, a further factorisation naturally follows to give a marginal posterior over each ith column of $\beta^{(k)}$ as follows:

$$Q(\beta_i^{(k)}) \sim N(\mu_{\beta_i}^{(k)}, \Sigma_{\beta_i}^{(k)}) \tag{D.3}$$

with update equations as follows:

$$\Sigma_{\beta_i}^{(k)} = \left(\mathbb{E}[\eta_i^{(k)}]\mathbf{I} + 2 \sum_{n=1}^{N} \sum_{t=1}^{T} \gamma_{n,t}^{(k)} \lambda(\xi_{n,t}) X_{n,t}^T X_{n,t} \right)^{-1} \tag{D.4}$$

$$\mu_{\beta_i}^{(k)} = \frac{1}{2} \Sigma_{\beta_i}^{(k)} \sum_{n=1}^{N} \sum_{t=1}^{T} \gamma_{n,t}^{(k)} Y_{n,t} X_{n,t} \tag{D.5}$$

where we have used shorthand $\eta_i^{(k)} = [\eta_{i,1}^{(k)}, \eta_{i,2}^{(k)}, ..., \eta_{i,Q}^{(k)}]$. Finally our prior over the hyperparameter η results in its marginal distribution being gamma distributed, with update equations as follows:

$$Q(\eta_{i,j}^{(k)}) \sim Gamma(\upsilon_{i,j}^{(k)}, \phi_{i,j}^{(k)}) \tag{D.6}$$

$$\upsilon_{i,j}^{(k)} = a_0 + \frac{(\mu_{\beta_{i,j}}^{(k)})^2 + \Sigma_{\beta_{i,j}}^{(k)}}{2} \tag{D.7}$$

$$\phi_{i,j}^{(k)} = b_0 + \frac{1}{2} \tag{D.8}$$

Furthermore, as we have used a parameterised lower bound approximation to the logistic link function, we have to tune our parameter $\xi_{n,t}$ at each timestep to maximally tighten the lower bound of this approximation, so at each iteration we similarly update as follows:

$$\xi_{n,t} = \left(\sum_{k=1}^{K} \gamma_{n,t}^{(k)} X_{n,t}^T (\mathbb{E}(\beta^{(k)})) X_t \right)^{\frac{1}{2}} \tag{D.9}$$

Appendix E

Rebalancing Adjustment

Where classes are unbalanced, logistic regression can be adjusted to re-weight samples such that an unbiased parameters estimate is inferred [124–126]. This involves a slight modification to the log likelihood term as follows:

$$log(\mathbb{L}) = \sum_{i=1}^{Q}\sum_{n=1}^{N}\sum_{t=1}^{T} w_{i,Z_{n,t}} log\Big[\sigma(Y_{n,t}^{(i)} X_{n,t}\beta^{(Z_{n,t})})\Big] \tag{E.1}$$

where we have introduced a weighted penalty term $w_{i,k}$ for each state and each dimension of the stimulus, and are using the notation $Y_{n,t}^{(i)} = \mathbb{I}(Y_{n,t} = i)$, the identity function taking values $[1, 0]$. As outlined elsewhere, this penalty term compensates for undersampling or over sampling of a particular class of data. In our case, we are concerned with undersampling or oversampling of a particular class of data by each state - so we impose an assumption of evenly sampled classes in each state, resulting in a weight computation for each class as follows:

$$w_{i,k} = \frac{\sum_{N=1}^{N}\sum_{t=1}^{T}\frac{1}{K}Y_{n,t}^{(i)}}{\sum_{N=1}^{N}\sum_{t=1}^{T}\gamma_{n,t}^{(k)}Y_{n,t}^{(i)}} \tag{E.2}$$

Carrying this through the our update equations following the same variational inference steps as above, we have adjusted update equations as follows:

$$Q(\beta_i^{(k)}) \sim N(\mu_{\beta_i}^{(k)}, \Sigma_{\beta_i}^{(k)}) \tag{E.3}$$

$$\Sigma_{\beta_i}^{(k)} = \left(\mathbb{E}[\eta_i^{(k)}]\mathbf{I} + 2\sum_{n=1}^{N}\sum_{t=1}^{T} w_{i,k}\gamma_{n,t}^{(k)}\lambda(\xi_{n,t})X_{n,t}^T X_{n,t}\right)^{-1} \tag{E.4}$$

$$\mu_{\beta_i}^{(k)} = \frac{1}{2}\Sigma_{\beta_i}^{(k)}\sum_{n=1}^{N}\sum_{t=1}^{T} w_{i,k}\gamma_{n,t}^{(k)}Y_{n,t}X_{n,t} \tag{E.5}$$

Appendix F

LFP Analysis Supplementary Information

The analysis conducted in chapter 2 is here expanded with additional data to support claims made in the main text.

F.1 Average LFP response over cues

Figure F.1 plots the average response per cortical subregion over successive cues.

The most prominent features immediately apparent in the data are not necessarily of a neural origin, but rather a sign of contamination by muscle artefacts. Immediately prior to each cue being revealed (with the time marked as zero corresponding to stimulus onset), the eye saccade made by the monkey to the cue location contributes a broadband increase in power across all channels, showing up on the average spectrogram as a vertical band approximately 50msec prior to the cue. Note that this highlights the eye saccades that are locked to the cue presentation; spontaneous eye saccades that are not aligned to the trial epochs are likely also present and will contaminate the reading from individual trials. Removal of such artefacts is a non-trivial problem [269, 270]; despite a variety of proposed methods for doing so we do not pursue this further other than to note their presence in the data.

Between the different recording sites, certain differences in signal spectral content are apparent. While all sites show a stimulus induced decrease in low frequency power, this drop is most apparent for DLPFC sites which exhibit a sharp decrease in power below 20Hz following cue presentation.

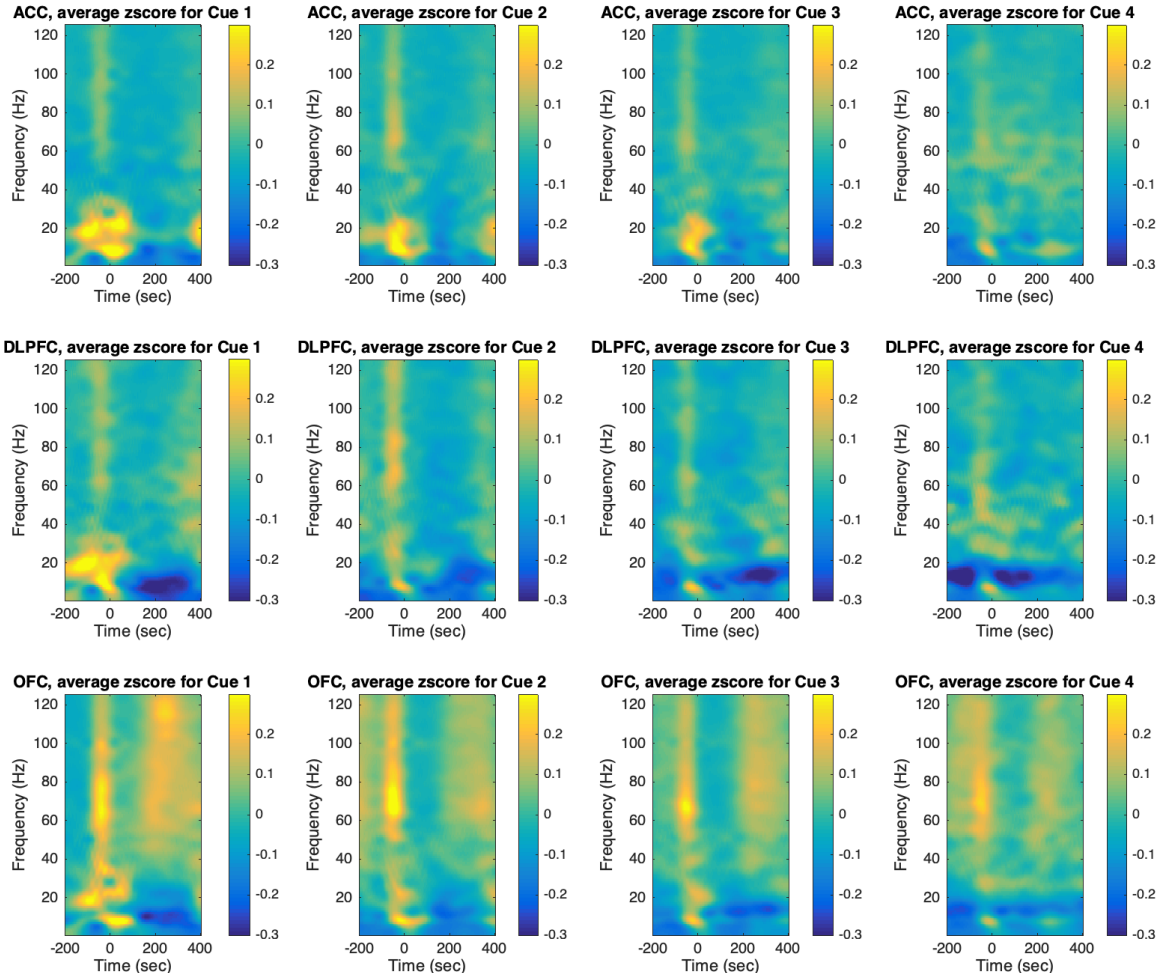


Figure F.1: Mean LFP power per frequency band aligned to the presentation of each cue. Common patterns emerge over successive cues; the most prominent LFP feature is actually a saccade artefact, which presents as a wideband increase in power immediately prior to cue presentation. Remaining features that are likely to be neural in origin include a beta band power decrease post stimulus in DLPFC and to a lesser extent OFC; and increases in high gamma approx 300msec post cue onset in OFC.

An increase in power in gamma and high gamma range (>50Hz) is apparent over multiple recording sites; the timing of this response which peaks approximately 300msec post stimulus likely reflects local neuronal discharge induced by the cue presentation, which has also been shown to peak approximately 300msec following the stimulus [30].

F.1.1 Characterising temporal differences in responses to subsequent cues

In addition to changes in the spectral power distribution between different recording sites, some differences can also be characterised over successive cue presentation. The working memory load (number of items the monkey must retain in memory) increases over successive cues, as well as a requirement to integrate or contrast values of presented cues. Differences are not overt over successive cues, however the pattern of stimulus induced low power decreases becomes less prominent over successive cues; high gamma activity decreases in OFC, while slightly increasing in ACC; and an increase in theta band power (4-8 Hz) over successive cues is apparent across all recording sites but especially in OFC and ACC.

F.2 Channel-wise univariate analysis

Figures F.2 and F.3 show that the time-frequency analysis results did not merely arise as a result of averaging over channels, and that the same analysis performed channel-by-channel similarly fails to detect significant correlates of cue parameters on all but one of the channels. This stands in stark contrast to the distributed patterns of information inferred by the multivariate model that strongly correlated with cue variables.

Figure F.4 complements the principal component analysis used to explore the state timecourses to show that the trajectories identified and discussed at length are a good characterisation of the data beyond the first two principal components. Later PCs are largely focussed on changes in the timing of one or other of these trajectories (or outlying noise trials which are not analysed further).

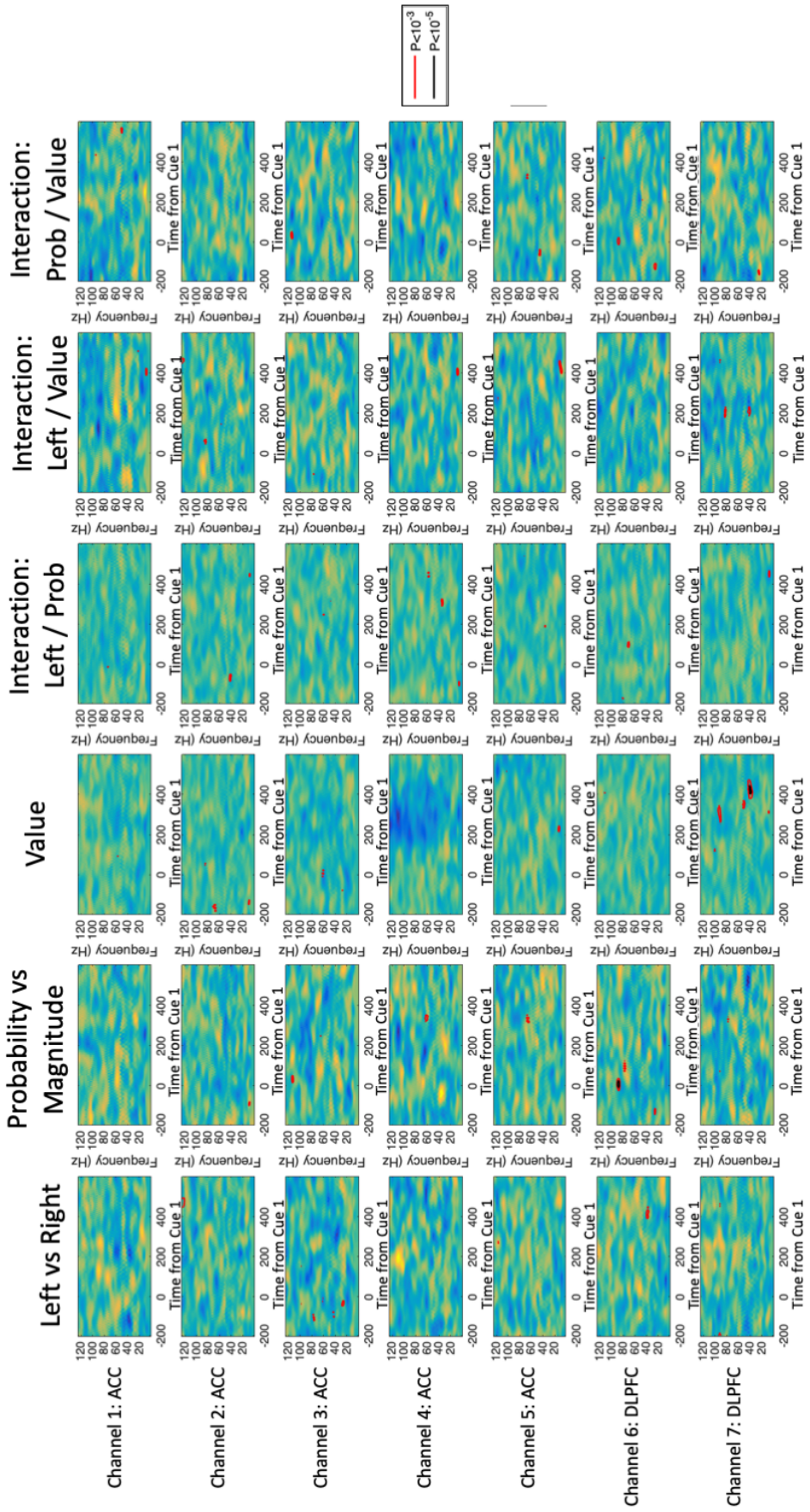


Figure F.2: Channel-by-channel time frequency analysis: the same analysis conducted on spectral power per subregion can be computed over each individual channel, confirming the absence of any robust relationship between cue features and single channel power. The only exception to this is channel 13, which shows significant encoding of cue value in high gamma power.

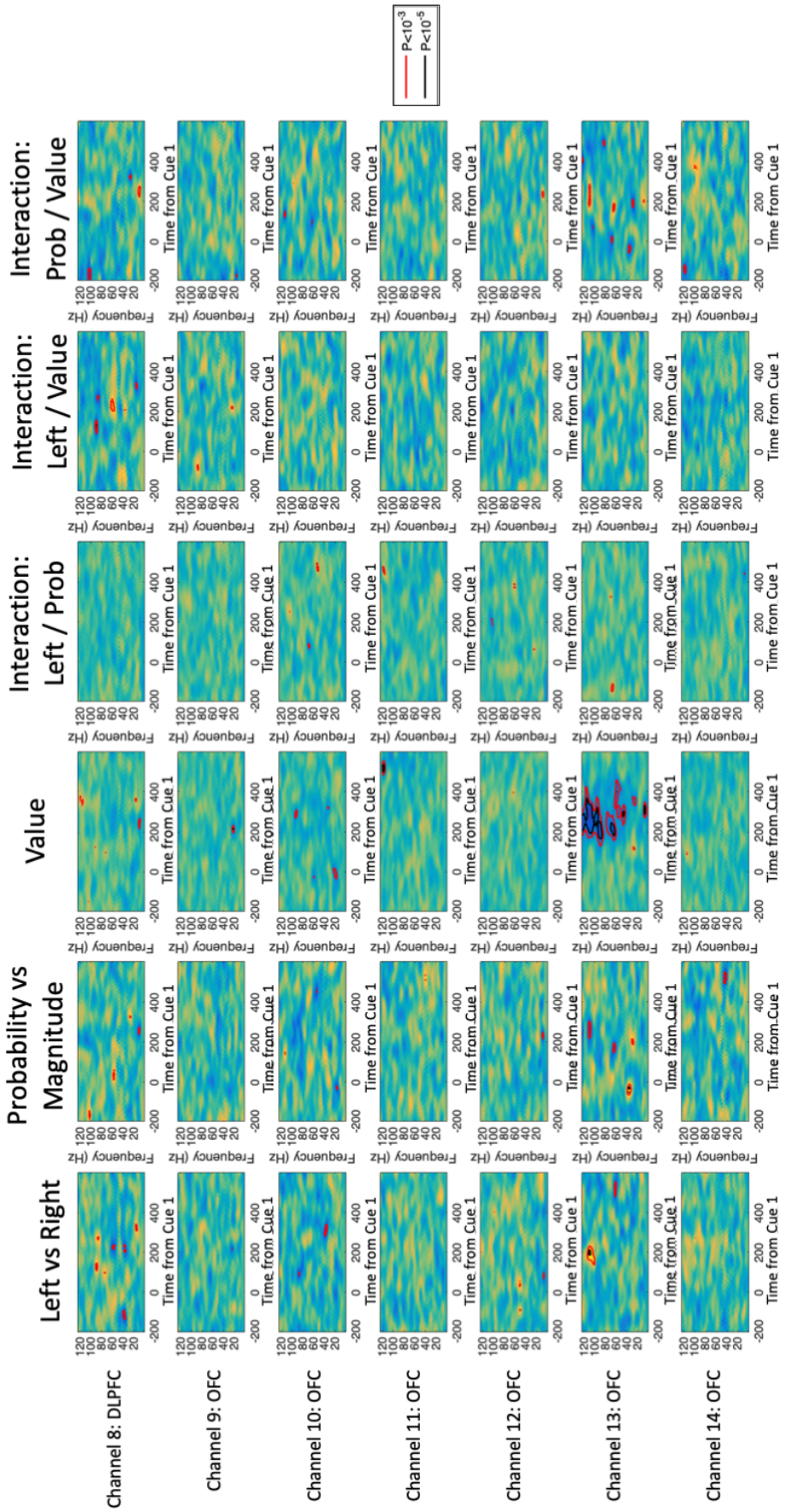


Figure F.3: Channel-by-channel time frequency analysis: the same analysis conducted on spectral power per subregion can be computed over each individual channel, confirming the absence of any robust relationship between cue features and single channel power. The only exception to this is channel 13, which shows significant encoding of cue value in high gamma power.

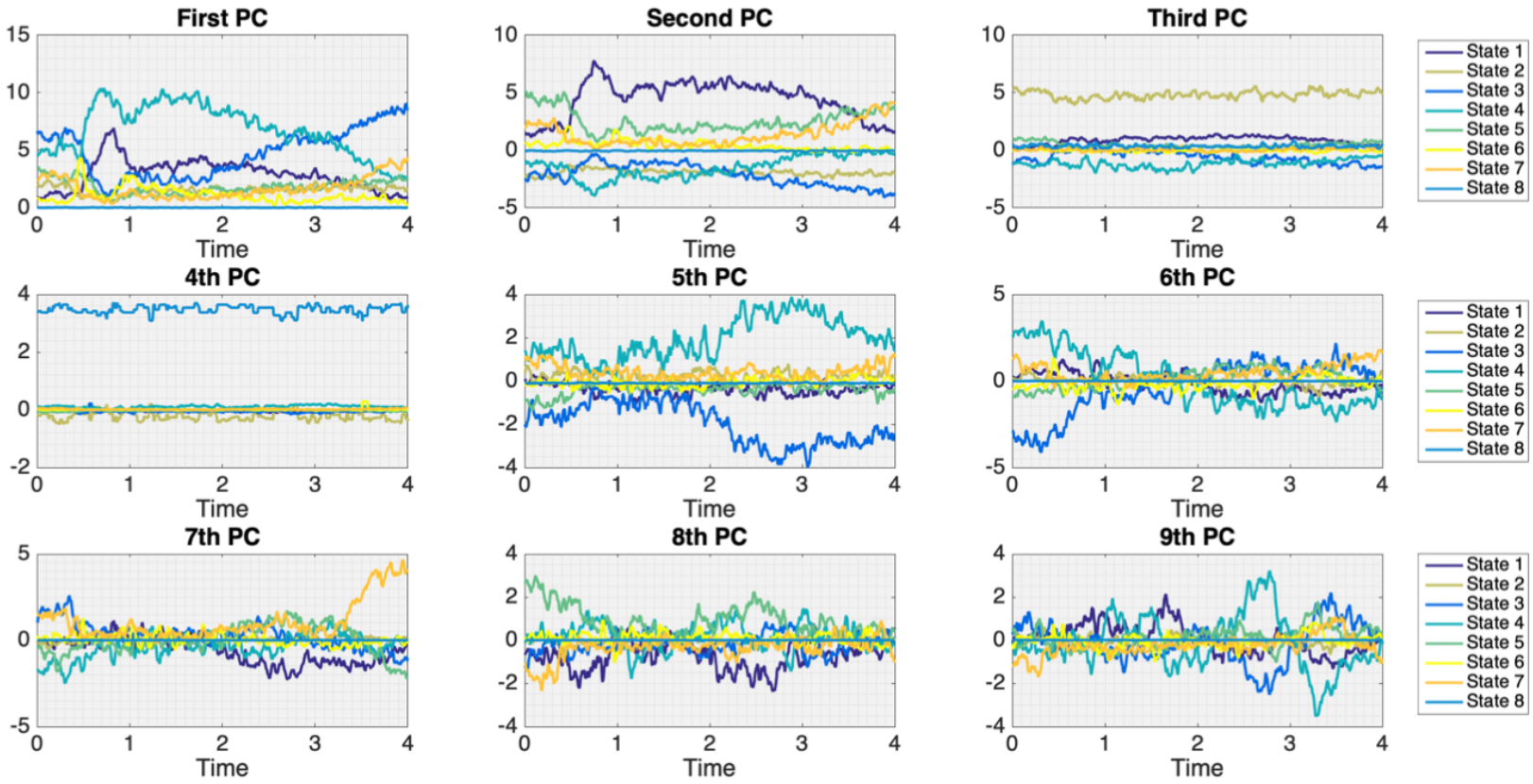


Figure F.4: State timecourse Principal Component Analysis: the first nine principal components explaining the variation in inferred state timecourses. There are two meta trajectories described by the first and second principal components; PCs 3 and 4 then describe noise artefacts. The remaining PCs describe variations in the timing of state transitions within the two meta-trajectories, for example in the timing of states 3 and 4 (explained by PCs 5 and 6) or states 1, 5 and 7 (explained by PCs 7 and 8).

Appendix G

Sequential Visual Stream Activation Supplementary Information

G.1 Performance of other methods

Tests conducted against alternative classification methods and summarised in figure G.1 reveal a comparable, if not equivalent, level of decoding accuracy achieved by each of them; our claims as to the advantages of the HMM are supported as it achieves the highest decoding accuracy of all methods. We plot here a number of different classifiers tested; note that the logistic regression GLM with L1 regularisation penalty of $\lambda = 0.01$ referred to in the text is the best performing L1 classifier, but is in fact surpassed in classification accuracy by logistic regression with $L2$ regularisation for a range of values of the regularisation value. Linear support vector machines were also compared and did not perform favourably on this data.

G.2 Frequency Domain Analysis

The frequency domain features plotted in figure 4.4 were computed using estimated state timecourses specific to each trial. Figure G.2 plots the same results computed using the mean state timecourses fit to all data; the plots are almost identical for frequency content below 10Hz, however for frequencies above this level the mean fitted state timecourses are only accurate to chance level. When using trial specific state timecourse estimates, models with high number of states were able to decode

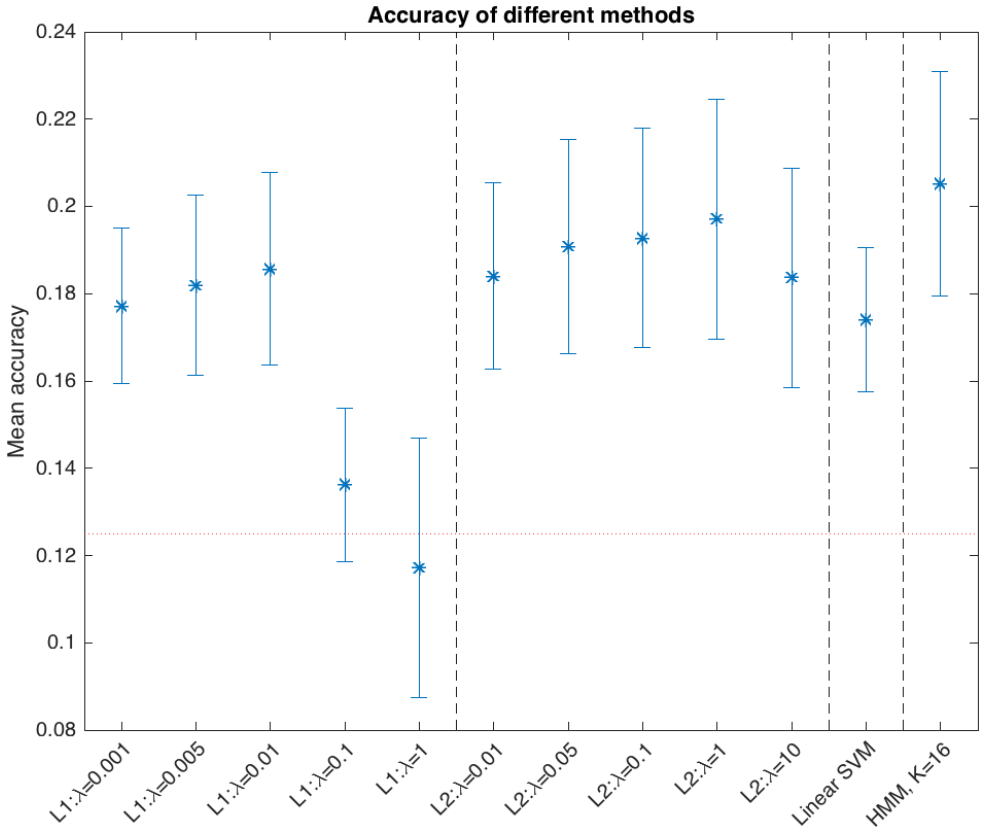


Figure G.1: Decoding accuracy of alternative classifiers: comparing a range of classifiers, our dynamic decoder achieves the greatest performance. The L1 logistic regression classifier is quite sensitive to the regularisation penalty (note the value selected for the comparison in the text was the best performing of these); its performance is surpassed slightly by L2 logistic regression classifiers for a range of values of the $L2$ penalty; linear support vector machines perform slightly worse than these; the dynamic decoder performs slightly better. Values plotted are mean +/- standard deviation over subjects of the total accuracy in the window from 0-500msec post stimulus presentation.

some information from frequencies greater than 10Hz. This suggests that decoding in this frequency range is entirely dependent on trial specific state timing.

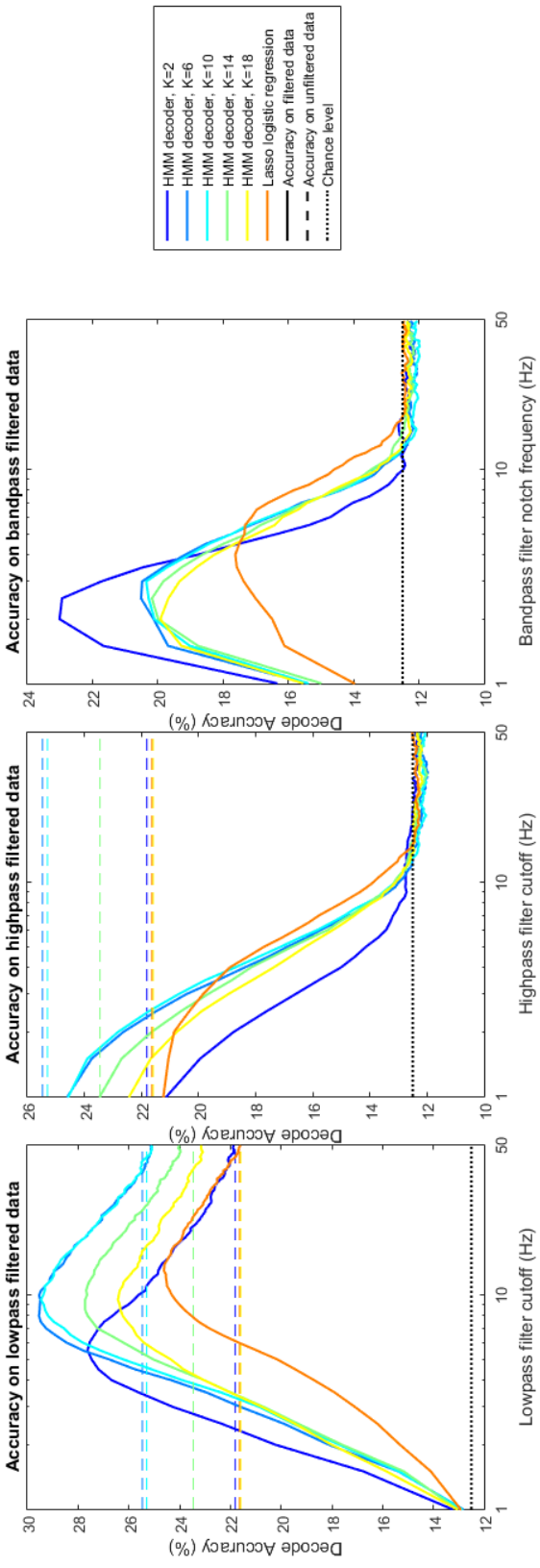


Figure G.2: Decoding accuracy as function of frequency content, using mean state timecourses identical over all test trials. Left panel: mean accuracy of a range of different models when test data is lowpass filtered. Middle panel: mean accuracy when test data is highpass filtered. Right panel: mean accuracy on bandpass filtered test data.

Appendix H

Replay Dynamics: Supplementary Information

In chapter 5 we identified that three of six HMM classifier states were associated with significant sequential reactivation patterns; we then identified that those activated in top-down sweeps, showing significant negative sequenceness patterns. Additionally, we tested the temporal pattern of correlation between each other pair of states, plotted in figure H.1; this shows that the top-down directionality does not hold as a general feature of these classifiers' outputs, but is specific to those states in which significant replay of sequences has been observed.

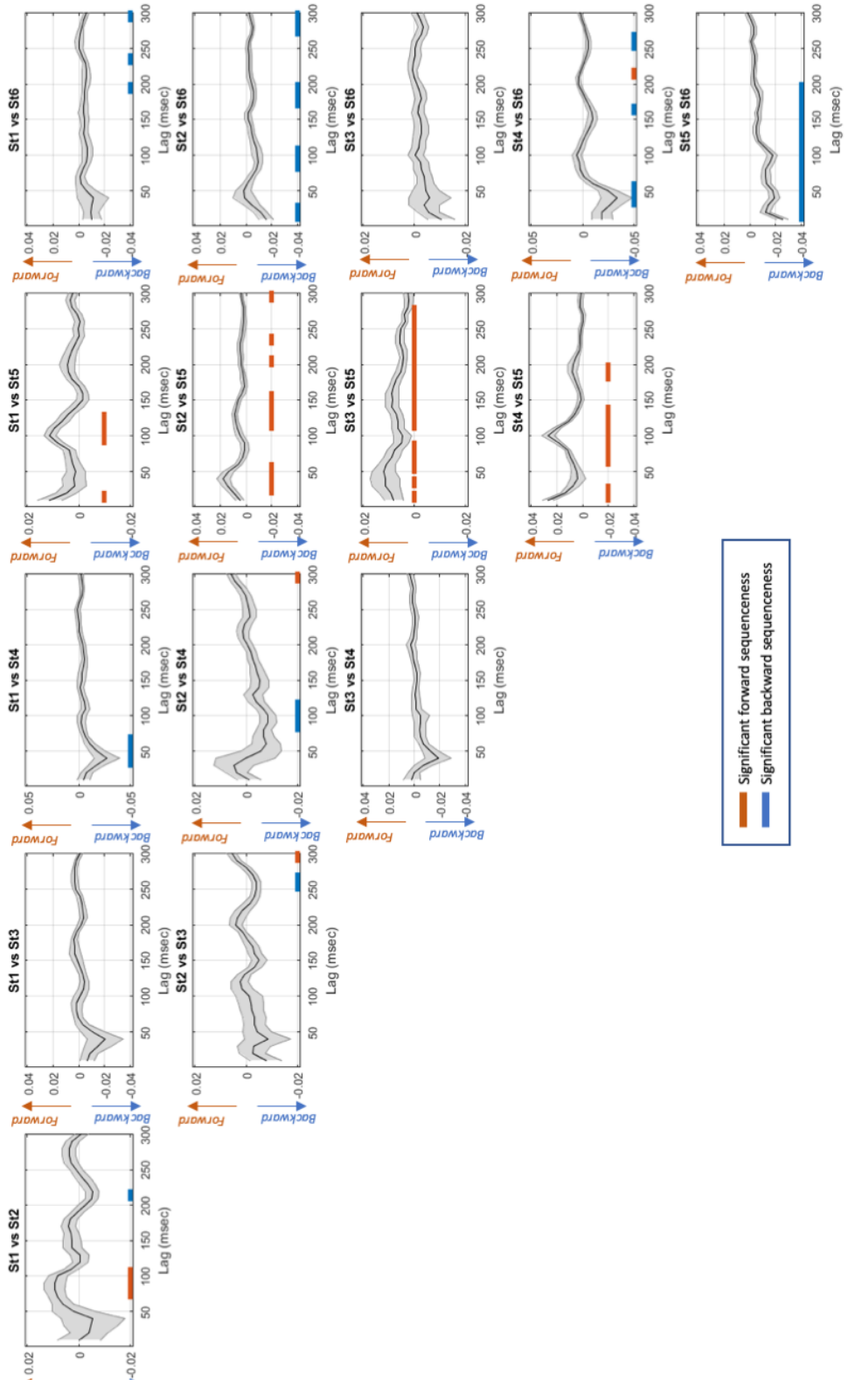


Figure H.1: Sequenceness results for each pair of states; significance lines denote timepoints where result is significantly greater than zero (in red) or less than zero (in blue) as computed by non-parametric signflipping tests with P=0.05 threshold. Recall that states 1, 4 and 6 were associated with significant replay of task relevant sequences; the significant top-down result does not extend to the other states as a general rule.

Appendix I

Replay and RSN Correlation: Supplementary Information

Figures I.1 to I.6 plot the power and coherence distributions that define the remaining states 1-6 inferred in chapter 6.

We have shown the strong correlation between RSNs that comprise the default mode network and replay, as measured using the HMM classifier that corresponded to data at approximately 200msec post stimulus. We can also show in figure I.7 that this result generalises if replay is measured using any of the other HMM classifier states that identified significant replay, i.e. decoding states 1, 4 or 6 from the model fit in chapter 5. No strong differences in the evoked state distributions are apparent.

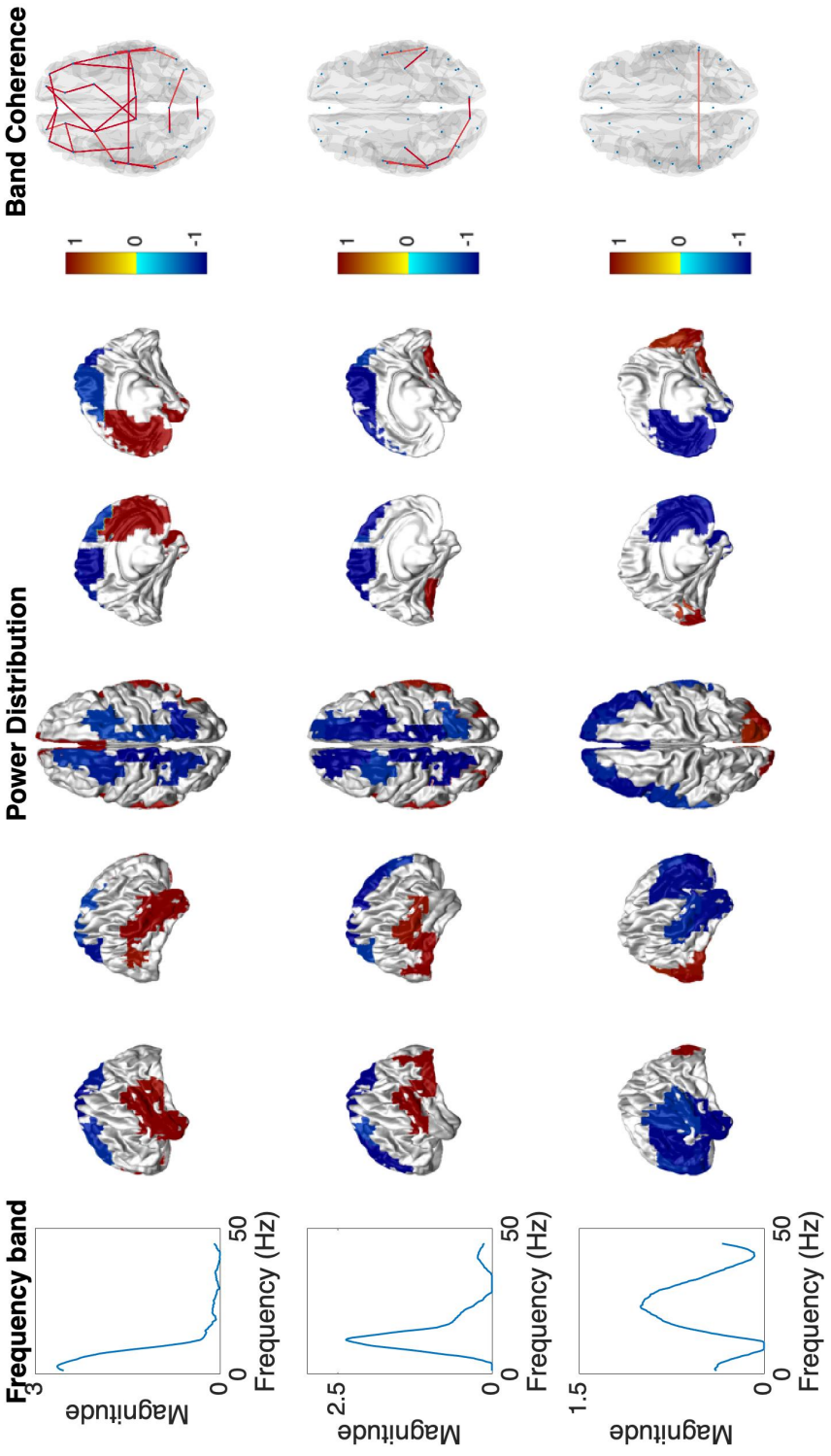


Figure I.1: Power and coherence profile of state 1.

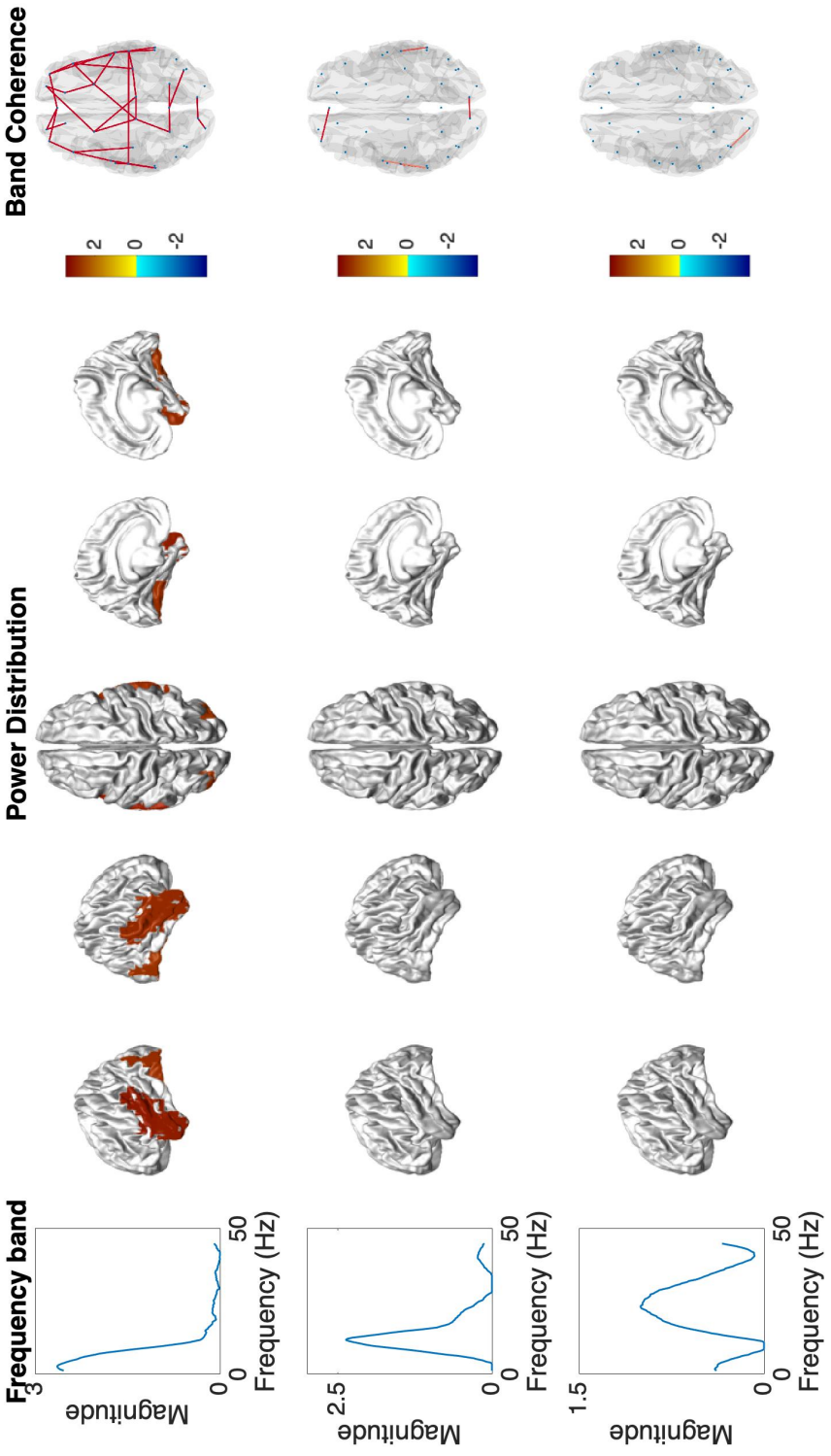


Figure I.2: Power and coherence profile of state 2.

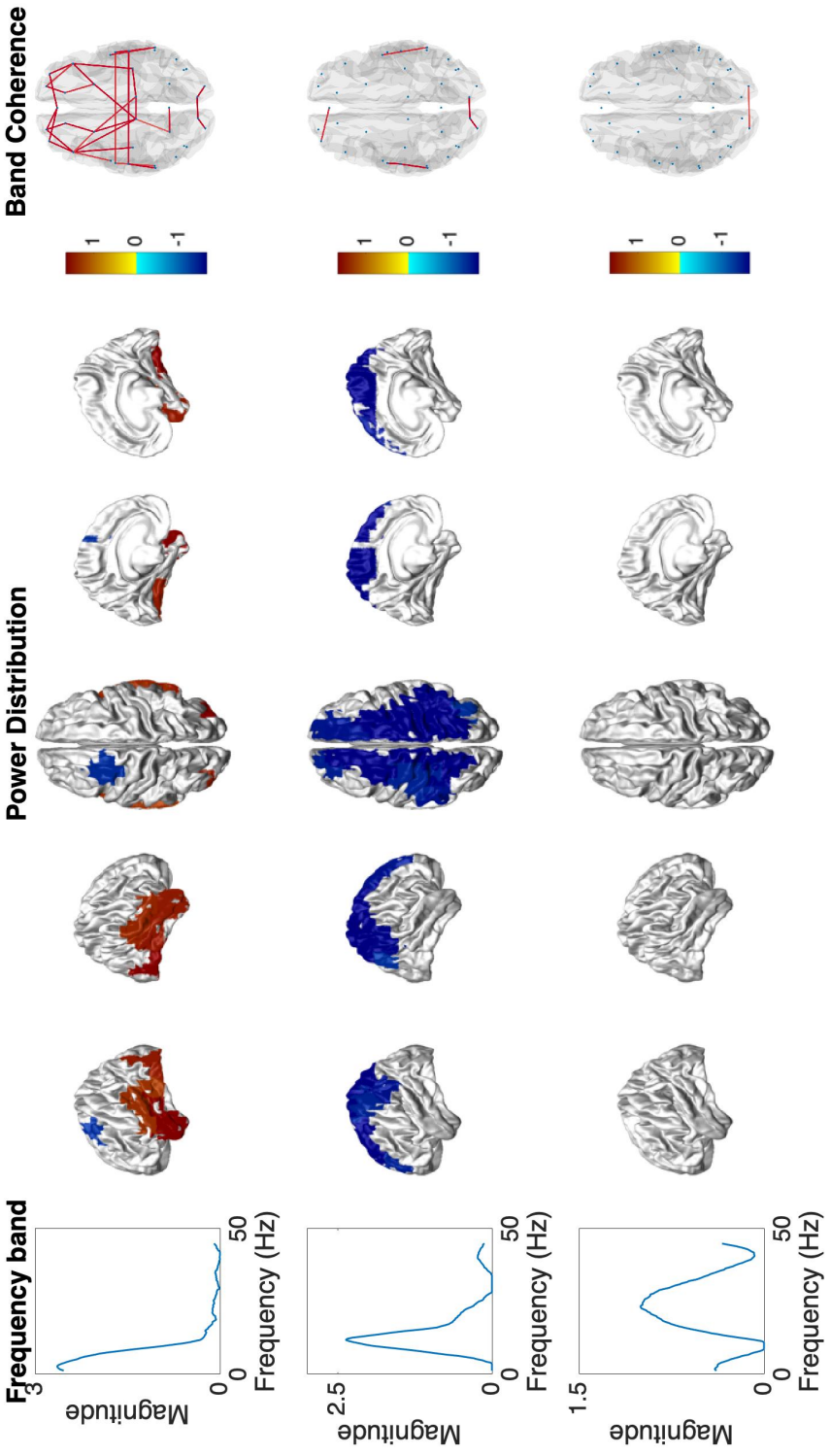


Figure I.3: Power and coherence profile of state 3.

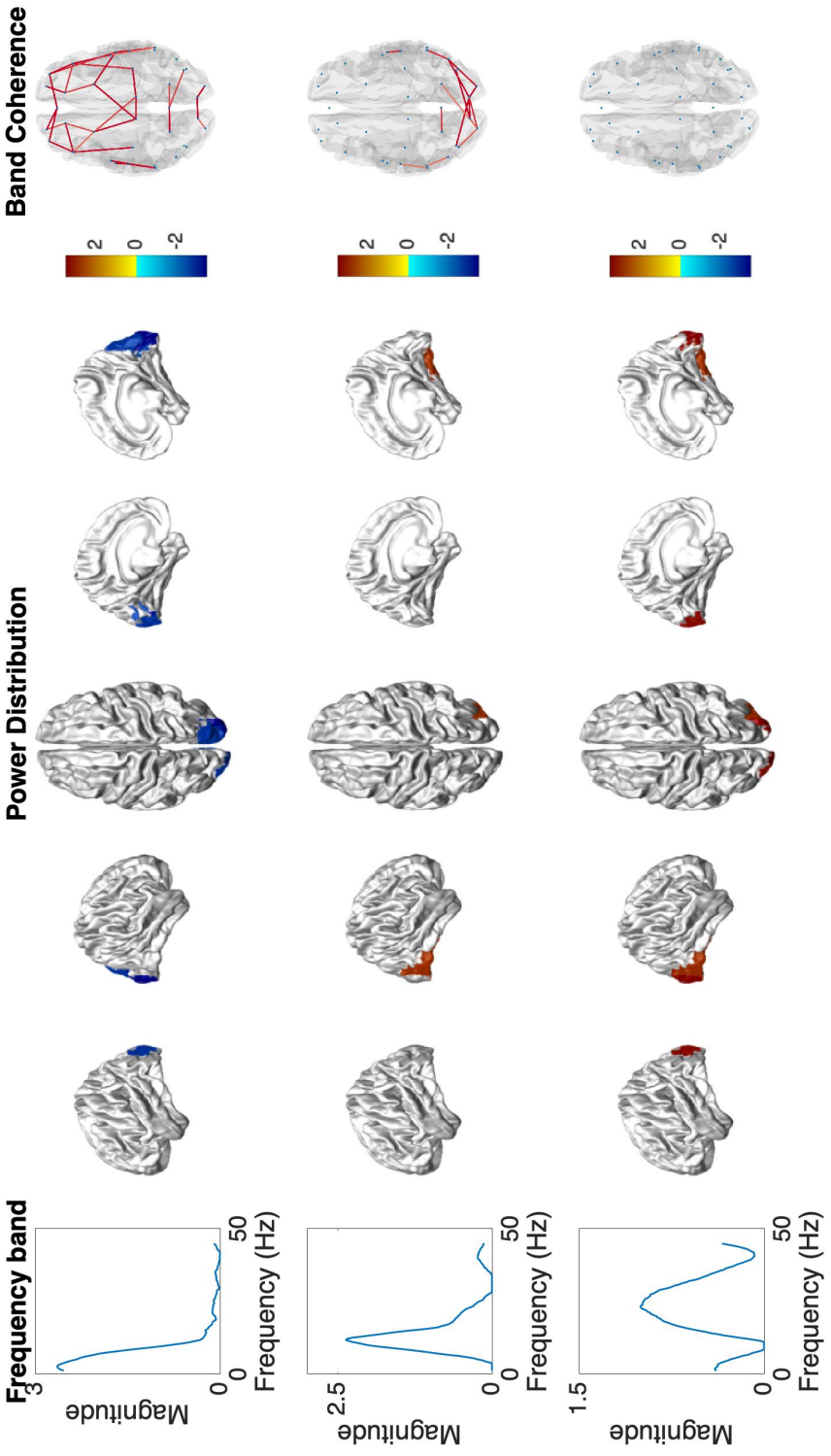


Figure I.4: Power and coherence profile of state 4.

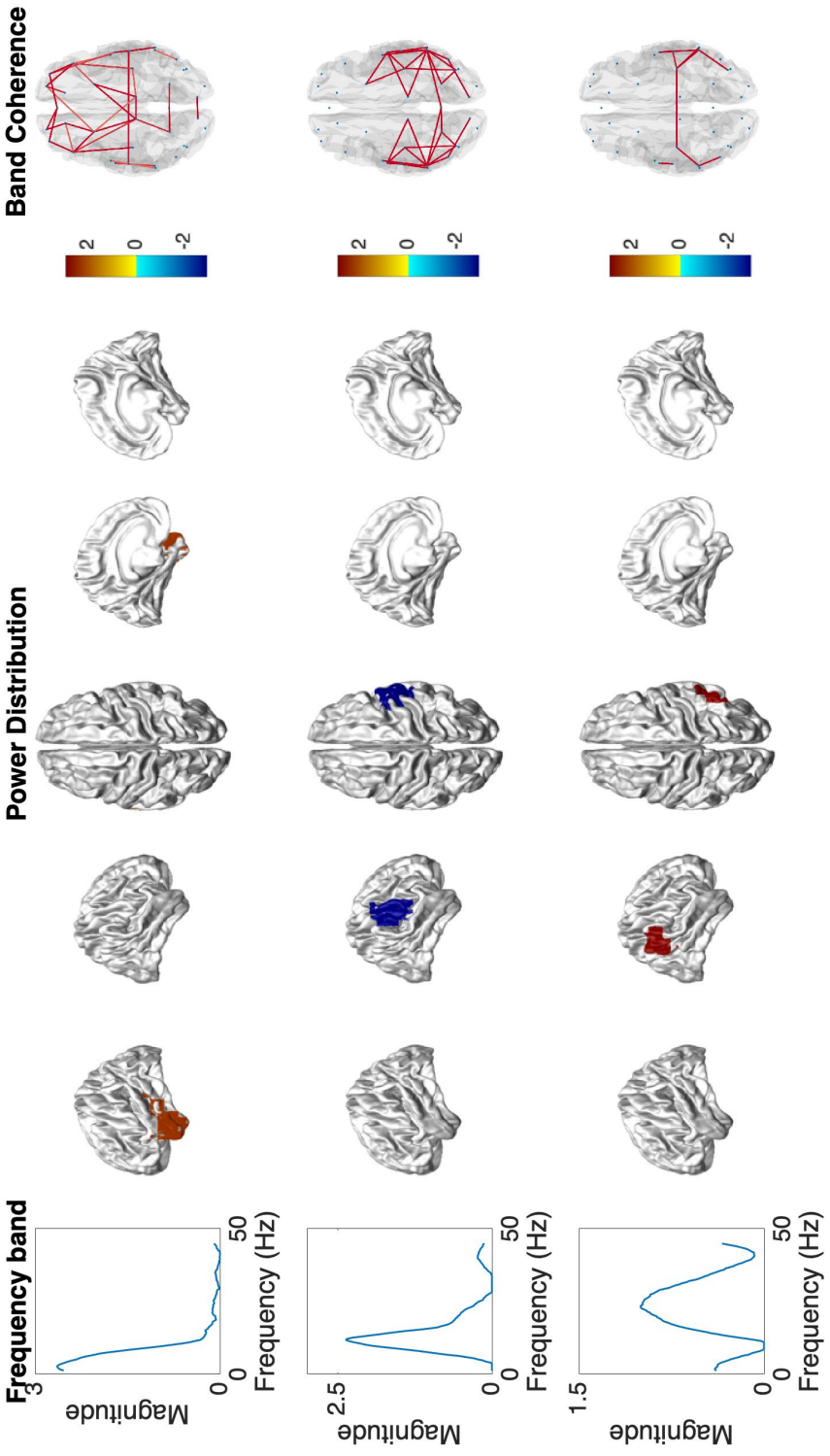


Figure I.5: Power and coherence profile of state 5.

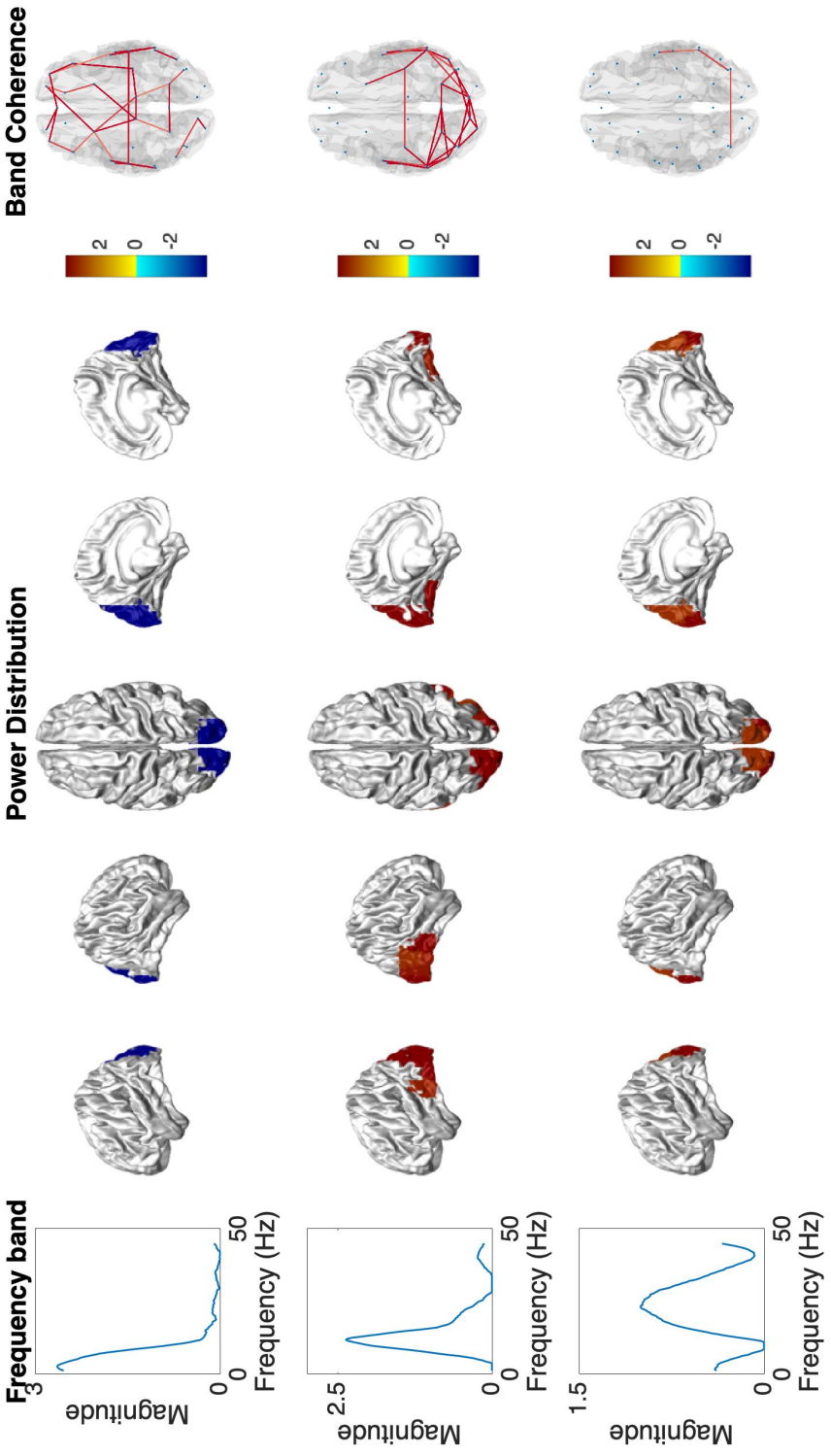


Figure I.6: Power and coherence profile of state 6.

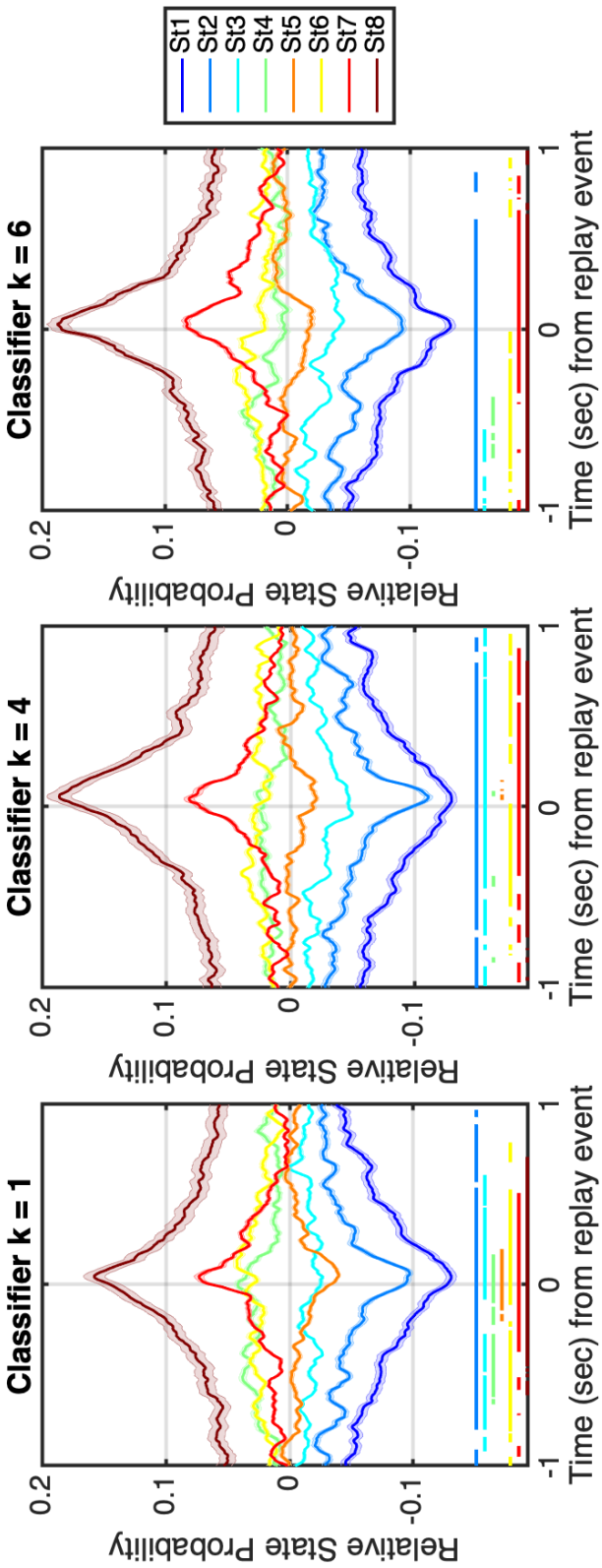


Figure I.7: Replay Evoked Response for each HMM classifier state: the results shown in chapter 6 pertained only to replay identified using HMM classifier state $k = 4$ For each of the states that were associated with significant replay detection, we here plot the same replay evoked response; this shows that it generalises regardless of HMM classifier state, supporting the model of top-down sweeps propagating rapidly through the visual hierarchy during activation of the default mode network and associated regions.

Bibliography

[1] Rajesh P N Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. ISSN 1546-1726. doi: 10.1038/4580. URL https://doi.org/10.1038/4580.

[2] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*, 11(2):127–138, 2010. ISSN 1471-003X. doi: 10.1038/nrn2787. URL http://www.nature.com/doifinder/10.1038/nrn2787.

[3] David C. Knill and Alexandre Pouget. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27 (12):712–719, 2004. ISSN 01662236. doi: 10.1016/j.tins.2004.10.007.

[4] Andreas K Engel, Pascal Fries, and Wolf Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews. Neuroscience*, 2(10):704–716, 2001. ISSN 1471-003X. doi: 10.1038/35094565. URL http://www.ncbi.nlm.nih.gov/pubmed/11584308.

[5] Wolfgang Klimesch. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606–617, 2012. ISSN 13646613. doi: 10.1016/j.tics.2012.10.007. URL http://dx.doi.org/10.1016/j.tics.2012.10.007.

[6] Ole Jensen, Jochen Kaiser, and Jean Philippe Lachaux. Human gamma-frequency oscillations associated with attention and memory, jul 2007. ISSN 01662236. URL https://www.sciencedirect.com/science/article/pii/S0166223607001051.

[7] Simon Hanslmayr, Joachim Gross, Wolfgang Klimesch, and Kimron L. Shapiro. The role of alpha oscillations in temporal attention. *Brain Research Reviews*, 67(1-2):331–343, 2011. ISSN 01650173. doi: 10.1016/j.brainresrev.2011.04.002. URL http://dx.doi.org/10.1016/j.brainresrev.2011.04.002.

[8] Bjørg Elisabeth Kilavik, Adrián Ponce-Alvarez, Romain Trachel, Joachim Confais, Sylvain Takerkart, and Alexa Riehle. Context-Related Frequency Modulations of Macaque Motor Cortical LFP Beta Oscillations. *Cerebral Cortex*, 22(9):2148–2159, oct 2011. ISSN 1047-3211. doi: 10.1093/cercor/bhr299. URL https://doi.org/10.1093/cercor/bhr299.

[9] Jeremy R Manning, Sean M Polyn, Gordon H Baltuch, Brian Litt, and Michael J Kahana. Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences*, 108(31):12893 LP – 12897, aug 2011. doi: 10.1073/pnas.1015174108. URL http://www.pnas.org/content/108/31/12893.abstract.

[10] Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews. Neuroscience*, 8(9):700–11, 2007. ISSN 1471-003X. doi: 10.1038/nrn2201. URL http://www.ncbi.nlm.nih.gov/pubmed/17704812.

[11] Gustavo Deco, Viktor K. Jirsa, and Anthony R. McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12(1):43–56, 2011. ISSN 1471003X. doi: 10.1038/nrn2961. URL http://dx.doi.org/10.1038/nrn2961.

[12] M Steriade, I Timofeev, N Dürmüller, and F Grenier. Dynamic properties of corticothalamic neurons and local cortical interneurons generating fast rhythmic (30-40 Hz) spike bursts. *Journal of neurophysiology*, 79(1):483–90, 1998. ISSN 0022-3077. doi: 10.1152/jn.1998.79.1.483. URL http://www.ncbi.nlm.nih.gov/pubmed/9425218.

[13] Amp Henrik Jahnsen and Rodolfo R Llinás. Electrophysiology of mammalian thalamic neurons in vitro. *Nature*, 279(June):406–408, 1982.

[14] Kathryn Kidd and George A Morgan. *Experimental Controls*, pages 1–2. American Cancer Society, 2010. ISBN 9780470479216. doi: 10.1002/9780470479216.corpsy0332. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0332.

[15] Susan M Courtney, Laurent Petit, Ma Maisog, Leslie G Ungerleider, and James V Haxby. An Area Specialized for Spatial Working Memory in Human Frontal Cortex. *Science*, 279(5355):1347–1351, 1998. ISSN 00368075. doi: 10.1126/science.279.5355.1347.

[16] Susan M Courtney. Development of orthogonal task designs in fMRI studies of higher cognition: The NIMH experience. *NeuroImage*, 62(2):1185–1189, 2012. doi: 10.1016/j.neuroimage.2012.01.007.Development.

[17] A L Hodgkin and A F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–44, aug 1952. ISSN 0022-3751.

[18] Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3):340–352, 2017. ISSN 15461726. doi: 10.1038/nn.4497. URL http://www.ncbi.nlm.nih.gov/pubmed/28230845.

[19] Constantinos Siettos and Jens Starke. Multiscale modeling of brain dynamics: from single neurons and networks to mathematical tools. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(5):438–458, 2016. ISSN 1939005X. doi: 10.1002/wsbm.1348.

[20] Stephen M. Smith, Diego Vidaurre, Christian F. Beckmann, Matthew F. Glasser, Mark Jenkinson, Karla L. Miller, Thomas E. Nichols, Emma C. Robinson, Gholamreza Salimi-Khorshidi, Mark W. Woolrich, Deanna M. Barch, Kamil Uğurbil, and David C. Van Essen. Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, 17(12):666–682, 2013. ISSN 13646613. doi: 10.1016/j.tics.2013.09.016.

[21] Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78, 1994. ISSN 1065-9471.

[22] David Sussillo, Rafal Jozefowicz, L. F. Abbott, and Chethan Pandarinath. LFADS - Latent Factor Analysis via Dynamical Systems. *arXiv*, 2016. URL http://arxiv.org/abs/1608.06315.

[23] David Sussillo and L F Abbott. Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron*, 63(4):544–557, 2009. ISSN 0896-6273. doi: 10.1016/j.neuron.2009.07.018. URL http://dx.doi.org/10.1016/j.neuron.2009.07.018.

[24] Zheng Wang and Anna W. Roe. Trial-to-trial noise cancellation of cortical field potentials in awake macaques by autoregression model with exogenous input (ARX). *Journal of Neuroscience Methods*, 194(2):266–273, jan 2011. ISSN 01650270. doi: 10.1016/j.jneumeth.2010.10.029. URL https://www.sciencedirect.com/science/article/pii/S0165027010006126.

[25] Steven L. Bressler, Mingzhou Ding, and Weiming Yang. Investigation of cooperative cortical dynamics by multivariate autoregressive modeling of event-related local field potentials. *Neurocomputing*, 26-27:625–631, 1999. ISSN 09252312. doi: 10.1016/S0925-2312(98)00162-3.

[26] Lionel Barnett and Anil K. Seth. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68, 2014. ISSN 1872678X. doi: 10.1016/j.jneumeth.2013.10.018. URL http://dx.doi.org/10.1016/j.jneumeth.2013.10.018.

[27] Juhan Aru, Jaan Aru, Viola Priesemann, Michael Wibral, Luiz Lana, Gordon Pipa, Wolf Singer, and Raul Vicente. Untangling cross-frequency coupling in neuroscience. *Current Opinion in Neurobiology*, 31(September 2014):51–61, 2015. ISSN 18736882. doi: 10.1016/j.conb.2014.08.002. URL http://dx.doi.org/10.1016/j.conb.2014.08.002.

[28] Ghahramani. An Introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(1):9–42, 2001. ISSN 0218-0014. doi: 10.1142/S0218001401000836.

[29] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint*, pages 1–14, 2013. ISSN 1312.6114v10. URL http://arxiv.org/abs/1312.6114.

[30] Laurence T. Hunt, W. M.Nishantha Malalasekera, Archy O. de Berker, Bruno Miranda, Simon F. Farmer, Timothy E.J. Behrens, and Steven W. Kennerley. Triple dissociation of attention and decision computations across prefrontal cortex. *Nature Neuroscience*, 21(10):1471–1481, 2018. ISSN 15461726. doi: 10.1038/s41593-018-0239-5. URL http://dx.doi.org/10.1038/s41593-018-0239-5.

[31] Yunzhe Liu, Raymond J. Dolan, Zeb Kurth-Nelson, and Timothy E J Behrens. Human replay spontaneously reorganises experience. *Cell*, 178(July):1–13, 2019. URL https://doi.org/10.1016/j.cell.2019.06.012.

[32] Diego Vidaurre, Laurence T. Hunt, Andrew J. Quinn, Benjamin A.E. E Hunt, Matthew J. Brookes, Anna C. Nobre, and Mark W. Woolrich. Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. *Nature Communications*, 9(2987), 2018. ISSN 20411723. doi: 10.1038/s41467-018-05316-z. URL http://dx.doi.org/10.1038/s41467-018-05316-zhttps://doi.org/10.1038/s41467-018-05316-z.

[33] Diego Vidaurre, Nicholas E. Myers, Mark Stokes, Anna C. Nobre, and Mark W. Woolrich. Temporally Unconstrained Decoding Reveals Consistent but Time-Varying Stages of Stimulus Processing. *Cerebral Cortex*, 29(2):863–874, 2019. ISSN 14602199. doi: 10.1093/cercor/bhy290.

[34] György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes. *Nature reviews. Neuroscience*, 13(6):407–20, jun 2012. ISSN 1471-0048. doi: 10.1038/nrn3241.

[35] Gaute T. Einevoll, Christoph Kayser, Nikos K. Logothetis, and Stefano Panzeri. Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nature Reviews Neuroscience*, 14(11):770–785, 2013. ISSN 1471003X. doi: 10.1038/nrn3599. URL `http://dx.doi.org/10.1038/nrn3599`.

[36] Karl J. Friston, André M. Bastos, Dimitris Pinotsis, and Vladimir Litvak. LFP and oscillations-what do they tell us? *Current Opinion in Neurobiology*, 31:1–6, 2015. ISSN 18736882. doi: 10.1016/j.conb.2014.05.004.

[37] R. T. Canolty, K. Ganguly, S. W. Kennerley, C. F. Cadieu, K. Koepsell, J. D. Wallis, and J. M. Carmena. Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. *Proceedings of the National Academy of Sciences*, 107(40):17356–17361, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1008306107. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.1008306107`.

[38] Rosalyn J. Moran, Fabienne Jung, Tetsuya Kumagai, Heike Endepols, Rudolf Graf, Raymond J. Dolan, Karl J. Friston, Klaas E. Stephan, and Marc Tittgemeyer. Dynamic causal models and physiological inference: A validation study using isoflurane anaesthesia in rodents. *PLoS ONE*, 6(8), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0022790.

[39] Adriano B L Tort, Mark A Kramer, Catherine Thorn, Daniel J Gibson, Yasuo Kubota, Ann M Graybiel, and Nancy J Kopell. Dynamic cross-frequency couplings of local field potential oscillations in rat striatum and hippocampus during performance of a T-maze task. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (51):20517–22, 2008. ISSN 1091-6490. doi: 10.1073/pnas.0810524105. URL `http://www.ncbi.nlm.nih.gov/pubmed/19074268{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2629291`.

[40] Junji Ito, Pedro Maldonado, and Sonja Gruen. Cross-frequency interaction of the eye-movement related LFP signals in V1 of freely viewing monkeys , 2013. URL `https://www.frontiersin.org/article/10.3389/fnsys.2013.00001`.

[41] Ryan C. Kelly, Matthew A. Smith, Robert E. Kass, and Tai Sing Lee. Local field potentials indicate network state and account for neuronal response variability. *Journal of Computational Neuroscience*, 29(3):567–579, 2010. ISSN 09295313. doi: 10.1007/s10827-009-0208-9.

[42] Yoshinao Kajikawa and Charles E. Schroeder. How local is the local field potential? *Neuron*, 72(5):847–858, 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.09.029. URL `http://dx.doi.org/10.1016/j.neuron.2011.09.029`.

[43] Steffen Katzner, Ian Nauhaus, Andrea Benucci, Vincent Bonin, Dario L. Ringach, and Matteo Carandini. Local Origin of Field Potentials in Visual Cortex. *Neuron*, 61(1):35–41, 2009. ISSN 08966273. doi: 10.1016/j.neuron.2008.11.016. URL http://dx.doi.org/10.1016/j.neuron.2008.11.016.

[44] D. Xing, C.-I Yeh, and R. M. Shapley. Spatial Spread of the Local Field Potential and its Laminar Variation in Visual Cortex. *Journal of Neuroscience*, 29(37):11540–11549, 2009. ISSN 0270-6474. doi: 10.1523/jneurosci.2573-09.2009.

[45] Philipp Berens, Georgios A Keliris, Alexander S Ecker, Nikos K Logothetis, and Andreas S Tolias. Feature selectivity of the gamma-band of the local field potential in primate primary visual cortex. *Frontiers in Neuroscience*, 2(2):199–207, dec 2008. ISSN 16624548. doi: 10.3389/neuro.01.037.2008. URL http://www.ncbi.nlm.nih.gov/pubmed/19225593http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2622750http://journal.frontiersin.org/article/10.3389/neuro.01.037.2008/abstract.

[46] Ian Nauhaus, Laura Busse, Matteo Carandini, and Dario L Ringach. Stimulus contrast modulates functional connectivity in visual cortex. *Nature Neuroscience*, 12(1):70–76, jan 2009. ISSN 1097-6256. doi: 10.1038/nn.2232. URL http://www.ncbi.nlm.nih.gov/pubmed/19029885http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2610236http://www.nature.com/articles/nn.2232.

[47] Chunmao Wang, Istvan Ulbert, Donald Schomer, Ksenija Marinkovic, and Eric Halgren. Responses of Human Anterior Cingulate Cortex Microdomains to Error Detection, Conflict Monitoring, Stimulus-Response Mapping, Familiarity, and Orienting. *Journal of Neuroscience*, 25(3):604–613, jan 2005. ISSN 0270-6474. doi: 10.1523/jneurosci.4151-04.2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15659596http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4151-04.2005.

[48] Gabriel Kreiman, Chou P. Hung, Alexander Kraskov, Rodrigo Quian Quiroga, Tomaso Poggio, and James J. DiCarlo. Object Selectivity of Local Field Potentials and Spikes in the Macaque Inferior Temporal Cortex. *Neuron*, 49(3):433–445, feb 2006. ISSN 08966273. doi: 10.1016/j.neuron.2005.12.019. URL http://www.ncbi.nlm.nih.gov/pubmed/16446146https://linkinghub.elsevier.com/retrieve/pii/S0896627305011311.

[49] C E Schroeder, C E Tenke, and S J Givre. Subcortical contributions to the surface-recorded flash-VEP in the awake macaque. *Electroencephalography and clinical neurophysiology*, 84(3):219–31, 1992. ISSN 0013-4694. URL http://www.ncbi.nlm.nih.gov/pubmed/1375881.

[50] Henrik Lindén, Tom Tetzlaff, Tobias C. Potjans, Klas H. Pettersen, Sonja Grün, Markus Diesmann, and Gaute T. Einevoll. Modeling the spatial reach of the LFP. *Neuron*, 72(5):859–872, dec 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.11.006. URL https://www.sciencedirect.com/science/article/pii/S0896627311010051?via{%}3Dihub.

[51] Szymon Leski, Henrik Lindén, Tom Tetzlaff, Klas H. Pettersen, and Gaute T. Einevoll. Frequency Dependence of Signal Power and Spatial Reach of the Local Field Potential. *PLoS Computational Biology*, 9(7):e1003137, jul 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003137. URL https://dx.plos.org/10.1371/journal.pcbi.1003137.

[52] Gary R. Holt and Christof Koch. Electrical Interactions via the Extracellular Potential Near Cell Bodies. *Journal of Computational Neuroscience*, 6(2): 169–184, 1999. ISSN 09295313. doi: 10.1023/A:1008832702585. URL `http://link.springer.com/10.1023/A:1008832702585`.

[53] Alexander Maier, Geoffer Adams, Christopher Aura, and David Leopold. Distinct Superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. *Frontiers in System Neuroscience*, 4(August): 2008–2011, 2010. doi: 10.3389/fnsys.2010.00031.

[54] Markus Siegel, Tobias H Donner, and Andreas K Engel. Spectral fingerprints of large-scale neuronal interactions. *Nature reviews. Neuroscience*, 13(2): 121–34, 2012. ISSN 1471-0048. doi: 10.1038/nrn3137. URL `http://www.ncbi.nlm.nih.gov/pubmed/22233726`.

[55] Tobias H. Donner and Markus Siegel. A framework for local cortical oscillation patterns. *Trends in Cognitive Sciences*, 15(5): 191–199, may 2011. ISSN 13646613. doi: 10.1016/j.tics.2011.03.007. URL `http://www.ncbi.nlm.nih.gov/pubmed/21481630https://linkinghub.elsevier.com/retrieve/pii/S1364661311000428`.

[56] V. Nacher, A. Ledberg, G. Deco, and R. Romo. Coherent delta-band oscillations between cortical areas correlate with decision making. *Proceedings of the National Academy of Sciences*, 110(37):15085–15090, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1314681110.

[57] P. Fries, T. Womelsdorf, R. Oostenveld, and R. Desimone. The Effects of Visual Stimulation and Selective Visual Attention on Rhythmic Neuronal Synchronization in Macaque Area V4. *Journal of Neuroscience*, 28(18): 4823–4835, 2008. ISSN 0270-6474. doi: 10.1523/jneurosci.4499-07.2008.

[58] Peter Lakatos, George Karmos, Ashesh D Mehta, Istvan Ulbert, and Charles E Schroeder. Entrainment of Neuronal Attentional Selection. *Science*, 320 (April):23–25, 2008. ISSN 10959203. doi: 10.1126/science.1154735. URL `http://www.ncbi.nlm.nih.gov/pubmed/18388295`.

[59] Maryam Saleh, Jacob Reimer, Richard Penn, Catherine L. Ojakangas, and Nicholas G. Hatsopoulos. Fast and Slow Oscillations in Human Primary Motor Cortex Predict Oncoming Behaviorally Relevant Cues. *Neuron*, 65 (4):461–471, 2010. ISSN 08966273. doi: 10.1016/j.neuron.2010.02.001. URL `http://dx.doi.org/10.1016/j.neuron.2010.02.001`.

[60] Thalía Harmony. The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neuroscience*, 7(December):1–10, 2013. doi: 10.3389/fnint.2013.00083.

[61] Alan Richardson-Klavehn, Hermann Hinrichs, Lars Buentjen, Friedhelm C Schmitt, Juergen Voges, Tino Zaehle, Klaus Kopitzki, Catherine M Sweeney-Reed, Robert T Knight, Christine Esslinger, and Hans-Jochen Heinze. Corticothalamic phase synchrony and cross-frequency coupling predict human memory formation. *eLife*, 3, dec 2014. ISSN 2050-084X. doi: 10.7554/elife.05352. URL `https://elifesciences.org/articles/05352`.

[62] Scott L. Brincat and Earl K. Miller. Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nature Neuroscience*, 18(4):576–581, 2015. ISSN 15461726. doi: 10.1038/nn.3954. URL http://dx.doi.org/10.1038/nn.3954.

[63] Andrew J. Watrous, Nitin Tandon, Chris R. Conner, Thomas Pieters, and Arne D. Ekstrom. Frequency-specific network connectivity increases underlie accurate spatiotemporal memory retrieval. *Nature Neuroscience*, 16(3):349–356, 2013. ISSN 10976256. doi: 10.1038/nn.3315. URL http://dx.doi.org/10.1038/nn.3315.

[64] Patrick S. Cooper, Aaron S.W. Wong, Montana McKewen, Patricia T. Michie, and Frini Karayanidis. Frontoparietal theta oscillations during proactive control are associated with goal-updating and reduced behavioral variability. *Biological Psychology*, 129:253–264, oct 2017. ISSN 03010511. doi: 10.1016/j.biopsycho.2017.09.008. URL http://www.ncbi.nlm.nih.gov/pubmed/28923361https://linkinghub.elsevier.com/retrieve/pii/S0301051117302430.

[65] James F Cavanagh and Michael J Frank. Frontal theta as a mechanism for cognitive control, 2014. ISSN 1879307X. URL http://dx.doi.org/10.1016/j.tics.2014.04.012.

[66] Randolph F. Helfrich and Robert T. Knight. Oscillatory Dynamics of Prefrontal Cognitive Control. *Trends in Cognitive Sciences*, 20(12):916–930, 2016. ISSN 1879307X. doi: 10.1016/j.tics.2016.09.007. URL http://dx.doi.org/10.1016/j.tics.2016.09.007.

[67] Yuval Nir, Lior Fisch, Roy Mukamel, Hagar Gelbard-Sagiv, Amos Arieli, Itzhak Fried, and Rafael Malach. Coupling between Neuronal Firing Rate, Gamma LFP, and BOLD fMRI Is Related to Interneuronal Correlations. *Current Biology*, 17(15):1275–1285, aug 2007. ISSN 09609822. doi: 10.1016/j.cub.2007.06.066. URL http://www.ncbi.nlm.nih.gov/pubmed/17686438https://linkinghub.elsevier.com/retrieve/pii/S0960982207016351.

[68] Supratim Ray and John H.R. Maunsell. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, 9(4), 2011. ISSN 15449173. doi: 10.1371/journal.pbio.1000610.

[69] Nicolas M. Brunet, Mark Roberts, Martin Vinck, Robert Oostenveld, Robert Desimone, Conrado A. Bosman, Peter De Weerd, and Pascal Fries. Stimulus repetition modulates gamma-band synchronization in primate visual cortex. *Proceedings of the National Academy of Sciences*, 111(9):3626–3631, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1309714111.

[70] K. J. Miller, G. Schalk, E. C. Leuthardt, D. W. Moran, N. R. Anderson, R. P. N. Rao, J. G. Ojemann, and J. W. Miller. Spectral Changes in Cortical Surface Potentials during Motor Movement. *Journal of Neuroscience*, 27(9):2424–2432, 2007. ISSN 0270-6474. doi: 10.1523/jneurosci.3886-06.2007.

[71] Erin L. Rich and Joni D. Wallis. Spatiotemporal dynamics of information encoding revealed in orbitofrontal high-gamma. *Nature Communications*, 8(1):1–13, 2017. ISSN 20411723. doi: 10.1038/s41467-017-01253-5. URL http://dx.doi.org/10.1038/s41467-017-01253-5.

[72] C. Magri, N. K. Logothetis, M. A. Montemurro, A. Gretton, A. Belitski, Y. Murayama, and S. Panzeri. Low-Frequency Local Field Potentials and Spikes in Primary Visual Cortex Convey Independent Visual Information. *Journal of Neuroscience*, 28(22):5696–5709, may 2008. ISSN 0270-6474. doi: 10.1523/jneurosci.0009-08.2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18509031http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0009-08.2008.

[73] J. Liu and W. T. Newsome. Local Field Potential in Cortical Area MT: Stimulus Tuning and Behavioral Correlations. *Journal of Neuroscience*, 26(30):7779–7790, jul 2006. ISSN 0270-6474. doi: 10.1523/jneurosci.5052-05.2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16870724http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5052-05.2006.

[74] János A Perge, Shaomin Zhang, Wasim Q Malik, Mark L Homer, Sydney Cash, Gerhard Friehs, Emad N Eskandar, John P Donoghue, and Leigh R Hochberg. Reliability of directional information in unsorted spikes and local field potentials recorded in human motor cortex. *Journal of Neural Engineering*, 11(4):046007, aug 2014. ISSN 17412552. doi: 10.1088/1741-2560/11/4/046007. URL http://www.ncbi.nlm.nih.gov/pubmed/24921388http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4142142http://stacks.iop.org/1741-2552/11/i=4/a=046007?key=crossref.50db475a7e9a37b3b63538a0e8e782e1.

[75] G. F. Woodman, M.-S. Kang, A. F. Rossi, and J. D. Schall. Nonhuman primate event-related potentials indexing covert shifts of attention. *Proceedings of the National Academy of Sciences*, 104(38):15111–15116, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0703477104.

[76] David C Godlove, Erik E Emeric, Courtney M Segovis, Michelle S Young, Jeffrey D Schall, and Geoffrey F Woodman. Event-Related Potentials Elicited by Errors during the Stop-Signal Task. I. Macaque Monkeys. *The Journal of Neuroscience*, 31(44):15640 LP – 15649, nov 2011. doi: 10.1523/JNEUROSCI.3349-11.2011. URL http://www.jneurosci.org/content/31/44/15640.abstract.

[77] Laurence T. Hunt, Timothy E.J. Behrens, Takayuki Hosokawa, Jonathan D. Wallis, and Steven W. Kennerley. Capturing the temporal evolution of choice across prefrontal cortex. *eLife*, 4(DECEMBER2015):1–25, 2015. ISSN 2050084X. doi: 10.7554/eLife.11945.001.

[78] M. Petti, S. Caschera, A. Anzolin, J. Toppi, F. Pichiorri, F. Babiloni, F. Cincotti, D. Mattia, and L. Astolfi. Effect of inter-trials variability on the estimation of cortical connectivity by Partial Directed Coherence. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3791–3794. IEEE, aug 2015. ISBN 978-1-4244-9271-8. doi: 10.1109/EMBC.2015.7319219. URL http://ieeexplore.ieee.org/document/7319219/.

[79] G. L. Colclough, M. W. Woolrich, P. K. Tewarie, M. J. Brookes, A. J. Quinn, and S. M. Smith. How reliable are MEG resting-state connectivity metrics? *NeuroImage*, 138:284–293, 2016. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.05.070. URL http://dx.doi.org/10.1016/j.neuroimage.2016.05.070.

[80] Gregor M. Hoerzer, Stefanie Liebe, Alois Schloegl, Nikos K. Logothetis, and Gregor Rainer. Directed coupling in local field potentials of macaque V4 during visual short-term memory revealed by multivariate autoregressive models. *Frontiers in Computational Neuroscience*, 4(May):1–13, 2010. doi: 10.3389/fncom.2010.00014.

[81] A. Bollimunta, Y. Chen, C. E. Schroeder, and M. Ding. Neuronal Mechanisms of Cortical Alpha Oscillations in Awake-Behaving Macaques. *Journal of Neuroscience*, 28(40):9976–9988, 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2699-08.2008. URL `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2699-08.2008`.

[82] James M. McFarland, Thomas T.G. Hahn, and Mayank R. Mehta. Explicit-duration hidden markov model inference of up-down states from continuous signals. *PLoS ONE*, 6(6), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0021606.

[83] Diego Vidaurre, Andrew J. Quinn, Adam P. Baker, David Dupret, Alvaro Tejero-Cantero, and Mark W. Woolrich. Spectrally resolved fast transient brain states in electrophysiological data. *NeuroImage*, 126:81–95, 2016. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.11.047. URL `http://dx.doi.org/10.1016/j.neuroimage.2015.11.047`.

[84] Andrew Quinn, Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Anna Christina Nobre, and Mark W Woolrich. Task-Evoked Dynamic Network Analysis through Hidden Markov Modelling. *Frontiers in Neuroscience*, 12 (August):1–17, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00603.

[85] Yan Xu, Simon Haykin, and Ronald J. Racine. Multiple window time-frequency distribution and coherence of EEG using Slepian sequences and Hermite functions. *IEEE Transactions on Biomedical Engineering*, 46(7):861–866, 1999. ISSN 00189294. doi: 10.1109/10.771197.

[86] Wong K, Mukamel E., Salazar A, Pierce E, Harrell P, Walsh J, Sampson A, and Purdon PL Brown E. Robust time-varying multivariate coherence estimation: application to electroencephalogram recordings during general anesthesia. *Conf Proc IEEE Eng Med Biol Soc*, 1(1):4725–4728, 2011. ISSN 0736-6205. doi: 10.1002/cyto.a.20594.Use.

[87] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002. ISSN 08966273. doi: 10.1016/S0896-6273(02)01091-7.

[88] T. Carlson, D. A. Tovar, A. Alink, and N. Kriegeskorte. Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10):1–1, 2013. ISSN 1534-7362. doi: 10.1167/13.10.1. URL `http://jov.arvojournals.org/Article.aspx?doi=10.1167/13.10.1`.

[89] John D Murray, Alberto Bernacchia, Nicholas A Roy, Christos Constantinidis, Ranulfo Romo, Xiao-Jing Wang, Stefano Fusi, and Julio C Martinez-Trujillo. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*, 114(2):1–6, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1619449114.

[90] E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, and T. Poggio. Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *Journal of Neurophysiology*, 100(3):1407–1419, 2008. ISSN 0022-3077. doi: 10.1152/jn.90248.2008. URL `http://jn.physiology.org/cgi/doi/10.1152/jn.90248.2008`.

[91] Mark G Stokes, Makoto Kusunoki, Natasha Sigala, Hamed Nili, David Gaffan, and John Duncan. Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78(2):364–375, 2013. ISSN 08966273. doi: 10.1016/j.neuron.2013.01.039. URL `http://dx.doi.org/10.1016/j.neuron.2013.01.039`.

[92] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013. ISSN 1476-4687. doi: 10.1038/nature12742. URL `http://dx.doi.org/10.1038/nature12742`.

[93] William E Skaggs, Bruce L McNaughton, Mathew A Wilson, and Carol A. Barnes. Theta Phase Precession in Hippocampal Neuronal Populations and the Compression of Temporal Sequences. *Hippocampus*, 6:149–172, 1996.

[94] Ole Jensen and John E. Lisman. Position Reconstruction From an Ensemble of Hippocampal Place Cells: Contribution of Theta Phase Coding. *Journal of Neurophysiology*, 83(5):2602–2609, 2000. ISSN 0022-3077, 1522-1598. doi: papers://FAFC0638-5DD4-4A81-A69F-F8A54DFE70C3/Paper/p7641.

[95] Ole Jensen and John E Lisman. Hippocampal CA3 region predicts memory sequences: accounting for the phase precession of place cells. *Learning & memory (Cold Spring Harbor, N.Y.)*, 3(2-3):279–287, 1996. ISSN 1072-0502. doi: 10.1101/lm.3.2-3.279.

[96] György Buzsáki and Edvard I Moser. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature neuroscience*, 16(2):130–8, 2013. ISSN 1546-1726. doi: 10.1038/nn.3304. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4079500{&}tool=pmcentrez{&}rendertype=abstract`.

[97] Mikael Lundqvist, Jonas Rose, Pawel Herman, Scott L L. Brincat, Timothy J J. Buschman, and Earl K K. Miller. Gamma and Beta Bursts Underlie Working Memory. *Neuron*, 90(1):152–164, 2016. ISSN 10974199. doi: 10.1016/j.neuron.2016.02.028. URL `http://dx.doi.org/10.1016/j.neuron.2016.02.028`.

[98] Ali Bahramisharif, Ole Jensen, Joshua Jacobs, and John Lisman. Serial representation of items during working memory maintenance at letter-selective cortical sites. *PLOS Biology*, 16(8):e2003805, 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2003805. URL `http://dx.plos.org/10.1371/journal.pbio.2003805`.

[99] Zeb Kurth-Nelson, Marcos Economides, Raymond J. Dolan, and Peter Dayan. Fast Sequences of Non-spatial State Representations in Humans. *Neuron*, 91(1):194–204, 2016. ISSN 10974199. doi: 10.1016/j.neuron.2016.05.028. URL `http://dx.doi.org/10.1016/j.neuron.2016.05.028`.

[100] Sebastian Michelmann, Howard Bowman, and Simon Hanslmayr. The Temporal Signature of Memories: Identification of a General Mechanism for Dynamic Memory Replay in Humans. *PLoS Biology*, 14(8):1–27, 2016. ISSN 15457885. doi: 10.1371/journal.pbio.1002528.

[101] Xin Di, Zening Fu, Shing Chow Chan, Yeung Sam Hung, Bharat B. Biswal, and Zhiguo Zhang. Task-related functional connectivity dynamics in a block-designed visual experiment. *Frontiers in Human Neuroscience*, 9(September):1–11, 2015. ISSN 1662-5161. doi: 10.3389/fnhum.2015.00543. URL http://journal.frontiersin.org/Article/10.3389/fnhum.2015.00543/abstract.

[102] David M. Groppe, Thomas P. Urbach, and Marta Kutas. Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, 48(12):1711–1725, 2011. ISSN 14698986. doi: 10.1111/j.1469-8986.2011.01273.x.

[103] N. Ding and J. Z. Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1):78–89, 2012. ISSN 0022-3077. doi: 10.1152/jn.00297.2011. URL http://jn.physiology.org/cgi/doi/10.1152/jn.00297.2011.

[104] Nicholas E. Myers, Gustavo Rohenkohl, Valentin Wyart, Mark W. Woolrich, Anna C. Nobre, and Mark G. Stokes. Testing sensory evidence against mnemonic templates. *eLife*, 4(DECEMBER2015):1–25, 2015. ISSN 2050084X. doi: 10.7554/eLife.09000.

[105] Matthew J. Brookes, Andrew M. Gibson, Stephen D. Hall, Paul L. Furlong, Gareth R. Barnes, Arjan Hillebrand, Krish D. Singh, Ian E. Holliday, Sue T. Francis, and Peter G. Morris. GLM-beamformer method demonstrates stationary field, alpha ERD and gamma ERS co-localisation with fMRI BOLD response in visual cortex. *NeuroImage*, 26(1):302–308, 2005. ISSN 10538119. doi: 10.1016/j.neuroimage.2005.01.050.

[106] O. Hauk, M. H. Davis, M. Ford, F. Pulvermüller, and W. D. Marslen-Wilson. The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, 30(4):1383–1400, 2006. ISSN 10538119. doi: 10.1016/j.neuroimage.2005.11.048.

[107] Theodore P. Zanto, Michael T. Rubens, Arul Thangavel, and Adam Gazzaley. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience*, 14(5):656–663, 2011. ISSN 10976256. doi: 10.1038/nn.2773. URL http://dx.doi.org/10.1038/nn.2773.

[108] Bernd Feige, Klaus Scheffler, Fabrizio Esposito, Di Salle Francesco, Jurgen Hennig, and Erich Seifritz. Cortical and subcortical correlates of electrocephalographic alpha rhythm modulation. *J Neurophysiol*, 93:2864–2872, 2005. doi: 10.1152/jn.00721.2004. URL http://www.researchgate.net/publication/8127707{_}Cortical{_}and{_}subcortical{_}correlates{_}of{_}electroencephalogra file/f2faf4f8575323260c.pdf.

[109] Giovanni M. Di Liberto, James A. O'Sullivan, and Edmund C. Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015. ISSN 09609822. doi: 10.1016/j.cub.2015.08.030. URL http://dx.doi.org/10.1016/j.cub.2015.08.030.

[110] Michael L. Platt and Paul W. Glimcher. Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238, 1999. ISSN 00280836. doi: 10.1038/22268.

[111] Yun Gao Yun Gao, M J Black, E Bienenstock, Wei Wu Wei Wu, and J P Donoghue. A quantitative comparison of linear and non-linear models of motor cortical activity for the encoding and decoding of arm motions. *First International IEEE EMBS Conference on Neural Engineering 2003 Conference Proceedings*, pages 189–192, 2003. ISSN 19483554 19483546. doi: 10.1109/CNE.2003.1196789. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1196789`.

[112] S. M. Montgomery and G. Buzsaki. Gamma oscillations dynamically couple hippocampal CA3 and CA1 regions during memory task performance. *Proceedings of the National Academy of Sciences*, 104(36):14495–14500, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0701826104. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0701826104`.

[113] K J Friston, Andrew Holmes, K J Worsley, J B Poline, C D Frith, and R S Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(189-210), 1995.

[114] Christian F. Beckmann, Mark Jenkinson, and Stephen M. Smith. General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, 20(2):1052–1063, 2003. ISSN 10538119. doi: 10.1016/S1053-8119(03)00435-X.

[115] Jeanette A. Mumford and Thomas Nichols. Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–1475, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.05.034. URL `http://dx.doi.org/10.1016/j.neuroimage.2009.05.034`.

[116] M. E. Van de Nieuwenhuijzen, A. R. Backus, A. Bahramisharif, C. F. Doeller, O. Jensen, and M. A.J. van Gerven. MEG-based decoding of the spatiotemporal dynamics of visual category perception. *NeuroImage*, 83:1063–1073, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.07.075. URL `http://dx.doi.org/10.1016/j.neuroimage.2013.07.075`.

[117] Abdulhamit Subasi and Ergun Erçelebi. Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine*, 78(2):87–99, 2005. ISSN 01692607. doi: 10.1016/j.cmpb.2004.10.009.

[118] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2013.10.067. URL `http://dx.doi.org/10.1016/j.neuroimage.2013.10.067`.

[119] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

[120] Adam P. Baker, Matthew J. Brookes, Iead A. Rezek, Stephen M. Smith, Timothy Behrens, Penny J.Probert Smith, and Mark Woolrich. Fast transient networks in spontaneous human brain activity. *eLife*, 2014(3):1–18, 2014. ISSN 2050084X. doi: 10.7554/eLife.01867.

[121] David J C Mackay. Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, jan 1995. ISSN 0954-898X. doi: 10.1088/0954-898X_6_3_011. URL https://doi.org/10.1088/0954-898X{_}6{_}3{_}011.

[122] Jan Drugowitsch. Variational Bayesian inference for linear and logistic regression. *arXiv preprint*, 2013. URL http://arxiv.org/abs/1310.5438.

[123] Christopher M Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 9780387310732.

[124] Charles F Manski and Steven R Lerman. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica*, 45(8):1977–1988, 1977. URL http://www.jstor.org/stable/1914121.

[125] Gary King and Langche Zeng. Logistic Regression in Rare Events Data. *Political Analysis*, 9(02):137–163, 2001. ISSN 1047-1987. doi: 10.1093/oxfordjournals.pan.a004868. URL https://www.cambridge.org/core/product/identifier/S1047198700003740/type/journal{_}article.

[126] Maher Maalouf and Mohammad Siddiqi. Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, 59:142–148, 2014. ISSN 09507051. doi: 10.1016/j.knosys.2014.01.012. URL http://dx.doi.org/10.1016/j.knosys.2014.01.012.

[127] FEJ Harrell and KL Lee. *A Comparison of the Discriminant of Discriminant Analysis and Logistic Regression Under Multivariate Normality*. 1985.

[128] Zeb Kurth-Nelson, Gareth Barnes, Dino Sejdinovic, Ray Dolan, and Peter Dayan. Temporal structure in associative retrieval. *eLife*, 2015(4):1–18, 2015. ISSN 2050084X. doi: 10.7554/eLife.04919.

[129] L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio. The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, 111(1):91–102, 2014. ISSN 0022-3077. doi: 10.1152/jn.00394.2013. URL http://jn.physiology.org/cgi/doi/10.1152/jn.00394.2013.

[130] Thomas A Carlson, Hinze Hogendoorn, Ryota Kanai, Juraj Mesik, and Jeremy Turret. High temporal resolution decoding of object position and category. *Journal of Vision*, 11(10):1–17, 2011. doi: 10.1167/11.10.9.Introduction.

[131] Nadine Dijkstra, Pim Mostert, Floris P. DeLange, Sander Bosh, and Marcel A. J. van Gerven. Differential temporal dynamics during visual imagery and perception. *Biorxiv*, pages 1–16, 2017. ISSN 2050-084X. doi: 10.1101/226217.

[132] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014. ISSN 10976256. doi: 10.1038/nn.3635. URL http://dx.doi.org/10.1038/nn.3635.

[133] Dean Wyatte, David J. Jilk, and Randall C. O'Reilly. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, 5(JUL):1–10, 2014. ISSN 16641078. doi: 10.3389/fpsyg.2014.00674.

[134] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154.2, 1962. ISSN 0022-3751. doi: 10.1523/JNEUROSCI.1991-09.2009.

[135] Yann Lecun, Le'on Bottou, Yoshua Bengio, and Parick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL `http://ieeexplore.ieee.org/document/726791/{#}full-text-section`.

[136] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, pages 1097–1105, 2012. doi: 10.1201/9781420010749.

[137] N. V.Kartheek Medathati, Heiko Neumann, Guillaume S. Masson, and Pierre Kornprobst. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Computer Vision and Image Understanding*, 150(April):1–30, 2016. ISSN 1090235X. doi: 10.1016/j.cviu.2016.04.009. URL `http://dx.doi.org/10.1016/j.cviu.2016.04.009`.

[138] Felleman DJ and Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[139] Erin Goddard, Thomas A. Carlson, Nadene Dermody, and Alexandra Woolgar. Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage*, 128:385–397, 2016. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.01.006. URL `http://dx.doi.org/10.1016/j.neuroimage.2016.01.006`.

[140] Melvyn A. Goodale and David A. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. ISSN 0269-1205. doi: 10.1093/litthe/11.1.80.

[141] J. R. King and S. Dehaene. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014. ISSN 1879307X. doi: 10.1016/j.tics.2014.01.002. URL `http://dx.doi.org/10.1016/j.tics.2014.01.002`.

[142] Matthew a Wilson and Bruce L McNaughton. Reactivation of Hippocampal Ensemble Memories During Sleep. *Science*, 265(14):676–679, 1994. ISSN 10959203. doi: 10.1126/science.8036517.

[143] C Pavlides and J Wilson. Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience*, 9(8):2907–2918, 1989.

[144] Joseph O'Neill, Timothy Senior, and Jozsef Csicsvari. Place-selective firing of CA1 pyramidal cells during sharp wave/ripple network patterns in exploratory behavior. *Neuron*, 49(1):143–155, jan 2006. ISSN 08966273. doi: 10.1016/j.neuron.2005.10.037. URL `http://www.ncbi.nlm.nih.gov/pubmed/16387646`.

[145] Hemant S. Kudrimoti, Carol A. Barnes, and Bruce L. McNaughton. Reactivation of Hippocampal Cell Assemblies: Effects of Behavioral State, Experience, and EEG Dynamics. *The Journal of Neuroscience*, 19(10):4090–4101, 1999. ISSN 0270-6474. doi: 10.1523/jneurosci.19-10-04090.1999.

[146] Y L Qin, B L McNaughton, W E Skaggs, and C A Barnes. Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *The hippocampal and parietal foundations of spatial cognition*, 352(1360): 305–319, 1997. ISSN 0962-8436. doi: 10.1098/rstb.1997.0139.

[147] Daoyun Ji and Matthew A Wilson. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, 10(1):100–107, 2007. ISSN 10976256. doi: 10.1038/nn1825.

[148] James J. Knierim. The hippocampus. *Current Biology*, 25(23):R1116–R1121, 2015. ISSN 09609822. doi: 10.1016/j.cub.2015.10.049.

[149] W B SCOVILLE and B MILNER. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):11–21, feb 1957. ISSN 0022-3050. doi: 10.1136/jnnp.20.1.11. URL http://www.ncbi.nlm.nih.gov/pubmed/13406589http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC497229http://www.ncbi.nlm.nih.gov/pubmed/13406589{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC497229.

[150] Timothy J. Teyler and Pascal DiScenna. The Hippocampal Memory Indexing Theory. *Behavioral Neuroscience*, 100(2):147–154, 1986. ISSN 07357044. doi: 10.1037/0735-7044.100.2.147. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0735-7044.100.2.147.

[151] Timothy J. Teyler and Jerry W. Rudy. The hippocampal indexing theory and episodic memory: Updating the index, dec 2007. ISSN 10509631. URL http://doi.wiley.com/10.1002/hipo.20350.

[152] J O'Keefe and J Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 34(1): 171–5, nov 1971. ISSN 0006-8993. URL http://www.ncbi.nlm.nih.gov/pubmed/5124915.

[153] JS Taube, RU Muller, and JB Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *The Journal of Neuroscience*, 10(2):436–447, feb 2018. ISSN 0270-6474. doi: 10.1523/jneurosci.10-02-00436.1990. URL http://www.ncbi.nlm.nih.gov/pubmed/2303851.

[154] Marianne Fyhn, Sturla Molden, Menno P Witter, Edvard I Moser, and May-Britt Moser. Spatial representation in the entorhinal cortex. *Science (New York, N.Y.)*, 305(5688):1258–64, aug 2004. ISSN 1095-9203. doi: 10.1126/science.1099901.

[155] Howard Eichenbaum. Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience*, 15(11):732–744, 2014. ISSN 14710048. doi: 10.1038/nrn3827.

[156] David J. Foster. Replay Comes of Age. *Annual Review of Neuroscience*, 40(1):581–602, 2017. ISSN 0147-006X. doi: 10.1146/annurev-neuro-072116-031538. URL http://www.annualreviews.org/doi/10.1146/annurev-neuro-072116-031538.

[157] WE Skaggs and BL McNaughton. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, 271: 1870–1873, 1996.

[158] Kenway Louie and Matthew A Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29 (1):145–156, jan 2001. ISSN 08966273. doi: 10.1016/S0896-6273(01)00186-6. URL `http://www.ncbi.nlm.nih.gov/pubmed/11182087`.

[159] Albert K Lee and Matthew A Wilson. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194, dec 2002. ISSN 08966273. doi: 10.1016/S0896-6273(02)01096-6. URL `http://www.ncbi.nlm.nih.gov/pubmed/12495631`.

[160] Szabolcs Káli and Peter Dayan. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience*, 7(3):286–294, 2004. ISSN 1546-1726. doi: 10.1038/nn1202. URL `https://doi.org/10.1038/nn1202`.

[161] Francesco P. Battaglia, Karim Benchenane, Anton Sirota, Cyriel M.A. Pennartz, and Sidney I. Wiener. The hippocampus: Hub of brain network communication for memory. *Trends in Cognitive Sciences*, 15(7): 310–318, 2011. ISSN 13646613. doi: 10.1016/j.tics.2011.05.008. URL `http://dx.doi.org/10.1016/j.tics.2011.05.008`.

[162] Anton Sirota, Jozsef Csicsvari, and Derek Buhl. Communication between neocortex and hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):2065–2069, 2003. ISSN 1574695X. doi: 10.1111/j.1574-695X.1993.tb00417.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/12576550`.

[163] David R Euston, Masami Tatsuno, Bruce L Mcnaughton, and R Euston. Fast-Forward Playback of Recent Memory Sequences in Prefrontal Cortex During Sleep. *Science*, 318(5853):1147–1150, 2007. URL `http://science.sciencemag.org/`.

[164] Adrien Peyrache, Mehdi Khamassi, Karim Benchenane, Sidney I Wiener, and Francesco P Battaglia. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7):919–926, 2009. ISSN 1097-6256. doi: 10.1038/nn.2337. URL `http://dx.doi.org/10.1038/nn.2337`.

[165] CA Pennartz, E Lee, J Verhuel, P Lipa, CA Barnes, and McNaughton. The Ventral Striatum in Off-Line Processing: Ensemble Reactivation during Sleep and Modulation by Hippocampal Ripples. *Journal of Neuroscience*, 24 (29):6446–6456, jul 2004. ISSN 0270-6474. doi: 10.1523/jneurosci.0575-04.2004. URL `http://www.ncbi.nlm.nih.gov/pubmed/15269254http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0575-04.2004`.

[166] Carien S Lansink, Pieter M Goltstein, Jan V Lankelma, Ruud N J M A Joosten, Bruce L McNaughton, and Cyriel M A Pennartz. Preferential reactivation of motivationally relevant information in the ventral striatum. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(25): 6372–82, jun 2008. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1054-08.2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18562607http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3844781`.

[167] Carien S. Lansink, Pieter M. Goltstein, Jan V. Lankelma, Bruce L. McNaughton, and Cyriel M.A. Pennartz. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology*, 7(8):e1000173, aug 2009. ISSN 15449173. doi: 10.1371/journal.pbio.1000173. URL `http://dx.plos.org/10.1371/journal.pbio.1000173`.

[168] Dhakshin S. Ramanathan, Tanuj Gulati, and Karunesh Ganguly. Sleep-Dependent Reactivation of Ensembles in Motor Cortex Promotes Skill Consolidation. *PLoS Biology*, 13(9):1–25, 2015. ISSN 15457885. doi: 10.1371/journal.pbio.1002263.

[169] Gyorgy Buzsaki, Zsolt Horvath, Ronald Urioste, Jamille Hetke, and Kensall Wise. High-frequency network oscillation in the hippocampus. *Science*, 256 (5059):1025–1027, 1992. ISSN 0036-8075. doi: 10.1126/science.1589772. URL `http://science.sciencemag.org/`.

[170] György Buzsáki and Fernando Lopes da Silva. High frequency oscillations in the intact brain. *Progress in Neurobiology*, 98(3):241–249, 2012. ISSN 03010082. doi: 10.1016/j.pneurobio.2012.02.004.

[171] Laure Buhry, Amir H. Azizi, and Sen Cheng. Reactivation, Replay, and Preplay: How It Might All Fit Together. *Neural Plasticity*, 2011:1–11, 2011. ISSN 2090-5904. doi: 10.1155/2011/203462.

[172] David J. Foster and Matthew A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440 (7084):680–683, 2006. ISSN 14764687. doi: 10.1038/nature04587.

[173] Jozsef Csicsvari, Joseph O'neill, Kevin Allen, and Timothy Senior. Place-selective firing contributes to the reverse-order reactivation of CA1 pyramidal cells during sharp waves in open-field exploration. *European Journal of Neuroscience*, 26(3):704–716, 2007. ISSN 0953-816X.

[174] Kamran Diba and György Buzsáki. Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, 10(10):1241–1242, oct 2007. ISSN 1097-6256. doi: 10.1038/nn1961. URL `https://www.ncbi.nlm.nih.gov/pubmed/17828259https://www.ncbi.nlm.nih.gov/pmc/PMC2039924/`.

[175] Anoopum S Gupta, Matthijs A A van der Meer, David S Touretzky, and A David Redish. Hippocampal replay is not a simple function of experience. *Neuron*, 65(5):695–705, 2010. ISSN 0896-6273.

[176] Mattias P Karlsson and Loren M Frank. Awake replay of remote experiences in the hippocampus. *Nature Neuroscience*, 12(7):913–918, jul 2009. ISSN 1097-6256. doi: 10.1038/nn.2344. URL `http://www.nature.com/articles/nn.2344`.

[177] Thomas J Davidson, Fabian Kloosterman, and Matthew A Wilson. Hippocampal Replay of Extended Experience. *Neuron*, 63(4):497–507, aug 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.07.027. URL `http://www.ncbi.nlm.nih.gov/pubmed/19709631http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4364032`.

[178] X. Wu and D. J. Foster. Hippocampal Replay Captures the Unique Topological Structure of a Novel Environment. *Journal of Neuroscience*, 34(19):6459–6469, may 2014. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3414-13.2014. URL http://www.ncbi.nlm.nih.gov/pubmed/24806672http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4012305http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3414-13.2014.

[179] Brad E. Pfeiffer and David J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, apr 2013. ISSN 00280836. doi: 10.1038/nature12112. URL http://www.nature.com/doifinder/10.1038/nature12112.

[180] George Dragoi and Susumu Tonegawa. Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, 469(7330):397–401, 2011. ISSN 00280836. doi: 10.1038/nature09633. URL http://dx.doi.org/10.1038/nature09633.

[181] George Dragoi and Susumu Tonegawa. Distinct preplay of multiple novel spatial experiences in the rat. *Proceedings of the National Academy of Sciences*, 110(22):9100–9105, may 2013. ISSN 0027-8424. doi: 10.1073/pnas.1306031110. URL http://www.ncbi.nlm.nih.gov/pubmed/23671088http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3670374.

[182] Andres D. Grosmark and György Buzsáki. Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science*, 351(6280):1440–1443, mar 2016. ISSN 10959203. doi: 10.1126/science.aad1935. URL http://science.sciencemag.org/content/351/6280/1440.

[183] George Dragoi and Susumu Tonegawa. Selection of preconfigured cell assemblies for representation of novel spatial experiences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):1–7, dec 2014. ISSN 14712970. doi: 10.1098/rstb.2012.0522. URL http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2012.0522.

[184] H. Freyja Ólafsdóttir, Daniel Bush, and Caswell Barry. The Role of Hippocampal Replay in Memory and Planning. *Current Biology*, 28(1):R37–R50, 2018. ISSN 09609822. doi: 10.1016/j.cub.2017.10.073.

[185] Delia Silva, Ting Feng, and David J Foster. Trajectory events across hippocampal place cells require previous experience. *Nature Neuroscience*, 18(12):1772–1779, dec 2015. ISSN 15461726. doi: 10.1038/nn.4151. URL http://www.nature.com/articles/nn.4151.

[186] J. C. Jackson, A. Johnson, and A. D. Redish. Hippocampal Sharp Waves and Reactivation during Awake States Depend on Repeated Sequential Experience. *Journal of Neuroscience*, 26(48):12415–12426, nov 2006. ISSN 0270-6474. doi: 10.1523/jneurosci.4118-06.2006. URL http://www.ncbi.nlm.nih.gov/pubmed/17135403http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4118-06.2006.

[187] Joseph O'Neill, Timothy J Senior, Kevin Allen, John R Huxter, and Jozsef Csicsvari. Reactivation of experience-dependent cell assembly patterns in the hippocampus. *Nature Neuroscience*, 11(2):209–215, feb 2008. ISSN 10976256. doi: 10.1038/nn2037. URL http://www.nature.com/articles/nn2037.

[188] Mayank R Mehta. Cortico-hippocampal interaction during up-down states and memory consolidation. *Nature Neuroscience*, 10(1):13–15, 2007. URL http://www.nature.com/natureneurosciencehttp://www.nature.com/neuro/journal/v10/n1/abs/nn0107-13.html;jsessionid=AEE78F8805D0F729393401DCC9F2D411.

[189] Qiaoli Huang, Jianrong Jia, Qiming Han, and Huan Luo. Fast-backward replay of sequentially memorized items in humans. *eLife*, 7:1–21, 2018. ISSN 2050084X. doi: 10.7554/eLife.35164.

[190] L. Deuker, J. Olligs, J. Fell, T. A. Kranz, F. Mormann, C. Montag, M. Reuter, C. E. Elger, and N. Axmacher. Memory Consolidation by Replay of Stimulus-Specific Neural Activity. *Journal of Neuroscience*, 33(49):19373–19383, 2013. ISSN 0270-6474. doi: 10.1523/jneurosci.0414-13.2013.

[191] Ida Momennejad, A Ross Otto, Nathaniel D Daw, and Kenneth A Norman. Offline Replay Supports Planning: fMRI Evidence from Reward Revaluation. *bioRxiv*, page 196758, jan 2017. doi: 10.1101/196758. URL http://biorxiv.org/content/early/2017/10/02/196758.abstract.

[192] Hui Zhang, Juergen Fell, and Nikolai Axmacher. Electrophysiological mechanisms of human memory consolidation. *Nature Communications*, 9(1), 2018. ISSN 20411723. doi: 10.1038/s41467-018-06553-y. URL http://dx.doi.org/10.1038/s41467-018-06553-y.

[193] Y Yamaguchi. A theory of hippocampal memory based on theta phase precession. *Biological Cybernetics*, 89(1):1–9, 2003. ISSN 0340-1200. doi: 10.1007/s00422-003-0415-9. URL http://links.isiglobalnet2.com/gateway/Gateway.cgi?GWVersion=2{&}SrcAuth=mekentosj{&}SrcApp=Papers{&}DestLinkType=FullRecord{&}DestApp=WOS{&}KeyUT=000184014100001{%}5Cnpapers3://publication/doi/10.1007/s00422-003-0415-9.

[194] Gabrielle Girardeau and Michaël Zugaro. Hippocampal ripples and memory consolidation. *Current Opinion in Neurobiology*, 21(3):452–459, 2011. ISSN 09594388. doi: 10.1016/j.conb.2011.02.005.

[195] Nicolas W Schuck and Yael Niv. Sequential replay of nonspatial task states in the human hippocampus. *Science*, 364(June):1–9, 2019. doi: 10.1126/science.aaw5181.

[196] K.L. Hoffman and B. L. McNaughton. Coordinated Reactivation of Distributed Memory Traces in Primate Neocortex. *Science*, 297(September):2070–2073, 2002. doi: 10.1126/science.1073538.

[197] Valentin Amrhein, Sander Greenland, and Blake Mcshane. Retire Statistical Significance. *Nature*, 567:305–307, 2019.

[198] Logan Grosenick, Stephanie Greer, and Brian Knutson. Interpretable Classifiers for fMRI Improve Prediction of Purchases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(6):539–548, 2008. ISSN 15580210. doi: 10.1109/TNSRE.2008.926701.

[199] Holger Mohr, Uta Wolfensteller, Steffi Frimmel, and Hannes Ruge. Sparse regularization techniques provide novel insights into outcome integration

processes. *NeuroImage*, 104:163–176, jan 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.10.025. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811914008490.

[200] G. Elliott Wimmer, Raymond J. Dolan, and Timothy E.J. Behrens. Human replay predicts behaviour (unpublished), 2019.

[201] Pierre Lavenex and David G Amaral. Hippocampal-Neocortical Interaction: A Hierarchy of Associativity. *Hippocampus*, 10:420–430, 2000.

[202] Juan Linde-Domingo, Matthias S Treder, Casper Kerrén, and Maria Wimber. Evidence that neural information flow is reversed between object perception and object reconstruction from memory. *Nature Communications*, 10(1), 2019. ISSN 20411723. doi: 10.1038/s41467-018-08080-2. URL http://dx.doi.org/10.1038/s41467-018-08080-2.

[203] Matthew J Brookes, Joanne R Hale, Johanna M Zumer, Claire M Stevenson, Susan T Francis, Gareth R Barnes, Julia P Owen, Peter G Morris, and Srikantan S Nagarajan. Measuring functional connectivity using MEG: methodology and comparison with fcMRI. *NeuroImage*, 56(3):1082–1104, jun 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.02.054. URL https://www.ncbi.nlm.nih.gov/pubmed/21352925https://www.ncbi.nlm.nih.gov/pmc/PMC3224862/.

[204] F. de Pasquale, S. Della Penna, A. Z. Snyder, C. Lewis, D. Mantini, L. Marzetti, P. Belardinelli, L. Ciancetta, V. Pizzella, G. L. Romani, and M. Corbetta. Temporal dynamics of spontaneous MEG activity in brain networks. *Proceedings of the National Academy of Sciences*, 107(13):6040–6045, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0913863107. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0913863107.

[205] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995. doi: 10.1002/mrm.1910340409. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910340409.

[206] M H Lee, C D Smyser, and J S Shimony. Resting-state fMRI: A review of methods and clinical applications. *American Journal of Neuroradiology*, 34(10):1866–1872, oct 2013. ISSN 01956108. doi: 10.3174/ajnr.A3263. URL http://www.ncbi.nlm.nih.gov/pubmed/22936095http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4035703.

[207] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, jan 2003. ISSN 0027-8424. doi: 10.1073/pnas.0135058100. URL http://www.ncbi.nlm.nih.gov/pubmed/12506194http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC140943http://www.pnas.org/cgi/doi/10.1073/pnas.0135058100.

[208] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, and Christian F Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy*

of Sciences, 106(31):13040–13045, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0905267106. URL https://www.pnas.org/content/106/31/13040.

[209] C J Honey, O Sporns, L Cammoun, X Gigandet, J P Thiran, R Meuli, and P Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106 (6):2035–2040, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0811168106. URL https://www.pnas.org/content/106/6/2035.

[210] Alexandre Gramfort, Bertrand Thirion, Michel Thiebaut de Schotten, Mert R. Sabuncu, B T Thomas Yeo, Jonathan Smallwood, Henrik Walter, Gaël Varoquaux, Danilo Bzdok, Avram J. Holmes, Julius M Kernbach, and Daniel S Margulies. Subspecialization within default mode nodes characterized in 10,000 UK Biobank participants. *Proceedings of the National Academy of Sciences*, 115(48):12295–12300, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1804876115.

[211] M. J. Brookes, M. Woolrich, H. Luckhoo, D. Price, J. R. Hale, M. C. Stephenson, G. R. Barnes, S. M. Smith, and P. G. Morris. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proceedings of the National Academy of Sciences*, 108(40): 16783–16788, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1112685108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1112685108.

[212] Florian Gompf, Anja Pflug, Helmut Laufs, and Christian A Kell. Non-linear Relationship between BOLD Activation and Amplitude of Beta Oscillations in the Supplementary Motor Area during Rhythmic Finger Tapping and Internal Timing. *Frontiers in Human Neuroscience*, 11:582, 2017. ISSN 1662-5161. doi: 10.3389/fnhum.2017.00582. URL https://www.frontiersin.org/article/10.3389/fnhum.2017.00582.

[213] S. M. Smith, K. L. Miller, S. Moeller, J. Xu, E. J. Auerbach, M. W. Woolrich, C. F. Beckmann, M. Jenkinson, J. Andersson, M. F. Glasser, D. C. Van Essen, D. A. Feinberg, E. S. Yacoub, and K. Ugurbil. Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 109(8):3131–3136, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1121329109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1121329109.

[214] Diego Vidaurre, Stephen M. Smith, and Mark W. Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):201705120, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1705120114. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1705120114.

[215] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and Steven E Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–78, 2011. ISSN 1097-4199. doi: 10.1016/j.neuron.2011.09.006. URL http://www.ncbi.nlm.nih.gov/pubmed/22099467{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3222858.

[216] Marcus E. Raichle Michael D. Fox, Abraham Z. Snyder, Justin L. Vincent, Maurizio Corbetta, David C. Van Essen. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005. ISSN 10991395. doi: 10.1002/poc.610010207.

[217] Yuan Zhou, Karl J. Friston, Peter Zeidman, Jie Chen, Shu Li, and Adeel Razi. The Hierarchical Organization of the Default, Dorsal Attention and Salience Networks in Adolescents and Young Adults. *Cerebral Cortex*, 28(2):726–737, 2018. ISSN 14602199. doi: 10.1093/cercor/bhx307.

[218] Daniel S. Margulies, Satrajit S. Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M. Huntenburg, Georg Langs, Gleb Bezgin, Simon B. Eickhoff, F. Xavier Castellanos, Michael Petrides, Elizabeth Jefferies, and Jonathan Smallwood. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113 (44):12574–12579, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1608282113. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1608282113`.

[219] Xiaoxiao Xu, Hong Yuan, and Xu Lei. Activation and Connectivity within the Default Mode Network Contribute Independently to Future-Oriented Thought. *Scientific Reports*, 6(January):1–10, 2016. ISSN 20452322. doi: 10.1038/srep21001. URL `http://dx.doi.org/10.1038/srep21001`.

[220] D. Mantini, M. G. Perrucci, C. Del Gratta, G. L. Romani, and M. Corbetta. Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104(32):13170–13175, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0700668104. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0700668104`.

[221] Marcus E Raichle, Ann Mary Macleod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001. URL `papers2://publication/uuid/B158EF84-1CE7-4EB3-9395-A2A1239966AA`.

[222] Masaki Fukunaga, Silvina G. Horovitz, Peter van Gelderen, Jacco A. de Zwart, J. Martijn Jansma, Vasiliki N. Ikonomidou, Renxin Chu, Roel H.R. Deckers, David A. Leopold, and Jeff H. Duyn. Large-amplitude, spatially correlated fluctuations in BOLD fMRI signals during extended rest and early sleep stages. *Magnetic Resonance Imaging*, 24(8):979–992, 2006. ISSN 0730725X. doi: 10.1016/j.mri.2006.04.018.

[223] L. J. Larson-Prior, J. M. Zempel, T. S. Nolan, F. W. Prior, A. Z. Snyder, and M. E. Raichle. Cortical network functional connectivity in the descent to sleep. *Proceedings of the National Academy of Sciences*, 106(11):4489–4494, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0900924106. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0900924106`.

[224] Michael D Greicius, Vesa Kiviniemi, Osmo Tervonen, Vilho Vainionpää, Alan L Reiss, and Vinod Menon. Persistent Default-Mode Network Connectivity During Light Sedation. *Human brain mapping*, 29(7):839–847, 2008. doi: 10.1002/hbm.20537.Persistent.

[225] Yajing Pang, Qian Cui, Xujun Duan, Heng Chen, Ling Zeng, Zhiqiang Zhang, Guangming Lu, and Huafu Chen. Extraversion modulates functional connectivity hubs of resting-state brain networks. *Journal of Neuropsychology*, 11(3):347–361, sep 2017. ISSN 17486653. doi: 10.1111/jnp.12090. URL `http://www.ncbi.nlm.nih.gov/pubmed/26566723http://doi.wiley.com/10.1111/jnp.12090`.

[226] Yvette I Sheline, Deanna M Barch, Joseph L Price, Melissa M Rundle, S Neil Vaishnavi, Abraham Z Snyder, Mark A Mintun, Suzhi Wang, Rebecca S Coalson, and Marcus E Raichle. The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences*, 106 (6):1942–1947, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0812686106.

[227] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–47, 2015. ISSN 1545-4126. doi: 10.1146/annurev-neuro-071013-014030. URL `http://www.ncbi.nlm.nih.gov/pubmed/25938726`.

[228] Fabrizio Esposito, Alessandro Bertolino, Tommaso Scarabino, Valeria Latorre, Giuseppe Blasi, Teresa Popolizio, Gioacchino Tedeschi, Sossio Cirillo, Rainer Goebel, and Francesco Di Salle. Independent component model of the default-mode brain function: Assessing the impact of active thinking. *Brain Research Bulletin*, 70(4-6):263–269, 2006. ISSN 03619230. doi: 10.1016/j.brainresbull.2006.06.012.

[229] Arnaud D'Argembeau, David Stawarczyk, Steve Majerus, Fabienne Collette, Martial der Linden, and Eric Salmon. Modulation of medial prefrontal and inferior parietal cortices when thinking about past, present, and future selves. *Social neuroscience*, 5:187–200, 2009. doi: 10.1080/17470910903233562.

[230] Jessica R. Andrews-Hanna, Jay S. Reidler, Christine Huang, and Randy L. Buckner. Evidence for the Default Network's Role in Spontaneous Cognition. *Journal of Neurophysiology*, 104(1):322–335, 2010. ISSN 0022-3077. doi: 10.1152/jn.00830.2009.

[231] F. H. Lopes da Silva, T. H.M.T. van Lierop, C. F. Schrijer, and W Storm van Leeuwen. Organization of thalamic and cortical alpha rhythms: Spectra and coherences. *Electroencephalography and Clinical Neurophysiology*, 35(6):627–639, dec 1973. ISSN 00134694. doi: 10.1016/0013-4694(73)90216-2. URL `https://www.sciencedirect.com/science/article/pii/0013469473902162`.

[232] G. Pfurtscheller, A. Stancák, and Ch Neuper. Event-related synchronization (ERS) in the alpha band - An electrophysiological correlate of cortical idling: A review. *International Journal of Psychophysiology*, 24(1-2):39–46, 1996. ISSN 01678760. doi: 10.1016/S0167-8760(96)00066-9.

[233] Ole Jensen, Bart Gips, Til Ole Bergmann, and Mathilde Bonnefond. Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in Neurosciences*, 37(7):357–369, 2014. ISSN 1878108X. doi: 10.1016/j.tins.2014.04.001. URL `http://dx.doi.org/10.1016/j.tins.2014.04.001`.

[234] M S Worden, J J Foxe, N Wang, and G V Simpson. Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(6):RC63, mar 2000. ISSN 1529-2401. URL `http://www.ncbi.nlm.nih.gov/pubmed/10704517`.

[235] Tonia A. Rihs, Christoph M. Michel, and Gregor Thut. Mechanisms of selective inhibition in visual spatial attention are indexed by α-band EEG synchronization. *European Journal of Neuroscience*, 25(2):603–610, jan 2007. ISSN 0953816X. doi: 10.1111/j.1460-9568.2007.

05278.x. URL http://www.ncbi.nlm.nih.gov/pubmed/17284203http://doi.wiley.com/10.1111/j.1460-9568.2007.05278.x.

[236] John J. Foxe and Adam C. Snyder. The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers in Psychology*, 2(JUL):1–13, 2011. ISSN 16641078. doi: 10.3389/fpsyg.2011.00154.

[237] Ole Jensen, Jack Gelfand, John Kounios, and John E Lisman. Oscillations in the alpha band (9-12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral cortex (New York, N.Y. : 1991)*, 12 (8):877–82, 2002. ISSN 1047-3211. URL http://www.ncbi.nlm.nih.gov/pubmed/12122036.

[238] Anil M. Tuladhar, Niels Ter Huurne, Jan Mathijs Schoffelen, Eric Maris, Robert Oostenveld, and Ole Jensen. Parieto-occipital sources account for the increase in alpha activity with working memory load. *Human Brain Mapping*, 28(8):785–792, aug 2007. ISSN 10659471. doi: 10.1002/hbm.20306. URL http://doi.wiley.com/10.1002/hbm.20306.

[239] Frédéric Roux and Peter J. Uhlhaas. Working memory and neural oscillations: Alpha-gamma versus theta-gamma codes for distinct WM information? *Trends in Cognitive Sciences*, 18(1):16–25, 2014. ISSN 13646613. doi: 10.1016/j.tics.2013.10.010.

[240] Ole Jensen and Ali Mazaheri. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in human neuroscience*, 4 (November):186, 2010. ISSN 1662-5161. doi: 10.3389/fnhum.2010.00186. URL http://www.ncbi.nlm.nih.gov/pubmed/21119777{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2990626.

[241] K. E. Mathewson, G. Gratton, M. Fabiani, D. M. Beck, and T. Ro. To See or Not to See: Prestimulus Alpha Phase Predicts Visual Awareness. *Journal of Neuroscience*, 29(9):2725–2732, mar 2009. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3963-08.2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19261866http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2724892http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3963-08.2009.

[242] N. A. Busch, J. Dubois, and R. VanRullen. The Phase of Ongoing EEG Oscillations Predicts Visual Perception. *Journal of Neuroscience*, 29(24):7869–7876, jun 2009. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0113-09.2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19535598http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0113-09.2009.

[243] Ole Jensen, Mathilde Bonnefond, and Rufin VanRullen. An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends in Cognitive Sciences*, 16(4):200–206, 2012. ISSN 1879307X. doi: 10.1016/j.tics.2012.03.002. URL http://dx.doi.org/10.1016/j.tics.2012.03.002.

[244] S. Haegens, V. Nacher, R. Luna, R. Romo, and O. Jensen. Alpha Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences*, 108(48):19377–19382, nov 2011. ISSN 0027-8424. doi: 10.1073/pnas.1117190108. URL http://www.ncbi.nlm.nih.gov/pubmed/22084106http://www.

pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3228466http://www.pnas.org/cgi/doi/10.1073/pnas.1117190108.

[245] Bradley Voytek, R.T. Canolty, Avgusta Shestyuk, N.E. Crone, Josef Parvizi, and R.T. Knight. Shifts in gamma phase-amplitude coupling frequency from theta to alpha over posterior cortex during visual tasks. *Frontiers in Human Neuroscience*, 4(191):1–9, oct 2010. ISSN 13882457. doi: 10.1016/S1388-2457(10)60372-0. URL http://journal.frontiersin.org/article/10.3389/fnhum.2010.00191/abstracthttps://linkinghub.elsevier.com/retrieve/pii/S1388245710603720.

[246] Thomas Miconi and Rufin VanRullen. The Gamma Slideshow: Object-Based Perceptual Cycles in a Model of the Visual Cortex. *Frontiers in Human Neuroscience*, 4(November):1–11, 2010. doi: 10.3389/fnhum.2010.00205.

[247] John E. Lisman and Ole Jensen. The Theta-Gamma Neural Code. *Neuron*, 77(6):1002–1016, 2013. ISSN 08966273. doi: 10.1016/j.neuron.2013.03.007. URL http://dx.doi.org/10.1016/j.neuron.2013.03.007.

[248] J J Chrobak and G Buzsáki. Gamma oscillations in the entorhinal cortex of the freely behaving rat. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 18(1):388–98, jan 1998. ISSN 0270-6474.

[249] Gennady G. Knyazev, Jaroslav Y. Slobodskoj-Plusnin, Andrey V. Bocharov, and Liudmila V. Pylkova. The default mode network and EEG alpha oscillations: An independent component analysis. *Brain Research*, 1402: 67–79, 2011. ISSN 00068993. doi: 10.1016/j.brainres.2011.05.052. URL http://dx.doi.org/10.1016/j.brainres.2011.05.052.

[250] Raphael Kaplan, Mohit H. Adhikari, Rikkert Hindriks, Dante Mantini, Yusuke Murayama, Nikos K. Logothetis, and Gustavo Deco. Hippocampal Sharp-Wave Ripples Influence Selective Activation of the Default Mode Network. *Current Biology*, 26(5):686–691, 2016. ISSN 09609822. doi: 10.1016/j.cub.2016.01.017. URL http://dx.doi.org/10.1016/j.cub.2016.01.017.

[251] Adam Johnson and A David Redish. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(45): 12176–89, 2007. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.3761-07.2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17989284.

[252] David J. Foster and Matthew A. Wilson. Hippocampal Theta Sequences. *Hippocampus*, 17:1093–1099, 2007. doi: 10.1002/hipo.

[253] Laurenz Muessig, Michal Lasek, Isabella Varsavsky, Francesca Cacucci, and Thomas Joseph Wills. Coordinated Emergence of Hippocampal Replay and Theta Sequences during Post-natal Development. *Current Biology*, 29(5):834–840.e4, mar 2019. ISSN 09609822. doi: 10.1016/j.cub.2019.01.005. URL https://www.sciencedirect.com/science/article/pii/S0960982219300065?via{%}3Dihub{#}bib12.

[254] Andrew M. Wikenheiser and A. David Redish. Hippocampal theta sequences reflect current goals. *Nature Neuroscience*, 18(2):289–294, 2015. ISSN 15461726. doi: 10.1038/nn.3909.

[255] Adam Johnson, Matthijs AA van der Meer, and A David Redish. Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology*, 17(6):692–697, dec 2007. ISSN 09594388. doi: 10.1016/j.conb.2008.01.003. URL `http://www.ncbi.nlm.nih.gov/pubmed/18313289http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3774291https://linkinghub.elsevier.com/retrieve/pii/S0959438808000056`.

[256] Barry D Van Veen, Wim Van Drongelen, and Akifumi Suzuki. Localization of Brain Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering. *IEEE Transactions on Biomedical Engineering*, 44(9):867–880, 1997. doi: 10.1109/10.623056.

[257] Mark Woolrich, Laurence Hunt, Adrian Groves, and Gareth Barnes. MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. *NeuroImage*, 57(4):1466–1479, 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.04.041. URL `http://dx.doi.org/10.1016/j.neuroimage.2011.04.041`.

[258] G. L. Colclough, M. J. Brookes, S. M. Smith, and M. W. Woolrich. A symmetric multivariate leakage correction for MEG connectomes. *NeuroImage*, 117:439–448, 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.03.071. URL `http://dx.doi.org/10.1016/j.neuroimage.2015.03.071`.

[259] Roberto D Pascual-Marqui, Rolando J Biscay, Jorge Bosch-Bayard, Pascal Faber, Toshihiko Kinoshita, Kieko Kochi, Patricia Milz, Keiichiro Nishida, and Masafumi Yoshimura. Innovations orthogonalization: a solution to the major pitfalls of EEG/MEG "leakage correction". *bioRxiv*, page 178657, jan 2017. doi: 10.1101/178657. URL `http://biorxiv.org/content/early/2017/08/23/178657.abstract`.

[260] Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Andrew J Quinn, Fidel Alfaro-Almagro, Stephen M Smith, and Mark W Woolrich. Discovering dynamic brain networks from big data in rest and task, 2018. ISSN 10959572. URL `http://dx.doi.org/10.1016/j.neuroimage.2017.06.077`.

[261] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, 9(6): e98679, jun 2014. URL `https://doi.org/10.1371/journal.pone.0098679`.

[262] Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008. ISSN 17425468. doi: 10.1088/1742-5468/2008/10/P10008.

[263] Mark M. Churchland, Byron M. Yu, John P. Cunningham, Leo P. Sugrue, Marlene R. Cohen, Greg S. Corrado, William T. Newsome, Andrew M. Clark, Paymon Hosseini, Benjamin B. Scott, David C. Bradley, Matthew A. Smith, Adam Kohn, J. Anthony Movshon, Katherine M. Armstrong, Tirin Moore, Steve W. Chang, Lawrence H. Snyder, Stephen G. Lisberger, Nicholas J. Priebe, Ian M. Finn, David Ferster, Stephen I. Ryu, Gopal Santhanam, Maneesh Sahani, and Krishna V. Shenoy. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13 (3):369–378, 2010. ISSN 1097-6256. doi: 10.1038/nn.2501. URL `http://dx.doi.org/10.1038/nn.2501`.

[264] N. K. Logothetis, O. Eschenko, Y. Murayama, M. Augath, T. Steudel, H. C. Evrard, M. Besserve, and A. Oeltermann. Hippocampal-cortical interaction during periods of subcortical silence. *Nature*, 491(7425):547–553, 2012. ISSN 00280836. doi: 10.1038/nature11618. URL `http://dx.doi.org/10.1038/nature11618`.

[265] Gyorgy Buzsaki. Memory consolidation during sleep: a neurophysiological perspective. *Journal of sleep research*, 7 Suppl 1:17–23, 1998. URL `http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2869.7.s1.3.x/full`.

[266] Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature reviews. Neuroscience*, 11(2):114–26, 2010. ISSN 1471-0048. doi: 10.1038/nrn2762. URL `http://www.ncbi.nlm.nih.gov/pubmed/20046194`.

[267] Matthew J Beal. *Variational algorithms for approximate bayesian inference*. PhD thesis, University of London, 2003.

[268] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2005. ISBN 9780521642989. doi: 10.1198/jasa.2005.s54.

[269] Vasile V Moca and Raul C Mureşan. Discriminating legitimate oscillations from broadband transients. *BMC Neuroscience*, 14(S1):P286, jul 2013. ISSN 1471-2202. doi: 10.1186/1471-2202-14-s1-p286. URL `https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-14-S1-P286`.

[270] Rodrigo Montefusco-Siegmund, Pedro E. Maldonado, and Christ Devia. Effects of ocular artifact removal through ICA decomposition on EEG phase. In *International IEEE/EMBS Conference on Neural Engineering, NER*, pages 1374–1377. IEEE, nov 2013. ISBN 9781467319690. doi: 10.1109/NER.2013.6696198. URL `http://ieeexplore.ieee.org/document/6696198/`.