State-Switching Models of Human Brain Activity Using Recurrent Neural Networks

Alexander Skates

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

Abstract

It has been shown that spatiotemporal dynamics of neuronal activity can be well described using state-related behaviour, comprising a discrete set of reoccurring quasi-stable states associated with distinct patterns of spatial and functional connectivity. Most methods of analysis will either assume stationarity of these states, as in ICA; or constrain the dynamics to be Markovian, as in hidden Markov models (HMMs). These tools lack the capability to explicitly model the higher order temporal dependencies that can occur over timescales of various scales.

In this thesis, we introduce a model that combines probabilistic state-space models with recurrent neural networks (RNNs), enabling us to relax the Markovian constraint of HMMs and learn temporal features of the data occurring over longer timescales. The model takes the form of a recurrent state-switching network, which models the uncertainty in time-varying state labels via discrete random variables. We introduce a variational Bayesian framework for computationally efficient inference of the model that also generalises to a variety of time series models.

Using simulations, data taken from the resting state magnetoencephalography (MEG) scans of 55 participants, and data taken from the MEG recordings of a face-viewing task undertaken by 19 participants, we demonstrate that we can reliably infer a set of states that fits the data better than where the Markovian constraint is enforced, however we do not see significantly different temporal behaviour emerging. We additionally demonstrate that unlike the Markovian model, the recurrent model can internally represent the temporal dynamics of the data.

State-Switching Models of Human Brain Activity Using Recurrent Neural Networks

Alexander Skates

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

ABSTRACT

It has been shown that spatiotemporal dynamics of neuronal activity can be well described using state-related behaviour, comprising a discrete set of reoccurring quasi-stable states associated with distinct patterns of spatial and functional connectivity. Most methods of analysis will either assume stationarity of these states, as in ICA; or constrain the dynamics to be Markovian, as in hidden Markov models (HMMs). These tools lack the capability to explicitly model the higher order temporal dependencies that can occur over timescales of various scales.

In this thesis, we introduce a model that combines probabilistic state-space models with recurrent neural networks (RNNs), enabling us to relax the Markovian constraint of HMMs and learn temporal features of the data occurring over longer timescales. The model takes the form of a recurrent state-switching network, which models the uncertainty in time-varying state labels via discrete random variables. We introduce a variational Bayesian framework for computationally efficient inference of the model that also generalises to a variety of time series models.

Using simulations, data taken from the resting state magnetoencephalography (MEG) scans of 55 participants, and data taken from the MEG recordings of a face-viewing task undertaken by 19 participants, we demonstrate that we can reliably infer a set of states that fits the data better than where the Markovian constraint is enforced, however we do not see significantly different temporal behaviour emerging. We additionally demonstrate that unlike the Markovian

model, the recurrent model can internally represent the temporal dynamics of the data.

ACKNOWLEDGMENTS

I consider myself very fortunate to have had the opportunity to pursue a DPhil under the supervision of Professor Mark Woolrich. The rigour with which he insists upon ensures that even offhand discussions end up wholly enlightening, and I have lost count of the number of times I've ended a "casual conversation" with him by having to take a photograph of his whiteboard. His passion for his work, and by extension my work, has given me a great deal of encouragement throughout the last three and a half years, and I'm not sure I could have finished without his help. I must also extend my thanks and gratitude to my other two supervisors, Professor Steve Smith, and Professor Markus von Kienlin. Steve, we didn't meet often, but every time we did I was always struck by your clarity of thought, and how you could immediately identify issues that I had been struggling with for weeks. Markus, I must thank you for your kindness and patience; I know my topic ended up somewhat tangential to the original aims of my original programme, but you nonetheless were wholly supportive and the engaged every time we spoke.

There are many more people to whom I owe a great deal of credit for making it through the last four years. All of the people in the OHBA Methods Group have provided abundant advice, inspiration, and constructive criticism, but in particular I owe a great technical debt to Andrew and Diego. It is also necessary to specifically thank Ryan for many crucial discussions in the implementation of our work. I have been lucky enough to work in an incredible building, full of interesting and intelligent people, all of whom have made coming in to work at OHBA a great experience. It's been a pleasure working with

you all, and I hope OHBA continues to crush the competition in all future sports days. Not many can legitimately claim to be the best, but we have the egg-and-spoon race, wheelbarrow race, and tug-of-war victories to prove it. To my housemates, for keeping me sane, letting me vent, and occasionally baking delicious cakes — especially Susan and Joe (who is luckily just as bad at squash as I am). My family has also been incredibly supportive towards my decision to go back to school. I think I'm done with education... for now at least.

I owe perhaps the greatest debt to my wonderful partner Emma, who has been utterly integral to my mental health over the duration of this degree. She is a beacon of patience, compassion, and good humour, and more importantly has taught me that good food is the solution to most of life's problems, or at least reinforced my pre-existing belief that it is.

I could not have ended up doing this research without the opportunity provided by the SABS DTC, nor without financial support from Roche, the EPSRC, and the MRC. It almost seems crazy that they would give me money to spend all this time thinking about such cool things.

Finally, I would be remiss not to mention the many pseudo-anonymous contributors to Stack Overflow, who taught me that no matter how obscure your problem, someone else has probably encountered it before. Though whether they themselves found a solution is perhaps another matter.

CONTENTS

ACRONYMS

BOLD Blood-oxygen-level dependent

CNN Convolutional neural network

CFC Cross-frequency coupling

CDF Cumulative distribution function

DMN Default mode network

EEG Encephalography

FC Functional connectivity

fMRI Functional magnetic resonance imaging

GRU Gated recurrent unit

GLM Generalised linear model

HPI Head position indicator

HMM Hidden Markov model

HSMM Hidden semi-Markov model

HOHMM Higher-order hidden Markov model

ICA Independent component analysis

KL Kullback-Leibler

LSTM Long-short term memory

MEG Magnetoencephalography

MAR Multivariate auto-regressive

RNN Recurrent neural network

RSN Resting state network

sICA Spatial independent component analysis

SGD Stochastic gradient descent

SQUID Superconducting quantum interference device

tICA Temporal independent component analysis

Part I

THESIS

1

INTRODUCTION

The human brain is a fantastically complex organ, comprised of over 100 billion neurons, each of which may make up to as many as 1000 synaptic connections. Somehow, this tangled web of interconnected processing units gives rise to perception, learning, thinking. Understanding the complex organisation of the brain at both micro- and macroscopic scales from neuroimaging data represents a key challenge for systems neuroscience, and one that has already uncovered some key insights into the ways in which the brain operates.

To this end, there exist a number of modalities in which we can measure (non-invasively) the activity of the brain. This chapter will briefly review one of the main techniques that have been utilised to record this activity, magnetoencephalography (MEG). We will also touch on some others, such as functional magnetic resonance imaging (fMRI) and encephalography (EEG), but we have a particular emphasis on MEG as this is the modality that the data analysed within this thesis has been recorded using.

1.1 A BRIEF PRIMER ON MEG

The basis for communication between neurons is both electrical and chemical in nature; *action potentials* are electrical signals propagated rapidly along neurons, which is then converted to a chemical message to be passed to other neurons across *synapses*. It is the magnetic fields associated with post-synaptic potentials that an MEG system is able

to detect. MEG is not sensitive to activity on the level of individual neurons, but rather necessitates neural assemblies comprising between 10,000 to 50,000 neurons [84]. The total post-synaptic potentials of a population of neurons is equivalent to the sum of potentials of the constituent neurons. It is therefore important that in order to measure the phenomena at a distance, the responsible neurons must be activated in a coherent and synchronous manner. These assemblies form macroscopic equivalent current dipoles that are able to pass unencumbered through the skull and are measurable at the scalp through the use of MEG [49]. For this reason, MEG is most sensitive to currents located in cortical regions close to the skull and sensitivity falls away with depth from the sensors [54]. However, MEG is not entirely blind to deeper activity, as experiments and simulations show [2, 24], though localisation accuracy is likely to be much poorer in comparison to superficial sources.

The magnetic fields elicited by neuronal activity are exceedingly small (from 10^{-15} - 10^{-11} tesla, or T), while background noise generated by external electrical and magnetic equipment can be much larger by comparison (around 10^{-7} T). Specialist equipment is therefore required to measure these magnetic fields, called Superconducting Quantum Interference Devices (SQUIDs), which combine high sensitivity with the ability to reject external noise. State-of-the-art MEG systems include around 300 such sensors contained within a cryogenic vessel called a dewar containing liquid helium. The system is additionally located within a magnetically shielded room such that environmental noise is reduced as much as possible. Additional suppression of noise is achieved with the use of gradiometers, which measure the magnetic field gradient instead of the magnetic field. As the magnetic field strength of a current dipole follows the inverse

square law, the gradient near a source is large, whereas a more distant source has a much lower gradient.

In order to accurately estimate the spatial locations of discrete neural sources, it is necessary to first solve the *forward problem*, which involves determining the associated magnetic field pattern that can be generated by a known distribution of current dipoles in the brain. This problem is relatively straightforward, as we can treat the entire head space as homogenous (in contrast to EEG where the conductive properties of the skull mean electric potential are distorted) and we can therefore model these dipole patterns in relatively straightforward manner. The reverse is not quite so simple; given a measured magnetic field distribution across sensors, reconstructing where the current sources are localised within the brain has no unique solution. This problem is known as the *inverse problem*. It is solved by introducing constraints on the spatial distributions of the current dipoles to exclude all but the most suitable solution.

There are a number of solutions that attempt to solve this problem, for example minimum norm estimation, which deals with the non-uniqueness of the inverse problem by introducing a spatial prior in the form of a source covariance matrix that favours solutions that minimise the L2 norm of the source estimate, simultaneously balancing fitting the data and minimising the contributions of noise [50]. Beamforming is another common source estimation approach. At each point in a predefined grid, a narrow-band spatial filter estimates the contribution of the source model while suppressing contributions from other brain regions. Beamformers make an assumption that data is generated from a discrete set of current dipoles, and there is no high temporal correlation between any of these dipoles — in this case they are interpreted by the beamformers as emerging from a single source [54]. It is therefore an important caveat that any source reconstructed

data is unlikely to be a *true* representation of the neural activity underlying the sensor recordings, however it is likely to be at least a good approximation.

1.1.1 *Other modes of acquisition*

Contrary to the uncertainty obtained through MEG, fMRI offers unparalleled whole-brain spatial resolution. Unlike MEG, which provides a direct measure of brain activity, fMRI provides a method to demonstrate changes within brain metabolism [87]. Essentially, fMRI uses blood oxygenation as a proxy for neuronal activation, where oxygen usage is increased in areas of the brain that are in use. The measure is referred to as the blood-oxygen-level dependent (BOLD) contrast. fMRI allows the characterisation of the spatial dynamics of brain activity with sub-millimetre precision, however the latency and temporal blurring of the haemodynamic response results in poor temporal resolution, on the order of seconds. PET also measures the haemodynamic changes induced by neuronal activity, and again suffers from the same issue of poor temporal resolution.

In EEG, measurements of the electrical currents in the brain are inferred through extracranial measurement of electric potential differences [8]. As with MEG, the neural currents must be estimated by projecting the extracranial voltage recordings onto the brain — this indirect measure of activity results in poor spatial resolution, but as with MEG, the recordings are limited only by the sampling rate, and therefore millisecond resolution can be achieved. While each of these modalities has advantages and disadvantages that come part-in-parcel with non-invasive imaging, it is increasingly seen that these limitations can be overcome to some degree by combining multiple modalities, though these can also come with non-trivial technical challenges [61].

1.2 SPONTANEOUS NEURAL OSCILLATIONS

The post-synaptic potentials that convey information to the brain exhibit a repetitive temporal structure, formed of spikes, or bursts of spikes, called oscillations [76]. The oscillatory activity of the human brain is dominated by a number of distinct frequency ranges and spatial patterns, and the frequency content and rhythm of the oscillations vary as a function of a given subject's mind state, attention, and behaviour. Broadly, this oscillatory activity is categorised into one of five frequency bands: delta (0.2-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-31 Hz), and gamma (32-90 Hz). [81]. Any kind of task stimuli can elicit both *evoked* and *induced* activity in these oscillations, where evoked signals are tightly time and phase-locked to a stimulus onset (and visualised by averaging responses to stimuli, time-locked to stimulus onsets), and induced signals are sustained over the course of stimulus presentation, however they are not time or phase-locked. The *total activity* is the combined evoked and induced activity [52].

In contrast to previous beliefs that spontaneous oscillations observed at rest were simply noise, they have become popular for the study of functional connectivity, both in the context of resting state analysis, and more traditional task paradigms.

Neural oscillations of different frequencies can also interact with one another [65]. This interaction is referred to as cross-frequency coupling (CFC), and exists in a number of types of interactions, for example phase-amplitude coupling, where the amplitude of high-frequency oscillations is modulated by the phase of low-frequency oscillations, or phase-phase coupling where there is amplitude-independent phase locking between n cycles of high frequency oscillations and m cycles of low frequency oscillations (which account for the name $n : m$ phase synchrony) [90]. Recent studies suggest that CFC may well

serve a functional role. For a comprehensive review on the role of cross-frequency coupling, refer to [19].

1.3 THE FUNCTIONAL ANALYSIS OF REST

In the seminal rs-fMRI study by Biswal et al. [10], Biswal studied how different regions of the brain communicated in the absence of a task, focusing on low frequency fluctuations (< 0.1 Hz) in the BOLD signal, and found that the left somatosensory cortex was highly correlated with homologous areas in the contralateral hemisphere. Unfortunately, his studies were initially disregarded by the scientific community at large and attributed to another signal source, however they are now widely replicated and accepted as a valid way to map functional brain networks. rs-fMRI in particular has seen a huge amount of interest. The study of resting state has the advantage over a task-based paradigm due to its ease of acquisition, and the minimal effort it requires from the patients. Contrary to a task-based study, wherein subjects are instructed to undertake specific tasks to identify regions that are functionally involved in that task, resting state is acquired in the absence of a stimulus or task and can therefore be utilised for patients who may struggle with task instructions; as in the case of paediatric patients or patients suffering from neurologic or psychiatric conditions [111]. Despite it's overwhelming popularity in fMRI, there are nonetheless an increasing amount of MEG-based studies examining the resting state, e.g. [15, 86]. Resting state MEG particularly complements rs-fMRI owing to the high temporal and spectral resolution which is unavailable in fMRI, and the fact that it is a more direct neuronal measure, free from confounds related to the vascular response (which could be affected for example in disease).

1.4 THESIS OVERVIEW

The aim of this thesis is to exploit recent advances in the field of machine learning, combined with more traditional probabilistic state-space modelling, to obtain a framework for which we can construct and infer upon models capable of learning complex spatiotemporal dynamics. We would like to particularly use these models to describe the interactions between different cortical regions, as recorded using a MEG scanner, into functional connectivity networks and to characterise the behaviour of such networks over time. It is hoped that the model might reveal interesting temporal dynamics of neural connectivity, but also that the framework outlined here might be applied to questions and models beyond what we have outlined here.

The thesis is broadly structured into two sections; first we aim to provide some background and define our model and the inference techniques and the novel algorithm we shall use to train the model.

- **Chapter 2** reviews what exactly functional connectivity is. We explain the two main approaches to analysing functional connectivity; statically and dynamically. We then review the most widely-used approaches to analysing such types of functional connectivity, including the HMM, which provides the primary methodological motivation for this work. The chapter concludes by briefly defining deep learning and describes gated recurrent units, a particular architecture of artificial neural networks.

- **Chapter 3** introduces the model we will use. We explain the motivation for both the choice of temporal model and the spatial model. We explain how we are able to make use of a multivariate Normal distribution to directly encode the functional connectivity structure of each network we are attempting to learn.

- **Chapter 4** reviews the ideas behind variational Bayesian inference, and stochastic gradient variational Bayes. We explain how to amortise inference such that a neural network can be trained to learn a mapping from the data space to some latent space, and the specific tricks that must be employed to train such a network. Finally, we outline how we can construct our inference framework from these constituent pieces.

Having outlined the technical details of the method, we will first evaluate its ability to model various spatiotemporal dynamics through training it upon a number of types of simulated data, and then demonstrate its applicability to both resting state MEG recordings as well as task data.

- **Chapter 5** applies the model to a number of simulations, both Markovian and non-Markovian in nature. We compare the results to that of a HMM and demonstrate comparable performance upon both Markovian and non-Markovian simulated sequences.

- **Chapter 6** makes use of the model to learn a set of spatial networks in the resting state MEG scans of 55 subjects. We test the robustness of the model, and compare the inferred temporal dynamics learned by our recurrent model to those inferred by a HMM.

- **Chapter 7** involves applying the model to task data, where 19 subjects were recorded while performing a face viewing task. We should that we can learn a set of task-relevant connectivity networks, and that the corresponding activations of a number of these networks are significantly associated with specific periods of the task. We again make comparisons between the recurrent model and the HMM.

Finally, in **Chapter 8**, we will present some concluding remarks, as well as a brief outline of potential future work.

2

CONNECTIVITY, FAST AND SLOW

Modern ideas about the characterisation of human brain function find their inception in the early 19th century. They can be thought specifically to have roots in early attempts to refute the field of phrenology, a theory of the mind postulated by the German physician Franz Joseph Gall, based on the idea that the brain is comprised of a series of distinct mental "organs", each of which associated with specific traits that together determine human personality. While phrenology was ultimately deemed unscientific and invalid by a scientific committee of the Athénée at Paris [1] and is still known to be a pseudo-science to this day, the notion that distinct brain functions could be localised still persists.

The idea first entered the scientific mainstream through the efforts of Paul Broca [14], who noticed that a patient with focal brain lesions in the frontal lobe showed highly specific impairments, notably forms of aphasia (a speech and language disorder that can result in the loss of ability to talk, read, and/or write), in this case an inability to say anything except for the word"tan". He declared to the Anthropological Society in Paris that the left frontal lobe was therefore the seat of speech. Shortly after, a similar connection was made in a case of aphasia by Carl Wernicke [124], this time owing to lesions within the temporal and parietal lobes. While the concept of *functional localisation* remained fiercely debated in subsequent years, it was generally agreed that at least some basic sensory and motor functions were localised in

1 The fact that Gall gave Napoleon Bonaparte a rather unflattering examination of the skull may well have contributed to the conclusions reached by the committee...

specific brain regions. However, the questions still remained: exactly how specialised were these regions of the brain, and exactly what functions were carried out in those specialised regions — was it only basic sensory and motor functions, or did it also apply to higher level cognitive functions?

With regards to the first question, we find that this is a matter of degrees, rather than simply all or none, and we arrive at the concept of *functional segregation*. The general idea of functional segregation is that a given cortical area is specialised for a particular aspect of cognitive processing, distributed across anatomically segregated regions in the cortex. A given function may then be supported by a complex infrastructure involving a number of specialised areas *functionally integrated* at a variety of levels [114].

As to the second question, we find that in fact a myriad of functions may take place within a single brain region. Contrary to the declaration of Broca that the left frontal lobe was specialised for aspects of speech, recent neuroimaging research has suggested that certain areas of the brain that have been associated with language processing appear to be recruited as well in other cognitive domains, suggesting that language may draw on some set of neural mechanisms shared across other functional specialisations [46], and language is not an isolated example in this aspect. It is evident that the concepts of functional segregation and functional integration are intimately linked, and the state of affairs is somewhat more complex than Broca believed.

2.1 FUNCTIONAL CONNECTIVITY

Throughout much of the history of neuroimaging, the field has primarily concerned itself with the topic of functional segregation, while functional integration has been more difficult to assess. Since the end

of the 20th century, we have nonetheless seen a shift in emphasis of studies in the field towards functional integration, in particular the concept of connectivity. This idea takes place on a massive scale: the number of connections within the brain sits at well over 100 trillion by many estimates. These connections provide a scaffold for neural structures to become transiently sychronised or functionally connected [17]. Such a complex network of interacting components facilitates the emergence of dynamically coupled neuronal assemblies, acting as the basic substrate for cognitive functions; perception, thinking, learning — all of which require precise integration of neural activity at highly specific spatiotemporal scales [116]. This synchronisation occurs over a range of times scales, and over both short and long intercerebral distances.

There are two main types of connectivity we can consider: functional, and effective. *Functional connectivity* focuses specifically on correlations between spatially remote neurophysiological events [43], where a deviation from statistical independence between neural areas implies functional cooperation. However, functional connectivity refers only to statistical dependencies between neural regions, and it does not incorporate any knowledge or assumptions about the structures or mechanisms of the neural structures in question. In order to gain a more causal insight, it is necessary to instead consider *effective connectivity*. Effective connectivity refers to the direct or indirect influence that one neural system exerts over another, either at a cortical level, or a synaptic level. Conceptually, we can therefore infer effective connectivity using models that describe causal links and that are based on proven measures of functional connectivity [60]. We are primarily concerned here with functional connectivity, as it underpins the analysis of neural interactions in functional neuroimaging data.

Studies of functional connectivity have taken place in a wide variety of modalities: PET, fMRI, EEG/MEG — and at a variety of levels: the activity of individual neurons, neural assemblies, or the activity of the whole brain. Non-invasive techniques have largely proved to be the most useful methods for obtaining the mapping of functional connections within the whole brain, primarily due to the fact that connectivity can be so widespread. The first such mappings were acquired using PET [41, 42], however were limited by the short half life of the radio-label used, ^{15}O (2.05 minutes). Soon after, the first fMRI study of functional connectivity was made by a graduate student at the Medical College of Wisconsin [10] in what is now recognised as a seminal paper, paving the way for innumerable other subsequent studies. fMRI provided improved spatial and temporal resolution over PET, which was reflected in the signal. By correlating all voxels in the brain with a single voxel in the sensorimotor cortex (an approach now commonplace, and referred to as seed-based correlation analysis), Biswal et al. found that even in the absence of a task they were able to see correlations between the left and right sensorimotor cortices. While this result received little attention at the time, this method of placing seeds in various areas of the brain soon after received widespread attention after revealing a previously-unknown brain network that appeared to play a key role in baseline, or default, behaviour of the brain [97]. Other such large-scale networks of correlated brain regions were soon after discovered during rest, corresponding to a variety of functions including vision, auditory, and working memory [6, 27, 38, 39, 109]. These networks exhibited the same functional structure and organisation as seen during tasks, and this connectivity structure proves to be robust even under anaesthesia or during sleep [104, 121]. Owing to their occurrence during rest, these networks are referred to as *resting state networks* (RSNs).

2.2 FUNCTIONAL CONNECTIVITY ANALYSIS

Approaches to studying functional connectivity lie along a spectrum of temporal resolution; on one end, static models make the assumption that connectivity between regions is constant over some arbitrarily long window, and the other end, dynamic connectivity models may reveal time-resolved connectivity at a timepoint-by-timepoint level (for instance, via sliding window approaches). Falling somewhere in the middle are models that make use of discrete, temporally contiguous brain states, each of which characterise a particular connectivity structure (e.g. HMMs or sliding windows + clustering). In these state-based models of connectivity, the dependence is typically assumed to be static *within states*, and the connectivity only changes upon transitioning to a new state.

What follows in this section will be a look at all three of these types of models of connectivity, the advantages and pitfalls of using such a model, and then outline some example methods for analysing connectivity in this way.

2.2.1 *Static connectivity*

Static measures of connectivity typically assume the functional connectivity in the brain remains constant over a period of a scanning session, or over the duration of a task. The primary 'classical' measures of functional connectivity include Pearson's correlation coefficient and cross-correlation. These are all linear methods and include an implicit assumption of Gaussianity, though there also exists some non-linear measures, primarily information theoretic in nature.

Pearson's correlation coefficient measures the linear correlation (in the time domain) between two signals $X = (x_t)_{t=1}^{T}$ and $Y =$

$(y_t)_{t=1}^{T}$. For demeaned and normalised signals with zero mean and unit variance, it is defined as:

$$R_{XY} = \frac{1}{T}\sum_{t=1}^{T} x_t \cdot y_t, \tag{2.1}$$

and R_{XY} takes values of between -1 and 1, where -1 denotes total inverse correlation between X and Y, 1 is total direct correlation between the two signals, and a value of 0 is linear independence.

The cross-correlation function measures the correlation between two signals as a function of the time-lag between the two signals:

$$C_{XY}(\tau) = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau} x_{(t+\tau)} \cdot y_t. \tag{2.2}$$

When $\tau = 0$, we can see that the cross-correlation is simply the Pearson's correlation coefficient. Like the Pearson's correlation coefficient, the cross-correlation takes values between -1 and 1, with the same interpretations of the values.

These are two of the most straightforward measures of functional connectivity, but there exist a multitude more; coherence, phase synchronisation, phase-slope index, and Granger causality, to name a few. Most of these measures are undirected, but Granger causality and related metrics are capable of providing estimates of directed connectivity for a single pair, include the directed influence of X on Y, as well as the directed influence of Y on X.

2.2.2 *Dynamic connectivity*

While it certainly cannot be contested that studies operating under the premise of static connectivity have afforded a wealth of information in the understanding of large-scale distributed brain function, ultimately such an analysis is only a temporal average of a very complex

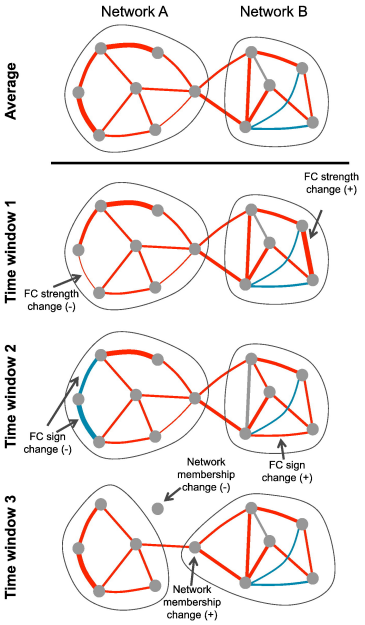

Figure 2.1: Illustration of functional connectivity. Each diagram shows possible changes in connectivity that can occur over a period of time. The connectivity can be altered in magnitude, as shown in the second row, the sign can change (as seen in the third row), or can even be lost entirely if the magnitude falls above or below a threshold, causing a membership change of each network, as seen in the final row. All of these changes are lost when averaging over time. Figure reproduced from [62].

spatial and temporal process. We know from empirical results that brain activity is largely dynamic and condition dependent [96], and therefore examining them as if they are invariant physiological and functional networks provides a useful, but possibly over-simplified characterisation of the brain's functional networks. Such networks have been demonstrated to change in the presence of tasks [34, 37, 40, 112] or physiological state changes such as sleep or sedation [47, 58, 59]. Additionally, it is not just inter-subject connectivity that is highly variable, but even within-subject connectivity has been shown to vary considerably, even across different scans in the same session [57, 77, 108].

To give a simple example; though of course a single neuron is spatially constrained in terms of it's connections, a typical voxel or parcel contains many millions of neurons. And though a cluster of

neurons within a given region may be connected to one region, there may be other clusters with strong connections to entirely different regions. If there is some periodicity of the two firing patterns between the different regions, the consideration of different time-scales or windows may show an entirely altered dominant set of connections between such sub-populations of neurons. A visual representation of some different ways in which averaging over time can misrepresent the underlying dynamics is shown in Figure 2.1. We can therefore obtain a greater understanding of the true connectivity by taking a more temporally-motivated approach to our analysis, rather than simply averaging out all of the connections.

Below, we review two popular strategies designed to characterise non-static spatiotemporal connectivities, ranging from the instantaneous reorganisations of FC that can be provided by sliding-window correlations, to the switching models of stationary connectivity described by hidden Markov models and other extensions.

2.2.2.1 *Temporal ICA*

Independent component analysis is a method for separating a signal into additive subcomponents, by assuming the subcomponents are non-Gaussian, and are maximally independent from each other. ICA can be applied to neural data, maximising independence in either space or in time; the former is referred to as 'spatial ICA' (sICA), while the latter is 'temporal ICA' (tICA). Spatial ICA has proved common in fMRI analysis, owing to the high spatial and relatively low temporal dimensionality of fMRI scans — it is also typically more robust, as it is thought that the underlying neural processes are more non-Gaussian in space than in time. However, for obtaining functionally independent functional networks, where no spatial independence is enforced (such that states can be spatially overlapping), temporal ICA

is ideal. Motivated by this precise reason, that a given brain region may be involved in multiple functional networks, Smith et al. made use of temporal ICA to identify a number of distinct networks they termed "temporal functional nodes" (TFMs) during resting state [110]. They found that these networks differed from networks identified using spatial ICA or seed-based correlation approaches. tICA has also been applied in MEG [15, 86] and EEG [21] in conjunction with source localisation to identify resting state network. Unlike fMRI, both MEG and EEG offer modalities well suited to the method due to their high temporal resolution and their typically high number of temporal samples.

2.2.2.2 *Sliding window approaches*

Sliding windows may well be considered to form the cornerstone of the dynamic functional connectivity analysis toolbox, and are the most widely-utilised strategy for estimating dynamic FC [1, 4, 20, 51, 63, 66, 68, 113, 127]. A sliding window analysis consists of calculating a given FC measure, for example the Pearson correlation coefficient, over consecutive windowed segments of data. The window has a fixed length, and is shifted in time by a fixed amounts (this amount varies from a single time point to the length of a complete window) such that there is a defined overlap between successive windows. This results in series of time-varying functional connectivity estimates, which can be then used to assess fluctuations in FC within sessions. This sliding window approach can be then used to detect transient and reproducible patterns of connectivity by way of clustering the resulting correlation/covariance matrices computed within each window. The centroids of each cluster can then be taken as the average state connectivity (Figure 2.3).

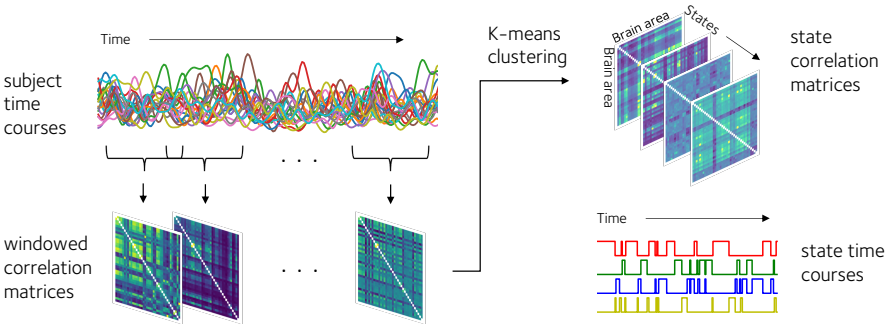

Figure 2.2: Detection of states containing similar functional connectivity using a sliding window approach. Correlation matrices are computed in windowed portions of a subjects scan, and are aggregated across all subjects. K-means clustering is applied to detect the centroids of each cluster of repeating connectivity, and then the corresponding time-courses for when each state is active can be derived from the corresponding cluster labels for each covariance matrix.

Despite it's popularity, the sliding window approach is not a panacea. It requires that certain parameters should be selected a priori: the size of the window, the window step size, and the weighting scheme for the data within each window (e.g. should the window be tapered?). The window size is a crucial parameter — there is a trade-off to be had here, whereby a smaller window size increases the capability of tracking faster temporal changes, at the cost of reducing statistical power and introducing sensitivity to noise/spurious fluctuations, while a longer window increases the robustness to noise but comes at the expense of averaging out the faster changes that occur [75, 105, 106]. Typical values for window sizes in fMRI are around 30-60 seconds [62], while MEG are on the order of 10 seconds [91]. The step size of the window is commonly selected as a single time point [1, 62]. There also exist a variety of window types that can be applied; while a number of studies simply employ a block rectangular window, there are also numerous such studies that make use of a weighted/tapering scheme, for example using Gaussian functions [92].

2.2.2.3 *Hidden Markov models*

While it is certainly possible to use sliding windows to obtain a discrete set of reproducible states corresponding to distinct patterns of functional connectivity, it is also possible to obtain such a description by taking a more principled, model-driven approach, via hidden Markov models (HMMs). HMMs assume that the time series data can be described by a fixed number of unknown brain states. Each state is characterised by a distinct set of parameters that describe a particular observation model, such as a multivariate Gaussian distribution [3]. In the case of the Gaussian distribution, the covariance structure of each state can be used to capture the state-specific connectivity patterns. Unlike the sliding window approach, we are able to obtain FC estimates from the data, weighted by the probability of the brain being in a given state at that time — therefore deriving FC in a far more efficient manner.

Hidden Markov models are a class of generative models that describe a system where the observable data $X = (x_t)_{t=1}^{T}$ (e.g. the MEG recordings) are governed by some unknown hidden process given by a sequence of states $Z = (z_t)_{t=1}^{T}$, of which the dynamics are described by a Markov chain[2] [9, 33]. The probabilistic model for a HMM with K states is given by the joint distribution

$$p(X, Z, \Theta) = p(z_1; \pi)\, p(x_1|z_1; \theta_x) \prod_{t=2}^{T} p(z_t|z_{(t-1)}; A)\, p(x_t|z_t; \theta_X), \quad (2.3)$$

where $\Theta = (\pi, A, \theta_x)$ are the parameters of the HMM. The temporal dynamics of the hidden state sequences are dictated by the initial probability distribution π, and a $K \times K$ matrix A called the transition probability matrix, where each row of the matrix $A_{k,\cdot}$ gives the probability

2 Whilst the brain is unlikely to be a truly Markovian system, the HMM has been applied extensively to many other problems that are not explicitly Markovian, such as speech recognition (for a comprehensive review, refer to [44]).

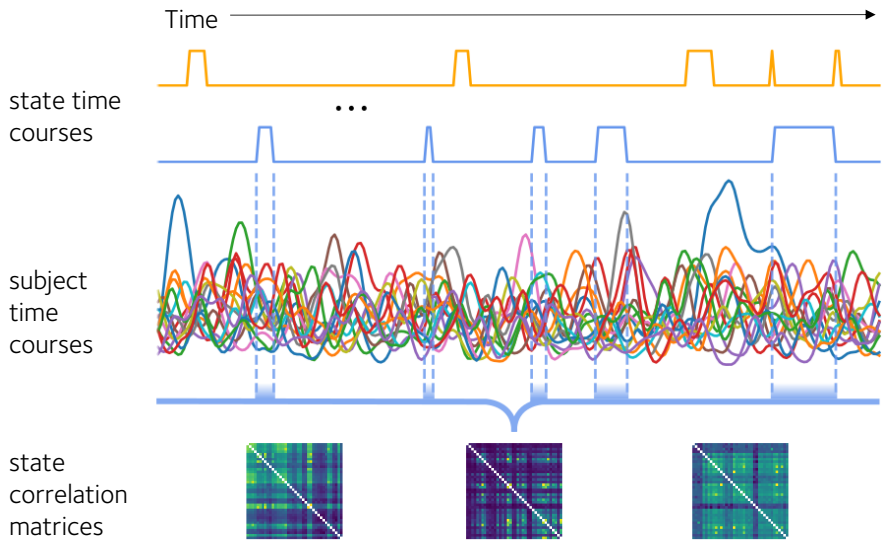

Figure 2.3: Detection of states containing similar functional connectivity using a HMM. The contribution of every point in time to each state is weighted by the probability of that state being active at that point in time. We can therefore efficiently use the entire session to compute each state's parameters. In the case of the multivariate Gaussian HMM, we learn the Covariance matrices, which directly encode the functional connectivity.

of transitioning from a given state $p(z_t|z_{(t-1)} == k)$. The observed dynamics are described by a set of distributions called the observation models, or *emission distributions*, with parameters $\theta_X = (\theta_1, \cdots, \theta_K)$, where θ_k is the parameter set for the k-th state emission distribution.

HMMs have been used in the context of functional connectivitiy on a number of occasions. Baker et al. [3] applied a HMM with zero-mean multivariate Gaussian observation models to broadband amplitude envelopes of source reconstructed MEG data. They identified a set of 8 transient states that were consistent with previously established resting state networks, bursting on time scales on the order of 100 ms — much faster than previously demonstrated results. HMMs have also been used to demonstrate the hierarchical nature of these states; Vidaurre et al. made use of a HMM to show that these states are organised into one of two metastates consistently across subjects, where transitions within a given metastate were far more likely to occur. Eavani et al [32] also made use of a HMM for resting state fMRI

data, using covariance matrices formed from a set of sparse rank-one basis matrices. Quinn et al. [95] also demonstrated that connectivity patterns in the presence of explicit cognitive tasks by using states with distinct power amplitudes and correlations (e.g. multivariate Gaussian observation models), and additionally with distinct spectral and cross-spectral properties.

While HMMs are adept at handling states occurring on very fast timescales, the Markov assumption does come with some limitations when it comes to modelling longer state durations. Notably, due to the structure of the transition matrix, state lifetimes (defined as the amount of time a state remains stable before transitioning to a new state) are distributed according to a geometric distribution, such that the probability of a given state lifetime decreases monotonically as the duration of this lifetime increases. While this does not mean that the empirical distribution of state lifetimes inferred from MEG or fMRI data is necessarily geometrically distributed, it does mean the model will place more weight on shorter lifetimes, which may not be appropriate in the context of functional connectivity.

To this end, the hidden semi-Markov model (HSMM) has emerged as a very interesting extension of the HMM, which is essentially a HMM except state lifetimes are no longer modelled by the transition probability matrix, but instead explicitly modelled using a distribution over state lifetimes, either parametric or nonparametric in form. Trujillo-Barreto et al. [115] made use of HSMM with a truncated log-Normal distribution for state lifetimes and a multivariate Normal observation distribution for analysing resting state EEG and reported higher state lifetimes than those seen by Baker et al. (on the order of 200-400 ms). Shappell et al. also make use of a HSMM in both task and resting-state fMRI [107]. They emphasise the importance of an appropriately specified lifetime distribution, and suggest where it

may be difficult to estimate (as in the case of resting state scans), a non-parametric distribution may be appropriate.

The dynamic programming algorithm at the core of HMM inference, the forwards-backwards algorithm, is also adaptable to a number of other extensions and variations upon the hidden Markov model, for example hierarchical hidden models [36] that explicitly model higher level states that might perform the same function as those identified by Vidaurre et al. in [119], or higher-order HMMs (HOHMMs) where a n-th order HMM's current state no longer depends on the state immediately preceding it, but all states going back n time-steps.

2.3 "DEEP" LEARNING

Hidden Markov models make use of the Markov property to facilitate computationally tractable learning of sequences, however this is by no means a requirement. Recent years have seen an explosion in the development of various artificial neural network architectures, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs). The term "deep" learning comes from the fact that such networks are typically multilayered, thus "deep". They have been shown capable of exponential gains in efficiency over more traditional machine learning models, i.e. given the same quantity of training data, deep neural networks can learn much more complex features than traditional linear/kernel methods [7, 74, 83]. They have also proven to be able to model highly complex natural sequences, such as voice, text, polyphonic music, or handwriting — therefore they should provide ideal tools for modelling neural activity. In particular, we are interested in building upon the foundations of the HMM for obtaining time-varying estimates of FC, so it stands to reason the most relevant tools are recurrent neural networks.

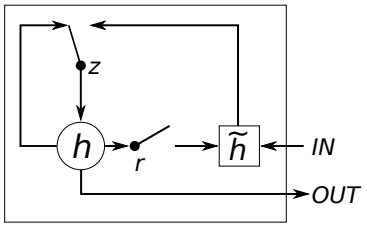

Figure 2.4: Graphical representation of 1 cell GRU unit. A GRU unit allows a network to store information by way of two gates; an update gate and a reset gate. The two gates z_t and r_t are nonlinear summation units (usually logistic sigmoids so that the gate remains between 0 and 1 — off or on) that control via multiplication whether to allow information to persist, or how to combine the new input with previous values. The activation h_t of the GRU is a linear interpolation of the previous activation $h_{(t-1)}$ and the candidate activation $\tilde{h}_t$. Image taken from [22].

2.3.0.1 *Recurrent neural networks*

Recurrent neural networks are a class of artificial neural networks that contain cyclic connections, enabling them to make use of sequential data. Theoretically, this leads to the persistence of information within the RNN – a "memory", of sorts. In practice, the influence of any given input on the output of a network either decays to zero or becomes exponentially large with increasing iterations of the network; an issue known as the vanishing gradient problem.

The addition of "gates" provide one way of addressing this issue, as in the case of *Gated Recurrent Units* (GRUs) or *Long-Short Term Memory* (LSTM) cells. With the addition of multiplicative "gates" that control access to a vector of memory cells, instructing them when to write, delete and output information, the GRU and LSTM cells are able to mitigate the effects of the vanishing gradient problem [22, 55].

The GRU architecture is implemented by the following equations:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{(t-1)} + b_z), \tag{2.4}$$

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{(t-1)} + b_r), \tag{2.5}$$

$$\tilde{h}_t = \tanh(W_{xr}x_t + r_t \odot W_{hr}h_{(t-1)} + b_h) \tag{2.6}$$

$$h_t = (1 - z_t) * h_{(t-1)} + z_t * \tilde{h}_t \tag{2.7}$$

where σ denotes the logistic sigmoid function. z and r, are the update gates and the reset gates, respectively, and the $\odot$ symbol represents element-wise multiplication. W refers to the weight matrix corresponding to its respective subscript. An example GRU unit is illustrated in Figure 2.4. A GRU unit allows information to persist through the use of it's two gates; the update gate decides how much of the previous memory to retain, while the reset gate defines exactly how much of the new input is combined with the 'memory'. The main distinction between an LSTM and a GRU is the addition of a persistent cell state in the case of an LSTM, as well as the addition of a read and write gate that controls when to write to the cell, and when to read from it. While the core idea of memory is the same within the two types of RNNs, the LSTM is more complex, and therefore may be more difficult to train.

As with other types of RNNs, GRUs and LSTMs are differentiable function approximators that can be trained with the use of gradient descent in order to minimize some objective function.

2.4 SUMMARY

There exist many methods of quantifying functional connectivity, both statically and dynamically. Whilst static estimates of functional connectivity have provided useful insights, dynamic functional connectivity

is being increasingly used. In particular, model-driven estimates provide flexible yet statistically principled approaches. We have outlined a few, notably the HMM and it's derivatives (HSMMs, HOHMMs, etc.). We have also provided a brief introduction to recurrent neural networks. In the following chapter, we will outline how we can combine the ideas underpinning HMMs and RNNs.

3

GENERATIVE MODEL

3.1 INTRODUCTION

Our aim is, in general terms, to obtain a parsimonious set of "states" describing spontaneous neural dynamics, such that each state describes a distinct pattern of connectivity. Contrary to the assumption of stationarity (where a stationary process is one in which the statistical properties are constant over time) made by methods such as ICA or time-averaging, studies have shown that the temporal dynamics underpinning the recordings obtained through tools like MEG are rich and varied over a range of time-scales, from hundreds of milliseconds to tens of seconds. In order to get high quality estimation of functional connectivity, this dynamic temporal nature should be explicitly accounted for.

We therefore present here a model that learns both the spatial patterns of activity and also the temporal dynamics in a wholly unsupervised manner. Unlike in ICA, we can explicitly specify some transition distributions and a more complex observation distribution that is not simply an activation map. Additionally, while we do not constrain the temporal dynamics, it is important to note that prior knowledge of the form of these dynamics can easily be integrated with slight modifications to the model structure.

3.1.1 *Probabilistic model*

The simplest method of achieving such a model is by way of a Hidden Markov Model (HMM), upon which we shall build our model. To revisit the HMM; a basic HMM consists of a *hidden* process $Z = (z_t)_{t=1}^T$ with dynamics dictated according to a first order Markov chain with a discrete state space of size K, and an *observed* process $X = (x_t)_{t=1}^T$ occurring sequentially over a set of discrete time points $\{1, \cdots, T\}$, with each $x_t \in \mathbb{R}^d$. The following set of conditional independence assumptions are made to model both of these processes:

$$p(z_t|z_{1:(t-1)}) = p(z_t|z_{(t-1)}) \tag{3.1}$$

$$p(x_t|x_{1:(t-1)}, z_{1:(t-1)}) = p(x_t|z_t), \tag{3.2}$$

where $p(z_t|z_{(t-1)})$ and $p(x_t|z_t)$ are referred to as the *transition distribution* and *emission distribution*, respectively. The temporal dynamics of the HMM are dictated by a $K \times K$ matrix A called the transition probability matrix, where each row of the matrix $A_{i,\cdot}$ gives a transition distribution $p(z_t|z_{(t-1)} == i)$, and each element of the matrix $A_{i,j}$ gives a probability $p(z_t == j|z_{(t-1)} == i)$. In general, an emission distribution can take the shape of any static distribution.

3.2 LATENT SPACE

It is first necessary to decide on the form of the latent variables Z. As with the HMM, we propose a state-switching model wherein the latent space is the discrete set of state labels $\{k\}_{k=1}^K$, making each z_t a categorical variable. A state time-course then consists of a sequence of mutually exclusive states switching between each other. This kind

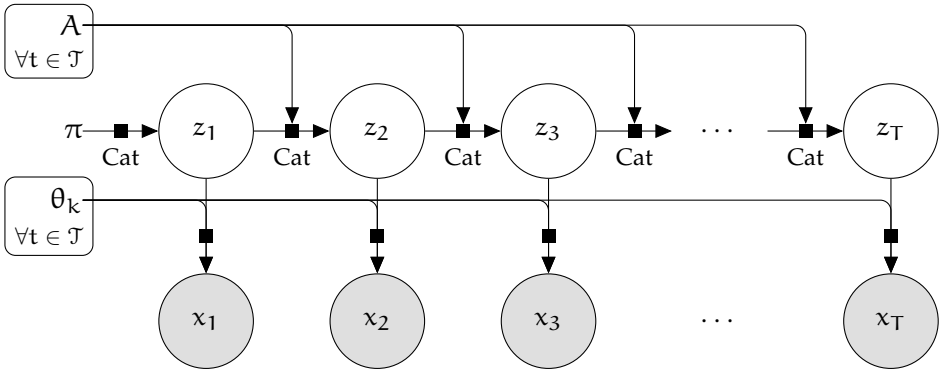

Figure 3.1: Graphical representation of a hidden Markov model. Latent variables are denoted by white nodes, observed variables by shaded nodes, and parameters are not enclosed within nodes. Arrows indicate conditional dependence.

of switching behaviour has been applied across multiple modalities; while there are some methodological differences, HMMs have been employed in modelling the temporal variability of MEG [3] as well as fMRI [32, 88]. Furthermore, while the idea of neural activity originating in a finite number of distinct networks that are entirely mutually exclusive is perhaps a simplistic picture of the true dynamics, there is some conjecture that the states identified by using the HMM can be interpreted as the "dominant" states occurring at a given point in time [3]. Though relaxation of the mutual exclusivity assumption would also provide a useful model, unfortunately given the observation model we have chosen particularly, this leads to a model that is no longer well determined. We have therefore chosen to assume the states are mutually exclusive.

3.3 SPATIAL MODEL

We will be modelling the activity of "source space" MEG data. When the analysis of MEG data takes place at the sensor level, we are restricted to a spatial analysis that conveys limited information about the cortical regions that are involved in the underlying processes we are

examining. Additionally, sensor positions are variable across subjects, which means sensor-space analysis may result in spurious results. By solving the inverse problem, we are able to project the measurements from sensor space to obtain voxel time-courses containing estimates of the source currents. We can therefore potentially obtain more accurate information regarding which brain regions are functionally connected. This source space data for each subject $S^{(i)}$ comprises of the activity of D voxels over $T^{(i)}$ time points, such that we have for each subject a matrix $X^{(i)} \in \mathbb{R}^{D \times T^{(i)}}$.

3.3.1 *Multivariate Gaussian distribution*

Recall that functional connectivity refers to the statistical dependencies between each voxel. Under an assumption of Gaussianity, we can therefore investigate functional connectivity with only second-order dependencies in the form of of covariances or correlations. While the Gaussian assumption is most certainly a simplification of the distribution that underpins the true network dynamics, it is not an unreasonable one; and indeed, it lies at the heart of many source localisation methodologies employed in MEG, in addition to a number of approaches that have made use of the covariance structure of the multivariate Gaussian distribution to model functional connectivity [26, 82, 117].

Formally, the observation model associated with each state $z_t = k$ will be given by a multivariate Normal distribution with zero mean and covariance matrix Σ_k. We have

$$p(x_t|z_t = k) = \mathcal{N}(x_t; \mathbf{0}, \Sigma_k) \tag{3.3}$$

The observation distributions are thus defined entirely by a set of covariance matrices, $\Sigma = \{\Sigma_k\}_{k=1}^{K}$, each Σ_k corresponding to one of

the K states. We make use of a model with zero mean such that all the information is contained within these covariance matrices and they can then be directly interpreted as the functional connectivity. The estimation of such covariance matrices through the optimisation of an objective function can be a difficult problem, owing to the constraint that a covariance matrix must be positive semi-definite.

3.3.2 *Covariance matrix parametrisation*

It is therefore necessary to reparametrise the covariance matrix in such a way that it is amenable to unconstrained optimisation. Because Σ_k is a symmetric positive semi-definite covariance matrix, it may therefore be factored by

$$\Sigma_k = L_k L_k^T \tag{3.4}$$

where L is a lower triangular matrix known as the *Cholesky* matrix. This method does have some issues, owing namely to the fact that a Cholesky factor is not unique; if L_k is a Cholesky factor of Σ_k, then so is any matrix obtained by multiplying a subset of the rows of L_k by -1. Numerical issues may arise in the optimisation of objective functions where multiple optimal solutions lie close together in the parameter space. Nonetheless, the Cholesky factorisation presents a computationally simple way to parametrise the covariance matrix, and therefore suits our purposes.

3.4 TEMPORAL MODEL

Owing to the temporal resolution available through the use of MEG, we hope to be able to explicitly capture the underlying dynamics that

drive functional connectivity at a variety of timescales. It is therefore necessary that we have an appropriate temporal model. In order to make learning tractable, previous models we have referenced have made use of simplistic models for temporal dynamics, relying on the use of Markov (and occasionally semi-Markov) chains. We would like to introduce a model that enables us to relax this Markovian (or semi-Markovian) constraint to account for longer histories — to build a generative model with a "memory" to some extent.

3.4.1 *Relaxing the Markov constraint*

One method to address the limitation of first order Markovian dynamics is to extend the framework to encompass higher order HMMs (HOHMMs) that allow the latent sequence Z to depend on time points extending further back into the past. We say a HOHMM has an *order* of q if the dynamics are dictated by the set of conditional independence assumptions

$$p(z_t|z_{1:(t-1)}) = p(z_t|z_{(t-q):(t-1)}) \tag{3.5}$$

$$p(x_t|x_{1:(t-1)}, z_{1:(t-1)}) = p(x_t|z_t). \tag{3.6}$$

While this framework allows us to relax the first order Markov assumption, it comes at a cost; the number of parameters of the model scale exponentially with the order of the model. Take a HOHMM with K states and order q; the transition distribution can be described by K^q probability distributions with K mass points each. Hence it involves a total number of $(K-1)K^q$ parameters, which very quickly becomes unmanageable.

3.4.2 *Deep models*

A more efficient method for representing the dynamics could perhaps come in the guise of an RNN. Both RNNs and the more traditional state space models (of which the HMM and HOHMM are both examples) are widely used model classes that are able to model temporal sequences of vectors $\{x_t\}$. Both models work on the basis that there exists some hidden state dynamics governing the observed behaviour of the system in question. However, while the hidden variables of state space models are probabilistic in nature, the hidden states of RNNs are deterministically determined.

An RNN makes use of this deterministic hidden state $h \in \mathbb{R}^p$ to model sequences, by means of the application of a parametrised nonlinear function f over the duration of the sequence. At each time step t in the sequence, the value of x_t is passed into the networks, and its hidden state is updated to

$$h_t = f(x_t, h_{(t-1)}). \tag{3.7}$$

A common choice for this function f is a gated activation function such as a long-short term memory (LSTM) cell [56], or a gated recurrent unit (GRU) [22].

While RNNs are powerful tools for modelling, the deterministic nature of their hidden state limits the capabilities of modelling stochastic sequences, as it becomes necessary for all the output distribution to model the entirety of the uncertainty in the system. Indeed, many such studies show dramatic improvements in the performance of RNNs that explicitly include uncertainty in their hidden states [5, 12, 23, 35, 45, 48]. We can also benefit by similarly combining the two model classes.

Our solution is to use a state space model, however instead of the transition function being given by a transition probability matrix, we instead introduce an RNN as a transition function. In practice, this means that rather than having to optimise a set of free parameters corresponding to each individual transition weight, we instead optimise a parametrised function that learns a mapping from previous latent states to future ones, similar to that presented in [72].

This parametric function has the advantage of having the number of parameters be constant with respect to the order of our model, enabling us to reasonably take into account a 'memory' of potentially hundreds of time steps. The limiting factor is no longer the number of parameters of the function but of the ability of the network to retain information, often called the *receptive field* of the network.

We denote the distribution $p(z_t|z_{1:(t-1)})$ accordingly:

$$p(z_t|z_{1:(t-1)}) = g(h_t) \tag{3.8}$$

$$h_t = f(h_{(t-1)}, z_{(t-1)}), \tag{3.9}$$

$$g(\cdot) = \text{Softmax}(\text{linear}(\cdot)) \tag{3.10}$$

where h_t is the hidden state of the network, f is a GRU, and g is some function mapping from the hidden state to the possible outputs (for example a linear affine transformation followed by a softmax function). We make use of a GRU rather than an LSTM or some other model as they are less complex computationally and therefore are typically thought to be simpler to train. The choice of layers and hidden states of the network is less straightforward, and will largely be decided by experimentation, though it is also possible to make use of more systematic hyperparameter optimisation methods.

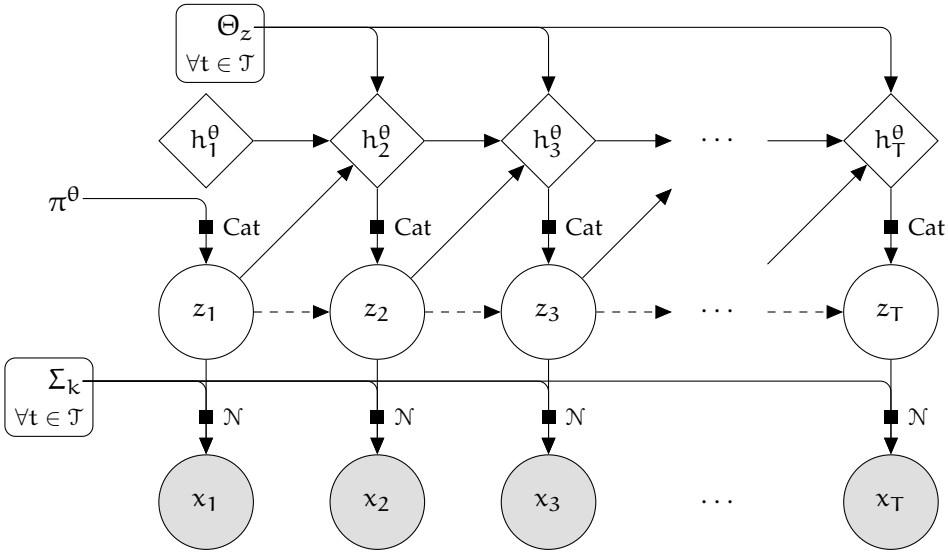

Figure 3.2: Graphical representation of the full model structure. Latent/hidden variables are shown in white, while observed variables are given by the shaded nodes. Diamond nodes indicate deterministic variables. Solid lines indicate the explicit conditional dependency structure of the model, while the dashed lines show the indirect conditional dependencies that exist owing to the deterministic nature of the hidden states. Owing to the recursive nature of RNNs, h_t^θ contains information about not only $z_{(t-1)}$, but all previous states, which is contained in $h_{1:(t-1)}^\theta$, (up to the limit of the receptive field of the RNN).

3.5 FULL MODEL SPECIFICATION

The full model is therefore given by the following:

$$p(X, Z; \theta) = p(z_1; \theta_z)\, p(x_1|z_1; \theta_x) \prod_{t=2}^{T} p(z_t|z_{1:(t-1)}; \theta_z)\, p(x_t|z_t; \theta_x) \tag{3.11}$$

$$p(z_1; \theta_z) = \text{Cat}(\pi^\theta) \tag{3.12}$$

$$p(z_t|z_{1:(t-1)}; \theta_z) = \text{Cat}(\alpha_t^\theta) \tag{3.13}$$

$$\alpha_t^\theta = g(h_t) \tag{3.14}$$

$$h_t = \text{GRU}(h_{(t-1)}, z_{(t-1)}; \Theta_z) \tag{3.15}$$

$$p(x_t|z_t) = \mathcal{N}(\mathbf{o}, \Sigma_{z_t}) \tag{3.16}$$

$$g(\cdot) = \text{Softmax}(\text{linear}(\cdot)) \tag{3.17}$$

where $\theta_z = \{\pi_\theta, \Theta_z\}$ are the parameters of the initial probability distribution and the GRU weights, respectively, and $\theta_x = \{\Sigma_k\}_{k=1}^K$ are the covariance matrices of each state.

4

INFERENCE

Given some MEG data, we wish to infer an underlying state sequence that is consistent with our probabilistic model. There are two main avenues to solving this problem; the first is to simply get point-estimates for each of the latent variables (using either maximum a posteriori or maximum likelihood estimates), however this is not ideal as we are unable to quantify the uncertainty associated with the resulting variables. The second is to do this within a Bayesian framework by finding a (joint) posterior distribution of each latent variable. However, calculating a full posterior analytically is not often possible, especially given here we are using a highly non-linear mapping between the latent states. We shall make use of the latter method here, focusing specifically on the use of variational Bayesian inference to achieve this in a computationally-efficient manner.

What follows in this chapter is an overview of variational Bayesian methods, and how it must be adapted to utilise it within a derivative-based optimisation scheme. We will then describe the details of the implementation of our model.

4.1 BAYESIAN IDEAS

There are two primary interpretations of statistical inference; that of frequentist inference, and that of Bayesian inference. The frequentist understanding of probability is that it is a limiting frequency; therefore a frequentist feels most comfortable assigning probability in the

context of experiments that are random and well-defined. Bayesian inference on the other hand, uses probability as a tool to represent an individual's degree of belief in a statement. Unlike the frequentist, a Bayesian is able to assign probability to an event where the uncertainty comes not from randomness, but it is also due to lack of knowledge.

Bayesian inference then specifies how one should update one's beliefs upon observing data. We can derive this from *Bayes' Theorem*. Bayes' theorem expresses the conditional probability of an event Z after X is observed, called the *posterior probability*, in terms of the *marginal likelihood* of X, the prior probability of Z, and the the *likelihood* of X given Z occurred.

$$\overbrace{p(Z|X)}^{\text{Posterior}} = \frac{\overbrace{p(X|Z)}^{\text{Likelihood}}\ \overbrace{p(Z)}^{\text{Prior}}}{\underbrace{P(X)}_{\text{Marginal likelihood}}}, \tag{4.1}$$

where the term in the denominator, known as the marginal likelihood, is given as:

$$p(X) = \int_{-\infty}^{\infty} p(X|Z)\, p(Z) dZ. \tag{4.2}$$

Let us consider a simple example of Bayes' theorem in action: imagine a model of the world, given by $p(\text{world})$. We would therefore implicitly expect to see certain types of behaviour under this prior model. Upon making an observation of the real world, we obtain a likelihood of this observation under our prior model, denoted as $p(\text{observation}|\text{world})$. We then obtain a new model of the world with the aid of Bayes' theorem, $p(\text{world}|\text{observation})$, called the posterior distribution. In the case that the observation is in line with the expected behaviour of the world under the prior model, then

$\frac{p(\text{observation}|\text{world})}{p(\text{observation})}$ is close to 1, and our posterior would remain similar to the prior.

4.2 VARIATIONAL METHODS

First let us set up the problem. Consider a joint probability density of some observed variables $X = \{x_m\}_{m=1}^{M}$, with $x_i \in \mathcal{X}$ and latent variables $Z = \{z_n,\}_{n=1}^{N}$, $z_i \in \mathcal{Z}$,

$$p(X, Z) = p(X|Z)\, p(Z). \tag{4.3}$$

The model first draws the latent variables from a prior density $p(Z)$, and then relates them to the observations through the likelihood $p(X|Z)$. While this is trivial in practise, obtaining the reverse is much more challenging. This process of inference involves conditioning the latent space upon the data in order to compute the posterior $p(Z|X)$. As this is not explicitly defined in Equation 4.3, we could try to compute it using Bayes rule, as in 4.1, but this is made difficult because of the marginal likelihood (also called the evidence). Unfortunately, for complex models, this computation is intractable and thus necessitates approximate inference procedures.

The idea behind variational inference is that while $p(Z|X)$ may have an arbitrarily complex form (even in the case of simple priors and likelihoods), instead of trying to compute it exactly, we posit a family of approximate densities $\mathcal{Q}$ over the latent variables, and then try to find the member of that family $q(Z)$ that best approximates $p(Z|X)$ [67, 122]. We can write this as

$$q^*(Z) = \min_{q(Z)\in\mathcal{Q}} G(q(Z)\|p(Z|X)),$$

where G is some dissimilarity measure between the two densities. How you pick $\mathcal{Q}$ and G can lead to different methods of inference. For example, Laplacian methods require Gaussian approximations to the posterior using the distance between the Hessians of the pdfs, while variational inference chooses G to be the Kullback-Leibler (KL) divergence.

4.2.1 *Relevant concepts from information theory*

Before we continue with the overview of variational inference, it will be useful to take a slight detour to first define some terms taken from Information Theory.

4.2.1.1 *Entropy*

While the term *entropy* originated in statistical thermodynamics, we are more interested in the information theoretic idea of entropy. More specifically, when we speak of entropy, we refer in particular to Shannon entropy.

Given a probability distribution with density or mass function p, the entropy is defined as

$$H(p) = -\int_X p(X) \log p(X) dX. \tag{4.4}$$

The value of $-\log p(X)$ is called the *information content*, or *surprisal*, and quantifies the amount of information gained when X is sampled using p (or you could think of it in terms of how 'surprised' you would be by the sample). The entropy is the expected rate at which information is produced by X, and can be intuitively thought of as a measure of the dispersal, or uncertainty, associated with X. Note that in the case of $p(X) = 0$, the convention is for the value of $0 \cdot \log 0$ is

taken to be 0. Entropy has the unit of measure *nats* when the natural logarithm is used, and *bits* when the logarithm is base 2.

To give an example, consider flipping an unfair coin that has a 99% chance of heads, and a 1% chance of tails. If we flip a coin once and get an outcome of heads, we are not very surprised, and hence the information content is low. The information entailed by heads is given as $-\log_2 0.99 = 0.014$ bits, whereas the information for tails is 6.64 bits. To obtain the entropy, we must calculate the average rate of information, which is given by

$$H(x) = -\sum_i p(x) \log_2 p(x) = -(0.99 \cdot \log_2 0.99 + 0.01 \cdot \log_2 0.01) = 0.08 \text{ bits.}$$

This is quite low, as the entropy is dominated by the heads outcome.

In the case the the coin we flip is fair, e.g. $p(0) = p(1) = \frac{1}{2}$, the entropy will be *maximised* (with a value of 1), as we have no reason to expect any particular outcome more than the other. Conversely, the entropy will be *minimised* in the case of an extremely unfair coin where $p(0) = 1$ and $p(1) = 0$, as in the case where the coin we flip has heads on both sides. This is due to the fact that there will be no surprise at all — we are certain of the outcome of when we sample x according to p.

4.2.1.2 *Cross-entropy*

Cross-entropy is, unsurprisingly, strongly related to the concept of entropy; whereas the entropy is the expectation under p of the surprisal $-\log p(X)$, *cross-entropy* is defined as the expectation under p of the surprisal for a different distribution $-\log q(X)$:[1]

$$H_q(p) = -\int_X p(X) \log q(X) dX. \tag{4.5}$$

1 Note that this is non-standard notation, and would typically be denoted $H(p, q)$. However, this is the same notation as used in joint entropy, and furthermore gives the impression that cross-entropy is symmetric (which it is not).

Going back to our example with coin flips; assume the coin we are flipping is fair, with the outcome given by $p(0) = p(1) = 0.5$, and yet we believe it to be unfair and given by $q(0) = 0.99$ and $q(1) = 0.01$. While the entropy is given by $H(p) = 1$, the cross-entropy is given by

$$H_q(p) = -(0.5 \cdot \log_2 0.99 + 0.5 \cdot \log_2 0.01) = 3.329.$$

That is, if we know the coin is fair, then the average surprise we experience is 1, however if we believe the coin to be strongly biased in favour of heads, the average surprise we experience is 3.329, and we are left scratching our heads and wondering how we are getting so many tails!

As we can see from this example, if your belief of the world diverges from the truth, your expectations will often be incorrect, and consequently you will be surprised more often. This can be demonstrated using *Gibb's inequality:*

$$H_q(p) \geqslant H(p), \tag{4.6}$$

with equality only where $p(X) = q(X)$.

4.2.1.3 *KL divergence*

Given that we know that when your beliefs do not match reality, you are going to be on average more surprised by events, the obvious question is then: by how much? We can calculate this using the *KL divergence* of p with respect to q, given as [2]

$$D_q(p) = H_q(p) - H(p), \tag{4.7}$$

2 This is also nonstandard notation; typically the KL divergence is denoted by $D_{kl}(p\|q)$, or $KL(p\|q)$.

or expanded to give

$$D_q(p) = -\int_X p(X) \log \frac{q(X)}{p(X)} dX. \tag{4.8}$$

By 4.6, we have that $D_q(p) \geqslant 0$, with $D_q(p) = 0$ if and only if $p(X) = q(X)$ for all values of X.

The KL divergence can also be used as a difference metric between two distributions p and q [73], although it is not a *distance* metric as it is not symmetric. For an intuitive interpretation from a Bayesian perspective, $D_q(p)$ can be thought of as the information gained when we move from a prior q to the posterior p.

4.2.1.4 *Forward and reverse KL divergence*

We noted that the KL divergence is not symmetric, i.e. $D_q(p) \neq D_p(q)$. The former, we call the *forward* KL divergence, whilst the latter is referred to as the *reverse* KL divergence.

Examining the forward KL divergence $D_q(p)$ given in 4.8, we see that there is a large positive contribution to the KL divergence where $q(X)$ is near zero, unless $p(X)$ is also close to zero. Consequently, an optimal variational distribution that minimises the forward KL divergence will attempt to stretch so as to maximally cover all of $p(X)$.

Conversely, the reverse KL divergence $D_p(q)$ gets a large contribution from regions in X space where $p(X)$ is near zero, unless $q(X)$ is also close to zero. The effect on an optimal variational distribution in this case that it will tend to underestimate p, by avoiding regions where $p(X)$ is small.

In the case where $p(X)$ is multimodal and $q(X)$ is unimodal, we see very different behaviour of $q(X)$. The forward KL will encourage q to place mass across all peaks, which invariably ends up with the point of maximum density for $q(X)$ may reside in an area that has low

density in the original distribution. Meanwhile, the reverse KL will result in q matching at least one of the modes of $p(X)$.

4.2.2 *Variational inference*

Returning to the problem at hand; we have some intractable Bayesian inference problem $p(Z|X)$ we are trying to compute, where X is our observed data and Z are the hidden variables. We are trying to find an approximate distribution q such that it minimises the *reverse* KL divergence — mostly due to the zero-forcing nature of the reverse KL divergence (where it discourages situations where $q(X)$ is high but $p(X|Z)$ is low). We can apply Bayes' rule to give this in a form that is easier to work with:

$$\begin{aligned} D_{p(Z|X)}(q(Z)) &= \int_Z q(Z) \log \frac{q(Z)}{p(X|Z)} dZ \\ &= \int_Z q(Z) \log \frac{q(Z)}{p(X,Z)} dZ + \int_Z q(Z) \log p(X) dZ \qquad (4.9) \\ &= \int_Z q(Z) \log \frac{q(Z)}{p(X,Z)} dZ + \log p(X). \qquad (4.10) \end{aligned}$$

Rearranging, we get:

$$\begin{aligned} \log p(X) &= D_{p(Z|X)}(q(Z)) - \int_Z q(Z) \log \frac{q(Z)}{p(X,Z)} dZ \\ &= D_{p(Z|X)}(q(Z)) + \mathcal{L}(q, Z; X), \qquad (4.11) \end{aligned}$$

where $\mathcal{L}$ is called the (negative) *variational free energy*. The log evidence $\log p(X)$ is constant with respect to q, therefore by maximising the variational free energy $\mathcal{L}$, we also minimise the KL divergence $D_p(q)$ as required.

Examining the negative variational free energy in more detail gives some intuition about the optimal variational distribution q:

$$\begin{aligned}
\mathcal{L}(q, Z; X) &= -\int_Z q(Z) \log \frac{q(Z)}{p(X, Z)} dZ \\
&= -\int_Z q(Z) \log \frac{q(Z)}{p(X|Z)\, p(Z)} dZ \\
&= -\int_Z q(Z) \log \frac{1}{p(X|Z)} dZ - \int_Z q(Z) \log \frac{q(Z)}{p(Z)} dZ \\
&= \int_Z q(Z) \log p(X|Z) dZ - \int_Z q(Z) \log \frac{q(Z)}{p(Z)} dZ \\
&= \mathbb{E}_{q(Z)}[\log p(X|Z)] - D_{p(Z)}(q(Z)). \qquad (4.12)
\end{aligned}$$

or alternatively expressed:

$$\mathcal{L}(q, Z; X) = \mathbb{E}_{q(Z)}[\log p(X, Z) - \log q(Z)] \qquad (4.13)$$

We have now shown that we can rewrite $\mathcal{L}$ to give us a sum between the expected log-likelihood of the data, given the model q, and the KL divergence between the prior $p(Z)$ and the variational distribution $q(Z)$. The first term, the log-likelihood of the data, will encourage densities that place their mass on configurations that explain the data, while the negative KL divergence between the prior and the variational distribution will encourage densities close to the prior.

Now, with the aid of the variational free energy, we can omit explicitly computing the marginal likelihood entirely and save ourselves from something of a headache.

4.2.3 *Factorised approximations*

Remember that Z is made up of N hidden variables $Z_1, \cdots, Z_N$. In general, this will mean that 4.11 will have multiple integrals (or in the case of discrete variables, multiple summations). One straightforward

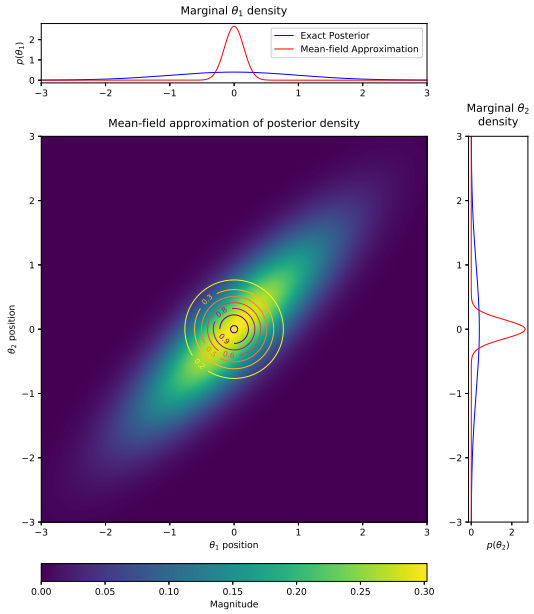

Figure 4.1: Visualisation of a mean-field approximation for a two-dimensional Gaussian posterior. The color plot shows the posterior density, whilst the contour plot shows the effect of mean-field factorisation. Marginal densities are shown for each dimension, for both the exact posterior (in blue), and the mean-field approximation (in red).

simplifying assumption that we can make for our density q is that we can partition the variables into independent parts. This assumption is called the *mean field approximation*, and looks like this:

$$p(Z|X) \approx q(Z) = q(z_1, \cdots, z_N) = \prod_{i=1}^{N} q_i(z_i). \tag{4.14}$$

We have factorised here such that we have single variable per partition, but a partition can contain any number of variables. Each partition has an associated variational density q_j chosen to maximise the lower bound in 4.12, and should take a parametric form appropriate to the corresponding random variable(s), for example a continuous variable should have a continuous density, such as a Gaussian or Gamma distribution, whereas a categorical variable should have a categorical distribution.

The mean field approximation does come with some trade-offs however. While $q(Z)$ is able to approximate the joint distribution, the individual $q_j(z_j)$ are poor approximations to the true marginals $p_j(z_j)$, and should not be expected to necessarily resemble them. Additionally, the mean-field approximation is unable to capture correlations between partitions.

These can both be demonstrated in Figure 4.1. Consider a highly correlated bivariate Gaussian posterior, as shown by the color plot. A mean-field approximation to this posterior would involve the product of two independent univariate Gaussian distributions, as shown by the contour plot. While the mean is the same for both the posterior and the approximation, the covariance structure differs in that it is completely decoupled. We also see the the marginal variances greatly differ from those of the true density — a consequence of optimising using $D_p(q)$, where $q(Z)$ is penalised for placing mass on areas where $p(Z)$ has little mass.

4.3 STOCHASTIC GRADIENT VARIATIONAL INFERENCE

In order to optimise 4.12, we need to be able to calculate the gradient with respect to the parameters of both $p(X, Z)$ and $q(Z)$. We will refer to these parameters as θ and ϕ, respectively, such that we have:

$$\theta = (\theta_X, \theta_Z) \tag{4.15}$$

$$\phi = (\phi_1, \cdots, \phi_N) \tag{4.16}$$

$$p(X, Z; \theta) = p(X|Z; \theta_X)p(Z; \theta_Z) \tag{4.17}$$

$$q(Z; \phi) = \prod_{i=1}^{N} q_i(z_i; \phi_i). \tag{4.18}$$

We use $\theta(\theta_X, \theta_Z)$ as the parameters of the generative model, where θ_X are any parameters of the observation model (mean, covariance,

etc.), and θ_Z are the parameters of the latent distribution. We then can rewrite 4.12 to explicitly include these parameters:

$$\mathcal{L}(\theta, \phi; X) = \mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)] - D_{p(Z;\theta_Z)}(q(Z;\phi)), \quad (4.19)$$

or alternatively:

$$\mathcal{L}(\theta, \phi; X) = \mathbb{E}_{q(Z;\phi)}[\log p(X, Z;\theta_X, \theta_Z) - \log q(Z;\phi)]. \quad (4.20)$$

After calculating the gradient with respect to both θ and ϕ, we can make use of a first-order iterative optimisation method. For traditional mean-field variational inference, where both the prior densities and their corresponding variational distribution are in the same exponential family (we say they are *conditionally conjugate*), the expectations in 4.19 can be calculated analytically, yielding a closed form objective. However, the calculation of the gradient may scale with the size of X, and may therefore be particularly expensive if X is large. Additionally, if we are unable to analytically compute the expectations, it is necessary to instead use statistical estimators (not to be confused with numerical estimation). Both of these issues can be addressed with the use of stochastic optimisation, where instead of computing exact gradients, we instead take noisy estimates of them.

Calculating the gradient of 4.19 with respect to θ is straightforward. We can make use of Leibniz's rule for the gradient of the expectation of the likelihood:

$$\begin{aligned}\nabla_{\theta_X}\mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)] &= \nabla_{\theta_X}\int_Z q(Z;\phi)\log p(X|Z;\theta_X)dZ \\ &= \int_Z q(Z;\phi)\nabla_{\theta_X}\log p(X|Z;\theta_X)dZ \\ &= \mathbb{E}_{q(Z;\phi)}[\nabla_{\theta_X}\log p(X|Z;\theta_X)]., \end{aligned} \tag{4.21}$$

where ∇_{θ_X} represents the gradient of this term with respect to θ_X, and is zero. The gradient of the KL term with respect to θ_Z may be computed analytically, however in the case where we cannot analytically calculate the KL divergence, we are instead required to use the gradients of 4.20, which is a similar procedure as above.

In other words, the gradient of the expectation is equal to the expectation of the gradient for the parameters of the generative model θ. This becomes a more difficult problem when calculating the gradient with respect to ϕ, which parametrises the expectation. Again we use Leibniz's rule:

$$\nabla_\phi\,\mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)] = \nabla_\phi\int_Z q(Z;\phi)\log p(X|Z;\theta_X)dZ \tag{4.22}$$

$$= \int_Z \log p(X|Z;\theta_X)\nabla_\phi\, q(Z;\phi)dZ, \tag{4.23}$$

however 4.23 does not take the form of an expectation, meaning we cannot simply use a Monte Carlo estimator. We must therefore find an alternative way of calculating $\nabla_\phi\mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)]$. There are two main ways that this can be achieved, the use of pathwise gradient estimators, or the use of score function estimators. We will outline the pathwise gradient estimators below, and the score function estimator method is detailed in the appendix.

4.3.1 *Pathwise gradient estimators*

We can obtain lower variance estimators with some additional assumptions, by way of pathwise gradient estimators.

Pathwise gradient estimators make use of the *reparametrisation trick* [11, 71, 93, 99, 102] to evaluate 4.22 by reparametrising Z in terms of an auxiliary noise variable ϵ with a known distribution $p(\epsilon)$ that does not depend on the variational parameters ϕ. We choose some deterministic transformation $t(\epsilon; \phi)$ such that $Z = t(X, \epsilon; \phi)$ is still distributed according to $q(Z; \phi)$.

We can then make use of this to obtain Monte Carlo estimates of the gradient of the expectations of some function $f(Z)$

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q(Z;\phi)}[f(Z)] &= \nabla_\phi \int_Z q(Z;\phi) f(Z) dZ \\
&= \nabla_\phi \int_Z p(\epsilon) f(t(X, \epsilon; \phi)) d\epsilon \\
&= \int_Z p(\epsilon) \nabla_\phi f(t(X, \epsilon; \phi)) d\epsilon \\
&= \int_Z p(\epsilon) \nabla_Z f(Z) \nabla_\phi t(X, \epsilon; \phi) d\epsilon \\
&= \mathbb{E}_{p(\epsilon)}[\nabla_Z f(Z) \nabla_\phi t(X, \epsilon; \phi)]. \qquad (4.24)
\end{aligned}$$

This yields the estimators for the gradient:

$$\nabla_\phi \mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)] \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_Z \log p(X|Z^{(s)};\theta_X) \nabla_\phi t(X, \epsilon^{(s)}; \phi), \qquad (4.25)$$

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X) - q(Z;\phi)] \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_Z (&\log p(X|Z^{(s)};\theta_X) \\
&- q(Z^{(s)};\phi)) \nabla_\phi t(X, \epsilon^{(s)}; \phi) \qquad (4.26)
\end{aligned}$$

where $Z^{(s)} = t(X, \epsilon^{(s)}; \phi)$ and $\epsilon^{(s)} \sim p(\epsilon)$.

A simple example is the case of a univariate Gaussian: let $Z \sim \mathcal{N}(\mu, \sigma^2)$. A valid reparametrisation is given by $Z = \mu + \sigma \cdot \epsilon$, where ϵ is an auxiliary noise variable drawn from the standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$, and we have for some function $f(Z)$

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(Z;\mu,\sigma^2)}[f(Z)] &= \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[f(\mu + \sigma \cdot \epsilon)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} f(\mu + \sigma\epsilon^{(s)}), \quad \text{where } \epsilon^{(s)} \sim \mathcal{N}(0, 1).\end{aligned} \tag{4.27}$$

There are three primary assumptions made in order to make use of the reparametrisation trick:

1. Z should be a continuous random variable, and there should exist a known *differentiable* reparametrisation $Z = t(\epsilon; \phi)$;
2. It is easy to generate samples from the distribution $p(\epsilon)$;
3. $f(Z)$ is differentiable with respect to Z.

The question is then: for what types of variational distributions are we able to choose such a differentiable transformation $t(\cdot; \phi)$ and auxiliary variable $\epsilon \sim p(\epsilon)$? As outlined in [71], there are three primary approaches:

1. **Tractable inverse CDF:** In this case, a valid transformation is to let $t(\cdot; \phi)$ be the inverse cumulative distribution function (CDF) of $q(Z; \phi)$, and $\epsilon \sim \mathcal{U}(\mathbf{0}, \mathbf{I})$.
2. **Location-scale family:** Similarly to how the univariate Gaussian was reparametrised, any "location-scale" family of distributions can be expressed in terms of the standard distribution with location = 0, and scale = 1. Set $p(\epsilon)$ to be the standard distribution, then $t(\cdot; \phi) = \text{location} + \text{scale} \cdot \epsilon$.

3. **Composition:** In this case, the transformation $t(\cdot;\phi)$ is obtained by expressing Z as different transformations of auxiliary variables, as in the case of e.g. the log-normal distribution, where $Z = \exp(\mu + \sigma \cdot \epsilon)$, and $\epsilon \sim \mathcal{N}(0, 1)$.

In the case that a differentiable transformation $t(\cdot;\phi)$ that falls into one of the above categories does not exist, it may be possible to instead use approximations to the inverse CDF.

4.3.2 *Mini-batch stochastic gradient descent*

Once we have calculated $\nabla_{\theta,\phi}\mathcal{L}(\theta, \phi; X)$, we can make use of the gradient to iteratively approach a local or global minimum[3]. A standard gradient descent algorithm would process all of the data when updating the parameter values, e.g.

$$\phi = \phi - \alpha\nabla_{\phi}\mathcal{L}(\theta, \phi, X). \tag{4.28}$$

For a sequence X of length T, we can subsample smaller sequences of the data $X^{(s)} = (x_t)_{t=i}^{(i+T)}$, where $i \in \{1, T - L\}$, and then update the weights after analysing each sample. Therefore stochastic gradient descent (SGD) looks like this:

$$\phi = \phi - \alpha\nabla_{\phi}\mathcal{L}(\theta, \phi, X^{(s)}). \tag{4.29}$$

Now, unlike in standard gradient descent, we no longer have to see all of the training data between updates, which can be time prohibitive for large data sets. Additionally, unlike standard gradient descent, we do not have to keep the entire dataset in memory. However, the noisy gradient approximations can result in slow convergence [18],

3 We have been discussing maximising the *negative* variational free energy 4.19, however we can equivalently minimise the non-negative variational free energy

although the extra stochasticity conferred by sampling can also help move away from local minima. Taking multiple samples can allow for closer approximations of the gradient without sacrificing much in the way of computational efficiency. We call this *minibatch* stochastic gradient descent:

$$\nabla_\phi \mathcal{L}(\theta, \phi, X^{(1:b)}) = \frac{1}{b} \sum_{s=1}^{b} \nabla_\phi \mathcal{L}(\theta, \phi, X^{(s)}), \tag{4.30}$$

$$\phi = \phi - \alpha \nabla_\phi \mathcal{L}(\theta, \phi, X^{(1:b)}). \tag{4.31}$$

Minibatch stochastic gradient descent does not guarantee convergence; learning rate choices can be difficult, leading to very slow convergence or even divergence depending on whether it is too small or too large, and learning rate schedules have to be specified in advance. Additionally, it can be prone to becoming trapped in local minima [29]. A number of extensions to SGD exist, including Adam [69] and Adagrad [31]. These make use of techniques like momentum [94], and adaptive and per-parameter learning rates to improve the performance of the "vanilla" SGD optimiser.

4.3.3 *KL annealing*

In order to facilitate training the generative model, we can also make use of a technique called KL annealing [13]. This involves adding a variable weight to the KL term in the cost function during training. This weight is set to zero at the start of training, such that the inference network learns to encode as much weight into Z as possible. As training then progresses, this weight is gradually increased according to some annealing schedule, and the prior starts to encourage the inference network to smooth out the inferred state sequences. Eventually this weight reaches 1 and the weighted cost function is equivalent

to the true variational free energy. A typical annealing scheme makes use of a sigmoid schedule.

4.3.4 *Amortized inference*

Even with efficient sampling procedures, inference can still be computationally expensive. Consider an example where each data point $X \in \mathcal{X}$ is governed by a corresponding latent variable $Z \in \mathcal{Z}$ with variational parameter ϕ. Traditional VI requires that it is necessary to optimise each ϕ for each local variable Z, which means that it does not scale well to large datasets — in particular when the optimisiation is embedded in a global parameter update loop. Amortisation proposes that rather than optimising each local variational parameter ϕ, instead we optimise a set of global parameters and we can *amortise* the computational cost by framing the per-data-point optimisation process instead as a regression problem; rather than solving for our optimal proposal $q^*(Z)$ directly, we make use of a parametric function (sometimes called a *recognition model*) $f_\phi \in \mathcal{F}(\Omega) : \mathcal{X} \to \Omega$ that predicts the value of $q^*(Z)$. We similarly construct a parametric function $g_\theta \in \mathcal{G}(\mathcal{P}) : Z \to \mathcal{P}$, where $\mathcal{P}$ is a family of distributions over X, $\mathcal{G}$ is a family of functions indexed by parameters θ to define the generative model.

Once the functions are estimated, it is then possible to "infer" new latent variables by passing new data points through the function f_ϕ, or to generate new data by passing new latent states through the function g_θ.In the case that f_ϕ takes the form of a neural networks, it may also be referred to as a *recognition network* or *inference network*.

4.4 IMPLEMENTATION OF INFERENCE NETWORK

Armed with our toolbox of approximate inference techniques, we are now well-equipped to deal with most problems requiring scalable probabilistic inference, even those with non-parametric and highly non-linear distributions governing the dynamics. It remains now to construct a working inference routine from our constituent pieces. Given data $X = (x_t)_{t=1}^T$ and latent variables $Z = (z_t)_{t=1}^T$, recall our generative model:

$$p(X, Z; \theta) = p(z_1; \theta_Z) p(x_1 | z_1; \theta_X) \prod_{t=2}^{T} p(z_t | z_{1:(t-1)}; \theta_Z) p(x_t | z_t; \theta_X), \tag{4.32}$$

where we use a GRU to parametrise the distribution $p(z_t | z_{1:(t-1)}; \theta_Z)$. As we have previously mentioned, the use of a highly nonlinear mapping between the latent states means that inference of the true posterior is intractable, and we are required to use variational inference. We shall now define a variational distribution $q(Z|X; \phi)$.

4.4.1 *Factorisation over time*

Considering the temporal structure of our probabilistic model in 4.32, it is appropriate to similarly factor the distributions over Z over time. However, a naïve mean-field approach whereby each local distribution is treated as independent would likely result in a poor fit with slow convergence. We shall therefore make an initial attempt at ensuring the approximate posterior integrates information over longer periods

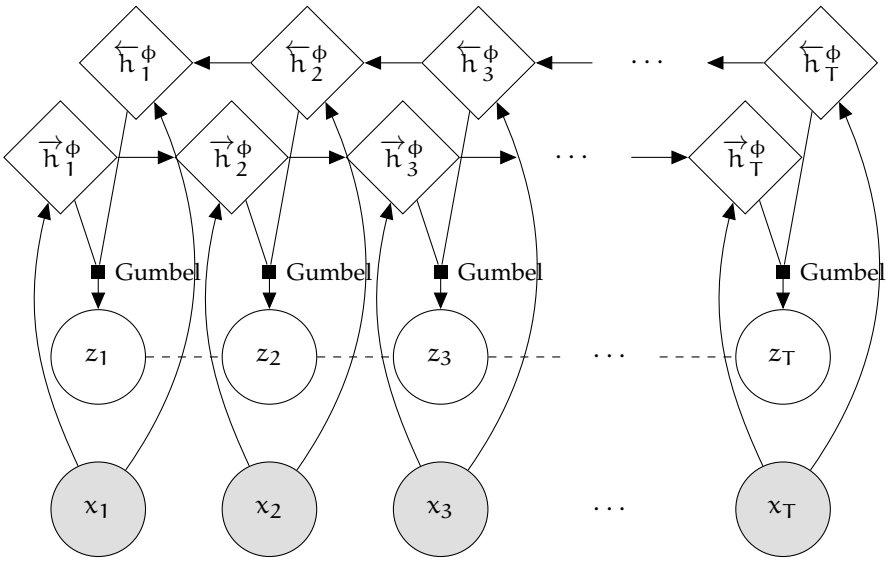

Figure 4.2: Graphical representation of the inference structure. Latent/hidden variables are shown in white, while observed variables are given by the shaded nodes. Diagonal nodes indicate deterministic variables. Solid lines indicate the explicit conditional dependency structure of the model, while the dashed lines show the indirect conditional dependencies that exist owing to the deterministic nature of the hidden states. The inference network makes use of a bidirectional recurrent neural network that makes use of information from both earlier and later in time.

of time. We are able to simply condition each z_t on the entire sequence of X

$$q(Z;\phi) = \prod_{t=1}^{T} q(z_t|X;\phi), \tag{4.33}$$

this way we can integrate information contained in the entirety of the sequence in order to best estimate the value of each z_t.

4.4.2 *Recurrent recognition model*

Owing to the form of the generative model, the posterior distribution at each time-point takes the form of a categorical distribution:

$$q(z_t|X;\phi) = \mathrm{Cat}(\alpha_t^{\phi}). \tag{4.34}$$

It is now necessary to parametrise this distribution. One way of obtaining each α_t^ϕ is with a neural network.

While neural networks are common in amortisation, owing to the sequential nature of the conditional distribution we are learning, a recurrent neural network would be an ideal choice for a recognition model, in particular a GRU. We will make use of two distinct GRUs to parametrise this function; the first makes use of information from the past, $(x_i)_{i=1}^t$, while the second makes use of information from the future, $(x_i)_{i=t}^T$. This structure, where two recurrent networks running in opposite directions are combined, is known as a bidirectional RNN.

The forwards and backwards are respectively given by the equations

$$\overrightarrow{h}_t = \text{GRU}(\overrightarrow{h}_{(t-1)}, x_t), \tag{4.35}$$

$$\overleftarrow{h}_t = \text{GRU}(\overleftarrow{h}_{(t+1)}, x_t), \tag{4.36}$$

where $\overrightarrow{h} \in \mathbb{R}^{h_1}$ denotes the hidden states of the forwards network and $\overleftarrow{h} \in \mathbb{R}^{h_2}$ the backwards network. The recursion on these hidden states is what allows information to persist through time.

Combining the two networks, we arrive at the equations for the variational distribution:

$$q(Z|X;\theta) = \prod_{t=1}^{T} g([\overrightarrow{h}_t, \overleftarrow{h}_t]). \tag{4.37}$$

Here $g(\cdot)$ is some parametric function that maps the concatenated hidden states to $\mathcal{Q}$. In deciding how to parametrise g, we are faced with the question of what form do we want $q(z_t|X;\phi)$ to take? To start with, $z_t \in \{k\}_{k=1}^K$ is essentially a state assignment. Much like the distribution $p(z_t|z_{1:(t-1)};\theta_z)$, that means this variational distribution is a categorical distribution. We therefore construct $g(\cdot)$ as follows. First, it is necessary to make use of a linear affine layer to map from

$\mathbb{R}^{(h_1+h_2)} \to \mathbb{R}^K$. We can then make use of a softmax function to map to the parameters of a categorical distribution.

$$\alpha_t^\phi = \text{softmax}([\overrightarrow{h}_t, \overleftarrow{h}_t]). \tag{4.38}$$

4.4.2.1 *The Gumbel-softmax distribution*

As was mentioned in section 4.3, we are unable to directly calculate the gradient with respect to the variational parameters ϕ, so approximate methods must be used. We would like to apply a low variance estimator if at all possible, and so the pathwise gradient estimators (recall: the "reparametrisation trick") are a good first choice. The issue is that, as we noted at the end of the last section, the distribution we would like to use is a discrete one. The question of how to apply a differentiable transformation of a categorical variable that is discontinuous by nature is then a requisite one for our purposes.

The Gumbel-softmax [64, 80] approximation is a way to overcome the limitations of the reparametrisation trick when applied to discrete data, by making use of a continuous approximation that depends on a temperature parameter τ, where the zero-temperature degenerates to the discrete case.

As is hinted by the name, this approximation makes use of the Gumbel distribution. A random variable γ has a standard Gumbel distribution if it can be given by $\gamma = -\log(-\log(\epsilon))$ with $\epsilon \sim \mathcal{U}(0, 1)$. For some discrete random variable z, we can parametrise it in terms of Gumbel random variables:

$$z = \underset{k}{\text{argmax}}(\log \pi_k + \gamma_k), \tag{4.39}$$

with $P(z = k) \propto \pi_k$ and $\{\gamma_k\}$ a set of i.i.d. standard Gumbel random variables.

Unfortunately, argmax is neither a continuous nor a differentiable operator. We can instead use the softmax function as both a continuous and differentiable approximation, where we have the softmax given by

$$f(\pi_i;\tau) = \frac{\exp(\pi_i/\tau)}{\sum_k \exp(\pi_k/\tau)}. \tag{4.40}$$

While we will not draw random variables that take the form of *one-hot* vectors (where all π_k values are zero, save for a single index with the value one) as with the application of the argmax operator, the softmax will result in values that lie within the $K-1$ simplex, e.g. $z \in \Delta^{K-1}$. We therefore have the random variable z given as

$$z = \{z_k\} = \{f(\log \pi_k + \gamma_k;\tau)\} = \left\{ \frac{\exp((\log \pi_k + \gamma_k)/\tau)}{\sum_i \exp((\log \pi_i + \gamma_i)/\tau)} \right\}, \tag{4.41}$$

with density

$$p(z;\pi,\tau) = \Gamma(K)\tau^{(K-1)} \prod_{k=1}^{K} \left(\frac{\pi_k z_k^{(-\tau-1)}}{\sum_{i=1}^{K} \pi_i z_i^{-\tau}} \right), \quad z \in \Delta^{K-1}. \tag{4.42}$$

While samples from the Gumbel-softmax distribution (also called the Concrete distribution, for being *con*tinuous and dis*crete*) are differentiable, they are not identical from the samples drawn from a Categorical distribution for non-zero temperatures. Where this is a requirement, we can make use of the *straight-through* estimator, where samples are discretised using argmax, but the approximation in 4.41 is used when calculating the gradient. This might be appropriate where it is necessary to use the sample as an index.

Returning to 4.37, we now have a parametrised distribution $q(Z|X;\phi)$, and a method to approximately sample from it in a differentiable manner, albeit through a surrogate distribution $\hat{q}$:

$$\hat{q}(z_t|X;\phi) = \text{Gumbel}(\alpha_t^{\phi}) \approx \text{Cat}(\alpha_t^{\phi}) = q(z_t|X;\phi) \tag{4.43}$$

4.4.3 *Objective function*

As with more standard stochastic gradient variational inference, we train both our inference and generative models jointly by maximising the variational free energy. By substituting 4.32 and 4.33 into the equation for variational free energy, we can obtain a timestep-wise variational lower bound:

$$\begin{aligned}
\mathcal{L}(\theta, \phi; X) &= -\int_Z q(z_t;\phi) \log \prod_{t=1}^{T} p(x_t|z_t;\theta_X) dz_t \\
&\quad + \int_Z q(z_{1:t};\phi) \log \frac{\prod_{t=1}^{T} q(z_t;\phi)}{p(z_1;\theta_z) \prod_{t=2}^{T} p(z_t|z_{1:(t-1)};\theta_z)} dz_{1:t} \\
&= -\sum_{t=1}^{T} \int_Z q(z_t;\phi) \log p(x_t|z_t;\theta_X) dz_t + \int_Z q(z_1;\phi) \frac{q(z_1;\phi)}{p(z_1;\theta_Z)} dz_1 \\
&\quad + \sum_{t=2}^{T} \int_Z q(z_{1:t};\phi) \log \frac{q(z_t;\phi)}{p(z_t|z_{1:(t-1)};\theta_Z)} dz_{1:t} \\
&= -\sum_{t=1}^{T} \mathbb{E}_{q(z_t);\phi)}[\log p(x_t|z_t;\theta_X)] + D_{p(z_1;\theta_Z)}(q(z_1;\phi)) \\
&\quad + \sum_{t=2}^{T} \int_Z q(z_{1:(t-1)};\phi) D_{p(z_t|z_{1:(t-1)};\theta_Z)}(q(z_t;\phi)) dz_{1:(t-1)} \\
&= -\sum_{t=1}^{T} \mathbb{E}_{q(z_t);\phi)}[\log p(x_t|z_t;\theta_X)] + D_{p(z_1;\theta_Z)}(q(z_1;\phi)) \\
&\quad + \sum_{t=2}^{T} \mathbb{E}_{\substack{q(z_1;\phi) \\ \cdots \\ q(z_{(t-1)};\phi)}} [D_{p(z_t|z_{1:(t-1)};\theta_Z)}(q(z_t;\phi))].
\end{aligned} \tag{4.44}$$

While we are unable to analytically calculate the expectation terms in 4.44, we can approximate them using:

$$\hbar\mathbb{E}_{q(z_t);\phi)}[\log p(x_t|z_t;\theta_X)] \approx \frac{1}{S} \sum_{s=1}^{S} \log p(x_t|z_t^{(s)};\theta_X) \tag{4.45}$$

$$\mathbb{E}_{\substack{q(z_1;\phi) \\ \cdots \\ q(z_{(t-1)};\phi)}} [D_{p(z_t|z_{1:(t-1)};\theta_Z)}(q(z_t;\phi))]. \approx \frac{1}{S} \sum_{s=1}^{S} D_{p(z_t|z_1^{(s)},\cdots,z_{(t-1)}^{(s)};\theta_Z)}(q(z_t;\phi)), \tag{4.46}$$

where samples $z_i^{(s)} \sim q(z_i; \phi)$ are drawn in topological or temporal order; starting with $p(z_1; \theta_Z)$ which has no parents, we sample sequentially through time by conditioning each variable on values sampled in previous steps, proceeding this way until all $t-1$ values have been sampled, and we can then calculate the KL divergence analytically using the distributions conditioned on our samples. We can therefore approximate the distribution $p(z_t|z_1, \cdots, z_{(t-1)}; \theta_Z)$ in linear time. We call this *ancestral* (or forward) sampling.

4.5 FULL INFERENCE SPECIFICATION

4.5.1 *Inference model*

Our full inference model is therefore the following:

$$q(Z; \phi) = \prod_{t=1}^{T} q(z_t|X; \phi) \tag{4.47}$$

$$q(z_t|X; \phi) = \text{Cat}(\alpha_t^\phi) \tag{4.48}$$

$$\hat{q}(z_t|X; \phi) \approx q(z_t|X; \phi) \tag{4.49}$$

$$\hat{q}(z_t|X; \phi) = \text{Gumbel}(\alpha_t^\phi) \tag{4.50}$$

$$\alpha_t^\phi = g([\overrightarrow{h}_t, \overleftarrow{h}_t]) \tag{4.51}$$

$$\overrightarrow{h}_t = \text{GRU}(\overrightarrow{h}_{(t-1)}, x_t), \tag{4.52}$$

$$\overleftarrow{h}_t = \text{GRU}(\overleftarrow{h}_{(t+1)}, x_t), \tag{4.53}$$

$$g(\cdot) = \text{Softmax}(\text{linear}(\cdot)) \tag{4.54}$$

where ϕ are the weights of the GRU networks. Note here that Z is therefore only variable here were are being probabilistic over, as it is difficult to introduce priors over RNN weights.

The final cost function is as follows:

$$\mathcal{L}(\theta, \phi; X) = -\frac{1}{S}\sum_{s=1}^{S}\sum_{t=1}^{T}\left[-\frac{1}{2}\log|\Sigma_{z_t^{(s)}}| + x_t^\intercal \Sigma_{z_t^{(s)}}^{-1} x_t\right] + \sum_{k=1}^{K}\alpha_{1,i}^{\phi(s)}\log\frac{\alpha_{1,i}^{\phi(s)}}{\pi_i^{\theta(s)}} + \frac{1}{S}\sum_{s=1}^{S}\sum_{t=2}^{T}\sum_{k=1}^{K}\left[\alpha_{t,i}^{\phi(s)}\log\frac{\alpha_{t,i}^{\phi(s)}}{\alpha_{t,i}^{\theta(s)}}\right], \quad (4.55)$$

where the first term is the log likelihood for a multivariate Gaussian distribution with zero mean, calculated using the covariance matrix indexed by the value of z_t sampled via the reparametrisation trick. The remaining terms are the KL divergence between the parameters of both the variational distributions and the prior distributions, indexed over each latent dimension, such that $\alpha_{t,i}$ is the probability of being in state i at time t. For the sampling, it is sufficient to simply use a value of $S = 1$ [71].

4.5.2 *Stochastic gradient optimisation algorithm*

Algorithm 1 Sequential Autoencoding Variational Bayes

Precondition: Generative distribution p_θ, generative network f_θ with parameters θ, variational distribution q_ϕ, inference network f_ϕ with parameters ϕ, observed data $X = (x_t)_{t=1}^T$.

$\theta, \phi \leftarrow$ Initialise parameters
repeat
  $x^s \leftarrow x_{i:i+L}$, $i \sim \text{Uniform}(T - L)$ ▷ Sample subsequence of x
  $\alpha_s^\phi \leftarrow f_\phi(x^s)$ ▷ Parameterise q_ϕ
  $\hat{z}^s \sim q_\phi(z^s; \alpha_s^\phi)$ ▷ Sample from q_ϕ using reparam trick
  $\alpha_s^\theta \leftarrow f_\theta(\hat{z}^s)$ ▷ Parameterise p_θ using samples from q_ϕ
  $g \leftarrow \nabla_{\theta,\phi}\mathcal{L}(x^s, \hat{z}^s, \alpha_s^\theta, \alpha_s^\phi)$ ▷ Calculate gradients
  $\theta, \phi \leftarrow$ Update parameters using g
until convergence of θ, ϕ
return θ, ϕ

Note that while we specify here that the generative distribution is parametrised by a GRU, in fact this algorithm can be applied to any sequential latent variable model that is entirely composed of differen-

tiable steps; a standard HMM for example, Linear Gaussian models, Autoregressive models, and so forth. In the case of alternative distributions that do not lend themselves towards the reparametrisation trick, alternative methods can be used including score function estimators.

5

SIMULATION RESULTS

In this chapter we aim to address a key point in the implementation of this model and the inference framework we have outlined — *do they actually work*? What follows in this chapter is the recurrent model and the inference algorithm demonstrated upon some toy data to illustrate their modelling capabilities. We shall also experiment with making use of a Markov chain in place of the recurrent temporal model to evaluate whether the addition of a more flexible generative model improves the performance of inference, particularly in the case of data that is non-Markovian (or at least only partially Markovian), and then make the comparison between these two models and a hidden Markov model.

5.1 SIMULATION DETAILS

We tested both the recurrent generative model and the Markovian model upon three different toy models. These three models shared a common observation model, and differed only in their underlying temporal model. We evaluated a number of types of generative temporal models; first using a basic Markov chain, where transitions are dictated by a transition probability matrix. Secondly, we made use of a hierarchical/switching Markov chain, characterised by two *meta-states* where the first meta-state characterised a sequence where only forward transitions were allowed (e.g. $1 \rightarrow 2 \rightarrow ... \rightarrow K$), and the other meta-state was the reverse ($K \rightarrow K-1 \rightarrow ... \rightarrow 1$).Thirdly, we

examined the idea that HMMs are biased towards shorter transitions, by making use of a number of semi-Markov chains, each characterised by a different state lifetime: 50, 100, 200, 400, and 1000 timesteps. Finally, we examined the effect of changing temporal scales by making use of a switching semi-Markov model, which switched between each of the previous state lifetimes after n time-steps. A more detailed account of each model generated from follows:

5.1.1 *Observation model*

Each of the models makes use of the same observation model, with the same parameters; a set of 6 10-dimensional Gaussian distributions with zero mean, and full covariance matrices. Each covariance was generated randomly. In order to ensure that these random matrices were still positive-semidefinite, we constructed them according to the following method:

$$W_k \in (0,1)^{D\times 1} \tag{5.1}$$

$$w_{k,i,1} \sim \mathcal{U}(0,1) \tag{5.2}$$

$$V_k \in (0,1)^{D\times 1} \tag{5.3}$$

$$v_{k,i,1} \sim \mathcal{U}(0,1) \tag{5.4}$$

$$\Sigma_k = W_k W_k^T + \text{diag}(V_k), \tag{5.5}$$

and writing out the observation model explicitly, we have:

$$p(x_t|z_t == k) = \mathcal{N}(\mathbf{0}, \Sigma_k). \tag{5.6}$$

5.1.2 *Temporal models*

We construct the generative model by pairing the above observation model with a temporal model. The temporal model is responsible for emitting sequences of what we shall term *emission states*, which in turn are responsible for indexing a set of parameters for the observation model. Each temporal model was responsible for generating a total of 100,000 observations.

5.1.3 *Markov chain*

By combining the observation model with a Markov chain, we obtain a HMM, where the transition through emission states is dictated by a transition probability matrix A. We generated a random transition probability matrix A by randomly sampling $K \times K$ values from a uniform distribution over $[0, 1]$, adding an additional weight to the diagonal to ensure some degree of stability, and subsequently applying the softmax function to each row to ensure it summed to 1.

5.1.4 *Semi-Markov chain*

A hidden semi-Markov model (HSMM) is similar in structure to a hidden Markov model, except for the fact that instead of parametrising a distribution over the probability of staying in the same state, as in the HMM, the HSMM parametrises a distribution over how long a state will remain fixed until transitioning into a new state. The choice of this distribution is entirely optional (though the distribution must have support over positive integers, as a model cannot have negative lifetimes, and it must be an integer as we are operating upon discrete time-steps, rather than using a continuous process).

We generated 5 semi-Markov chains, each with the same transition probabilities (given by a transition probability matrix with no self-transitions) and a different set of lifetime distributions, ranging from shorter lifetimes (50 time-steps) to much longer lifetimes (1,000 time-steps). The lifetimes for each state was sampled from the same distribution, a uniform distribution over $[\text{lifetime}, \text{lifetime} + 10]$, where lifetime is set to 50, 100, 200, 400, and 1,000 for each respective realisation of the semi-Markov chain.

5.1.5 *Hierarchical Markov chain*

In the case of a hierarchical models, the emission states serve as the bottom layer in a 'hierarchy' of states, where higher level states are responsible for dictating the behaviour of the lower level states. We refer to these higher-level states as meta-states.

We use here a 2 layer hierarchical Markov chain, with 2 meta-states, and 6 emission states. Essentially, each meta-state indexes a different transition probability matrix. We tried to ensure the dynamics of each meta-state are completely different, and so imposed an offset diagonal structure on the transition probability matrices such that the first meta-state characterised by forwards transitions, and the second meta-state was characterised by reversed transitions. We additionally had a transition probability matrix dictating the transitions between meta-states.

5.1.6 *Hierarchical semi-Markov chain*

Finally, we wanted to expand on the semi-Markov example to examine the effect of having multiple scales of lifetimes present in the data. We did this by combining the previous two models, such that we had 5

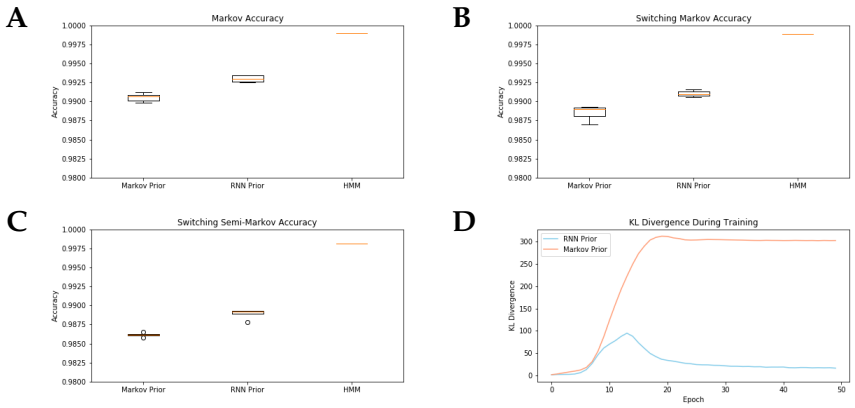

Figure 5.1: Accuracy of each of the models upon the different datasets. (A) the hidden Markov model dataset, (B) the hierarchical hidden Markov model dataset, and (C) the switching semi-Markov model dataset. (D) shows the KL divergence between the prior and variational distributions while training for the different prior models.

meta-states, each of which was associated with a different lifetime. We could therefore have both very fast transitions immediately followed by much longer transitions. We additionally set a lifetime for each meta-state of a total of 2,000 time-steps.

5.2 RESULTS

The RNN we used were 64 hidden state GRUs with 2 layers, and made use of dropout during training with a keep-probability of 0.5. The data was batched into sequences of length 200, and a minibatch size of 40 was used. We trained the model for 50 epochs, using the Adam optimiser [70], with a learning rate of 0.001 and all other parameters fixed. We also made use of KL-annealing using a logistic function with mean 10 and scaling parameter 0.3. We split the data into a training and a validation dataset, using a 80:20 ratio, and then performed the final inference upon the whole dataset. Dropout was turned off for the final inference. We also used a Gaussian HMM from the HMM-MAR toolbox [120] with 6 states.

We trained 5 realisations of both the recurrent model and the HMM upon each of these datasets to ascertain how well they could infer the true dynamics that generated the data. We additionally aimed to see the effect of replacing the recurrent prior with a Markov chain, on the basis that given the training data is (at least partially) Markovian, it should still provide comparable results to the use of the recurrent prior. Figure 5.1 shows the results of running each of these realisations, showing that each of the models obtain very accurate results across all datasets. It is clear that even if the dynamics are not directly Markovian, the HMM is nonetheless able to infer the correct state sequence to a very high degree of accuracy, consistently more accurately than both the Markovian and recurrent prior models. Even in the case where the data has both very long and very short transitions as in the case of the switching semi-Markov model in Figure 5.1C, the HMM is nonetheless able to infer the correct sequence, showing the duration of state lifetimes have little effect on the accuracy of the inference. In the case of the recurrent model, addition of a more flexible temporal prior appears to aid inference even in the case where the original prior is consistent with the underlying data. We can see the effect of the choice of prior on the KL divergence during training in Figure 5.1D. While both sequences share the same characteristic shape, the recurrent prior reaches a much smaller peak before starting to converge, and does so earlier. It may well be that identifying a better generative model earlier in the training has the effect of identifying better a local optimum.

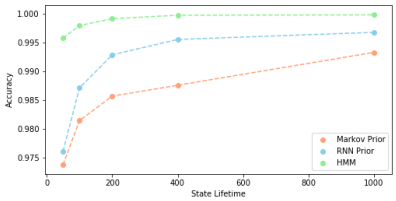

Figure 5.2: Accuracy for different length lifetimes. Shown is the accuracy for each model class in inferring the true state

5.2.1 *Effect of state lifetimes*

While we showed that both types models are able to learn the inferred sequence of a variety of different state lifetimes, we were curious to see the effect of different durations upon the final inference — were we able to show more accurate inference in the presence of long, or of short lifetimes. Figure 5.2 shows the result of this experiment. In all cases, the accuracy increases as the duration of state lifetimes also increase, although the effect is more pronounced in the recurrent model.

5.2.2 *Effect of RNN size*

We also wanted to investigate the effect hyperparameter selection upon the resulting inferred results, specifically the capabilities of the GRU as dictated by the hidden state size of the networks (not to be confused with the latent space size K, which was fixed over all experiments). We therefore trained a model with different numbers of hidden states (2, 16, 64, 128, 256) to evaluate the resulting accuracy, and the speed of convergence. Results are shown in Figure 5.3. As we can see from Figure 5.3B, we achieve very accurate ($> 99\%$) results for all models with the exception of using only 2 hidden states in the GRU which performs very poorly. While increasing the number of hidden states from 16 doesn't seem to increase the accuracy of the model significantly, we can see from Figure 5.3A that is does appear to increase convergence speed.

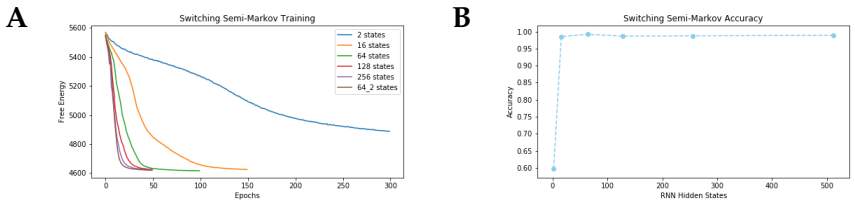

Figure 5.3: Hidden state size comparison upon switching semi-Markov chain data. (A) While increasing the number of hidden states appears to result in improved free energies, there appears to be some convergence, indicating both that it does not appear to result in drastic overfitting, but also that it does not result in any significant gains to accuracy. We also tried increasing the number of layers to 2, as shown by the loss for"64_2", which denotes a 2 layer network with 64 hidden states. (B) The agreement of the inferred states with the true states that generated the data for each number of hidden states.

5.3 CONCLUSION

We demonstrated in this chapter a number of points. We first empirically validated the recurrent network on a variety of Markovian and non-Markovian models. In all of the datasets we presented, we were able to achieve over 97% accuracy with the recurrent model, although we also demonstrated that the HMM provides a benchmark even when the data is not explicitly Markovian in nature.

While we do not have any decisive reason for the slight difference in performance, it may be perhaps due to an greater requirement for data on the part of the GRU in order to achieve direct correspondence with the HMM. We also see that it is not necessary to have a large network to perform approximate inference, as we achieve comparable results using a single layer GRU with 16 hidden states and a GRU with 256 hidden states. However, increasing the network size also increases the efficiency of learning, and larger networks require far fewer training epochs.

6

RESTING STATE RESULTS

What follows in this chapter is an overview of the results obtained by applying this model to learn the spatial maps and corresponding time courses that characterise the resting state neural activity of a set of 55 subjects. As the model acts in a wholly unsupervised manner, and as there is no objective ground truth, we make use of a HMM to provide a baseline for comparison, and show the improved ability of the recurrent model to explicitly capture the dynamics of the data.

6.1 DATA

The data used in this chapter was originally collected from the University of Nottingham as part of the UK MEG Partnership. Resting state data was collected from a total of 55 subjects between the ages of 18 and 48 (77 participants were originally scanned, but 22 had to be discarded due to excessive head motion or artefacts), with a mean age of 26.5 years old. 35 of the subjects were male. Both MEG and MRI data was acquired for each subject.

6.1.1 *Data acquisition*

MEG data acquisition was carried out using a 275-channel CTF whole-head system (MISL, Conquitlam, Canada) with third-order gradiometer correction applied. The subject was seated in the scanner and presented with a fixation target, whist data was recorded for 300 sec-

onds at a sampling rate of 1200 Hz, and subsequently down sampled to 600 Hz with a 300 Hz low-pass anti-aliasing filter. Three head position indicator (HPI) coils were attached to the participant's head as fiducial markers for tracking head position at the nasion, left and right preauricular points. These coils were periodically energised throughout the acquisition in order to enable localisation of the head, relative to the MEG sensor array. Structural MRI scans were used for the purposes of MEG coregistry. The scans were acquired at an 0.8mm isotropic resolution, using a Phillips 7T Achieva MRI system.

6.1.2 *Data preprocessing*

The MEG system geometry was coregistered to the subject's anatomical MRI using a Polhemus FASTRAK 3D digitiser system that recorded the position of the fiducial points, and the subject's head shape. The digitised head shape could then be registered to the structural MRI in order to determine the position of the MEG sensors relative to the subject's head shape. By registering the structural MRI to the MNI152 standard brain, all source space analysis could then be performed in MNI space.

The data was converted to SPM12 format, and down sampled to 250 Hz using an anti-aliasing low-pass filter, and then data was then bandpass filtered from 1-45 Hz. Artefact detection was carried out using an OSL (OHBA Software Library) function that identifies regions with particularly high variance, and those regions were omitted from further analysis.

This filtered data was then projected onto an 8mm grid spanning the whole brain using Linearly Constrained Minimum Variance (LCMV) beamforming [118, 125], and parcellated into 38 distinct brain regions. We were able to compute a single activity timecourse for each of these

regions by taking the first principal component of the activity in the voxels within each region. We then compensated for spatial leakage by using symmetric multivariate orthogonalisation, as developed by Colclough et al. [25]. This algorithm essentially projects all parcel time-courses onto a new orthogonal basis, such that zero-lag correlations are removed.

6.1.3 *Amplitude envelopes*

After orthogonalisation, the oscillatory amplitude envelopes for each parcel were computed by taking the magnitude of the Hilbert transform and down sampled to 40 Hz by temporally averaging within sliding windows of 100 ms in length, with 75% overlap between consecutive windows. The data was demeaned over time and normalised by the global standard deviation across all parcels, and batched across subjects such that we did not have to deal with jumps between two different subjects.

6.1.4 *Model training*

The data was split into sequences of 1000 timesteps in length, and minibatches consisting of 40 randomly sampled sequences (sampled without replacement) were used to input data into the model. The model was trained over the course of 100 epochs using the inference algorithm detailed in previous chapters. GRUs with 2 layers and 32 hidden cells were used in both the inference and the prior model, and the parameters were optimised with the Adam [69] optimiser, using a learning rate of 0.001. KL annealing was performed by scaling the KL-divergence by a logistic function with mean 30 and scale parameter 0.5.

6.2 ANALYSIS

Before we carry out any analysis, it is necessary to first ask the question: exactly how many states are we trying to infer? If we expect to find that our states bear some resemblance to the canonical resting state networks found in fMRI analysis, that may give us a good starting point — while the total number of reported networks in such studies vary depending on groups of subjects, analysis methods, or acquisition protocol [6, 10, 28, 30, 53, 103], there are a number of networks that are consistently reported; the default mode network, the primary sensorimotor network, the primary visual network, the extra-striate visual network, the left parietal and frontal network, the right parietal and frontal network, the frontal network, and the insular-temporal and anterior cingulate cortex network. This is by no means a comprehensive list however, and therefore does not provide a conclusive answer to the question.

An alternative method may be to take a data-driven approach to the specification of the total number of states. While it is true that as it is, our model requires a priori specification of the number of latent states to infer, there are a number of ways in which we can change this,, for example by making use of stick-breaking priors, where the model is free to increase the number of states as necessary [85], or by way of automatic relevance determination, where states may be eliminated if they are not supported by the data [78, 79]. Another alternative that requires no architectural considerations is to make use of the Bayesian model evidence to make an unbiased decision. Whilst it would be intractable to compute this exactly, we are able to obtain an approximation by way of the variational free energy.

6.2.1 *Latent dimensionality*

In practise, we intentionally avoided making use of a wholly unsupervised approach to the problem of dimensionality specification, as we are interested in interpreting and comparing the results with a previous study [3].

To see the effect of specifying different numbers of states, we ran the model with a range of values; 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, and 20 states. As different initialisations of gradient optimisation methods can lead to convergence to different local optima, we attempted to address this issue by repeating the training procedure a total of 8 different times for the 55 subjects. The results are shown in Figure 6.1. We saw a monotonic reduction in the average free energy as we increased the total number of states (and appeared to avoid overfitting on the basis of the held out validation set), although by the time we reached 8 or 9 states we started to see diminishing returns with the addition of each subsequent state. It also appeared that models with fewer number of states converged to more consistent free energy values; by the time we were inferring 9 states and higher, we saw a greater range of values, indicating that the solutions reached by the model across each realisation were somewhat more sensitive to initial conditions.

So as to better understand the consistency and stability of the inferred states across each run, we looked to implement a similar methodology to that described by Yang et al. [126] for each set of realisations. We first calculated the absolute value of what is termed the *RV coefficient* [100] (a multivariate generalisation of the squared Pearson correlation coefficient) between each pairwise spatial map across every realisation. We were then aiming to group states across different realisations according to those most spatially correlated, such that we could obtain a consistent set of spatial maps that were all

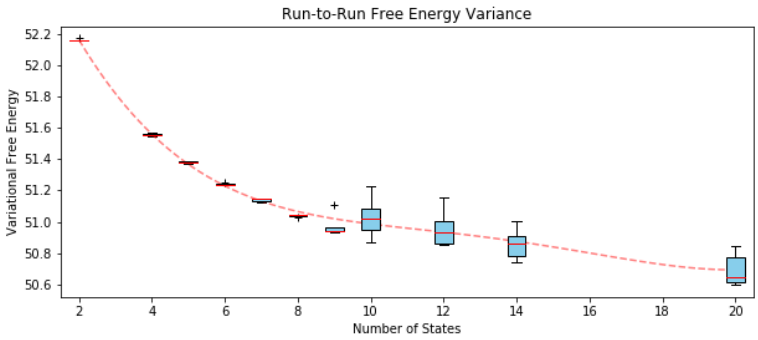

Figure 6.1: A total of 8 different realisations of the model were obtained for various values of K, each initialised with a different random seed. The average variational free energy was then computed for the entire data set, as shown by the box plots. The red trend line is for illustrative purposes only.

highly correlated. It was therefore necessary to define some threshold for the correlation coefficient to determine at which point two states were not considered "the same".

We did this by plotting the histogram of all the spatial correlations for the set of realisations. An example of this can be found in Figure 6.2A for $K = 8$. The histograms shows a roughly bimodal distribution, with modes near 0 and 1, representing a large number of uncorrelated states, and a smaller number of highly correlated states. We set the threshold at a point between the two modes; in our case we used a value of 0.75, although we repeated the analysis with values between 0.6 and 0.8 and obtained comparable results.

We then took an iterative approach to matching states, where the most similar pair of neural states (that is, the pair with the highest correlation coefficient of covariance matrices) from different realisations were assigned to the same group. The process could then be repeated until we had assigned each state to a group, or there were no longer any remaining state with correlation coefficients greater than our chosen threshold. The remaining states were then all grouped together into what we will consider a non-significant group. Note that each realisation could contribute at most one state to each group.

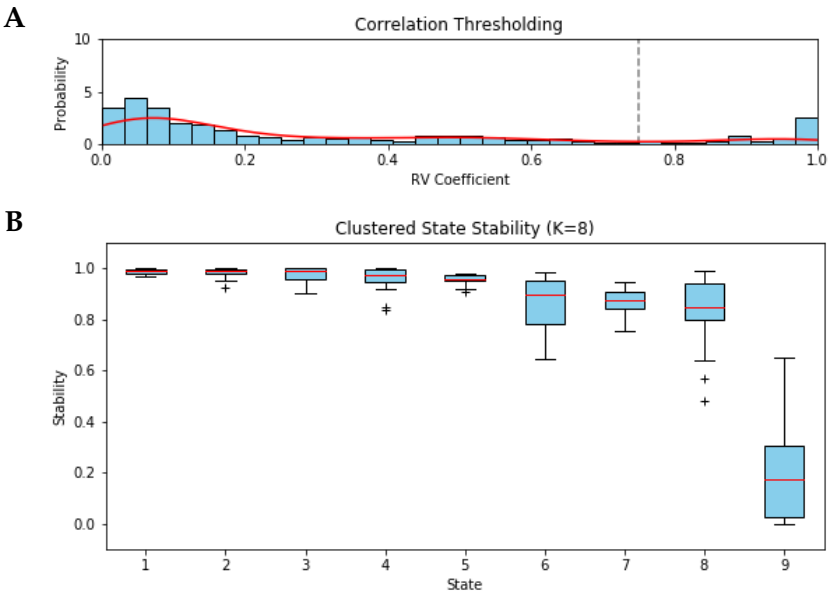

Figure 6.2: State stability for K = 8. (A) Histogram for all 2D spatial correlation coefficients between states across each different realisations.The grey line indicates the chosen threshold of 0.75. (B) Resulting clustered states, ordered by average stability. We find 8 stable groups emerging, with an additional group comprising mostly uncorrelated states.

After grouping the states across all realisations, the "stability" could be assessed by calculating the average correlation between each within-group pair of brain states. For full results across all values of K, refer to the appendix, however the resulting groupings for K = 8 are shown in 6.2B — we identified a total of 8 stable states, along with an additional group of uncorrelated states (which upon visual inspection appear to resemble the other groups, likely the result of a poor initialisation ending up in a local minima). Higher than 9 states appear to result in increasing numbers of groups, for example K = 10 resulted in 13 groups, while K = 14 identified 19 total stable groups.

In order to provide a like-for-like comparison with the results of [3], we selected the run with 8 states. We then sorted and labelled each group in the order of descending stability. To ensure we were closest to achieving a global optimum, we selected the realisation with the lowest free energy for further comparisons. We also experimented

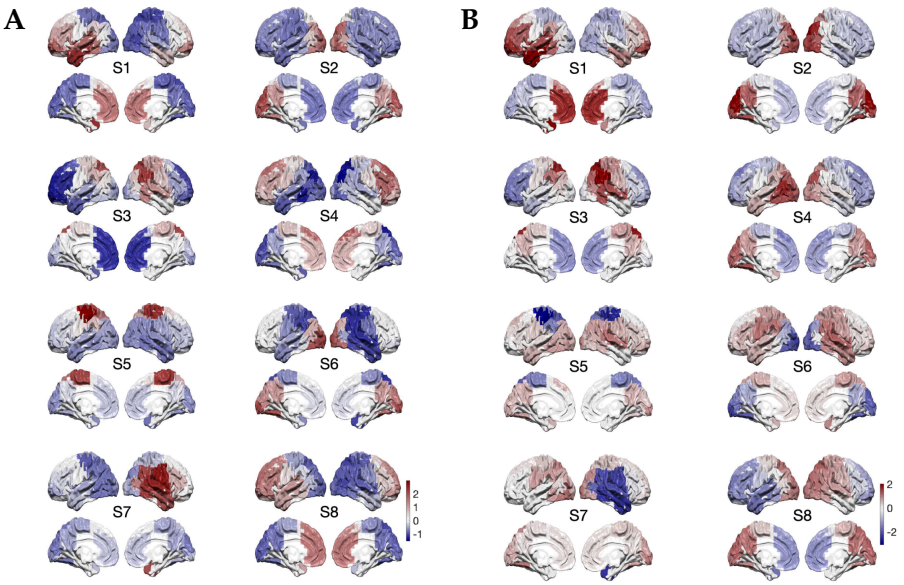

Figure 6.3: Unthresholded spatial distributions of power deviation from the mean, and spatial distributions of connectivity. (A) The activation maps inferred by the final model, ordered by stability. The surface rendering shows the mean activation by way of the z-score for each state across the left and right lateral, and the left and right medial positions. Brighter red regions indicate higher power than the average power across all states, while brighter blue regions indicate lower than average power. (B) Within-state functional connectivity as given by degree/connectedness. Brighter red regions indicate higher connectivity, while blue means less connectivity.

with using the average of each group for initialisation, but this did not prove to achieve any increase in performance.

6.2.2 *Spatial maps*

Once we have fitted the complete model, we are able to extract information about the spatial distribution directly from the observation models that parametrise each state. We can examine the variance associated with each parcel to generate state-wise activation maps, which represent the deviation of the power within each region in the brain from the average power across the whole brain. Similarly, we can estimate the functional connectivity by assessing how correlated a parcel's time course envelope is with the rest of the brain.

The power of the 8 brain states inferred from the entire dataset, which we denote as S1-S8 (sorted by order of stability), are shown by the unthresholded z-maps in Figure 6.3A. Group averaged state specific increases in oscillatory power within a given parcel are denoted by red regions, while blue indicates a decrease in power. State 1 shows increased activation in the frontal and left lateralised temporal nodes. Both state 2 and state 6 display increased activation in the visual cortex, with state 6 seeing particularly reduced activation in the temporal and sensorimotor regions. Both state 3 and 5 appear to be sensorimotor networks, with state 3 experiencing a notable reduction in parietal activation. Both state 4 and state 8 show decreases in activity in parietal regions in parallel with increases in the frontal regions. State 7 meanwhile appears to be a right auditory network.

There appear to be strong similarities with those states identified in [3]. With the exception of state 8, we seem replication of most states. State 4 appears resemble the DMN identified using the HMM, and we likewise see multiple visual states, a parietal state, left and right lateralise temporal states, and a sensorimotor state.

6.2.3 *Temporal dynamics*

While each state is associated with a distinct set of spatial and connectivity features, we also obtain a probabilistic time course containing the periods of activity for each state. This time course is the posterior probability, that is the probability that each state is active given the data. We can estimate a number of statistics about the temporal dynamics of each state using these probabilistic maps. The primary measures we will be reporting include the state fractional occupancy across subjects, which is the relative proportion of time spent in any given state; the average state lifetimes, or the average amount of time

elapsed between a states activation and subsequently switching to a new state; and finally the state interval time. The interval time for a state is simply the duration between successive state visits.

These posterior probabilities are also useful for decoding purposes to estimate the optimal sequence of states $Z^* = (z_t^*)_{t=1}^T$. This "optimal" sequence of states for a given data set consists of the most likely state sequence that generated the data, or

$$Z^* = \max_{z_1,\cdots,z_T} p(x_1,\cdots,x_T,z_1,\cdots,z_T). \tag{6.1}$$

In the case of the HMM, this can be achieved by way of a dynamic programming algorithm called the Viterbi algorithm, and owing to the Markovian conditions within which the HMM operates, we can guarantee that, *for a given HMM*, the Viterbi algorithm estimates the optimal path. Unfortunately, our model has no such constraint, and so to guarantee that a given sequence is the global optimal is not feasible. We are however, able to find an approximate sequence by passing the data through the complete inference network to obtain $q(z_t|X;\phi)$, and then simply taking the state that has the highest posterior probability of occurring at each time point.

We calculated the temporal statistics on the posterior distributions, and subsequently decoded the "optimal" state sequence. Temporal statistics are reported in Figure 6.4, as well as a portion of the state sequence from the first subject. From the statistics, we see that the states state time courses are characterised by relatively short lived (median lifetimes are around 50 ms and, average state lifetimes are all around 100 ms or less), rapidly transitioning sets of states. These durations are faster than even the quick transitions reported by Baker et al. [3], who saw state durations of between 100 and 200 ms, let alone those typically associated with resting state networks. Each distribution of lifetimes has an extremely long tail however, and some

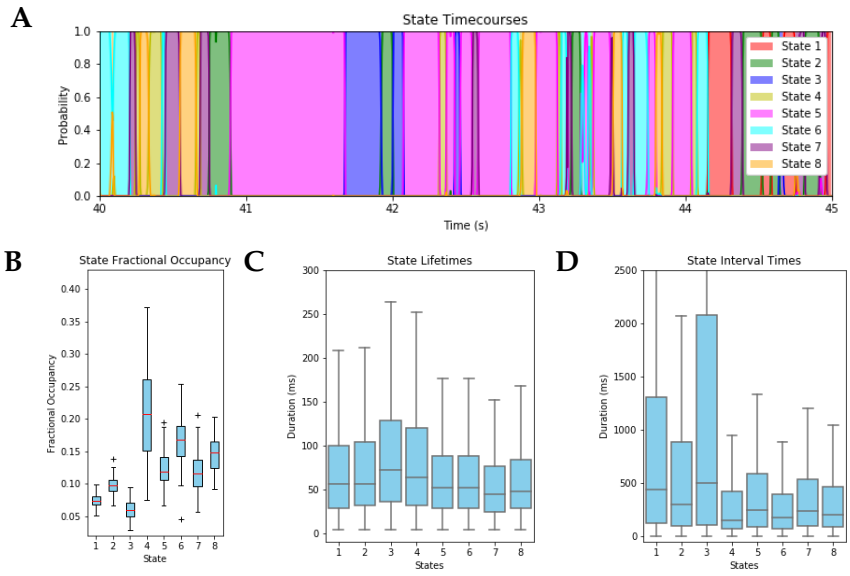

Figure 6.4: State time course example and temporal statistics of state activations computed on the state posterior time course for each subject. (A) Example state time course from between the first 40 and 45 seconds of the scan of the first subject. Each background color indicates the estimated active state, while posterior distributions for each state are denoted by the solid lines for each associated colour. (B) Fractional occupancy across subjects reported for each state. (C) State lifetimes, also known as dwell times, were computed by calculating how long each state visit lasts. (D) State interval times, defined as the time elapsed between successive state visits. The state lifetimes and interval times reported are across all subjects.

states have maximum lifetimes of over a second. State 4 in particular appears to have higher range of lifetimes, with lifetimes of over 2 seconds occurring. We find that approximately 20% of the time is spent in state 4 on average, and on average the lifetime is also greater for state 4. Conversely, states 1, 2 and 3 appear to be visited less frequently. Interestingly, we see another group of states, 5-8 that exhibit relatively high fractional occupancy (between 11 and 16%); despite their short lifetimes (median lifetimes around 50 ms), the states exhibit comparatively low interval times.

We then looked to see whether there were any particular transitions between pairs of states that occurred particularly frequently, or that were particularly rare — these might indicate a commonality between the states. We counted the number of transitions occurring between

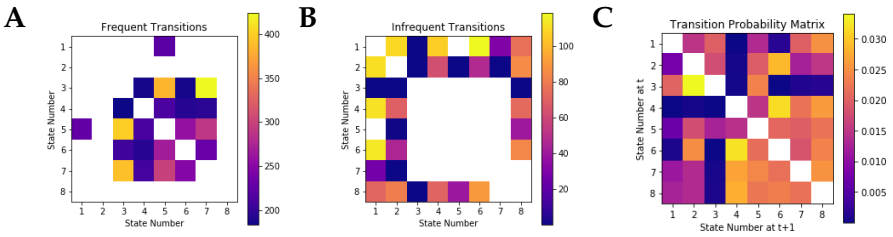

Figure 6.5: State transition frequencies. (A) Average number of transitions (across subjects) between pairs of states occurring significantly more than on average. (B) Average number of transitions between pairs of states occurring significantly less than on average. In both images, the position in the (i, j) element of the matrix denotes a transition from state i to state j. Only those with significance differences ($p < 0.0009$) between the means are shown in each figure. The colour map indicates the average number of each transition occurring in each subject. (C) State transition matrix from the HMM.

each pair of states across subjects, and then recorded transitions for which there was a significant difference in average (we did not test self-transitions, so a Bonferroni correction was applied to account for the 56 other possible pairs of states). This information is explicitly encoded within the transition matrix of the HMM, however a subsequent analysis of the HMM transition probability matrix did not appear to show this same structure, as shown in Figure 6.5C

We see that there appear to be two clusters of states; a core group that transition between each other frequently, comprised of states 3-7, and a peripheral group made up of states that are transitioned into much more rarely (states 1, 2, and 8). The core group is made up of a sensorimotor networks (state 5), a visual network, an left lateralised network, a frontal network, and state 3, which given its increased interval time, may be somewhat analogous to a default mode. Conversely, the peripheral group is made up of two states associated with the frontal lobe; a frontal DMN and what may be referred to as an anterior DMN. Given that the frontal lobe is responsible for executive functions, it is unsurprising that they display reduced transition frequency in the context of a resting state scan.

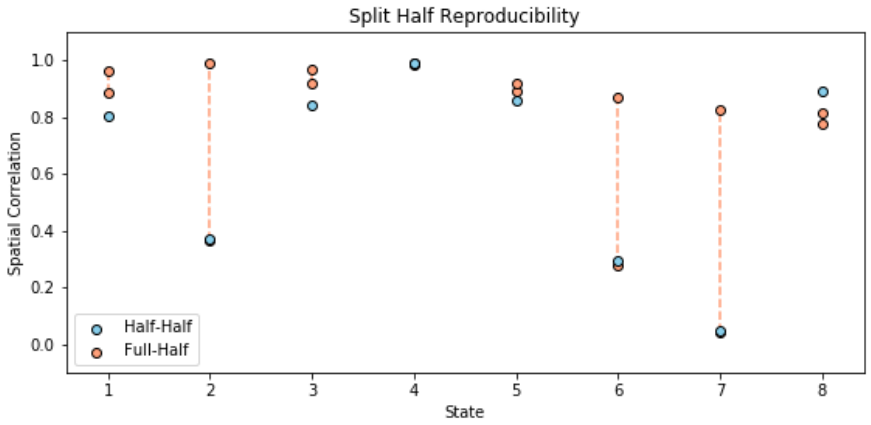

Figure 6.6: Shown is the split-half reproducibility of the recurrent model. Each state from both halves of the data was paired to the corresponding state in the full data, and the spatial correlation was recorded. Both the correlation between the spatial maps of each half of the data was calculated (Half-Half), as well as the correlation between the spatial maps inferred on half the data and the full data (Full-Half).

6.2.4 *Reproducibility*

While we have shown that our model returns a number of states fairly reliably when applied to the whole data set, however we would also like to assess the reliability of the model when applied to entirely different sets. One way to do this is by way of split-half testing; we split the complete set of 55 subjects into 2 non-overlapping groups of subjects (27 and 28 subjects each). Each half was then used to train a different model with 8 states. Initialisations were obtained using the methodology defined above, where we initialised the final models with the top 8 most stable states within each respective half of the data set. We then independently paired the spatial maps of each state to those trained on the full data, and reported the spatial correlations between each corresponding state, as shown in Figure 6.6.

These results suggest that of the 8 states we have inferred, at least 5 of them (states 1, 3, 4, 5, and 8) occurred reliably even in a reduced subset of the data. We can therefore have some degree of confidence

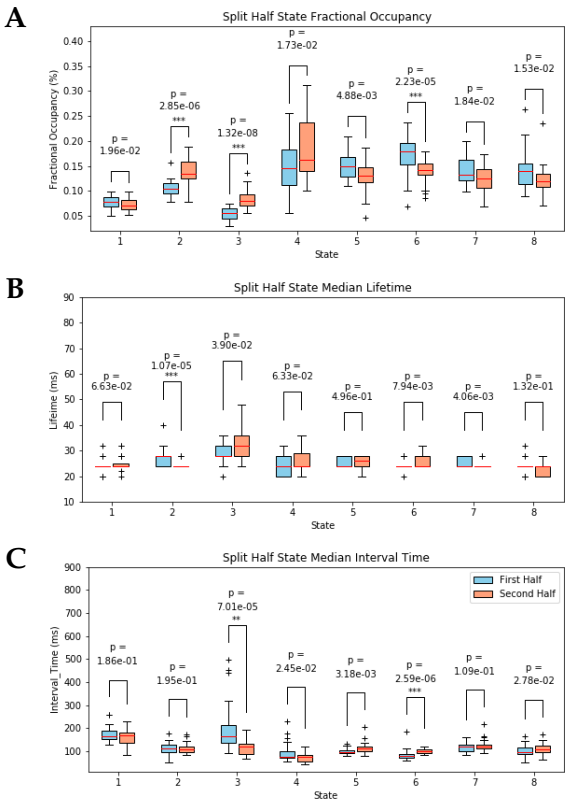

Figure 6.7: Temporal statistics of state activations across subjects compared between the two halves, in particular (A) the fractional occupancy, (B) the median lifetimes, and (C) the median interval times. We report the median as the long tailed nature of the distribution makes visualisation difficult.

that these were not simply spurious states. We also see that each state from the full data was represented well (over 0.8) in at least one half of the data set, though the between-half reproducibility of states 2, 6 and 7 are both fairly low at 0.37, 0.29 and 0.05, respectively. This could indicate that there is more natural variation across subjects within these states, and therefore the group average was not well represented within the subset of subjects used in the respective half. This spatial reproducibility largely indicates that we have not simply found a local minima, however perhaps the model performance would be increased if we could account for the natural variation that occurs within the states.

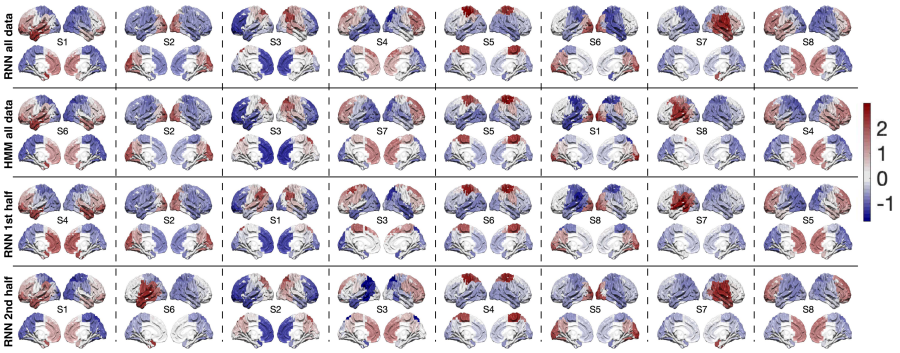

Figure 6.8: Mean activation z-maps for the recurrent model derived from the entire data set, from the HMM, and from each recurrent model run on non-overlapping halves of the data. Spatially similar states have been aligned such that they lie in the same column, and are ordered from left to right in order of average stability of the states in the full recurrent model.

As shown in Figure 6.7, we see that many of the temporal features of each state are highly reproducible (we again applied a correction such that $\alpha = 0.05/24 = 0.002$). We see some differences in the features of the second and sixth states, both of which were states that exhibited significant spatial variance. Additionally, a comparison of the lifetimes across all states with the full dataset shows that the model is finding states with on the order of half of the lifetimes found in the model trained upon the full dataset. The cause of this is highly likely to be noisier estimates of the posterior distribution in the presence of less data, which results in smoother state time courses in the full data set. While this does potentially flag up some issues for smaller datasets, we have nonetheless shown that we have identified a relatively parsimonious set of states that not only have reproducible spatial features, but also have reproducible temporal features, and then inferred the most probable state at each time.

6.2.5 *Comparison with HMM*

Given our model has similarities with the HMM, it is natural that we make a direct comparison with a Gaussian HMM from the HMM-

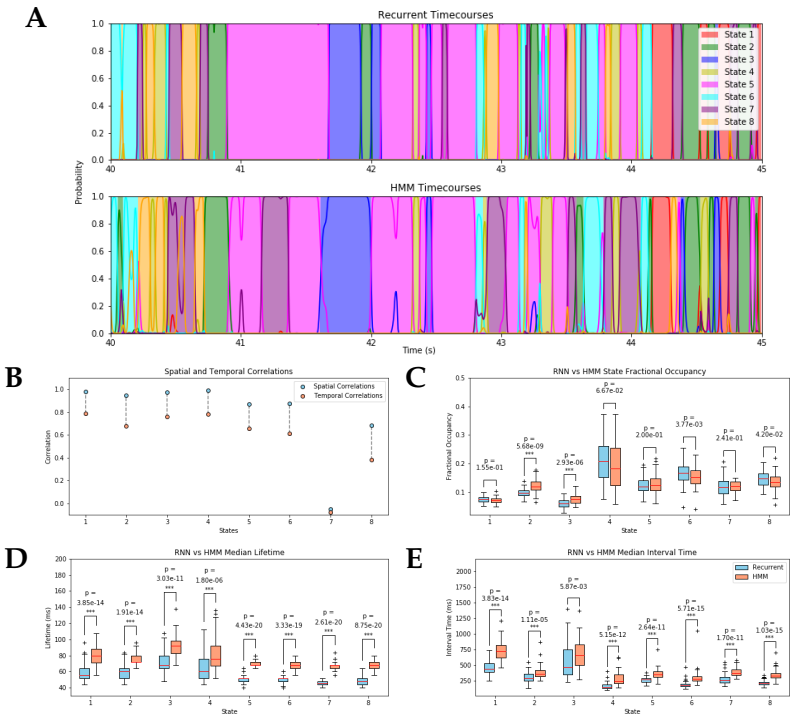

Figure 6.9: Comparison of the temporal features of the recurrent model with the HMM. (A) Comparison of two inferred state sequences; the recurrent model on top, the HMM on the bottom. (B) Spatial and temporal correlation between recurrent model and HMM. (C-E) Temporal features of both the recurrent model and the HMM compared. All p-values are reported, and an α of 0.002 was used for significance.

MAR toolbox. [120] Using the amplitude-envelope GHMM detailed in [3], we fully trained 5 realisations of an 8-state HMM with random initialisations. We performed inference upon the temporally concatenated amplitude envelopes

The ensuing spatial maps can be seen in Figure 6.8, although for larger images (and for the functional connectivity maps) refer to the appendix. Unsurprisingly given the observation models are the same, the two models found very visually similar activation patterns. Spatial correlations across all states are high, with the exception of state 7, the right lateralised network. The HMM seemed to detect the left lateralised network. A casual glance also shows that in the 6th state, the HMM identified reduced activation in this very same region, as well as the motor cortex. Interestingly, both of these states were picked

up by the split half models, perhaps giving reasonable due cause for the exploration of higher numbers of states. Spatial correlations, shown in Figure 6.9B, show high spatial correspondence across all states, with the exception of in state 7.

Considering the temporal correlations between the probabilistic time courses tells largely the same story; each recurrent model state time course is correlated with its HMM counterpart, with the exception of state 7. State 8 also exhibits reduced spatial and temporal correlations, as again the HMM appears to have identified the mirror image of the recurrent state.

A comparison of the time course for the recurrent model that we showed earlier with the HMM sequence for the same period illustrates this point in Figure 6.9A; there is high correspondence between the state sequences, with the exception of state 7. The activations of state 7, shown by the light maroon colour, are completely different in the two sequences. A comparison of the temporal features across these states in 6.9B-D demonstrates that the states have comparable fractional occupancies, with the exceptions of states 2 and 3. We do however find highly significant differences between the median lifetimes and interval times between each pair of states, with the recurrent model finding state lifetimes lasting around 20 ms less than the HMM, and shorter interval times by a similar margin.

6.2.6 *Generative modelling capabilities*

We have our models, and they produce comparable results when performing inference — so why bother? One question we had was given we have a more flexible model in the recurrent model, one that is not bound by the structure of the transition distribution, does this enable it to explicitly "learn" temporal dynamics of the data? The

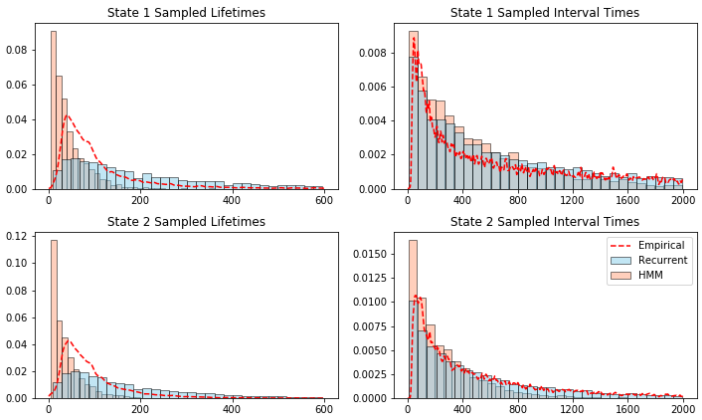

Figure 6.10: Temporal features of synthetic sequences generated by using the recurrent and HMM. Orange histograms show the features from sequences generated from the HMM, while light blue histograms show the features from the recurrent model-generated sequences. The red dashed line shows the empirical distributions derived from the data.

most trivial way to test to see whether it has learnt these dynamics is to generate new sequences, and then evaluate the temporal properties of the generated sequences. This is simply a case of feeding the generative model, whether it be a GRU or a transition probability matrix, a random seed state and sampling from the distribution of next states. We can then proceed forwards in time in an iterative fashion by using the current state to generate a new state.

We generated a total of 800 minutes worth of data from both models. An example of the lifetimes and the interval times for the resulting samples for states 1 and 2 can be seen in 6.10, and the full features can be found in the appendix. Unsurprisingly, the lifetimes of all HMM states are that of an geometric distribution, as the state lifetimes of a markov chain are essentially Poisson processes, and the state lifetimes are therefore geometric distributions, and likewise with state interval times. We can see that our empirical distributions do not take this shape, although the HMM appears to provide a relatively good approximation to the interval times. We see that the sequence generated by the recurrent neural network does a better job of approximating

the empirical distribution of state lifetimes in both cases. While the mass is not quite so focused on the median, and allocated further out in the tail, we nonetheless have recovered the median. In each state, despite not specifying any explicit form for the state duration, we seem to have roughly recovered the empirical dynamics, where the lifetime distribution is a highly skewed with a very long tail. This understanding of the empirical distributions would be particularly helpful in the case of HSMMs, where we could use the features derived from this nonparametric empirical distribution to inform future model specifications, for example choice of lifetime distributions.

6.3 SUMMARY

In this chapter, we have introduced a robust way to estimate the total number of stable states that can be used to model the dynamics of the brain as a multivariate state switching model. Our model was able to identify a total of 8 stable networks that were reproducible across subsets of the data and in other models, however we found some indication that further such networks exist. In particular, we also identified additional states that appeared to be the mirror images of each of the lateralised networks identified by the model. From this we can conclude that 8 states is perhaps a conservative value for a complete set of descriptive states — although we did see that increasing the additional states increases the complexity of the model and increases the propensity for the optimisation space to have more local minima, as shown by the increasing variance on the complete free energy values across realisations in Figure 6.1. The same holds true when considering the inherent stability of the resulting states; more states result in increasing total groups of identified states across realisations.

In terms of the temporal dynamics, we saw that these states act over very fast timescales, with state lifetimes lasting on the order of 100 ms. We saw that we could broadly stratify the groups into 2 groups on the basis of transition frequency; one core group of five states (S3-7) that appear to be very strongly linked with each other, very frequently transitioning between one-another. The second group (S1, 2, 8) exhibits much lower frequency of occurrence..

One point we noticed about the temporal dynamics of the states is that the state lifetimes vary significantly; the maximum lifetime of state 4 for example, is over 4 times that of state 7, while the median lifetimes vary by a factor of two. This indicates that care should be taken when performing a conventional sliding-window based analysis on functional connectivity, as application of certain fixed window lengths may result in merging multiple states together — a model such as the recurrent model or the HMM could be applied initially to decode the data to better inform the window lengths.

We then compared the resulting spatial maps and state time courses of the recurrent model to those of the HMM. We saw that despite the memory-less temporal model used by the HMM, it was nonetheless able to successfully decode the data and achieve in comparable results to those of the recurrent model. When evaluating the performance of the two generative models for explicitly capturing the structure of the temporal dynamics within the transition function, the recurrent model proved the superior. While the HMM is limited by the form of the transition probability matrix, it is possible to combine the HMM with parametric distributions over the state lifetimes to learn the state dynamics, as in the case of hidden semi-Markov models [115]. However, the advantage of the recurrent model lies in the fact that we are able to model complex dynamics without specifying properties of the model a priori. In theory at least, recurrent neural

networks are universal function approximators, and therefore should be able to learn any arbitrary transition function. In practise, this is largely subject to having access to sufficient amounts of data and training time. We saw that for the 55 subjects here (and the particular hyperparameters used in the case of this model), we were able to approximately recover the distribution over state lifetimes — while the medians of the sampled sequences aligned with that of the empirical distribution seen in the data, more mass was allocated in the tail of the distribution in the sampled sequences than was seen in the data. It should be noted that we did not subject the model to an exhaustive hyperparameter search to find the optimal distribution, so it may be likely to achieve superior performance given the quantity of data we had available.

7

VISUAL TASK RESULTS

In the previous chapter, we applied the model to resting state data. As noted, there is no way of objectively assessing the "correctness" of the inferred results. In this chapter we aim to apply the model on a unsupervised dataset in the form of an MEG face processing task performed by 19 subjects. In lieu of known labels, we therefore used the task analysis to find behaviourally relevant states and time courses, and then examined the plausibility of the within-state activation and connectivity patterns given the particular task they were associated with. To begin with, we shall first explore the appropriate number of states with which to train the model with. We shall then proceed in a similar manner as the previous chapter involving resting state analysis, however we shall focus particularly on identifying task-associated states and their properties. Further, we will aim to identify key features that relate to task-processing, both temporal and spatial in nature.

7.1 DATA

We used open source data MEG data taken from a study performed upon members of the MRC Cognition and Brain Sciences Unit participants panel (Wakeman and Henson, 2015 [123]; Revision 0.1.1). A total of 19 participants were scanned, who were between the ages of 23 and 37, and of which 8 were female and 11 were male. Subjects were

scanned for both the MEG experiment and so as to obtain a structural MRI.

7.1.1 *Experimental design*

The experiment was conducted over the course of 6 MEG scans during which they completed a visual perception task. The stimuli consisted of two sets of greyscale photographs, half of which were of famous people selected in order to be recognisable to the majority of British adults, while the second half consisted of non famous people unknown to the participants. These photographs were cropped to show only the face. An additional set of scrambled faces was generated by applying a set of transformations to the famous and nonfamous faces, and then cropping to a mask made up by a combination of a famous and a non-famous face. Each trial began with a fixation cross for a duration of between 400 and 600 ms, after which the stimulus (face or scrambled face) was presented for a duration of between 800 and 1000 ms. To ensure attention, the participants were asked to press one of two keys with either their left or right finger on the basis of how symmetric an image was, and reaction time was recorded. For full details of the experimental paradigm, refer to the original study [123].

7.1.2 *Data acquisition*

Acquisition for the MEG data was performed using a 306-channel Elekta Neuromag Vectorview 306 system (Helsinki, Finland). Subjects were seated in the scanner, and the stimuli was presented over six runs of 7.5 minutes in length, and the data was acquired at a sampling rate of 1100 Hz with a 350 Hz lowpass filter. A 3D digitiser was used to record the locations of the HPI coils relative to the nasion, left and

right preauricular points. The resulting data was then processed with a MaxFilter 2.2 with the aim of removing spurious data that originated from external noise sources. MRI data was collected at a standard 1 mm isotropic resolution using a Siemens 3T TIM TRIO (Siemens, Erlangen, Germany). The structural scans were obtained for MEG coregistry, and used an MPRAGE sequence. They were additionally de-faced for anonymity, as this was a publicly available dataset.

7.1.3 *Data preprocessing*

As the structural scans did not have all Polhemus head shape points due to the de-facing, coregistration between the structural MRI and the MEG geometry was performed using only the Fiducial landmarks. The MEG data was then downsampled to 250 Hz and bandpass filtered from 1-45 Hz. The data was cleaned by performing artefact detection and rejection, which essentially identifies outliers within non-overlapping windows on the basis of high variance.

As the data was collected using an Elekta Neuromag system, which contains both Magnetometers and Planar-Gradiometers, it is necessary to correct for the different variances for each sensor type during beamformer estimation by normalising each type of sensor prior to beamforming. The data was projected onto an 8 mm grid spanning the MNI152 brain with the aid of an LCMV beamformer, using each session independently. We parcellated the source localised data into 39 distinct brain regions, and estimated the timecourse for each of these regions from the first principal component across voxels, with voxel contributions weighted by the parcellation. We corrected for the effects of spatial leakage using symmetric multivariate orthogonalisation [25].

7.1.4 *Amplitude envelopes*

We obtained the amplitude envelopes using the Hilbert transform, and then smoothing using a 100 ms moving average filter. We then demeaned and normalised the data using the global variance across all parcels, and temporally concatenated over subjects.

7.1.5 *Model training*

We split the data into sequences of length 1000 timesteps, and randomly sampled 40 of these sequences for each minibatch. The model was trained over the course of 100 epochs using the inference algorithm detailed in previous chapters. The model was trained unsupervised, i.e. the model had no knowledge of the different trials and the task structure. GRUs with 2 layers and 32 hidden cells were used in both the inference and the prior model, and the parameters were optimised with the Adam [69] optimiser, using a learning rate of 0.001. KL annealing was performed by scaling the KL-divergence by a logistic function with mean 30 and scale parameter 0.5.

7.2 ANALYSIS

As with the resting state analysis, before inferring our model, it is necessary to first specify the total number of states that we will use. We proceed in largely the same manner as with the resting state MEG data in chapter 6. We first trained models with 2, 4, 5, 6, 7, 8, 9, and 10 states over a total of 8 realisations each. Figure 7.1 shows the resulting average free energies across the whole of the data achieved for each number of states. As with the resting state data, we see that as the

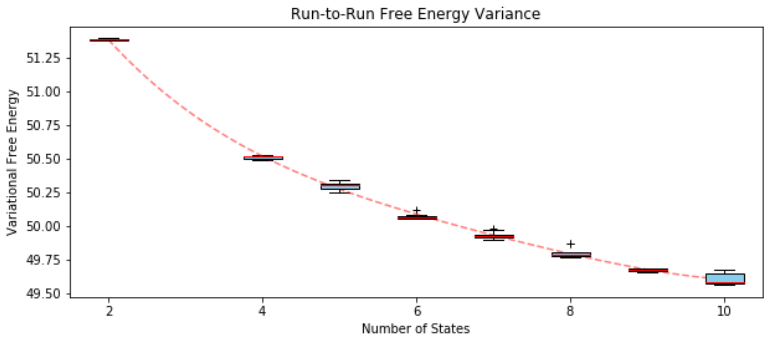

Figure 7.1: 8 different realisations of the model were obtained for various values of K, each initialised with a different random seed. The average variational free energy was then computed for the entire data set, as shown by the box plots. The red trend line is for illustrative purposes only.

number of states increases, the fit of the model is improved, although the improvement is reduced with each additional state.

While we could simply specify the model with the lowest free energy, in the interest of reproducibility we aim to obtain a "stable" set of parameters with which to initialise the observation models. As variational inference can be prone to getting stuck in local minima, we would like to ensure that states consistently appear across multiple realisations. We shall again cluster the states into those with the highest degree of spatial correlation across realisations using the RV coefficient for each number of states.

Upon grouping and sorting the states in the order of descending stability (whereby we consider state 1 to be the most stable state), we are left with an idea of the state landscape across each specification of K. We show the resulting states and their associated stability for $K = 8$ and $K = 9$ in Figure 7.2 (refer to the appendix for the full graphs), both of which identified roughly 9 stable states (although $K = 8$ identified an additional pair that are roughly 50% correlated with each other). Owing to the better model fit achieved by using 9 total states, as evident by the free energy, we decided to use 9 states

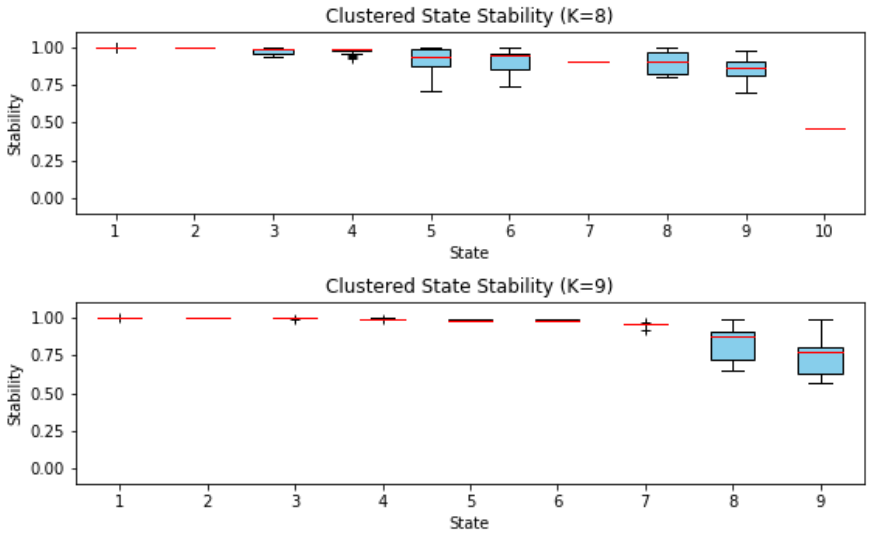

Figure 7.2: State stability shown for K = 8 and K = 9. Both values find 9 total stable states, although K = 8 also identifies an additional set of poorly matching states.

for the final analysis, and selected the run with the lowest free energy as the final realisation.

7.2.1 *Spatial maps*

We extracted the spatial maps from the observation distributions of each state. We have visualised the activation maps, which correspond to the average power associated with each region of the brain, in Figure 7.3A, while Figure 7.3B shows the functional connectivity in each state given by the parcels containing the strongest covariance. As before, we show the complete, unthresholded z-maps, as a choice of thresholding can be somewhat arbitrary, and the complete maps show a more complete picture than simply showing the strongest areas. The maximum and minimum values for the colour maps for both the activations and functional connectivity were set to be the same across all states to better show relative activity levels.

Similar to the resting state analysis in Chapter 6, we appear to have recovered a number of specialised connectivity networks. State 1

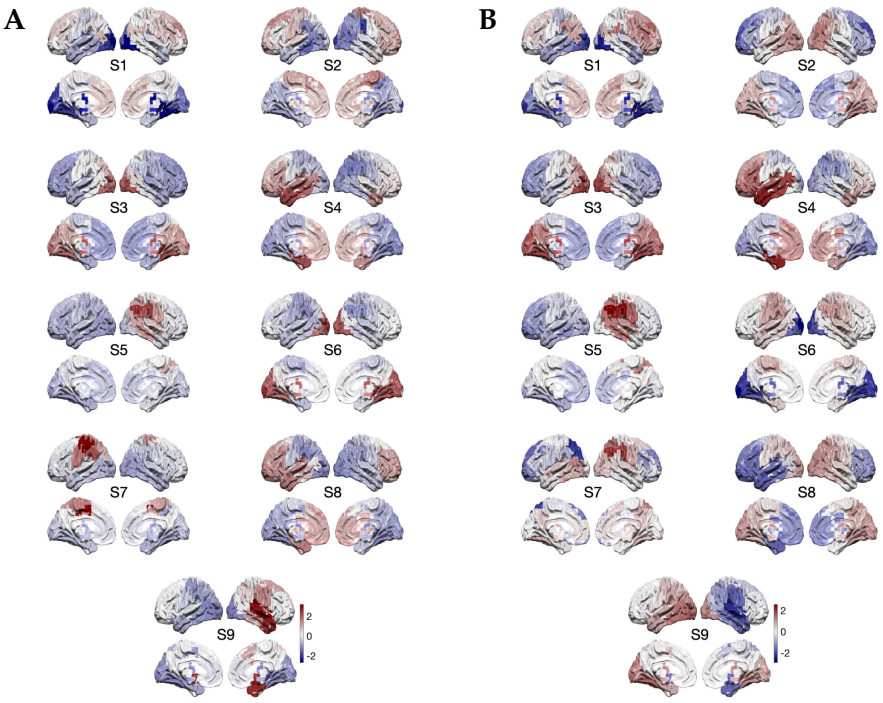

Figure 7.3: Unthresholded spatial distributions of power and connectivity. (A) The activation maps inferred by the final model, ordered by stability. The surface rendering shows the activation by way of the z-score for each state across the left and right lateral, and the left and right medial positions. Brighter red regions indicate stronger power, while brighter blue regions indicate lower than average power. (B) Within-state functional connectivity. Brighter red regions indicate higher connectivity, while blue means less connectivity.

appears to represent a default mode network, while state 2 exhibits power through the frontal and sensorimotor regions. States 5 and 7 both show increased power in the right and left lateralised parietal regions respectively, with the other lateralised hemisphere showing little to no activation. We also have state 9 that has power distributions focused in the right lateralised temporal lobe, and exhibits strong connectivity with the right temporal and occipital lobes.

As with the previous chapter, there is some indication of shared average power distributions across some states. For example, states 3 and 6 appear to be strongly visual states, with activation focused on the primary visual cortex, and deactivation across the frontal and temporal lobes. States 3 and 6 exhibit connectivity localised to the visual cortex,

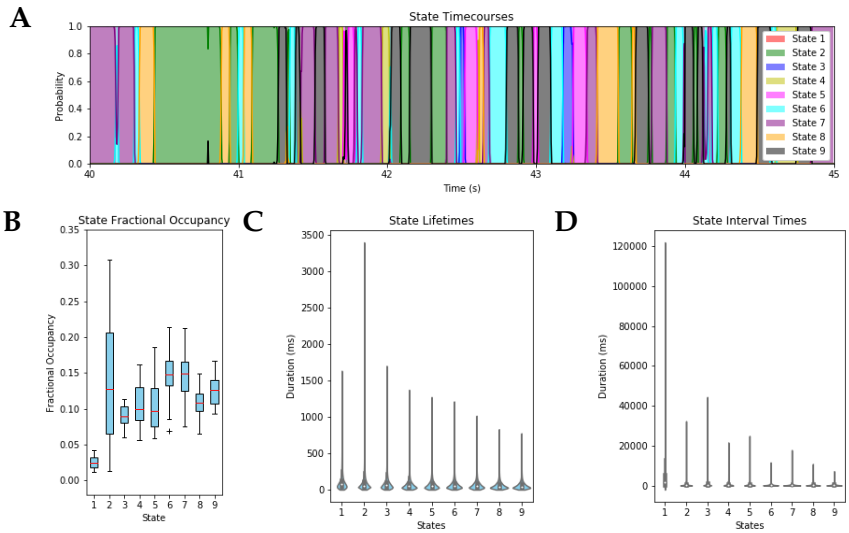

Figure 7.4: State Temporal Features. (A) Example state time course from between the first 40 and 45 seconds of the scan of the first subject. Each background color indicates the estimated active state, while posterior distributions for each state are denoted by the solid lines for each associated colour. (B) Average fractional occupancy across subjects reported for each state. (C) State lifetimes, also known as dwell times, were computed by calculating how long each state visit lasts. (D) State interval times, defined as the time elapsed between successive state visits. The state lifetimes and interval times reported are across all subjects.

and the parietal lobe respectively. Once again, similar activation patterns with markedly different characteristic connectivity patterns. States 4 and 8 are both active in the left lateralised temporal lobe and the frontal lobe, while connectivity within state 4 is high in the left temporal lobe/frontal lobe and low in the right parietal/occipital lobe. State 8 is again inverted; high in the right parietal/occipital lobe, and low in the left temporal and frontal lobes.

7.2.2 *Temporal dynamics*

Characterisations of the temporal properties of each of these states are shown in Figure 7.4. All of the nine states have a fractional occupancy of around 15%, with the exception of State 1. Median state lifetimes are in the region of 50-80 ms, and median interval times are around

400 ms, with the exception of state 1 which is on the order of between 1 and 2 seconds, with some intervals lasting as long as 2 minutes.

7.2.3 *Task-evoked occcupancies*

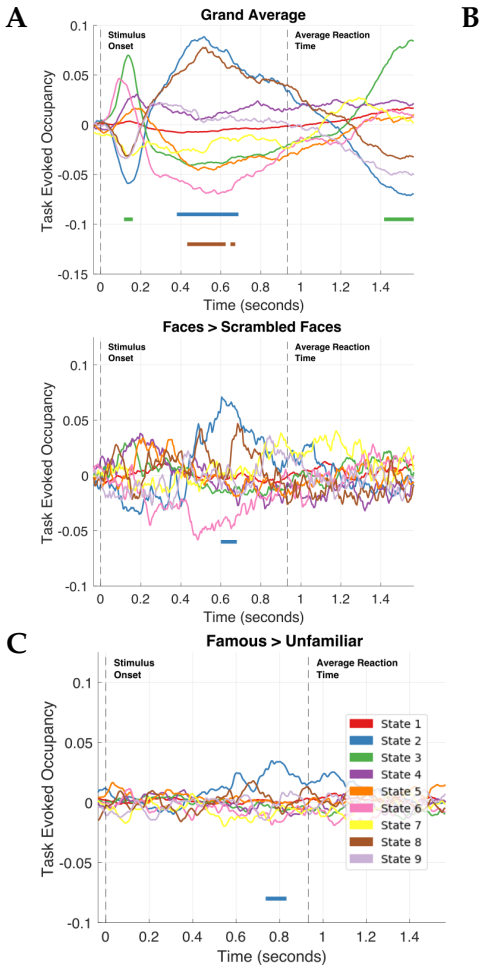

Figure 7.5: Group level task evoked factional occupancies. The results of the GLM analysis computed on the state posterior probabilities for (A) all trials, (B) the faces vs scrambled faces. Coloured bars below the fractional occupancies indicate periods of significant change.

We then wanted to investigate the link between the tasks and state time course. To obtain the task-dependencies of the state posterior probabilities, we followed the methods outlined by Quinn et al. [95]. This involved performing an event-related potential-like analysis upon the state time courses, resulting in an evoked fractional occupancy

giving the evoked fractional occupancy at any given point in the trial. From this evoked fractional occupancy, we can see the proportion of the trials for which the recurrent model was in any given state at each time step over the entirety of the trial.

Once we had the evoked fractional occupancies, we passed this into a two-level GLM. We first normalised the evoked occupancies by the baseline period before object onset (this is between 130 ms and 30 ms before the object onset). The first level of the GLM was to fit the evoked fractional occupancy for each participant across trials at each time point using a trial-wise design matrix. The second level computed the effect across participants, and modelled any between-subject variance as random effects. We made use of a design matrix with regressors for the mean and each of the trial types corresponding to famous faces, unfamiliar faces, and scrambled faces, and obtained parameter estimates for each of the regressors using 3 Contrast of Parameter Estimates (COPEs); one for the mean, and two for a contrast between the famous faces versus unfamiliar faces conditions, and the scrambled faces versus faces conditions. For the full details of the structure and implementation of the GLM, refer to [95].

The results from the GLM can be seen in Figure 7.5. Despite the increased number of states that we have used over the analysis by Quinn et al. (from 6 states to 9 states), we see largely similar results, particularly in the states that show significant increases in occupancy post-stimulus relative to baseline.

Following task onset, we very quickly see an increase in the occupancy within states 6 and 3, although only 3 shows significance. State 6 reaches a peak of about 5% evoked fractional occupancy within 90 ms post stimuli-onset — though we do not find this achieves significance, however state 3 shows a significant increase between 110 ms and 160 ms, peaking at 7% occupancy. We see no significant increases across

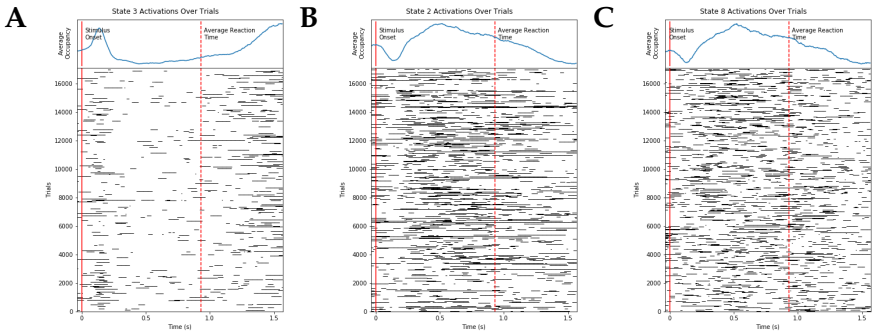

Figure 7.6: Trialwise state activation raster plots. State activations across every trial for (A) state 3, (B) state 2, and (C) state 8. The solid and dashed red lines indicate the onset of the stimulus and the average reaction time across all trials and subjects respectively.

the other contrasts, despite a slight increase for state 3 in a similar time period for the faces vs scrambled faces contrast. We also see a second, greater peak in the fractional occupancy of state 3 occurring at the end of the task, roughly 1.4 seconds after the stimulus onset, peaking at 9%. Recall that both states 3 and 6 showed clear power in the occipital cortex relative to the other states. Interestingly, state 6, which showed connectivity with the parietal lobe, peaked before state 3, which exhibited connectivity primarily within the occipital cortex — we might have expected this to occur the other way around due to visual information cascading through the rest of the brain. Nonetheless, it is likely that both of these states are involved in early visual processing. We can compare the trial-wise occupancy of each of these states in Figure 7.6B and Figure 7.6C, which show that state 6 is much more consistently active over each trial out of the two states. State 3, though more sparsely active, does appear to be more isolated in the periods of activity to immediately after stimulus onset and at the end of the task, while state 6 appears more intermittently. The raster plots for all other states can be found in the appendix.

Subsequent to this early occipital response, we see a marked reduced occupancy in both states relative to the baseline, and a large and sustained response on the part of states 2 and 8. Following the

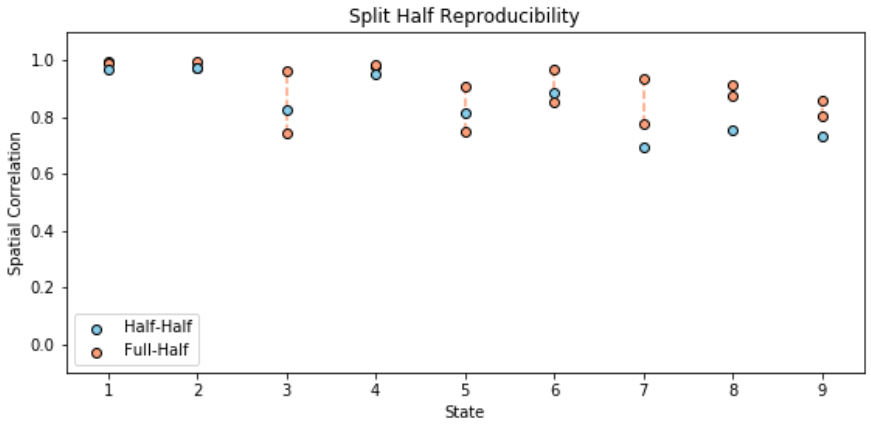

Figure 7.7: Split half reproducibility after matching all states to their nearest counterpart in the full data set. Both the correlation between the spatial maps of each half of the data was calculated (Half-Half), as well as the correlation between the spatial maps inferred on half the data and the full data (Full-Half).

reduction in the occipital response, state 2 peaks between 370 ms and 700 ms at around 9% occupancy, while state 8 reaches significance roughly 70 ms later, between 440 ms and 670 ms, at a slightly reduced 8%. By 1.2 seconds, both these states have started dropping below the baseline. We see from Figure 7.5B that state 2 has increased fractional occupancy in the faces compared with the scrambled faces, and shows a (delayed) slight increase in evoked fractional occupancy in the famous faces compared to the unfamiliar faces. Both states 2 and 8 were associated with high power in the frontal cortex, although we can see that the two other states associated with frontal cortex activity — states 1 and 4 — show no significant changes over the course of the task. The differentiating element between the two sets of states is that both states 2 and 8 show high connectivity with the visual cortex. The frontal cortex is largely responsible for decision making, therefore it appears that these two states are integrating information from the visual cortex in order to make a decision regarding which key to press.

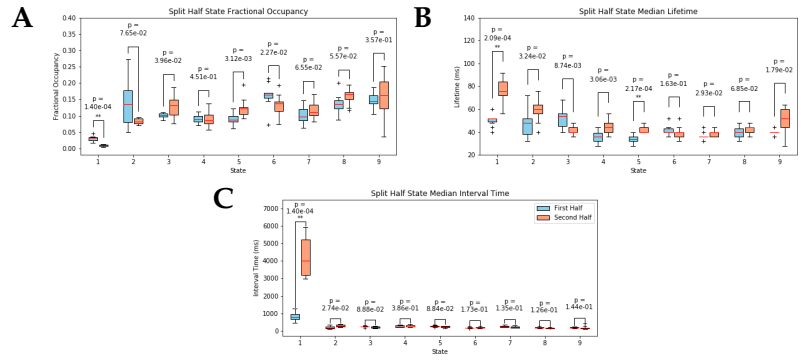

Figure 7.8: Temporal statistics of state activations across subjects compared between the two halves, in particular (A) the fractional occupancy, (B) the median lifetimes, and (C) the median interval times. We report the median as the long tailed nature of the distribution makes visualisation difficult.

7.2.4 *Split-half reproducibility*

We aimed to assess how consistent models would be if trained upon only a subset of the data, and so split the subjects into two groups containing a total of 10 and 9 participants in the first half and the second half, respectively. We assessed a number of metrics; the temporal metrics of the state fractional occupancies, their interval times, and their median lifetimes; and we also tested for the consistency of the spatial maps across the two split subjects, and across the full cohort, by way of measuring the corresponding spatial correlations.

The spatial correlations between the full data set and each subset of the subjects (Full-Half), as well as between the two halves (Half-Half), is shown in Figure 7.7. We can see that there are three states that are highly repeatable in particular; states 1, 2, and 4 all have correlations over .97. The remaining states show a lower degree of consistency, but all states remain highly correlated (greater than .7), suggesting that all of the identified states correspond to the same states, but there exists a higher degree of spatial variability across subjects.

The state posterior time course dynamics remain highly reproducible (Figure 7.8), for all states with the exception of state 1. Despite

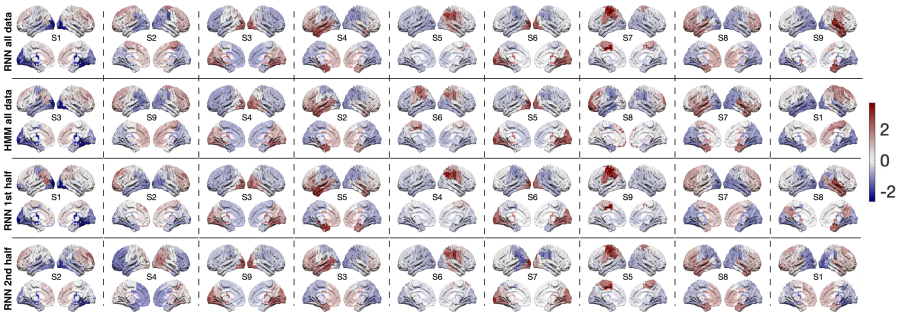

Figure 7.9: Activation z-maps for the recurrent model derived from the entire data set, from the HMM, and from each recurrent model run on non-overlapping halves of the data. Spatially similar states have been aligned such that they lie in the same column, and are ordered from left to right in order of average stability of the states in the full recurrent model.

being incredibly consistent spatially (correlations over 0.99), there nonetheless exists a large disparity between the dynamics of this state in each half; in the first half, the median lifetime is just under half in the first split (50 ms versus 80 ms), while the interval time differs by a factor of 4 (approximately 1 second versus 4 seconds). Both of these differences are highly significant (where a Mann-Whitney U test was applied with an $\alpha = 0.05/27 = 0.002$). The lifetimes of state 5 differs between the two sets also, though not quite as dramatically as the first state. It should be noted perhaps that neither state 5 nor state 1 proved to have particularly strong associations with the task, while we might expect states that were associated with the task (as in the case of states 2, 3 and 8) to be be consistent due to the repeated tasks undergone by each subject.

7.2.5 *Comparison with HMM*

As a second level of reproducibility checks, we look to see whether our results are repeatable across models (albeit ones that differ only in the temporal domain). We temporally concatenated the data, and trained the HMM in a completely unsupervised way such that it remained

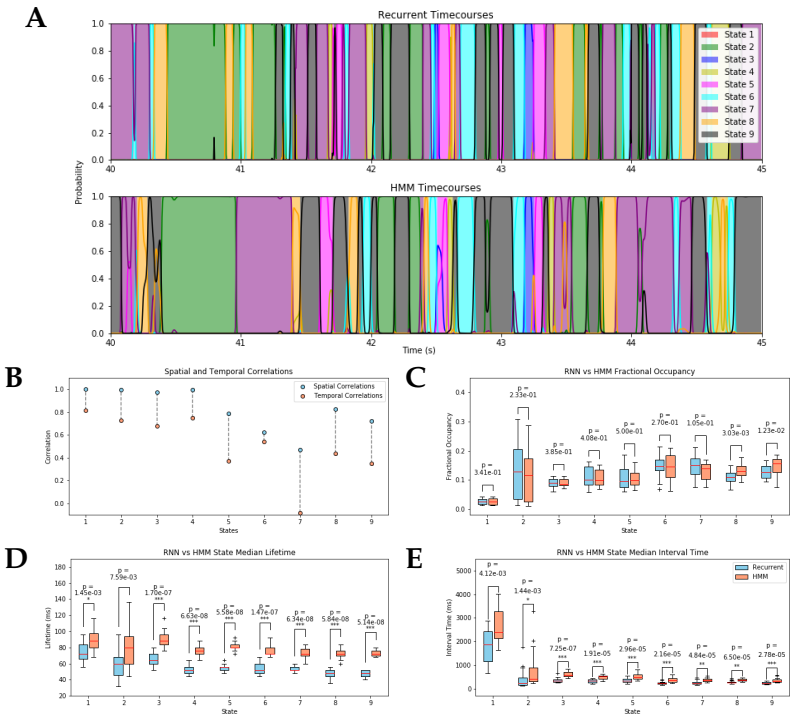

Figure 7.10: Comparison of the temporal features of the recurrent model with the HMM. (A) Comparison of two inferred state sequences; the recurrent model on top, the HMM on the bottom. (B) Spatial and temporal correlation between recurrent model and HMM. (C-E) Temporal features of both the recurrent model and the HMM compared. All p-values are reported, and an α of 0.002 was used for significance.

blind to the tasks, as with the recurrent model. We then randomly initialised the HMM using 9 states across a total of 5 realisations, and picked the one with the lowest free energy to ensure we were as close as possible to a global minima.

We present the spatial maps, aligned as best as possible to the spatial maps that were obtained from the recurrent model, in Figure 7.9. There is strong correspondence across *almost* all of the states identified by the two models, however there do appear to be a few differences; namely, while the recurrent model broke up each of the lateralised parietal lobes into distinct states (states 5 and 7), the HMM grouped them together into a single state (state 6). The extra state gained appears to be have power distributed primarily in the temporal lobes on both the right and left hemispheres.

Correlations between the state posterior distributions across time prove to be somewhat less correlated between the two models; whilst fractional occupancies are similar across all states (there are no significant differences, even between the cases where states are completely different), but then the fractional occupancies for all states bar 1 and 2 fell between 10% and 20%. Most states (with the exception of state one) also appear to share similar lifetimes and interval times — although the recurrent model detects lifetimes on average 22 ms shorter, and interval times 250 ms shorter.

7.3 SUMMARY

Using task data, we have demonstrated that even given completely different temporal models, the recurrent model and the hidden Markov model arrive at largely the same solution; that is to say, both models were able to identify a set of 9 stable neural states, all of which were largely reproducible across subjects, realisations, and models. Despite the fact that we impose no spatial constraints upon the states, we find the states identified in the model exhibit smooth contiguous increases and decreases of power and connectivity across the whole brain. Furthermore, unlike the states that were observed in the previous chapter, these states appear to be largely related to the task at hand, relating primarily to the visual cortex or the frontal lobes, reflecting the visual and perceptual elements of the task.

In particular, we identified 3 main states associated with the visual perception task given to the 19 subjects. We identified one state with an early occipital response occurring very shortly after the subjects were shown each stimulus, followed by a sustained frontal response in two distinct states. Of these three task associated states, only a single state showed any difference in the task-specific contrasts, showing increased

occupancy at around 700 ms in the famous faces vs unfamiliar faces, and 600 ms in the faces vs scrambled faces. This state (state 2) showed activation in the frontal lobe, and connectivity with the visual cortex, implying some degree of visual integration in the decision process.

As with the previous chapter, we also saw some degree of similar power distributions across pairs of states, where only the connectivity differed, specifically where we saw two behaviours for each; a 'disconnected' state where connectivity was specific only to the regions with a higher average power, and a 'connected' state, where the regions with higher power were functionally connected to the regions with lower average power, as in the case mentioned earlier with state 2.

8

CONCLUSIONS AND FUTURE WORK

This thesis has two main components; a methodological contribution regarding constructing flexible generative temporal models and performing inference with them, and a subsequent characterisation of the spatio-temporal dynamics of both resting state and task MEG data. We additionally compared the recurrent generative and inference models with a hidden Markov model. Below, we will summarise the major conclusions and discuss related future avenues for extending this work.

8.1 CONCLUSION

For the methodological contribution, the aim of this project was to explore the effect of relaxing the Markovian constraint, and whether moving to a more flexible generative model would result in significantly different temporal dynamics. To achieve these aims, we have introduced a novel extension of a hidden Markov model such that the state transitions are parametrised by a recurrent neural network. This model focused specifically on the modelling functional connectivity of the brain as a switching process, comprised of a discrete set of static connectivity states represented by a zero-mean multivariate Gaussian distribution, where the connectivity patterns of each state were encoded within each respective covariance matrix.

Standard inference routines are unfortunately not applicable to highly nonlinear functions like recurrent neural networks, so it was

therefore necessary to derive a new one. Unfortunately, existing stochastic gradient variational Bayes algorithms do not facilitate the learning of a temporal prior, rather they operate using fixed (typically uninformative) priors. The novel contribution of this algorithm was they key point of making use of ancestral sampling in order to train the temporal prior, which was optimised through the KL-divergence component of the free energy.

We primarily explored effect of relaxing the Markovian constraint through the exploration of the temporal and spatial dynamics of two sets of data: the resting state scans of 55 subjects, and the scans of 19 subject engaged in a face viewing task paradigm. Our analysis demonstrated that the temporal dynamics of the recurrent model were entirely consistent (though not a direct replication) with previously reported analyses. When we compared the inferred states and state time-courses from the recurrent model and the HMM, we found that the results were both highly spatially and highly temporally correlated, and shared similar temporal features. An additional visual comparison makes it clear that the two models are detecting comparable solutions.

In particular, we observed that the complex interactions that occur during rest can be decomposed into rapid switches between a set of states that bear much correspondence with resting state networks obtained through fMRI scans. We observed a "default mode" state, as well as other common RSNs: visual, sensorimotor, and frontal-parietal. We saw a similar pattern in the task recordings, although the observed states were far more task-relevant than those demonstrated during rest. Again we saw a state that appeared to bear strong resemblance to the DMN, both spatially and temporally, as well as some states that were highly significantly time-locked to the post-stimulus onset.

Unexpectedly, the recurrent model identified faster state switching than identified by the HMM, with lifetimes approximately 20 ms

shorter in duration. Owing to the HMM's inherent geometric distribution over state lifetimes, we expected it to have a slight bias towards shorter timescales, and thus we expected the recurrent model (which we hoped to not suffer from this bias) to show longer state lifetimes. Upon analysing the recurrent generative model in the resting state data, we did find concurrence with the inferred results (the medians were approximately equal).

What we have largely shown through this work is the flexibility of the HMM as a model. We see that it is highly robust to violations of the Markov assumption, even in the case of data with such complex dynamics as neural activity. Given the constraints imposed by both the discrete nature of the state space, and the static nature of the observation models, it may well be that by using the RNN we are essentially overcomplicating the model more than necessary.

8.2 FUTURE WORK

As always, the work here is far from complete, and there yet remains much work to be done in building on the results presented here, in terms of enabling the generative model to better capture the subject-level spatial variability in functional connectivity, exploring new model structured to exploit observations we have made over the course of this work, and in order to make the inference algorithm more robust.

Firstly, there are some issues with the way the model handles multiple subjects; namely, we temporally concatenate the data, and use a point estimate for the covariance matrices that define the functional connectivity. We are therefore computing group-averaged estimates of functional connectivity and the corresponding time-courses for these group averaged states. While it is true that we can still obtain subject-specific spatial maps through a post-hoc analysis, it increases the

utility (particularly in cases where the resulting networks are shown to be disrupted) of the model to explicitly account for subject-level variability. This could be implemented by way of a prior over the covariance matrices. In the HMM, Baker et al. made use of a Wishart prior. Unfortunately we are working within the limitations of the Pytorch framework, and they do not have efficient samplers for any distributions over positive-semi-definite matrices (e.g. Wishart or LKJ distributions). New features are being added every day though, and therefore this may be feasible in the near future.

Secondly, the actual inference routine needs some improvements to ensure it is sufficiently robust. While it worked for the amplitude-envelope data, simulations showed that the result of applying the model/inference to data with low amounts of signal reverted to showing a single state active, and the HMM was far more tolerable of low SNR. It may well be possible to improve this by way of a more intelligent sampling routine, for example importance weighting as in [16], or perhaps sequential Monte Carlo methods. We could also improve the initialisation of the model. As with all variational Bayesian approaches, even stochastic ones are highly sensitive to initialisations and will only ever converge to a local minimima. Given that the covariance matrices for each observation have no degree of uncertainty associated with them, this leads to a model that is highly sensitive to initialisation. It may well end up that by incorporating a spatial prior over each distribution, the model becomes less likely to end up in a poor local minima.

Finally, the inference algorithm presents us with the ability to quickly prototype and learn new models without all of the overheads typically associated with Bayesian inference. While traditional probabilistic models are often subject to numerous constraints to make learning tractable, or require complex mathematics to derive param-

eter optimisation schemes, with sequential SGVB it is all computed via automatic differentiation, meaning it is only necessary to derive some form of loss term for your model. Possible directions include the use of hierarchical models to explicitly represent the kind of temporal structure demonstrated in Chapter 6 by the dichotomy between the dominant and transient states, as well as the structure observed in [119]. There may also have been some evidence of the nonstationarity of states, as shown especially by the two visual states observed during the face viewing task. While they had comparable power distributions, their connectivity structures greatly differed. A hierarchical structure could likewise help here, where the first level states define the power distribution, and the lower level states define connectivity. Alternatively, we could abandon the Gaussian observation model entirely, and make use of some non-stationary model, as in the HMM-MAR [119]. The use of amortised inference could also prove valuable in the case of online learning, as future datasets simply require a forward pass through the inference network.

Part II

APPENDIX

A

APPENDIX A

A.1 SCORE FUNCTION ESTIMATORS

While pathwise gradient estimators are convenient for their low-variance, they are limited in the distributions they are amenable to. Fortunately, we are able to make use of score function estimators in other cases. The REINFORCE algorithm [101], which is also known as the score function estimator, makes use of the differentiation rule for the logarithm, colloquially referred to as the "log-derivative trick":

$$\nabla_\phi \, q(Z;\phi) = q(Z;\phi)\nabla_\phi \log q(Z;\phi). \tag{A.1}$$

We can use this to help calculate the gradient of the expectation of a function $f(Z)$:

$$\begin{aligned}
\nabla_\phi \, \mathbb{E}_{q(Z;\phi)}[f(Z)] &= \nabla_\phi \int_Z q(Z;\phi) f(Z) dZ \\
&= \int_Z \nabla_\phi (q(Z;\phi) f(Z)) dZ \\
&= \int_Z f(Z) \nabla_\phi \, q(Z;\phi) dZ \\
&= \int_Z f(Z) q(Z;\phi) \nabla_\phi \, \log q(Z;\phi) dZ \\
&= \mathbb{E}_{q(Z;\phi)}[f(Z) \nabla_\phi \, \log q(Z;\phi)],
\end{aligned} \tag{A.2}$$

where the term $\nabla_\phi \, q(Z;\phi)$ is called the *score function*, which has the nice property of having an expectation of zero. Since we now have a

way of moving the gradient to inside the expectation, we can compute unbiased estimates of 4.23 using Monte Carlo sampling:

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)] &= \int_Z \log p(X|Z;\theta_X)\nabla_\phi\, q(Z;\phi)dZ \\ &= \mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X)\nabla_\phi\, \log q(Z;\phi)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \log p(X|Z^{(s)};\theta_X)\nabla_\phi\, \log q(Z^{(s)};\phi),\end{aligned} \tag{A.3}$$

or in the case where we cannot calculate the analytic KL term, as in Equation 4.20,

$$\begin{aligned}h\nabla_\phi \mathbb{E}_{q(Z;\phi)}[\log p(X|Z;\theta_X) - q(Z;\phi)] &= \int_Z (\log p(X|Z;\theta_X) \\ &\qquad - q(Z;\phi))\nabla_\phi\, q(Z;\phi)dZ \\ &= \mathbb{E}_{q(Z;\phi)}[(\log p(X|Z;\theta_X) \\ &\qquad - q(Z;\phi))\nabla_\phi\, \log q(Z;\phi)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} (\log p(X|Z^{(s)};\theta_X) \\ &\qquad - q(Z^{(s)};\phi))\nabla_\phi\, \log q(Z^{(s)};\phi),\end{aligned} \tag{A.4}$$

where $Z^{(s)}$ are samples drawn from $q(Z;\phi)$, and S is the number of samples drawn. This assumes it is possible to draw samples cheaply, however it places minimal restrictions on the nature of $\log p(X|Z;\theta_X)$ — it does not even have to be differentiable to estimate the gradients of its expected value.

In practice, however, the variance of the estimator can be very large, as sampling rare values of Z can lead to large scores (and therefore high variance). There have been multiple strategies designed to address this issue; common methods are Rao-Blackwellization and control variates [89, 98].

B

APPENDIX B

B.1 RESTING STATE NETWORK STABILITY

We present here the number of stable states inferred for each specified value of K.

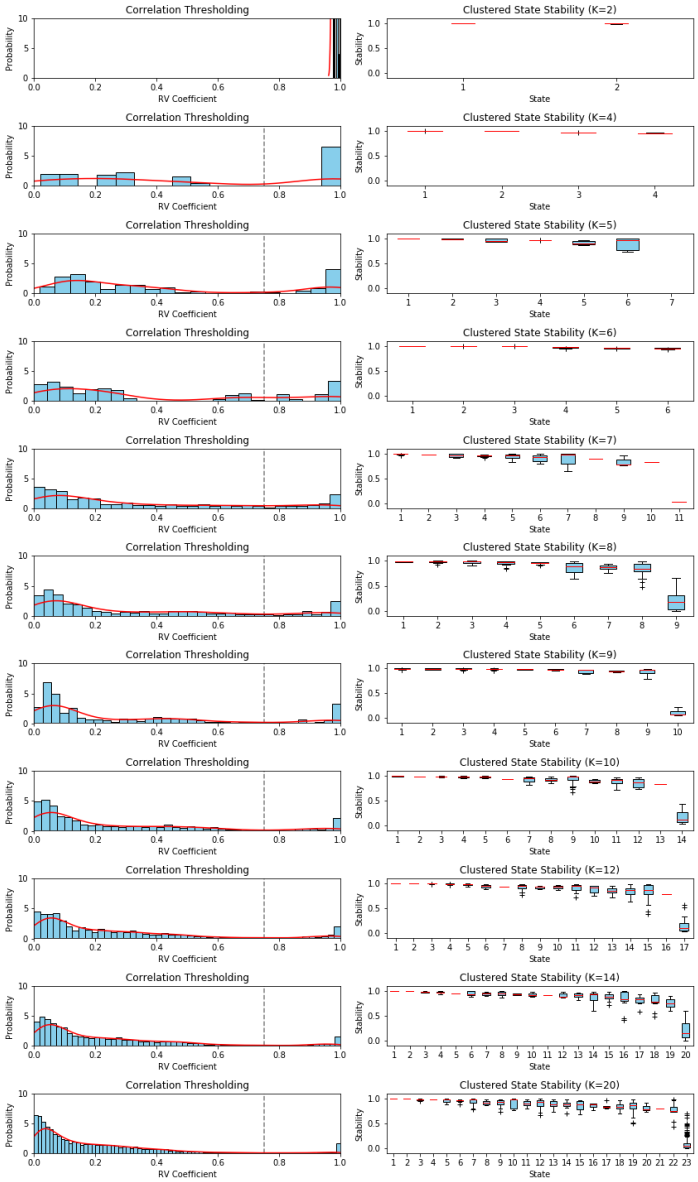

Figure B.1: Groupings of states by stability. The histogram shows all correlations across each run, and the dashed line indicates where the chosen threshold for grouping similar states was made.

B.2 RESTING STATE NETWORKS

In the following section, we show all spatial power distribution and connectivity maps for the full recurrent model, the HMM, and each split half.

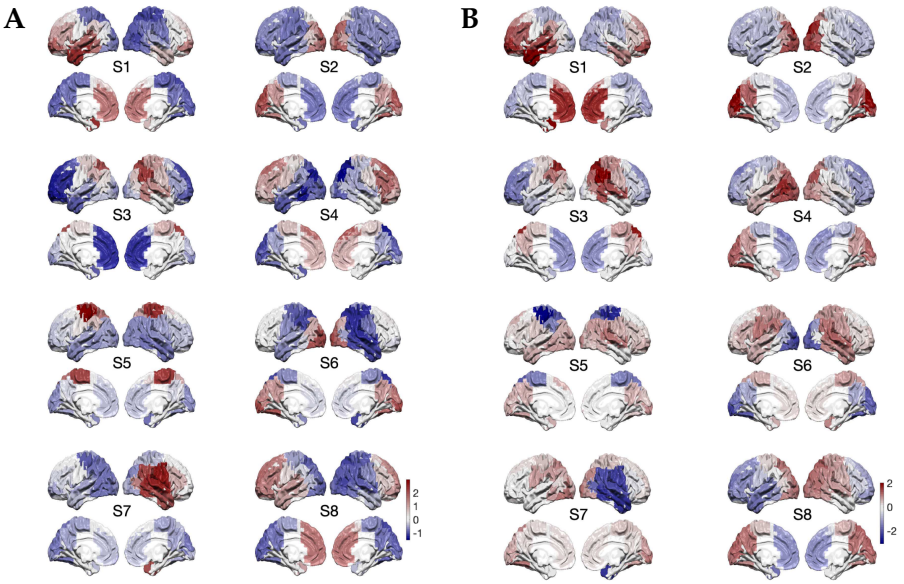

Figure B.2: Unthresholded spatial distributions of power and connectivity for the full data inferred by the recurrent model. (A) The mean activation maps inferred by the recurrent model applied to the entire data set, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

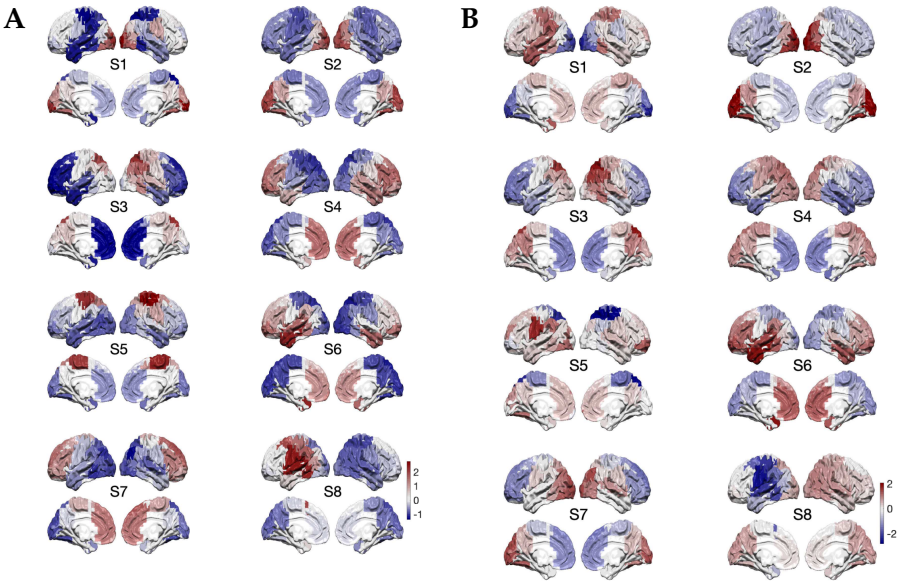

Figure B.3: Unthresholded spatial distributions of power and connectivity for the full data inferred by the HMM. (A) The mean activation maps inferred by the HMM applied to the entire data set, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

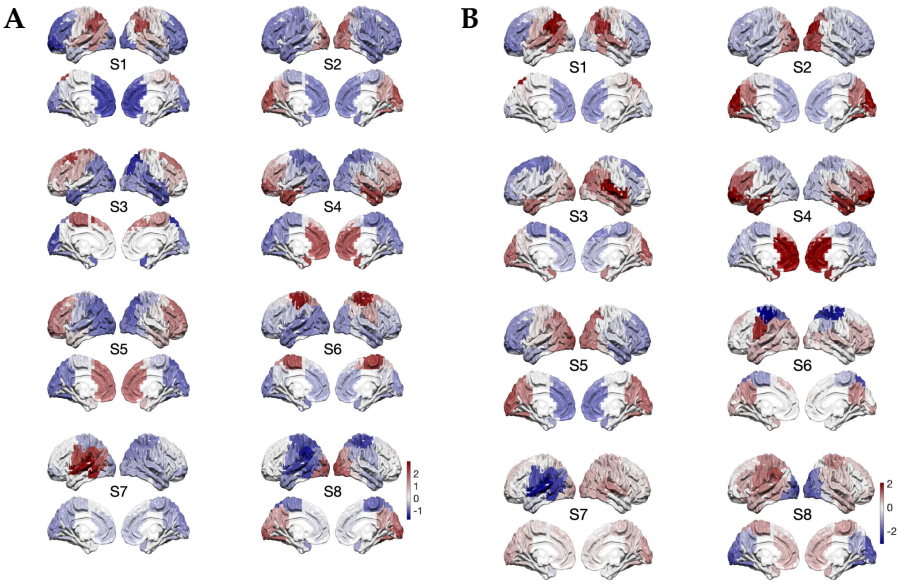

Figure B.4: Unthresholded spatial distributions of power and connectivity for the first half of the data inferred by the recurrent model. (A) The mean activation maps inferred by the recurrent model applied to the first half of the dataset, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

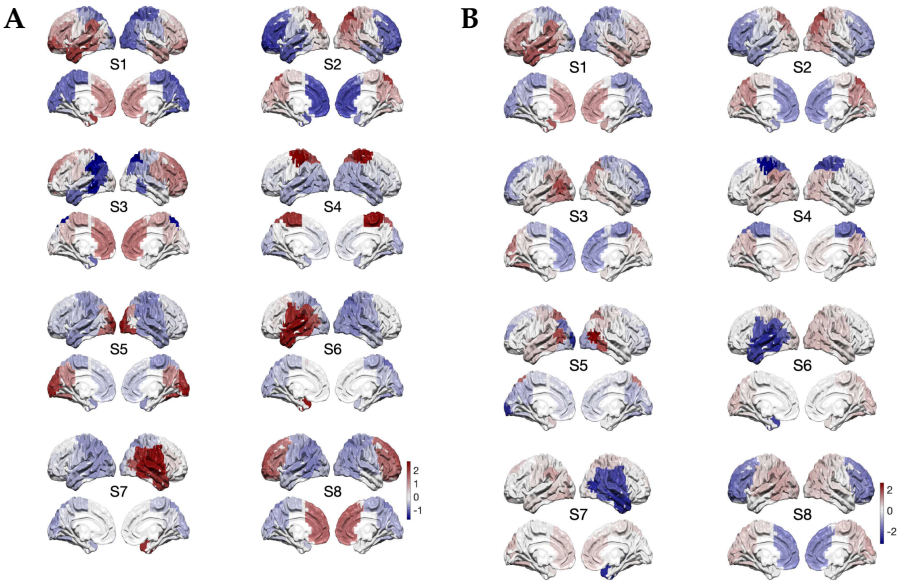

Figure B.5: Unthresholded spatial distributions of power and connectivity for the second half of the data inferred by the recurrent model. (A) The mean activation maps inferred by the recurrent model applied to the second half of the dataset, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

C

APPENDIX B

C.1 TASK STATE NETWORK STABILITY

We present here the number of stable states inferred for each specified value of K.

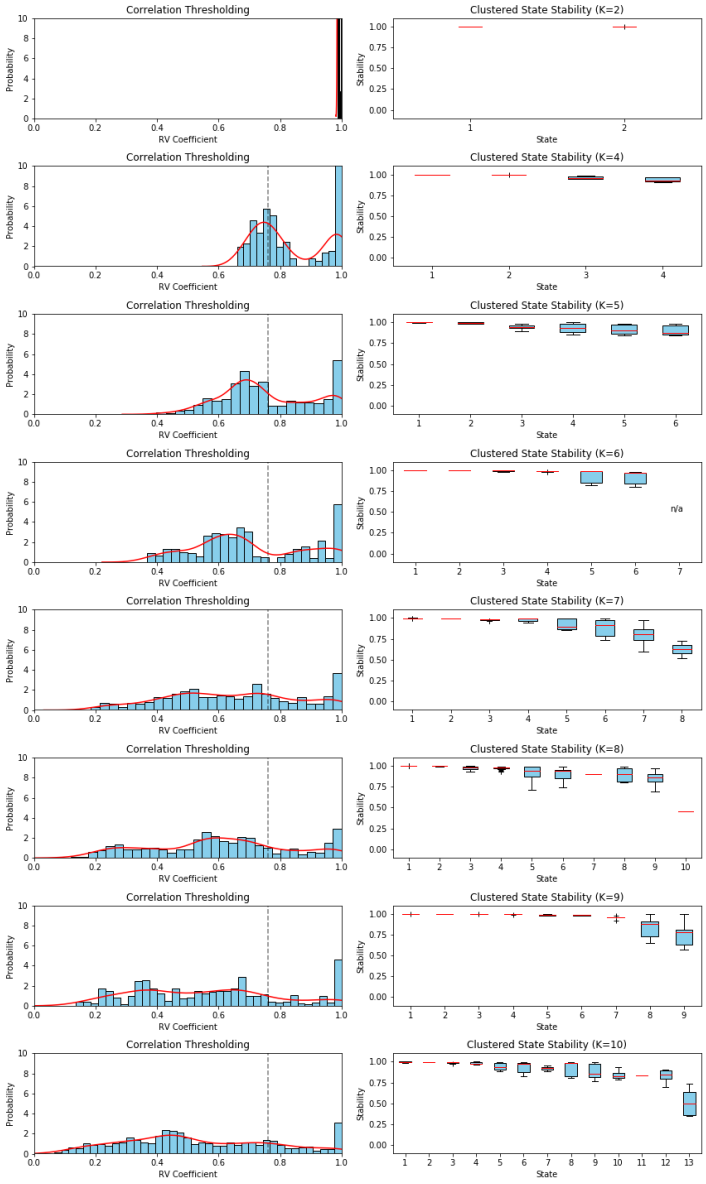

Figure C.1: Groupings of states by stability. The histogram shows all correlations across each run, and the dashed line indicates where the chosen threshold for grouping similar states was made.

C.2 TASK STATE NETWORKS

In the following section, we show all spatial power distribution and connectivity maps for the full recurrent model, the HMM, and each split half.

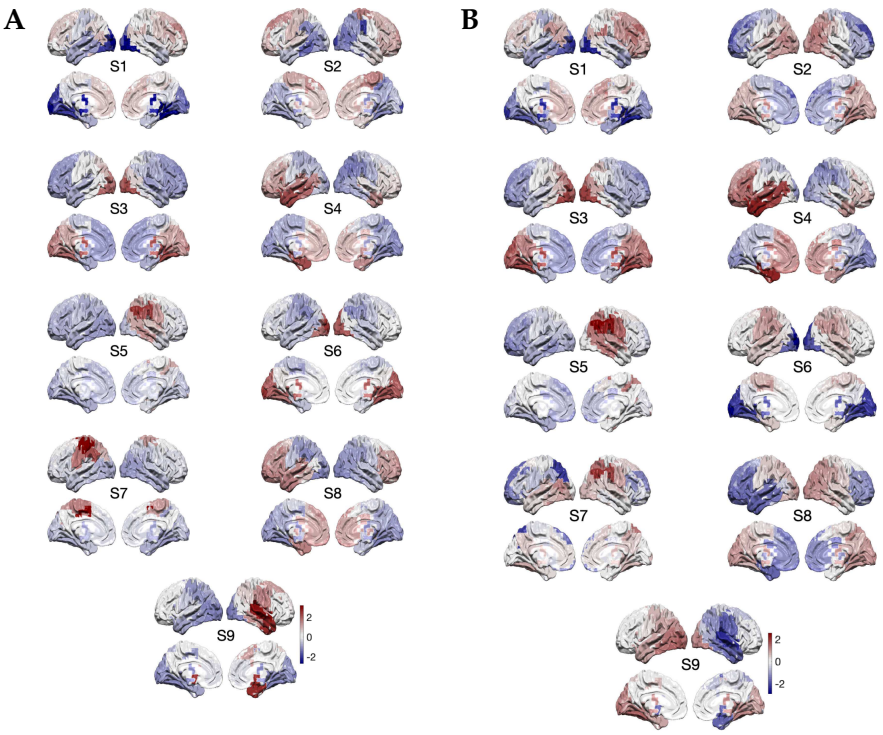

Figure C.2: Unthresholded spatial distributions of power and connectivity for the full data inferred by the recurrent model. (A) The mean activation maps inferred by the recurrent model applied to the entire data set, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

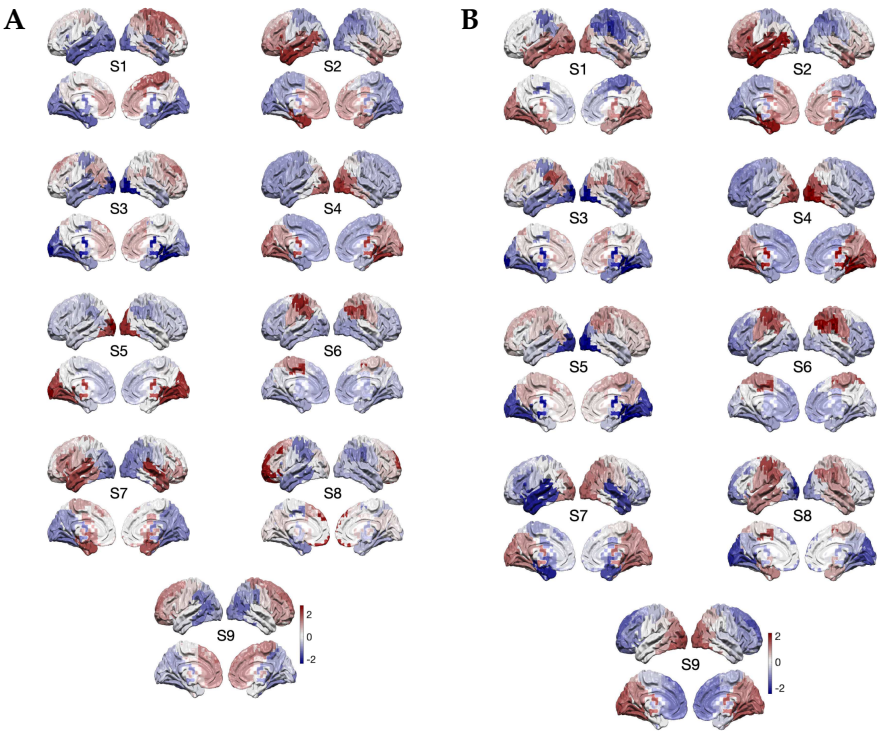

Figure C.3: Unthresholded spatial distributions of power and connectivity for the full data inferred by the HMM. (A) The mean activation maps inferred by the HMM applied to the entire data set, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

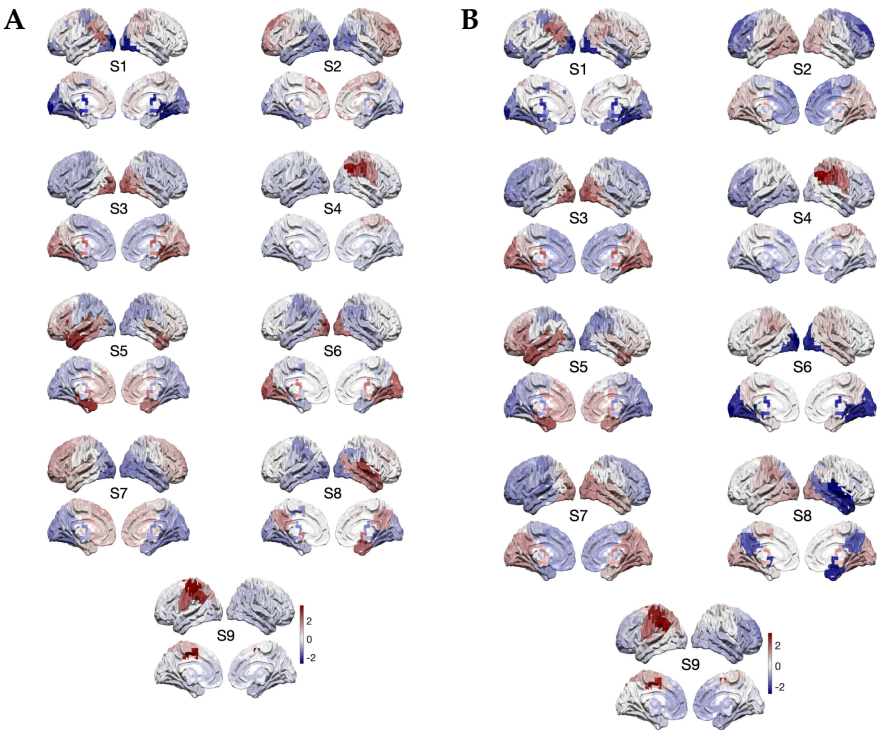

Figure C.4: Unthresholded spatial distributions of power and connectivity for the first half of the data inferred by the recurrent model. (A) The mean activation maps inferred by the recurrent model applied to the first half of the dataset, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

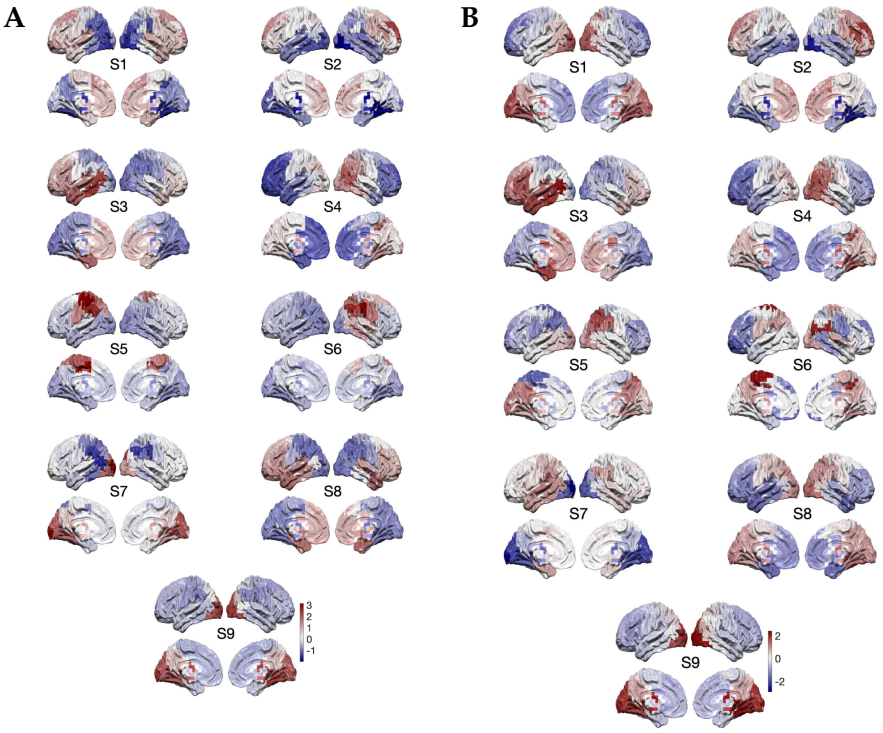

Figure C.5: Unthresholded spatial distributions of power and connectivity for the second half of the data inferred by the recurrent model. (A) The mean activation maps inferred by the recurrent model applied to the second half of the dataset, ordered by stability. (B) Within-state functional connectivity as given by degree/connectedness

C.3 TRIAL-WISE STATE ACTIVATIONS

In the following section, we display raster plots of each state activation across all subjects and all trials.

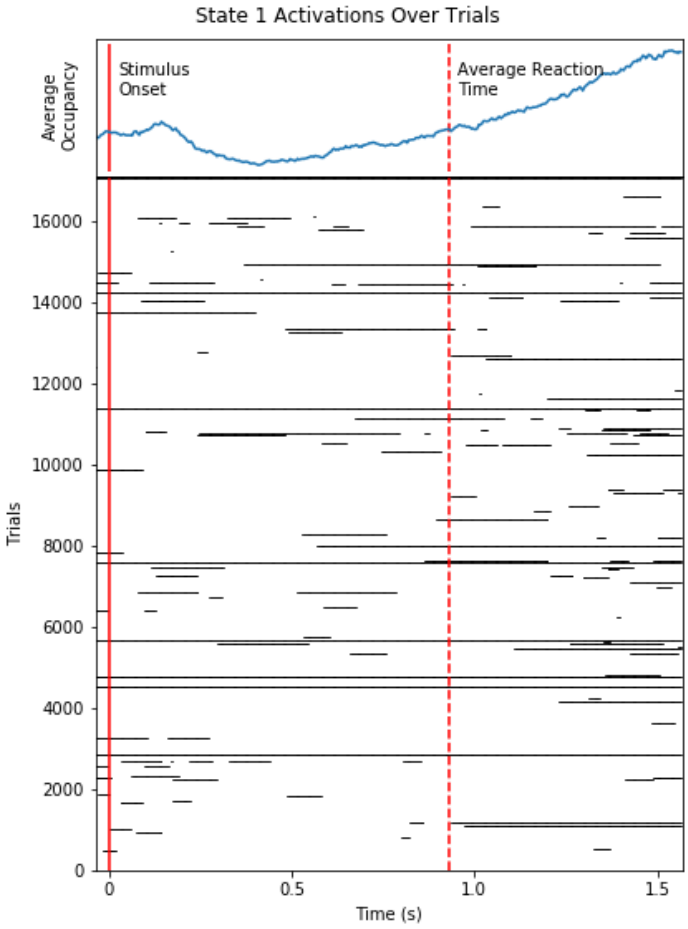

Figure C.6: Trialwise state activation raster plot for state 1.

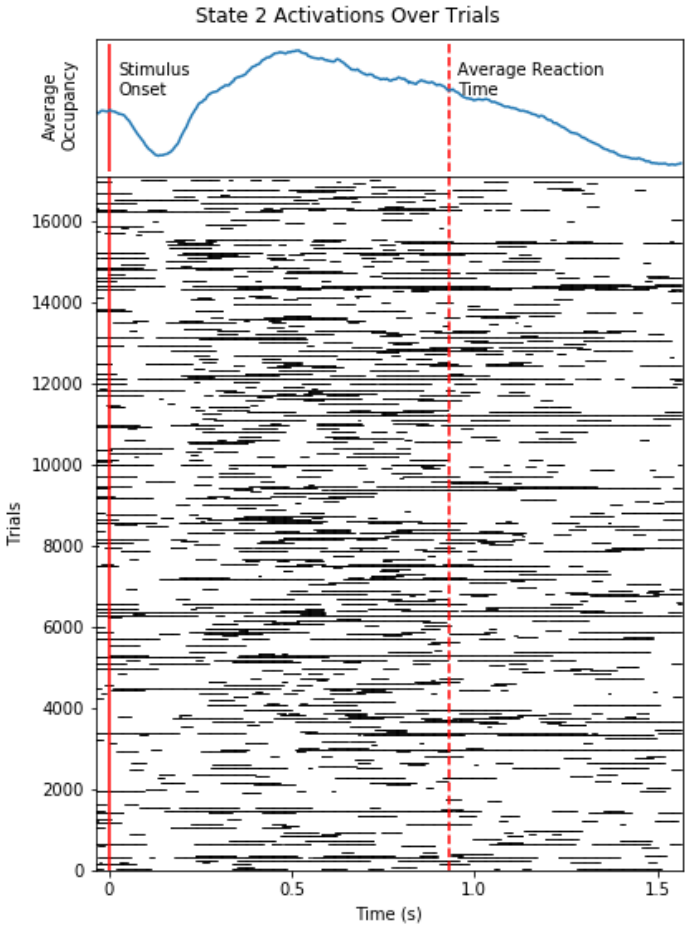

Figure C.7: Trialwise state activation raster plot for state 2.

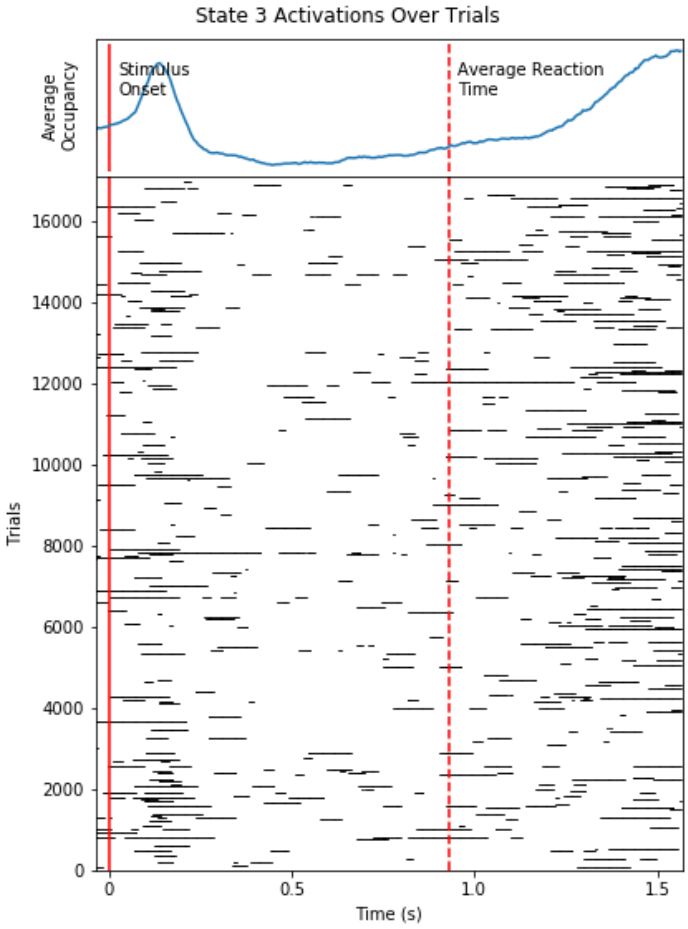

Figure C.8: Trialwise state activation raster plot for state 3.

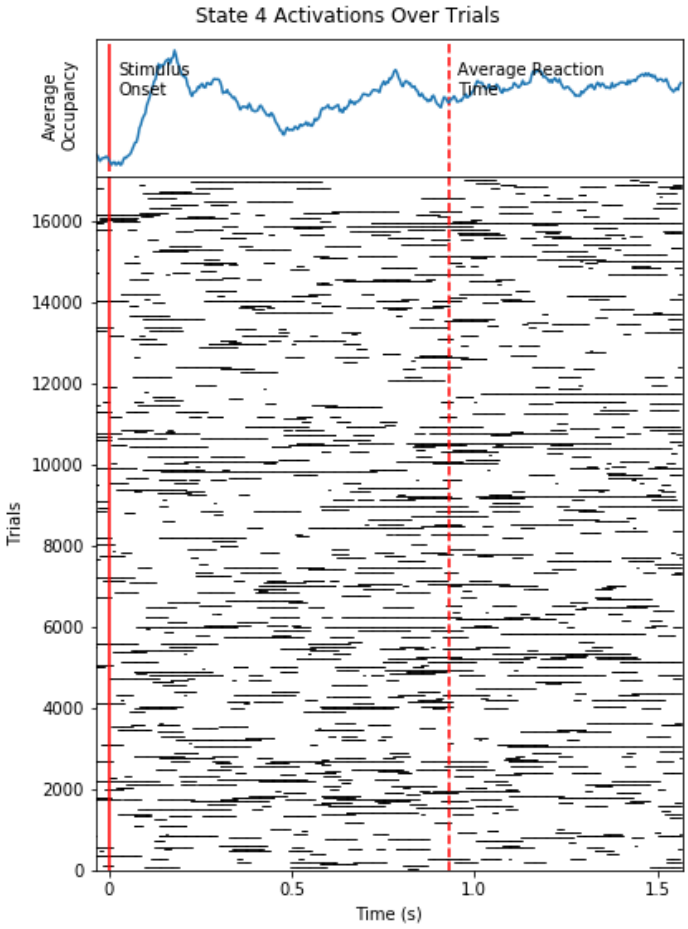

Figure C.9: Trialwise state activation raster plot for state 4.

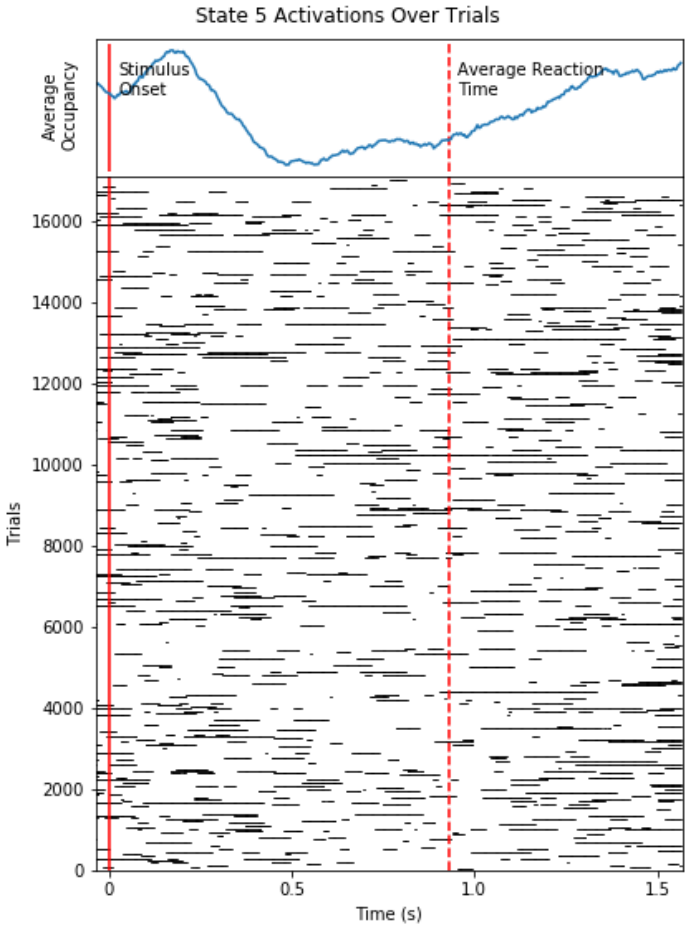

Figure C.10: Trialwise state activation raster plot for state 5.

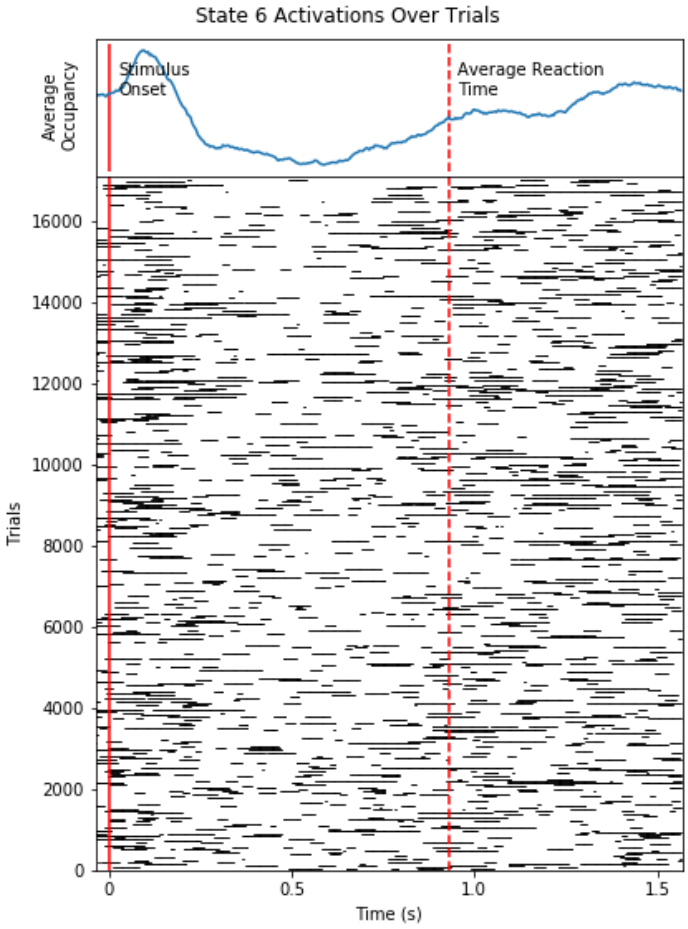

Figure C.11: Trialwise state activation raster plot for state 6.

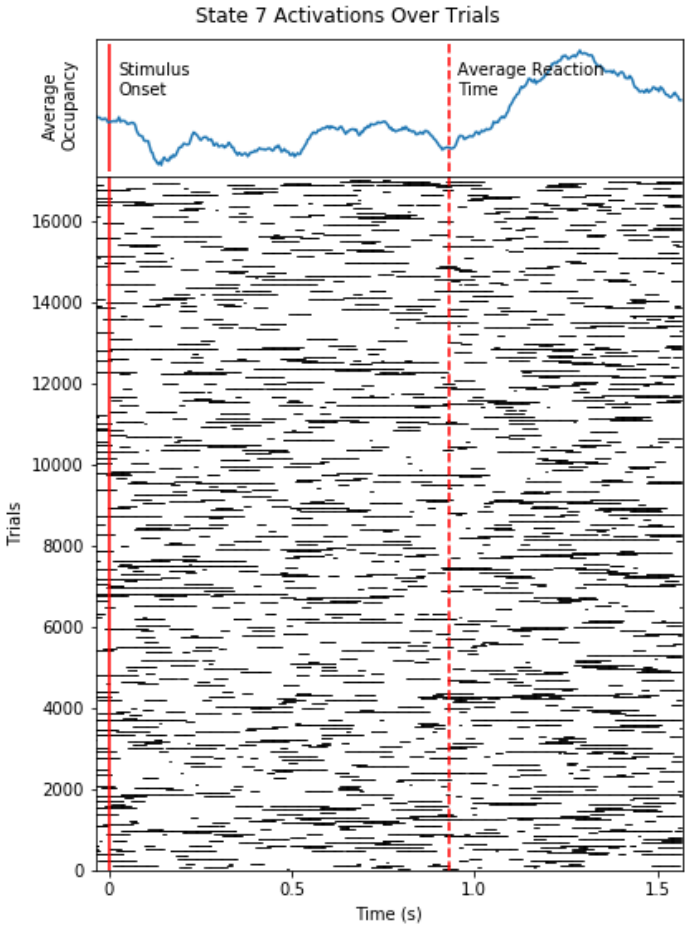

Figure C.12: Trialwise state activation raster plot for state 7.

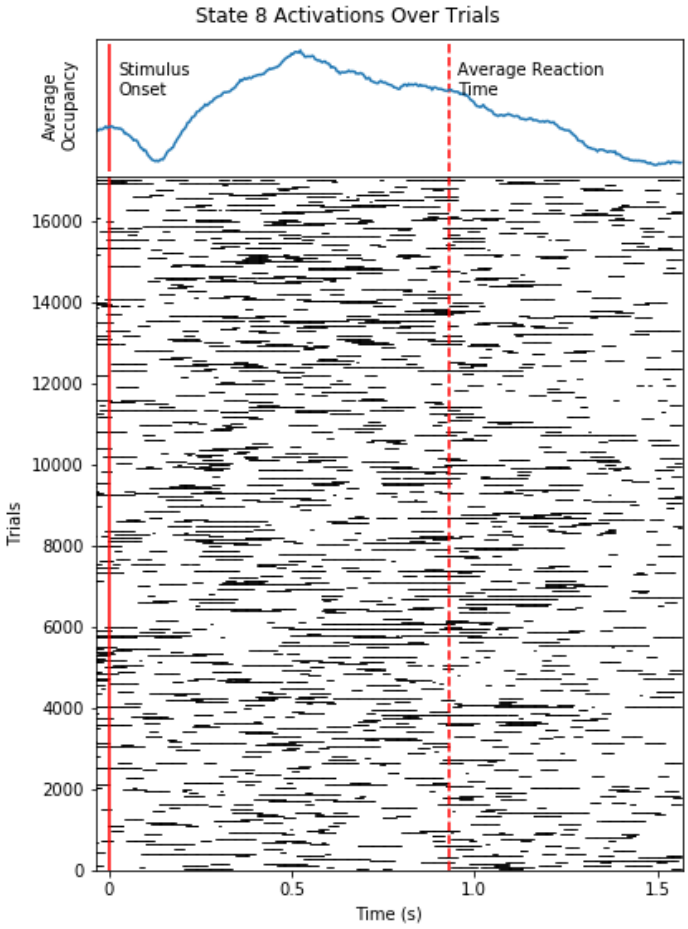

Figure C.13: Trialwise state activation raster plot for state 8.

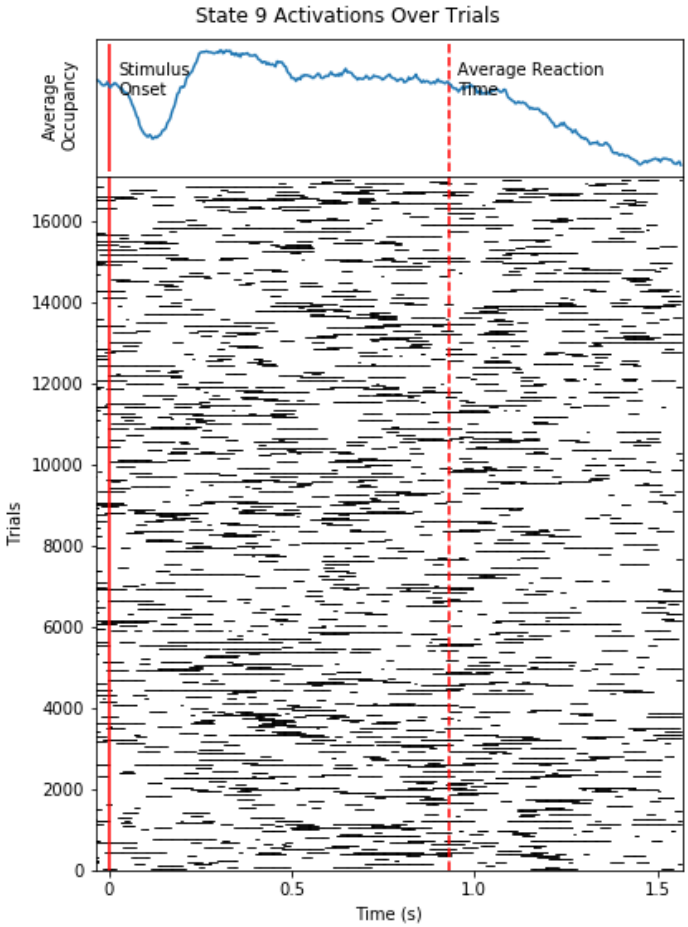

Figure C.14: Trialwise state activation raster plot for state 9.

BIBLIOGRAPHY

[1] Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. "Tracking whole-brain connectivity dynamics in the resting state." eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 24.3 (2014), pp. 663–676. ISSN: 1460-2199 (Electronic). DOI: 10.1093/cercor/bhs352.

[2] Y Attal, M Bhattacharjee, J Yelnik, B Cottereau, J Lefèvre, Y Okada, E Bardinet, M Chupin, and S Baillet. "Modelling and detecting deep brain activity with MEG and EEG." In: *IRBM* 30.3 (2009), pp. 133–138. ISSN: 1959-0318. DOI: https://doi.org/10.1016/j.irbm.2009.01.005. URL: http://www.sciencedirect.com/science/article/pii/S1959031809000414.

[3] 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." In: *eLife* 3 (2014). Ed. by Jody C Culham, e01867. ISSN: 2050-084X. DOI: 10.7554/eLife.01867. URL: https://dx.doi.org/10.7554/eLife.01867.

[4] Pablo Barttfeld, Lynn Uhrig, Jacobo D Sitt, Mariano Sigman, Béchir Jarraya, and Stanislas Dehaene. "Signature of consciousness in the dynamics of resting-state brain activity." In: *Proceedings of the National Academy of Sciences* 112.3 (2015), 887 LP –892. DOI: 10.1073/pnas.1418031112. URL: http://www.pnas.org/content/112/3/887.abstract.

[5] Justin Bayer and Christian Osendorfer. "Learning Stochastic Recurrent Networks." In: (2014). arXiv: 1411.7610. URL: http://arxiv.org/abs/1411.7610.

[6] Christian F Beckmann, Marilena DeLuca, Joseph T Devlin, and Stephen M Smith. "Investigations into resting-state connectivity using independent component analysis." eng. In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.1457 (2005), pp. 1001–1013. ISSN: 0962-8436 (Print). DOI: 10.1098/rstb.2005.1634.

[7] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. "The Curse of Highly Variable Functions for Local Kernel Machines." In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS'05. Cambridge, MA, USA: MIT Press, 2005, pp. 107–114. URL: http://dl.acm.org/citation.cfm?id=2976248.2976262.

[8] Hans Berger. "Über das Elektrenkephalogramm des Menschen." In: *Archiv für Psychiatrie und Nervenkrankheiten* 87.1 (1929), pp. 527–570. ISSN: 1433-8491. DOI: 10.1007/BF01797193. URL: https://doi.org/10.1007/BF01797193.

[9] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[10] 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." In: *Magnetic Resonance in Medicine* 34.4 (1995), pp. 537–541. ISSN: 07403194. DOI: 10.1002/mrm.1910340409. URL: http://www.ncbi.nlm.nih.gov/pubmed/8524021http://doi.wiley.com/10.1002/mrm.1910340409.

[11] Georges Bonnet. "Transformations des signaux aléatoires a travers les systèmes non linéaires sans mémoire." In: *Annales des Télécommunications* 19.9-10 (), pp. 203–220. ISSN: 0003-4347. DOI: 10.1007/bf03014720. URL: https://link.springer.com/article/10.1007/BF03014720.

[12] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. *Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription*. Tech. rep. 2012. URL: http://www-etud.iro.umontreal.ca/{~}boulanni/ICML2012.pdf.

[13] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. "Generating Sentences from a Continuous Space." In: (2015). arXiv: 1511.06349. URL: http://arxiv.org/abs/1511.06349.

[14] Pierre Broca. "Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech)." In: *Bulletin de la Société Anatomique* 6 (1861), pp. 330–357.

[15] Matthew J Brookes, Mark Woolrich, Henry Luckhoo, Darren Price, Joanne R Hale, Mary C Stephenson, Gareth R Barnes, Stephen M Smith, and Peter G Morris. "Investigating the electrophysiological basis of resting state networks using magnetoencephalography." In: *Proceedings of the National Academy of Sciences* 108.40 (2011), pp. 16783–16788. ISSN: 0027-8424. DOI: 10.1073/pnas.1112685108. URL: https://www.pnas.org/content/108/40/16783.

[16] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. "Importance Weighted Autoencoders." In: (). URL: https://arxiv.org/pdf/1509.00519.pdf.

[17] György Buzsáki. *Rhythms of the brain.* New York, NY, US: Oxford University Press, 2006, pp. xv, 448–xv, 448. ISBN: 0-19-530106-4 (Hardcover); 978-0-19-530106-9 (Hardcover). DOI: 10.1093/acprof:oso/9780195301069.001.0001.

[18] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. "Sample size selection in optimization methods for machine learning." In: *Mathematical Programming* 134.1 (2012), pp. 127–155. ISSN: 1436-4646. DOI: 10.1007/s10107-012-0572-5. URL: https://doi.org/10.1007/s10107-012-0572-5.

[19] Ryan T Canolty and Robert T Knight. "The functional role of cross-frequency coupling." eng. In: *Trends in cognitive sciences* 14.11 (2010), pp. 506–515. ISSN: 1879-307X. DOI: 10.1016/j.tics.2010.09.001. URL: https://www.ncbi.nlm.nih.gov/pubmed/20932795https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3359652/.

[20] Catie Chang and Gary H. Glover. "Time–frequency dynamics of resting-state brain connectivity measured with fMRI." In: *NeuroImage* 50.1 (2010), pp. 81–98. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2009.12.011. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811909012981.

[21] Jean-Lon Chen, Tomas Ros, and John H Gruzelier. "Dynamic changes of ICA-derived EEG functional connectivity in the resting state." eng. In: *Human brain mapping* 34.4 (2013), pp. 852–868. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.21475.

[22] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches." In: *CoRR* abs/1409.1 (2014). URL: http://arxiv.org/abs/1409.1259.

[23] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. *A Recurrent Latent Variable Model for Sequential Data*. 2015. URL: https://papers.nips.cc/paper/5653-a-recurrent-latent-variable-model-for-sequential-data.

[24] Emily B J Coffey, Sibylle C Herholz, Alexander M P Chepesiuk, Sylvain Baillet, and Robert J Zatorre. "Cortical contributions to the auditory frequency-following response revealed by MEG." In: *Nature Communications* 7.1 (2016), p. 11070. ISSN: 2041-1723. DOI: 10.1038/ncomms11070. URL: https://doi.org/10.1038/ncomms11070.

[25] G.L. Colclough, M.J. Brookes, S.M. Smith, and M.W. Woolrich. "A symmetric multivariate leakage correction for MEG connectomes." In: *NeuroImage* 117 (2015), pp. 439–448. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2015.03.071. URL: http://www.ncbi.nlm.nih.gov/pubmed/25862259http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4528074https://linkinghub.elsevier.com/retrieve/pii/S1053811915002670.

[26] Giles L Colclough, Stephen M Smith, Thomas E Nichols, Anderson M Winkler, Stamatios N Sotiropoulos, Matthew F Glasser, David C Van Essen, and Mark W Woolrich. "The heritability of multi-modal connectivity in human brain activity." In: *eLife* 6 (2017). Ed. by Jack L Gallant, e20178. ISSN: 2050-084X. DOI: 10.7554/eLife.20178. URL: https://doi.org/10.7554/eLife.20178.

[27] Maurizio Corbetta. "Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems?" In: *Proceedings of the National Academy of Sciences* 95.3 (1998), 831 LP –838.

DOI: 10.1073/pnas.95.3.831. URL: http://www.pnas.org/content/95/3/831.abstract.

[28] J S Damoiseaux, S A R B Rombouts, F Barkhof, P Scheltens, C J Stam, S M Smith, and C F Beckmann. "Consistent resting-state networks across healthy subjects." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.37 (2006), pp. 13848–13853. ISSN: 0027-8424 (Print). DOI: 10.1073/pnas.0601417103.

[29] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization." In: (2014). arXiv: 1406.2572. URL: http://arxiv.org/abs/1406.2572.

[30] Marilena De Luca, Stephen Smith, Nicola De Stefano, Antonio Federico, and Paul M Matthews. "Blood oxygenation level dependent contrast resting state networks are relevant to functional activity in the neocortical sensorimotor system." eng. In: *Experimental brain research* 167.4 (2005), pp. 587–594. ISSN: 0014-4819 (Print). DOI: 10.1007/s00221-005-0059-1.

[31] John Duchi JDUCHI and Yoram Singer. *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization * Elad Hazan*. Tech. rep. 2011, pp. 2121–2159. URL: http://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf.

[32] Harini Eavani, Theodore D Satterthwaite, Raquel E Gur, Ruben C Gur, and Christos Davatzikos. "Unsupervised learning of functional network dynamics in resting state fMRI." eng. In: *Information processing in medical imaging : proceedings of the ... conference* 23 (2013), pp. 426–437. ISSN: 1011-2499. URL: https://www.ncbi.nlm.nih.gov/pubmed/24683988https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974209/.

[33] Y Ephraim and N Merhav. "Hidden Markov Processes." In: *IEEE Trans. Inf. Theor.* 48.6 (2006), pp. 1518–1569. ISSN: 0018-9448. DOI: 10.1109/TIT.2002.1003838. URL: http://dx.doi.org/10.1109/TIT.2002.1003838.

[34] 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." eng. In: *Brain research bulletin* 70.4-6 (2006), pp. 263–269. ISSN: 0361-9230 (Print). DOI: 10.1016/j.brainresbull.2006.06.012.

[35] Otto Fabius and Joost R Van Amersfoort. *VARIATIONAL RECURRENT AUTO-ENCODERS*. Tech. rep. arXiv: 1412.6581v6. URL: https://arxiv.org/pdf/1412.6581.pdf.

[36] Shai Fine, Yoram Singer, and Naftali Tishby. "The Hierarchical Hidden Markov Model: Analysis and Applications." In: *Machine Learning* 32.1 (1998), pp. 41–62. ISSN: 1573-0565. DOI: 10.1023/A:1007469218079. URL: https://doi.org/10.1023/A:1007469218079.

[37] Alex Fornito, Ben J Harrison, Andrew Zalesky, and Jon S Simons. "Competitive and cooperative dynamics of large-scale brain functional networks supporting recollection." In: *Proceedings of the National Academy of Sciences* 109.31 (2012), 12788 LP –12793. DOI: 10.1073/pnas.1204185109. URL: http://www.pnas.org/content/109/31/12788.abstract.

[38] Michael D Fox and Marcus E Raichle. "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging." In: *Nat Rev Neurosci* 8.9 (2007), pp. 700–711. URL: http://dx.doi.org/10.1038/nrn2201.

[39] Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. "The human brain is intrinsically organized into dynamic, anticorrelated functional networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.27 (2005), 9673 LP –9678. DOI: 10.1073/pnas.0504136102. URL: http://www.pnas.org/content/102/27/9673.abstract.

[40] Peter Fransson. "How default is the default mode of brain function? Further evidence from intrinsic BOLD signal fluctuations." eng. In: *Neuropsychologia* 44.14 (2006), pp. 2836–2845. ISSN: 0028-3932 (Print). DOI: 10.1016/j.neuropsychologia.2006.06.017.

[41] K J Friston, C D Frith, P F Liddle, and R S Frackowiak. "Investigating a network model of word generation with positron emission tomography." In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 244.1310 (1991), pp. 101–106. ISSN: 0962-8452. DOI: 10.1098/rspb.1991.0057. URL: http://www.ncbi.nlm.nih.gov/pubmed/1679543https://royalsocietypublishing.org/doi/10.1098/rspb.1991.0057.

[42] K. J. Friston, C. D. Frith, P. F. Liddle, and R. S. J. Frackowiak. "Functional Connectivity: The Principal-Component Analysis of Large (PET) Data Sets." In: *Journal of Cerebral Blood Flow & Metabolism* 13.1 (1993), pp. 5–14. ISSN: 0271-678X. DOI: 10.1038/jcbfm.1993.4. URL: http://www.ncbi.nlm.nih.gov/pubmed/8417010http://journals.sagepub.com/doi/10.1038/jcbfm.1993.4.

[43] Karl J. Friston. "Functional and effective connectivity in neuroimaging: A synthesis." In: *Human Brain Mapping* 2.1-2 (1994), pp. 56–78. ISSN: 10659471. DOI: 10.1002/hbm.460020107. URL: http://doi.wiley.com/10.1002/hbm.460020107.

[44] Mark Gales and Steve Young. "The Application of Hidden Markov Models in Speech Recognition." In: *Found. Trends Signal Process.* 1.3 (2007), pp. 195–304. ISSN: 1932-8346. DOI: 10.1561/2000000004. URL: http://dx.doi.org/10.1561/2000000004.

[45] Zhe Gan, Chunyuan Li, Ricardo Henao, David Carlson, and Lawrence Carin. *Deep Temporal Sigmoid Belief Networks for Sequence Modeling*. Tech. rep. arXiv: 1509.07087v1. URL: https://arxiv.org/pdf/1509.07087.pdf.

[46] Michael S Gazzaniga and George R Mangun, eds. *The cognitive neurosciences, 5th ed.* Cambridge, MA, US: MIT Press, 2014, pp. xvi, 1106–xvi, 1106. ISBN: 978-0-262-02777-9 (Hardcover).

[47] Michael D Greicius, Vesa Kiviniemi, Osmo Tervonen, Vilho Vainionpaa, Seppo Alahuhta, Allan L Reiss, and Vinod Menon. "Persistent default-mode network connectivity during light sedation." eng. In: *Human brain mapping* 29.7 (2008), pp. 839–847. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.20537.

[48] Shixiang Gu, Zoubin Ghahramani, and Richard E Turner. *Neural Adaptive Sequential Monte Carlo*. Tech. rep. 2015. arXiv: 1506.03338v3. URL: https://arxiv.org/pdf/1506.03338.pdf.

[49] M S Hamalainen and R J Ilmoniemi. "Interpreting magnetic fields of the brain: minimum norm estimates." eng. In: *Medical & biological engineering & computing* 32.1 (1994), pp. 35–42. ISSN: 0140-0118 (Print). DOI: 10.1007/bf02512476.

[50] M S Hämäläinen and R J Ilmoniemi. "Interpreting magnetic fields of the brain: minimum norm estimates." In: *Medical & Biological Engineering & Computing* 32.1 (1994), pp. 35–42. ISSN: 1741-0444. DOI: 10.1007/BF02512476. URL: https://doi.org/10.1007/BF02512476.

[51] Daniel A. Handwerker, Vinai Roopchansingh, Javier Gonzalez-Castillo, and Peter A. Bandettini. "Periodic changes in fMRI connectivity." In: *NeuroImage* 63.3 (2012), pp. 1712–1719. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2012.06.078. URL: http://www.sciencedirect.com/science/article/pii/S1053811912007124?via{\%}3Dihub.

[52] Riitta Hari et al. "IFCN-endorsed practical guidelines for clinical magnetoencephalography (MEG)." eng. In: *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 129.8 (2018), pp. 1720–1747. ISSN: 1872-8952 (Electronic). DOI: 10.1016/j.clinph.2018.03.042.

[53] Martijn P van den Heuvel and Hilleke E Hulshoff Pol. "Specific somatotopic organization of functional connections of the primary motor network during resting state." eng. In: *Human brain mapping* 31.4 (2010), pp. 631–644. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.20893.

[54] A. Hillebrand and G.R. Barnes. "A Quantitative Assessment of the Sensitivity of Whole-Head MEG to Activity in the Adult Human Cortex." In: *NeuroImage* 16.3 (2002), pp. 638–650. ISSN: 1053-8119. DOI: 10.1006/NIMG.2002.1102. URL: https://www.sciencedirect.com/science/article/pii/S105381190291102X?via{\%}3Dihub.

[55] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[56] Sepp Hochreiter and Jj Urgen Schmidhuber. "LONG SHORT-TERM MEMORY." In: *Neural Computation* 9.8 (1997), pp. 1735–1780. URL: http://www7.informatik.tu-muenchen.de/{~}hochreithttp://www.idsia.ch/{~}juergen.

[57] 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." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.6 (2009), pp. 2035–2040. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.0811168106.

[58] Silvina G Horovitz, Masaki Fukunaga, Jacco A de Zwart, Peter van Gelderen, Susan C Fulton, Thomas J Balkin, and Jeff H Duyn. "Low frequency BOLD fluctuations during resting wakefulness and light sleep: a simultaneous EEG-fMRI study." eng. In: *Human brain mapping* 29.6 (2008), pp. 671–682. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.20428.

[59] Silvina G Horovitz, Allen R Braun, Walter S Carr, Dante Picchioni, Thomas J Balkin, Masaki Fukunaga, and Jeff H Duyn. "Decoupling of the brain's default mode network during deep sleep." In: *Proceedings of the National Academy of Sciences* 106.27 (2009), 11376 LP –11381. DOI: 10.1073/pnas.0901435106. URL: http://www.pnas.org/content/106/27/11376.abstract.

[60] Barry Horwitz. "The elusive concept of brain connectivity." In: *NeuroImage* 19.2 Pt 1 (2003), pp. 466–70. ISSN: 1053-8119. URL: http://www.ncbi.nlm.nih.gov/pubmed/12814595.

[61] René J Huster, Stefan Debener, Tom Eichele, and Christoph S Herrmann. "Methods for Simultaneous EEG-fMRI: An Introductory Review." In: *Journal of Neuroscience* 32.18 (2012), pp. 6053–6060. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0447-12.2012. URL: https://www.jneurosci.org/content/32/18/6053.

[62] R. Matthew Hutchison et al. "Dynamic functional connectivity: Promise, issues, and interpretations." In: *NeuroImage* 80 (2013), pp. 360–378. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.

2013.05.079. URL: http://www.sciencedirect.com/science/article/pii/S105381191300579X?via{\%}3Dihub{\#}bb0255.

[63] R Matthew Hutchison, Thilo Womelsdorf, Joseph S Gati, Stefan Everling, and Ravi S Menon. "Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques." In: *Hum Brain Mapp* 34.9 (2013), pp. 2154–2177.

[64] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical Reparameterization with Gumbel-Softmax." In: (2016). arXiv: 1611.01144. URL: https://arxiv.org/abs/1611.01144.

[65] Ole Jensen and Laura L Colgin. "Cross-frequency coupling between neuronal oscillations." eng. In: *Trends in cognitive sciences* 11.7 (2007), pp. 267–269. ISSN: 1364-6613 (Print). DOI: 10.1016/j.tics.2007.05.003.

[66] David T Jones et al. "Non-stationarity in the "resting brain's" modular architecture." eng. In: *PloS one* 7.6 (2012), e39731. ISSN: 1932-6203 (Electronic). DOI: 10.1371/journal.pone.0039731.

[67] Michael I Jordan and Lawrence K Saul. *An Introduction to Variational Methods for Graphical Models*. Tech. rep. 1999, pp. 183–233.

[68] Shella D. Keilholz, Matthew E. Magnuson, Wen-Ju Pan, Martha Willis, and Garth J. Thompson. "Dynamic Properties of Functional Connectivity in the Rodent." In: *Brain Connectivity* 3.1 (2013), pp. 31–40. ISSN: 2158-0014. DOI: 10.1089/brain.2012.0115. URL: http://online.liebertpub.com/doi/abs/10.1089/brain.2012.0115.

[69] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: (2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[70] Diederik P Kingma and Jimmy Ba. "Adam: {A} Method for Stochastic Optimization." In: *CoRR* abs/1412.6 (2014). URL: http://arxiv.org/abs/1412.6980.

[71] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes." In: *Iclr* Ml (2014), pp. 1–14. ISSN: 1312.6114v10. DOI: 10.1051/0004-6361/201527329. arXiv: 1312.6114. URL: https://arxiv.org/pdf/1312.6114.pdfhttp://arxiv.org/abs/1312.6114.

[72] Rahul G Krishnan, Uri Shalit, and David Sontag. *Structured Inference Networks for Nonlinear State Space Models*. Tech. rep. arXiv: 1609.09869v2. URL: www.aaai.org.

[73] S Kullback and R A Leibler. "On Information and Sufficiency." en. In: *Ann. Math. Statist.* 22.1 (1951), pp. 79–86. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729694. URL: https://projecteuclid.org:443/euclid.aoms/1177729694.

[74] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444. ISSN: 0028-0836. DOI: 10.1038/nature14539. URL: http://www.nature.com/articles/nature14539.

[75] Nora Leonardi and Dimitri Van De Ville. "On spurious and real fluctuations of dynamic functional connectivity during rest." eng. In: *NeuroImage* 104 (2015), pp. 430–436. ISSN: 1095-9572 (Electronic). DOI: 10.1016/j.neuroimage.2014.09.007.

[76] R Lestienne. "Spike timing, synchronization and information processing on the sensory side of the central nervous system." eng. In: *Progress in neurobiology* 65.6 (2001), pp. 545–591. ISSN: 0301-0082 (Print).

[77] Hesheng Liu, Steven M Stufflebeam, Jorge Sepulcre, Trey Hedden, and Randy L Buckner. "Evidence from intrinsic activity

that asymmetry of the human brain is controlled by multiple factors." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.48 (2009), pp. 20499–20503. ISSN: 1091-6490. DOI: 10.1073/pnas.0908073106. URL: https://www.ncbi.nlm.nih.gov/pubmed/19918055https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2777963/.

[78] David J C MacKay. "Bayesian Non-Linear Modeling for the Prediction Competition BT - Maximum Entropy and Bayesian Methods: Santa Barbara, California, U.S.A., 1993." In: ed. by Glenn R Heidbreder. Dordrecht: Springer Netherlands, 1996, pp. 221–234. ISBN: 978-94-015-8729-7. DOI: 10.1007/978-94-015-8729-7_18. URL: https://doi.org/10.1007/978-94-015-8729-7{_}18.

[79] David J C Mackay. "Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks." In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 469–505. ISSN: 0954-898X. DOI: 10.1088/0954-898X_6_3_011. URL: https://doi.org/10.1088/0954-898X{_}6{_}3{_}011.

[80] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables." In: (2016). arXiv: 1611.00712. URL: http://arxiv.org/abs/1611.00712.

[81] Pravat K Mandal, Anwesha Banerjee, Manjari Tripathi, and Ankita Sharma. "A Comprehensive Review of Magnetoencephalography (MEG) Studies for Brain Functionality in Healthy Aging and Alzheimer's Disease (AD)." eng. In: *Frontiers in computational neuroscience* 12 (2018), p. 60. ISSN: 1662-5188. DOI: 10.3389/fncom.2018.00060. URL: https://www.ncbi.nlm.nih.

gov/pubmed/30190674https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6115612/.

[82] Guillaume Marrelec, Alexandre Krainik, Hugues Duffau, Mélanie Pélégrini-Issac, Stéphane Lehéricy, Julien Doyon, and Habib Benali. "Partial correlation for functional brain interactivity investigation in functional MRI." In: *NeuroImage* 32.1 (2006), pp. 228–237. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2005.12.057. URL: http://www.sciencedirect.com/science/article/pii/S1053811906000103.

[83] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. "On the Number of Linear Regions of Deep Neural Networks." In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2924–2932. URL: http://dl.acm.org/citation.cfm?id=2969033.2969153.

[84] Shingo Murakami and Yoshio Okada. "Contributions of principal neocortical neurons to magnetoencephalography and electroencephalography signals." eng. In: *The Journal of physiology* 575.Pt 3 (2006), pp. 925–936. ISSN: 0022-3751 (Print). DOI: 10.1113/jphysiol.2006.105379.

[85] Eric Nalisnick and Padhraic Smyth. "Stick-Breaking Variational Autoencoders." In: (2016). arXiv: 1605.06197. URL: http://arxiv.org/abs/1605.06197.

[86] Allison C Nugent, Bruce Luber, Frederick W Carver, Stephen E Robinson, Richard Coppola, and Carlos A Jr Zarate. "Deriving frequency-dependent spatial patterns in MEG-derived resting state sensorimotor network: A novel multiband ICA technique." eng. In: *Human brain mapping* 38.2 (2017), pp. 779–791. ISSN: 1097-0193 (Electronic). DOI: 10.1002/hbm.23417.

[87] S Ogawa, T M Lee, A R Kay, and D W Tank. "Brain magnetic resonance imaging with contrast dependent on blood oxygenation." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 87.24 (1990), pp. 9868–9872. ISSN: 0027-8424. DOI: 10.1073/pnas.87.24.9868. URL: https://www.ncbi.nlm.nih.gov/pubmed/2124706https://www.ncbi.nlm.nih.gov/pmc/articles/PMC55275/.

[88] Jinli Ou, Li Xie, Changfeng Jin, Xiang Li, Dajiang Zhu, Rongxin Jiang, Yaowu Chen, Jing Zhang, Lingjiang Li, and Tianming Liu. "Characterizing and Differentiating Brain State Dynamics via Hidden Markov Models." In: *Brain Topography* 28.5 (2015), pp. 666–679. ISSN: 0896-0267. DOI: 10.1007/s10548-014-0406-2. URL: http://link.springer.com/10.1007/s10548-014-0406-2.

[89] John Paisley, David M Blei, and Michael I Jordan. *Variational Bayesian Inference with Stochastic Search*. Tech. rep. 2012. URL: https://icml.cc/2012/papers/687.pdf.

[90] J Matias Palva, Satu Palva, and Kai Kaila. "Phase synchrony among neuronal oscillations in the human cortex." eng. In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 25.15 (2005), pp. 3962–3972. ISSN: 1529-2401 (Electronic). DOI: 10.1523/JNEUROSCI.4250-04.2005.

[91] Francesco de Pasquale et al. "Temporal dynamics of spontaneous MEG activity in brain networks." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.13 (2010), pp. 6040–6045. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.0913863107.

[92] Maria Giulia Preti, Thomas AW Bolton, and Dimitri Van De Ville. "The dynamic functional connectome: State-of-the-art and perspectives." In: *NeuroImage* 160 (2017), pp. 41–54. ISSN:

1053-8119. DOI: 10.1016/J.NEUROIMAGE.2016.12.061. URL: https://www.sciencedirect.com/science/article/pii/S1053811916307881.

[93] R. Price. "A useful theorem for nonlinear devices having Gaussian inputs." In: *IEEE Transactions on Information Theory* 4.2 (1958), pp. 69–72. ISSN: 0018-9448. DOI: 10.1109/TIT.1958.1057444. URL: http://ieeexplore.ieee.org/document/1057444/.

[94] Ning Qian. "On the momentum term in gradient descent learning algorithms." In: *Neural Networks* 12.1 (1999), pp. 145–151. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(98)00116-6. URL: https://www.sciencedirect.com/science/article/pii/S0893608098001166.

[95] Andrew J Quinn, Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Anna C Nobre, and Mark W Woolrich. *Task-Evoked Dynamic Network Analysis Through Hidden Markov Modeling*. 2018. URL: https://www.frontiersin.org/article/10.3389/fnins.2018.00603.

[96] Mikhail Rabinovich and Pablo Varona. *Robust Transient Dynamics and Brain Functions*. 2011. URL: https://www.frontiersin.org/article/10.3389/fncom.2011.00024.

[97] M E Raichle, A M MacLeod, A Z Snyder, W J Powers, D A Gusnard, and G L Shulman. "A default mode of brain function." In: *Proceedings of the National Academy of Sciences of the United States of America* 98.2 (2001), pp. 676–82. ISSN: 0027-8424. DOI: 10.1073/pnas.98.2.676. URL: http://www.ncbi.nlm.nih.gov/pubmed/11209064http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC14647.

[98] Rajesh Ranganath, Sean Gerrish, and David M. Blei. "Black Box Variational Inference." In: (2013). arXiv: 1401.0118. URL: http://arxiv.org/abs/1401.0118.

[99] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models." In: (2014). arXiv: 1401.4082. URL: http://arxiv.org/abs/1401.4082.

[100] P Robert and Y Escoufier. "A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient." In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 25.3 (1976), pp. 257–265. ISSN: 00359254, 14679876. DOI: 10.2307/2347233. URL: http://www.jstor.org/stable/2347233.

[101] Francisco J R Ruiz, Michalis K Titsias, and David M Blei. "The Generalized Reparameterization Gradient." In: (). URL: https://arxiv.org/pdf/1610.02287.pdf.

[102] Tim Salimans and David A. Knowles. "Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression." In: (2012). DOI: 10.1214/13-BA858. arXiv: 1206.6679. URL: http://arxiv.org/abs/1206.6679http://dx.doi.org/10.1214/13-BA858.

[103] Raymond Salvador, John Suckling, Martin R Coleman, John D Pickard, David Menon, and Ed Bullmore. "Neurophysiological architecture of functional magnetic resonance images of human brain." eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 15.9 (2005), pp. 1332–1342. ISSN: 1047-3211 (Print). DOI: 10.1093/cercor/bhi016.

[104] Philipp G. Sämann, Renate Wehrle, David Hoehn, Victor I. Spoormaker, Henning Peters, Carolin Tully, Florian Holsboer, and Michael Czisch. "Development of the Brain's Default Mode

Network from Wakefulness to Slow Wave Sleep." In: *Cerebral Cortex* 21.9 (2011), pp. 2082–2093. ISSN: 1460-2199. DOI: 10.1093/cercor/bhq295. URL: http://www.ncbi.nlm.nih.gov/pubmed/21330468https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bhq295.

[105] S Shakil, S D Keilholz, and Chin-Hui Lee. "On frequency dependencies of sliding window correlation." In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2015, pp. 363–368. ISBN: VO -. DOI: 10.1109/BIBM.2015.7359708.

[106] Sadia Shakil, Jacob C Billings, Shella D Keilholz, and Chin-Hui Lee. "Parametric Dependencies of Sliding Window Correlation." eng. In: *IEEE transactions on bio-medical engineering* 65.2 (2018), pp. 254–263. ISSN: 1558-2531 (Electronic). DOI: 10.1109/TBME.2017.2762763.

[107] Heather Shappell, Brian S Caffo, James J Pekar, and Martin A Lindquist. "Improved state change estimation in dynamic functional connectivity using hidden semi-Markov models." In: *NeuroImage* 191 (2019), pp. 243–257. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2019.02.013. URL: http://www.sciencedirect.com/science/article/pii/S1053811919300990.

[108] Zarrar Shehzad et al. "The resting brain: unconstrained yet reliable." eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 19.10 (2009), pp. 2209–2229. ISSN: 1460-2199 (Electronic). DOI: 10.1093/cercor/bhn256.

[109] Stephen M Smith et al. "Correspondence of the brain's functional architecture during activation and rest." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.31 (2009), pp. 13040–13045. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.0905267106.

[110] Stephen M Smith et al. "Temporally-independent functional modes of spontaneous brain activity." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.8 (2012), pp. 3131–3136. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.1121329109.

[111] Andrea Soddu, Audrey Vanhaudenhuyse, Athena Demertzi, Marie-Aurelie Bruno, Luaba Tshibanda, Haibo Di, Boly Melanie, Michele Papa, Steven Laureys, and Quentin Noirhomme. "Resting state activity in patients with disorders of consciousness." eng. In: *Functional neurology* 26.1 (2011), pp. 37–43. ISSN: 0393-5264 (Print).

[112] Felice T Sun, Lee M Miller, Ajay A Rao, and Mark D'Esposito. "Functional connectivity of cortical networks involved in bimanual motor sequence learning." eng. In: *Cerebral cortex (New York, N.Y. : 1991)* 17.5 (2007), pp. 1227–1234. ISSN: 1047-3211 (Print). DOI: 10.1093/cercor/bhl033.

[113] Enzo Tagliazucchi, Frederic von Wegner, Astrid Morzelewski, Verena Brodbeck, and Helmut Laufs. "Dynamic BOLD functional connectivity in humans and its electrophysiological correlates." eng. In: *Frontiers in human neuroscience* 6 (2012), p. 339. ISSN: 1662-5161 (Electronic). DOI: 10.3389/fnhum.2012.00339.

[114] G Tononi, O Sporns, and G M Edelman. "A measure for brain complexity: relating functional segregation and integration in the nervous system." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 91.11 (1994), pp. 5033–5037. ISSN: 0027-8424. DOI: 10.1073/pnas.91.11.5033. URL: https://www.ncbi.nlm.nih.gov/pubmed/8197179https://www.ncbi.nlm.nih.gov/pmc/articles/PMC43925/.

[115] Nelson J Trujillo-Barreto, David Araya, and Wael El-Deredy. "The discrete logic of the Brain - Explicit modelling of Brain

State durations in EEG and MEG." In: *bioRxiv* (2019), p. 635300. DOI: 10.1101/635300. URL: http://biorxiv.org/content/early/2019/05/10/635300.abstract.

[116] Francisco Varela, Jean-Philippe Lachaux, Eugenio Rodriguez, and Jacques Martinerie. "The brainweb: Phase synchronization and large-scale integration." In: *Nature Reviews Neuroscience* 2.4 (2001), pp. 229–239. ISSN: 1471-003X. DOI: 10.1038/35067550. URL: http://www.ncbi.nlm.nih.gov/pubmed/11283746http://www.nature.com/articles/35067550.

[117] Gaël Varoquaux, Alexandre Gramfort, Jean Baptiste Poline, and Bertrand Thirion. "Brain covariance selection: better individual functional connectivity models using population prior." In: (2010). arXiv: 1008.5071. URL: http://arxiv.org/abs/1008.5071.

[118] B D Van Veen, W Van Drongelen, M Yuchtman, and A Suzuki. "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering." In: *IEEE Transactions on Biomedical Engineering* 44.9 (1997), pp. 867–880. ISSN: VO - 44. DOI: 10.1109/10.623056.

[119] Diego Vidaurre, Stephen M Smith, and Mark W Woolrich. "Brain network dynamics are hierarchically organized in time." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.48 (2017), pp. 12827–12832. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.1705120114.

[120] 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." In: *NeuroImage* 126 (2016), pp. 81–95. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2015.11.047.

URL: http://www.sciencedirect.com/science/article/pii/S1053811915010691.

[121] J. L. Vincent, G. H. Patel, M. D. Fox, A. Z. Snyder, J. T. Baker, D. C. Van Essen, J. M. Zempel, L. H. Snyder, M. Corbetta, and M. E. Raichle. "Intrinsic functional architecture in the anaesthetized monkey brain." In: *Nature* 447.7140 (2007), pp. 83–86. ISSN: 0028-0836. DOI: 10.1038/nature05758. URL: http://www.ncbi.nlm.nih.gov/pubmed/17476267http://www.nature.com/articles/nature05758.

[122] M J Wainwright, M I Jordan, Martin J Wainwright, and Michael I Jordan. "Graphical Models, Exponential Families, and Variational Inference." In: *Foundations and Trends R in Machine Learning* 1.2 (2008), pp. 1–305. DOI: 10.1561/2200000001.

[123] Daniel G Wakeman and Richard N Henson. "A multi-subject, multi-modal human neuroimaging dataset." In: *Scientific Data* 2 (2015), p. 150001. URL: https://doi.org/10.1038/sdata.2015.1http://10.0.4.14/sdata.2015.1.

[124] Carl Wernicke. *Der aphasische Symptomencomplex; eine psychologische Studie auf anatomischer Basis.* German. Breslau: Cohn & Weigert, 1874.

[125] Mark Woolrich, Laurence Hunt, Adrian Groves, and Gareth Barnes. "MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization." eng. In: *NeuroImage* 57.4 (2011), pp. 1466–1479. ISSN: 1095-9572 (Electronic). DOI: 10.1016/j.neuroimage.2011.04.041.

[126] Zhi Yang, Stephen LaConte, Xuchu Weng, and Xiaoping Hu. "Ranking and averaging independent component analysis by reproducibility (RAICAR)." In: *Human Brain Mapping* 29.6 (2008),

pp. 711–725. ISSN: 1065-9471. DOI: 10.1002/hbm.20432. URL: https://doi.org/10.1002/hbm.20432.

[127] Andrew Zalesky, Alex Fornito, Luca Cocchi, Leonardo L Gollo, and Michael Breakspear. "Time-resolved resting-state brain networks." In: *Proceedings of the National Academy of Sciences* 111.28 (2014), 10341 LP –10346. DOI: 10.1073/pnas.1400181111. URL: http://www.pnas.org/content/111/28/10341.abstract.