Time-Varying Brain Connectivity with Multiregression Dynamic Models

by

Ruth Harbord

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

MOAC Doctoral Training Centre, University of Warwick

September 2017

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Prof. Tom Nichols for giving me the opportunity to pursue this project and for his ongoing guidance and support.

I would also like to thank the NeuroStat group at the University of Warwick (now the NISOx group at the University of Oxford). I am particularly grateful to Dr. Simon Schwab for his help with developing the MDM-DGM code. I would also like to thank Dr. Lilia Costa for providing the original code.

My extended thanks go to Prof. Jim Smith for his advice on the MDM and Prof. David Rossell, for going above and beyond to help me understand non-local priors. I would also like to thank Dr. Janine Bijsterbosch and Dr. Sonia Bishop for sharing their data with me.

I would like to thank my examiners, Prof. Will Penny and Dr. Theo Damoulas, for their detailed and helpful feedback.

Within the MOAC Doctoral Training Centre, my heartfelt thanks go to Prof. Alison Rodger for giving me the chance to undertake a PhD and her continued faith in me. I gratefully acknowledge financial support from the Engineering and Physical Sciences Research Council that made my PhD possible. I would also like to thank Dr. Hugo van den Berg, for his ongoing advice and for many interesting conversations over the years. I would like to thank my MOAC cohort, in particular my office buddies Katherine Lloyd and Victor Quan, whose moral support has been invaluable.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. Where I have drawn on the work of others, this is clearly attributed.

The work presented (including data generated and data analysis) was carried out by the author except in the case outlined below:

Preprocessed functional MRI datasets were provided by Dr. Janine Bijsterbosch, The FMRIB Centre, University of Oxford.

Parts of this thesis have appeared in

Harbord, R. et al., (2016), Scaling up Directed Graphical Models for Resting-State fMRI with Stepwise Regression *22nd Annual Meeting of the Organization for Human Brain Mapping, Geneva, Switzerland.*

Harbord, R. et al., (2015), Dynamic Effective Connectivity Modelling of Pain with Multiregression Dynamic Models *21st Annual Meeting of the Organization for Human Brain Mapping, Honolulu, Hawaii.*

Code for the MDM-DGM search was originally provided by Dr. Lilia Costa and has been optimised and extended by the author in collaboration with Dr. Simon Schwab. Functions for both exhaustive and stepwise MDM-DGM searches are available in the `multdyn` package for R[1,2].

[1] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`

[2] Schwab, S., Harbord, R., Costa, L., and Nichols, T. *multdyn: Multiregression Dynamic Models*, 2017a. URL `https://CRAN.R-project.org/package=multdyn`. R package version 1.6

Abstract

Functional magnetic resonance imaging (fMRI) is a non-invasive method for studying the human brain that is now widely used to study functional connectivity. Functional connectivity concerns how brain regions interact and how these interactions change over time, between subjects and in different experimental contexts and can provide deep insights into the underlying brain function.

Multiregression Dynamic Models (MDMs) are dynamic Bayesian networks that describe contemporaneous, causal relationships between time series. They may therefore be applied to fMRI data to infer functional brain networks. This work focuses on the MDM Directed Graph Model (MDM-DGM) search algorithm for network discovery. The Log Predictive Likelihood (model evidence) factors by subject and by node, allowing a fast, parallelised model search. The estimated networks are directed and may contain the bidirectional edges and cycles that may be thought of as being representative of the true, reciprocal nature of brain connectivity.

In Chapter 2, we use two datasets with 15 brain regions to demonstrate that the MDM-DGM can infer networks that are physiologically-interpretable. The estimated MDM-DGM networks are similar to networks estimated with the widely-used partial correlation method but advantageously also provide directional information. As the size of the model space prohibits an exhaustive search over networks with more than 20 nodes, in Chapter 3 we propose and evaluate stepwise model selection algorithms that reduce the number of models scored while optimising the networks. We show that computation time may be dramatically reduced for only a small trade-off in accuracy. In Chapter 4, we use non-local priors to derive new, closed-form expressions for the model evidence with a penalty on weaker, potentially spurious, edges. While the application of non-local priors poses a number of challenges, we argue that it has the potential to provide a flexible Bayesian framework to improve the robustness of the MDM-DGM networks.

Abbreviations

aMCC	Anterior mid-cingulate cortex
Amyg	Amygdala
AntIns	Anterior insula
BE	Backward elimination
BOLD	Blood-Oxygenation-Level Dependant
DAG	Directed Acyclic Graph
DCG	Directed Cyclic Graph
DCM	Dynamic Causal Modelling
DLM	Dynamic Linear Model
DLPFC	Dorsolateral prefrontal cortex
fMRI	Functional Magnetic Resonance Imaging
FS	Forward selection
LPL	Log Predictive Likelihood
MDM	Multiregression Dynamic Model
MDM-DGM	Multiregression Dynamic Model Directed Graph Model
MDM-IPA	Multiregression Dynamic Model Integer Programming Algorithm
OFC	Orbitofrontal cortex
PAG	Periaqueductal gray
PostIns	Posterior insula
pMOM	Product moment
SI	Primary somatosensory cortex
SII	Secondary somatosensory cortex
SEM	Structural Equation Model
SVAR	Structural Vector Autoregression
VMPFC	Ventromedial prefrontal cortex

Chapter 1

Inferring Brain Connectivity with Functional MRI

1.1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is a neuroimaging modality that offers a number of advantages: it is non-invasive and allows whole-brain coverage. It has high spatial resolution and relatively high temporal resolution, meaning the brain may be imaged in almost real time. Subsequently, fMRI has become a widely-used technique for examining both normal and pathological human brain function.

A century of neuroscience research has established that at the macroscopic level the brain is organised into a collection of distinct anatomical regions, each with its own highly specialised function. This specialisation has been referred to as *functional segregation* (Friston et al., 2013). One example is the visual cortex, where it has long been established that different aspects of an image (e.g. colour, motion, orientation) are processed separately in different cortical areas (Zeki and Shipp, 1988). Communication within and between these localised, specialised regions has been termed *functional integration* and is achieved through the extensive afferent, efferent and intrinsic myelinated axonal nerve fibres that compose the structural architecture of the brain (Zilles and Amunts, 2015). Deeper insights may be obtained by considering the activation and communication of brain regions over time, during a particular task or at rest. Functional MRI, in combination with appropriate statistical models, provides a powerful tool for inferring time-varying patterns of brain connectivity.

1.2 Thesis Outline

A typical fMRI scan measures the activity of a set of anatomical regions over the course of a few minutes (see the next section for more details). A number of methodologies have been developed which aim to infer brain connectivity from this type of data. Network models treat the brain as a collection of nodes (anatomical regions) and edges (connections between regions) and provide a powerful framework to model both structural and functional connectivity (Sporns, 2014). Approaches based on network modelling range from simple descriptions of the data to detailed models with a definite biophysical interpretation (Smith, 2012).

As well as the ability to infer the presence of connections between anatomical regions,

deeper insights may be obtained with models that also estimate the orientation of connections (i.e. the direction of information flow). Methods which allow bidirectional edges and cycles (thereby modelling feedback loops) may be desirable as they are likely to be more representative of the underlying brain function. Some models of brain connectivity also provide an estimate of the strength of the influence of one region over another. If these estimates are dynamic, it is possible to model how the strength of the influence changes over time. It may also be advantageous to be able to estimate networks and (potentially dynamic) connectivity strengths for individual subjects, as well as at the group level.

Costa (2014) developed two algorithms for network discovery using fMRI, based on the Multiregression Dynamic Model (MDM) of Queen and Smith (1993). The MDM-Integer Programming Algorithm (MDM-IPA) and MDM-Directed Graph Model (MDM-DGM) searches may be used to infer networks and provide dynamic connectivity estimates both for individual subjects and at the group level. These connectivity estimates are the regression coefficients of a Dynamic Linear Model (West and Harrison, 1997). While the MDM-IPA constrains each network to be a directed acyclic graph (DAG), the MDM-DGM permits cycles and bidirectional edges. Using simulated data, Costa et al. (2015) showed the MDM-IPA could perform as well as, or better than, a number of competing methods for inferring the presence and direction of edges. The MDM-IPA and MDM-DGM algorithms were also applied to real fMRI data with 11 brain regions (Costa, 2014; Costa et al., 2015, 2017).

Focusing on the MDM-DGM, in this thesis we extend the work of Costa (2014) and Costa et al. (2015, 2017) in a number of directions. To further explore the behaviour of the MDM-DGM search on real data, we made use of two fMRI resting-state datasets. The original experiment was designed to explore differences in connectivity between brain regions, specifically relating to trait and induced anxiety. Previous analysis of this data was reported in Bijsterbosch et al. (2015). We began by validating the MDM-DGM search by comparing the estimated networks with partial correlation networks as partial correlation is an established method for edge detection (see section 1.5.1). Given the close correspondence of the networks estimated by these two methods, we used two advantageous features of the MDM-DGM, the ability to infer the orientation of edges and the ability to provide time-varying connectivity weights, in order to further explore the role of certain brain regions in trait and induced anxiety. Results are presented in Chapter 2.

To gain fundamental insights into brain function, it is desirable to work with networks with much larger numbers of brain regions than have typically been used in models of directed connectivity. However, as will be discussed in Chapter 3, the most significant limitation in terms of the computational complexity of the MDM-DGM algorithm is the size of the model space, which increases exponentially with the number of brain regions. Using stepwise regression methods, forward selection and backward elimination, the size of the model space increases quadratically with the number of regions. Stepwise

methods therefore have the potential to allow the extension of the MDM-DGM to much larger networks. The performance of these algorithms is assessed in Chapter 3.

MDM-DGM fMRI networks (which we will present in Chapter 2) tend to contain a number of connections which occur inconsistently across subjects and have low connectivity strengths. We hypothesised that these connections were potentially spurious. In order the increase the sparsity of the networks, we considered non-local priors (Johnson and Rossell, 2012; Rossell and Telesca, 2017) to include a penalty on the model evidence for unnecessarily complex models. One advantageous feature of non-local priors is that the expressions for the penalised model evidence are closed-form. However, this approach also presents a number of theoretical and computational challenges, as will be discussed in Chapter 4.

This chapter introduces the use of functional Magnetic Resonance Imaging to infer dynamic, directed brain activity. For the remainder of this section, some of the key concepts are introduced, including the basic physiology behind the fMRI signal, the nature of functional brain networks and the insights into brain function that may be possible with this type of inference. In section 1.3, some of the methods that have been developed to date are outlined, with a particular focus on Bayesian networks and state-space models. The Multiregression Dynamic Model and the Dynamic Linear Model, which will be the focus of this work, are described in detail in section 1.4. Section 1.5 reviews some commonly-used algorithms for network discovery with fMRI data and our search procedure, the MDM Directed Graph Model (MDM-DGM) search, is outlined in section 1.6. Further discussion of the MDM-DGM, with a particular focus on the inference of graphs with cycles, is provided in section 1.7.

1.2.1 BOLD fMRI and Resting-State Networks

Functional MRI measures the blood oxygenation level-dependent (BOLD) contrast. Increased neuronal activity causes increased cerebral blood flow (CBF), increased cerebral blood volume (CBV) and oxygen consumption ($CMRO_2$). The increase in cerebral blood flow is greater than the increased oxygen consumption, resulting in a decrease in the total amount of deoxygenated hemoglobin (dHb). As deoxyhemoglobin is paramagnetic, the magnetic resonance signal is reduced in its vicinity, so a decrease in dHb results in a positive BOLD contrast. For a detailed review of the origins of the BOLD signal, see Mark et al. (2015). The BOLD signal is a hemodynamic response to neuronal activity, occurring at much slower timescales (hundreds of milliseconds to seconds) than the activity of the underlying neurons which occurs at the millisecond scale (Shmuel and Maier, 2015). This hemodynamic response is known to vary between cortical regions, subjects and experimental paradigms (Handwerker et al., 2012) and the exact mechanisms behind the coupling of neural activity, metabolism and hemodynamics are still an active area of research (for reviews see Ugurbil (2016) and Keilholz et al. (2017)).

Brain activity accounts for 20 % of the body's energy consumption, and most of this

energy is used on spontaneous activity, rather than task-based responses (Fox and Raichle, 2007). This has motivated the field of resting-state fMRI, where patterns of activity that are biophysically-meaningful and reproducible across subjects may be extracted from the scans of people at rest rather than engaged in a particular task. Spontaneous infra-slow (< 0.1 Hz) fluctuations in the BOLD response were shown to be strongly correlated between the motor cortices by Biswal et al. (1995) and since then a number of resting-state networks have been identified. In these networks, brain regions can interact strongly even though they may be spatially-distant. Resting-state networks not only show a strong correspondence with task-activation networks (Smith et al., 2009), but have been used to successfully predict task-based activation for individual subjects (Tavor et al., 2016).

There is strong evidence that resting-state fMRI has a neural basis. Keilholz (2014) provides a review of the use of electrophysiological techniques such as EEG (electroencephalography) and invasive intracranial recordings to explore the relationship between patterns of resting-state BOLD connectivity and the underlying neuronal processes. The relationship between BOLD response and the underlying electrophysiology is complex and it is likely that the BOLD signal arises from multiple electrophysiological processes. Resting-state networks have been found to have distinct spectral 'fingerprints', composed of multiple frequency bands of the EEG signal (Mantini et al., 2007) and resting-state connectivity patterns (correlation matrices) from (intracranial) electrocorticography recordings have been shown to correlate with fMRI data (Foster et al., 2015).

Using resting-state fMRI and diffusion spectrum imaging, a non-invasive method for determining structural (anatomical) connectivity, as well as a computational model, Honey et al. (2009) showed that structural connectivity was a good predictor of resting-state connectivity. Cortical regions that were connected anatomically exhibited stronger and more consistent resting-state connectivity when compared with anatomically unconnected regions. However, the reverse was not true, resting-state connectivity was an unreliable predictor of the underlying structural connectivity.

1.2.2 Functional vs. Effective Connectivity

The discovery of synchronised temporal activity across spatially-distant brain areas has given rise to the field of resting-state fMRI connectivity with a large body of research dedicated to the development and validation of methods for resting-state connectivity analysis. These methods range from simple descriptions of the data (e.g. full and partial correlation methods) to detailed biophysical models, which map the underlying neural activity to the observed hemodynamic response (Smith, 2012). Some of these methods will be reviewed later in this chapter. Central to fMRI connectivity modelling is the distinction between *functional* and *effective* connectivity, first described by Friston et al. (1993). Functional connectivity considers statistical dependencies between two neuronal systems, while effective connectivity describes the influence of one neuronal

system over another (Friston et al., 2013). Functional connectivity models correlations between brain regions, so the estimated connectivity is *undirected*, while effective connectivity methods are able to make inferences about directed connectivity (Smith, 2012). While methods that include directional information clearly have the potential to provide a much richer understanding of the data (and the underlying brain function), these models require careful interpretation, may be computationally-intensive and rely on specific and often different definitions of causality (see, for example, Henry and Gates (2017) for review).

1.2.3 Dynamic Functional Connectivity

The original research into resting-state functional connectivity assumed that connectivity was static over the duration of a typical fMRI scan, i.e. over periods over several minutes. However, in recent years, evidence has emerged which suggests temporal dynamics on much shorter timescales (Chang and Glover, 2010). This has motivated the development of a number of methods which aim to quantify this *dynamic* functional connectivity. As the exact relationship between the underlying neural activity and the observed dynamics in the BOLD signal is still unclear, interpretation of dynamic functional connectivity requires caution. Careful data preprocessing is necessary to remove the effects of head motion without destroying any true dynamics present in the BOLD signal (Laumann et al., 2016). Inappropriate statistical analysis can also lead to the erroneous detection of dynamic functional connectivity. Sliding window methods divide the time series into (sometimes overlapping) intervals and infer dynamics by comparing the estimated functional connectivity between these intervals. Laumann et al. (2016) used a sliding-window approach to test for dynamic connectivity on BOLD time series that had been simulated to be stationary, finding that most 'dynamic' connectivity could be attributed to sampling variability. For a detailed review of dynamic functional connectivity analysis, including discussion of the sliding window method, see Hutchison et al. (2013). It should also be emphasised that detection of dynamics using a statistical method does not in itself provide any information about the dynamics of the underlying system (Hindriks et al., 2016).

1.3 Modelling Functional and Effective Connectivity

In this section, we review models for estimating directed connectivity from fMRI data. This section focuses on model equations and interpretation. The application of some of these models as a basis for network discovery algorithms is reserved until section 1.5.

1.3.1 Bayesian Networks

Inferred directed brain networks from fMRI data relies on a number of concepts from the field of graphical models. Many of the methods developed to date (including the Multiregression Dynamic Model) are Bayesian network methods. Basic concepts and notation are defined in this subsection.

Graphical Models

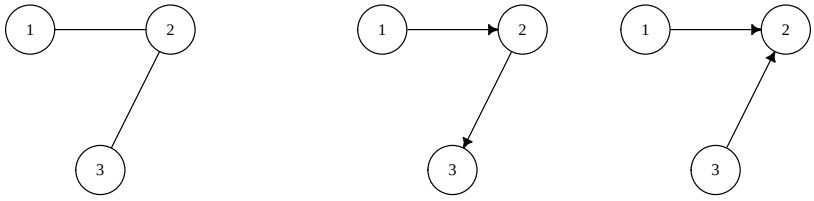
A graph may be defined by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is some set of vertices or *nodes* and $\mathcal{E}$ is the set of edges (or connections) between the nodes. Two nodes are *adjacent* if an edge exists between them, e.g. in Figure 1.1a, node 1 and node 2 are adjacent, as are node 2 and node 3, but node 1 and node 3 are not (i.e. there is no edge between node 1 and node 3). If each node represents a variable, then a graph may be used to represent interactions between a set of variables. These interactions may be undirected (as in Figure 1.1a) or directed (Figure 1.1b-1.1e), with arrows to denote the direction of influence. Directed edges can be used to represent causal relationships, so if there is an arrow from node 1 to node 2, representing the influence of the variable Y(1) on the variable Y(2), we have the interpretation that $Y(1)$ *causes* $Y(2)$. For any directed edge, the node from which the edge originates is called the *parent* and the node to which the arrow points is the *child* (Spirtes et al., 2000). If $Pa(i)$ and $Ch(i)$ denote the parents and children of node i respectively, then, for example, Figure 1.1c has $Pa(2) = \{1, 3\}$, $Ch(1) = \{2\}$ and $Ch(3) = \{2\}$. If a node has no parents, e.g. node 1 (and node 3), its parent set may be denoted by the empty set as $Pa(1) = \{\varnothing\}$.

A *directed acyclic graph* (DAG) is a directed graph which contains no cycles. The graphs in Figure 1.1b and 1.1c obey the DAG principle. Examples of non-DAGs, or *directed cyclic graphs* (DCGs), are shown in Figures 1.1d and 1.1e.

If there is a direct or indirect path from node i to node j, node j is a *descendant* of node i (Spirtes et al., 2000). The graph in Figure 1.1b is an example of a *chain* graph, where node 1 influences node 2 which influences node 3. There is an *indirect* influence of node 1 on node 3 and node 3 is a descendant of node 1. If the influence of node 2 on node 3 is known, the activity of node 3 may be explained without knowledge of node 1. This may be expressed more formally in terms of conditional independence. Let the graph in Figure 1.1b be represented by the set of variables $\mathbf{Y} = \{Y(1), Y(2), Y(3)\}$. The indirect nature of the influence of node 1 is described by the conditional independence relation

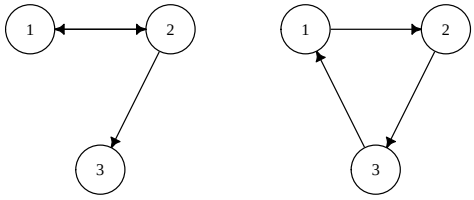
$$Y(3) \perp\!\!\!\perp Y(1) \,|\, Y(2)$$

which may be read as: node 3 is *conditionally independent* of node 1 given node 2. Given a set of conditional independence relations, a graph $\mathcal{G}$ may be represented by a probability distribution $\mathcal{P}$. A graph and its associated distribution satisfies the *Causal Markov Condition* if and only if for every vertex i in $\mathcal{V}$, i is independent of $\mathcal{V} \setminus \{Pa(i) \cup Ch(i)\}$ given $Pa(i)$. In other words, given its parents, node i is independent of its non-descendants (Spirtes et al., 2000; Mumford and Ramsey, 2014).



(a) **(b)** **(c)**

Undirected graph **Directed Acyclic Graphs**



(d) **(e)**

Directed Cyclic Graphs

Figure 1.1: Example graphs for a 3 node network.

The Markov condition means the probability distribution $\mathcal{P}$ factorises as

$$p[Y(1),\ldots,Y(n)] = \prod_{i=1}^{n} p[Y(i)\,|\,Pa(i)]$$

where n is the number of nodes in the graph $\mathcal{G}$. If the graph is acyclic, then, for some ordering of the nodes, its probability distribution will obey the Bayesian decomposition rule, such that it may be expressed as

$$p[Y(1),\ldots,Y(n)] = p[Y(1)]\,p[Y(2)\,|\,Y(1)],\ldots,p[Y(n)\,|\,Y(1),\ldots,Y(n-1)].$$

A graph with an associated probability distribution of this form comprise a Bayesian network, $\mathcal{B} = \{\mathcal{G},\mathcal{P}\}$. These concepts are illustrated in Figure 1.2 where a conditional probability may be obtained for each node in the graph as follows

Graph (i) $\quad p_{Y(1)} = p[Y(1)\,|\,\varnothing]$

Graph (ii) $\quad p_{Y(2)|Y(1)} = p[Y(2)\,|\,Pa(2)]$

Graph (iii) $\quad p_{Y(3)|Y(1),Y(2)} = p[Y(3)\,|\,Pa(3)].$

The probability distribution associated with the DAG in graph (iv) is therefore

Graph (iv) $\quad p[Y(1),\ldots,Y(3)] = p[Y(1)\,|\,\varnothing]\,p[Y(2)\,|\,Pa(2)]\,p[Y(3)\,|\,Pa(3)].$

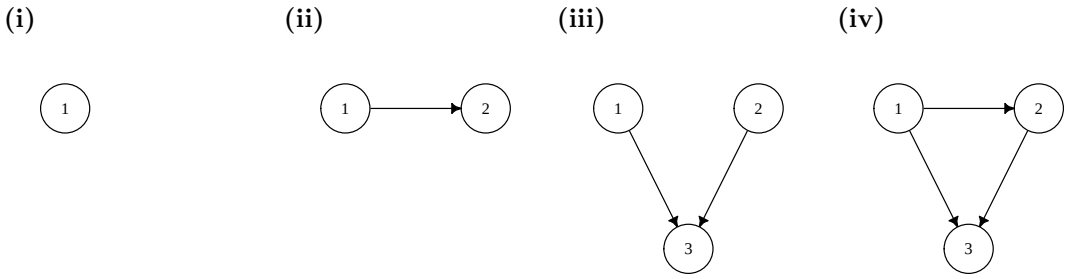


Figure 1.2: The probability distribution associated with a Bayesian network can be expressed in terms of a set of conditional probabilities.

Markov Equivalence

Consider the two graphs in Figure 1.3. Their associated joint probability distributions are

$$\textbf{Graph (i)} \qquad p[Y(1), Y(2), Y(3)] = p[Y(1)]\, p[Y(2) \,|\, Y(1)]\, p[Y(3) \,|\, Y(2)]$$

$$\textbf{Graph (ii)} \qquad p[Y(1), Y(2), Y(3)] = p[Y(1) \,|\, Y(2)]\, p[Y(2)]\, p[Y(3) \,|\, Y(2)].$$

Both graphs imply that conditional on node 2, node 1 is independent of node 3. Graphs with the same conditional independence structure are said to be *Markov equivalent*. Bayesian networks that are Markov equivalent will have the same skeleton (undirected graph) (Mumford and Ramsey, 2014). A Bayesian network cannot distinguish between Markov equivalent graphs.

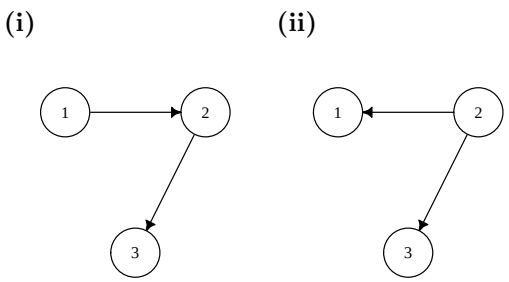


Figure 1.3: Example of Markov equivalent graphs. Graphs are said to be *Markov equivalent* if they share the same conditional independence relations.

Dynamic Bayesian Networks

An extension is a dynamic Bayesian network. Imagine there are T observations from n nodes so that at each time t there is an $n \times 1$ vector $\mathbf{Y}_t^\top = \{Y_t(1), Y_t(2), \ldots, Y_t(n)\}$. Using

a first-order Markovian transition model, the joint probability distribution factors as

$$p[\mathbf{Y}_1, \ldots, \mathbf{Y}_T] = p[\mathbf{Y}_1] \prod_{t=2}^{T} p[\mathbf{Y}_t \,|\, \mathbf{Y}_{t-1}]$$

$$= p[\mathbf{Y}_1] \prod_{t=2}^{T} \prod_{i=1}^{n} p[Y_t(i) \,|\, Pa(i)^t]$$

where $Pa(i)^t$ contains the parents for node i in time-slice t or $t-1$ (Bielza and Larrañaga, 2014). For a more in-depth discussion, see Costa (2014) and Bielza and Larrañaga (2014).

A detailed review of Bayesian networks for fMRI data is provided by Mumford and Ramsey (2014). Some network discovery algorithms which rely on Bayesian network principles will be reviewed in section 1.5.

1.3.2 Structural Equation Models

Consider the graph in Figure 1.4a. Let this graph represent a system about which some inference is to be made, i.e. let each node represent a brain region with some activity and each edge an interaction between the regions. The strength of the influence of one node on another is represented by the connection weights α and β. The graph in Figure 1.4a has a corresponding *structural equation model* (SEM), described by the equations in Figure 1.4b.

The variables of any SEM are split into *substantive* variables and error variables. The variables represented by each node are the substantive variables and the error terms represent the effect of any causes other than the substantive ones, such as the effect of exogenous variables. The equations in Figure 1.4b describe a linear SEM. In a linear SEM, every substantive variable is a linear function of the other substantive variables and its associated error (Spirtes et al., 2000). For example, the edge $Y(1) \rightarrow Y(2)$ (in Figure 1.4a) means that the variable $Y(1)$ appears in the right hand side of the equation for $Y(2)$:

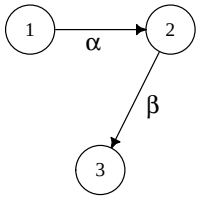
$$Y(2) = \alpha Y(1) + \epsilon_{Y(2)}$$

so that $Y(1)$ may be described as a direct, substantive cause of $Y(2)$.

More generally, consider a set of n brain regions, represented by a graph with n nodes. For each region, there is a BOLD time series with length T such that at time t, there is an $n \times 1$ vector of observations $\mathbf{Y}_t^\top = \{Y(1), \ldots, Y(n)\}$. This system may be written in matrix form as

$$\mathbf{Y}_t = \mathbf{G}_0 \mathbf{Y}_t + \boldsymbol{\epsilon}_t \tag{1.3.1}$$

where the $n \times n$ matrix $\mathbf{G}_0$ is called the *path coefficient* matrix and $\boldsymbol{\epsilon}$ is a $n \times 1$ error vector. Each component of $\mathbf{G}_0$ specifies an instantaneous effect between two regions



$$\begin{pmatrix} Y(1) \\ Y(2) \\ Y(3) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ 0 & \beta & 0 \end{pmatrix} \begin{pmatrix} Y(1) \\ Y(2) \\ Y(3) \end{pmatrix} + \begin{pmatrix} \epsilon_{Y(1)} \\ \epsilon_{Y(2)} \\ \epsilon_{Y(3)} \end{pmatrix}$$

(a) DAG with weighted edges **(b) SEM equations**

Figure 1.4: Inferring connectivity strengths with path analysis. A linear structural equation model consists of a graph with a corresponding set of equations. The path coefficient matrix is lower triangular and each entry represents a connectivity strength.

and the absence of a connection is represented by a zero (Penny et al., 2004; Chen et al., 2011). If the path coefficient matrix is lower triangular (as in Figure 1.4b), the structural equation model will represent a directed acyclic graph. An SEM that represents a directed acyclic graph is said to be *recursive*, while a *non-recursive* SEM may be used to model a directed cyclic graph (Spirtes, 1995).

In an SEM, the presence of an arrow from $Y(1)$ to $Y(2)$ means that $Y(1)$ *causes* $Y(2)$. These causal relationships are assumed *a priori* rather than inferred from the data so SEM methods involve estimating the set of connection strengths represented by the entries in the matrix $\mathbf{G}_0$ (Penny et al., 2004). It is possible to solve for $\mathbf{G}_0$ by rearranging equation 1.3.1 as

$$\mathbf{Y}_t = \mathbf{A}_0 \boldsymbol{\epsilon}_t$$

where the matrix $\mathbf{A}_0 = (\mathbb{I} - \mathbf{G}_0)^{-1}$ is called the *mixing* matrix. Under certain assumptions, this matrix may be estimated by independent component analysis (ICA) and, with appropriate permutation and normalisation, the matrix of connectivity strengths $\mathbf{G}_0$ may be obtained (Shimizu et al., 2006; Lacerda et al., 2012). Algorithms for both acyclic and cyclic graphs exist and will be discussed in more detail in section 1.5.

1.3.3 Structural Vector Autoregressive Models

Consider two time series from two brain regions $\mathbf{Y}(1)$ and $\mathbf{Y}(2)$. Let $\mathbf{Y}^{t-1}(r)$ be a vector containing all observations from region r up until time $t-1$, so that we may write $\mathbf{Y}^{t-1}(1)^\top = \{Y_1(1), \ldots, Y_{t-1}(1)\}$ and $\mathbf{Y}^{t-1}(2)^\top = \{Y_1(2), \ldots, Y_{t-1}(2)\}$. If a better prediction for $Y_t(2)$ at time t can be obtained using $\mathbf{Y}^{t-1}(1)$ than using only $\mathbf{Y}^{t-1}(2)$, then it may be said that $\mathbf{Y}(1)$ *Granger-causes* $\mathbf{Y}(2)$ (Granger, 1969; Mannino and Bressler, 2015). Granger causality is often implemented through *vector autoregressive* (VAR) models of the form

$$\mathbf{Y}_t = \sum_{k=1}^{K} \mathbf{G}_k \mathbf{Y}_{t-k} + \boldsymbol{\epsilon}_t.$$

where each coefficient matrix $\mathbf{G}_k$ describes effects k steps back in time and, as usual, $\boldsymbol{\epsilon}_t$ is an error vector.

An extension, which combines SEM and VAR models, is the *structural vector autoregressive* (SVAR) model, described by

$$\mathbf{Y}_t = \mathbf{G}_0 \mathbf{Y}_t + \sum_{k=1}^{K} \mathbf{G}_k \mathbf{Y}_{t-k} + \boldsymbol{\epsilon}_t$$

(Hyvärinen et al., 2010). Structural equation models describe *instantaneous* connectivity. Let the $(j,i)^{\text{th}}$ coefficient of the matrix $\mathbf{G}_0$ be denoted by $\mathbf{G}_0^{(j,i)}$. The definition of causality in an instantaneous SEM may be stated as follows: $Y(i)$ causes $Y(j)$ if $\mathbf{G}_0^{(j,i)} > 0$. Similarly, denote the $(j,i)^{\text{th}}$ coefficient of the matrix $\mathbf{G}_k$ by $\mathbf{G}_k^{(j,i)}$, where this coefficient represents the effect of the variable $Y_{t-k}(i)$ on the variable $Y_t(j)$. Using this framework, Hyvärinen et al. (2010) provide a definition of causality for the SVAR model such that $Y(i)$ is said to *cause* $Y(j)$ if at least one of the coefficients $\mathbf{G}_k^{(j,i)}$ is significantly non-zero for $k \geq 0$. Note that Granger causality does not assume any particular underlying causal mechanism (Mannino and Bressler, 2015).

Instantaneous or *within-sample* connectivity may be thought of as connectivity that occurs at much faster timescales than the temporal resolution of the data (Smith et al., 2013). Functional MRI causal searches often focus on contemporaneous connectivity. It was noted by Granger (1969) that causal effects may appear contemporaneous when the data are sampled at much slower rates than the underlying generational process. This is the case with fMRI data as the underlying neuronal processes occurs at much faster timescales than the measured BOLD signal, due to the relatively slow hemodynamic response (Henry and Gates, 2017).

1.3.4 State-Space Models

As the BOLD signal is an indirect measure of neuronal activity, models of directed connectivity which only consider the observed response may be unreliable, as the estimated 'causal' effects may arise due to variations in the timing of the hemodynamic response, rather than reflecting a true causal relationship between the brain regions of interest. As mentioned in section 1.2.1, the hemodynamic response is known to vary between cortical regions, as well as between subjects and populations (Handwerker et al., 2012). To overcome this, the state-space model framework, which is the focus of this section, defines directed connectivity in terms of a set of latent or *state* variables, which may then be related to the measured response. Assume each individual cortical region gives rise to an individual observation at time t so that for a system with n brain regions, $\boldsymbol{\theta}_t^\top = \{\theta_t(1), \ldots, \theta_t(n)\}$ is an $n \times 1$ vector representing some 'true' state of the system at time t. Let the state variables $\boldsymbol{\theta}_t^\top$ represent the activity of neurons in a cortical area, so that they may be said to represent some *quasi-neural* variable. The behaviour of these state variables is governed by a *state* equation. The state variables are then related to the measured variables (the BOLD response) through an *observation* equa-

tion. Let there be an n-dimensional random vector $\mathbf{Y}_t^\top = \{Y_t(1), Y_t(2), \ldots, Y_t(n)\}$ with a corresponding set of observed values $\mathbf{y}_t^\top = \{y_t(1), y_t(2), \ldots, y_t(n)\}$.

Let $\mathcal{N}(\cdot,\cdot)$ denote the multivariate normal distribution with some mean vector and covariance matrix. A linear dynamical system (LDS) may be described by

$$\textit{Observation equation} \qquad \mathbf{Y}_t = \mathbf{F}\boldsymbol{\theta}_t + \mathbf{v}_t \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \tag{1.3.2a}$$

$$\textit{State equation} \qquad \boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}) \tag{1.3.2b}$$

where $\mathbf{G}$ is a $n \times n$ *state transition* matrix, which describes directed interactions between hidden states. The coefficients of the matrix $\mathbf{G}$ may be interpreted as connectivity strengths, where the diagonal and off-diagonal elements control 'intrinsic' (within region) and 'extrinsic' (between region) effective connectivity respectively (Kahan and Foltynie, 2013). The $n \times n$ *observation* matrix $\mathbf{F}$ defines a linear relationship between the hidden states and the measured response. The n-dimensional vector $\mathbf{w}_t$ is the state evolution noise and the n-dimensional vector $\mathbf{v}_t$ is the observation noise. Both are assumed to be zero-mean (multivariate) Gaussian with (time-invariant) covariances $\mathbf{W}$ and $\mathbf{V}$ respectively (Roweis and Ghahramani, 1999).

It is possible to define some set of n known, exogenous inputs to the system at time t, via an n-dimensional vector $\mathbf{u}_t = \{u_t(1), \ldots, u_t(n)\}$. These external inputs might be, for example, the timings of a stimulus presented to a participant in a task-based fMRI study; these external variables are known because they are controlled by the experimenter. Let the strength of influence of these inputs be determined by an $n \times n$ coefficient matrix $\mathbf{D}$. The state equation 1.3.2b may now be extended to become

$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \mathbf{D}\mathbf{u}_t + \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}).$$

If the coefficient matrices $\mathbf{G}$, $\mathbf{D}$ and $\mathbf{W}$ are allowed to vary with time, but it is assumed only a small ($<< T$) number of distinct matrices exist, indicated by some index s with some associated transition probability $p(s_t = i \,|\, s_{t-1} = j)$, the state equation becomes that of the Switching Linear Dynamic System (SLDS) model, developed for BOLD data by Smith et al. (2010) and Smith et al. (2012):

$$\boldsymbol{\theta}_t = \mathbf{G}^{s_t}\boldsymbol{\theta}_{t-1} + \mathbf{D}^{s_t}\mathbf{u}_t + \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{W}^{s_t}).$$

The altered neuronal activity $\boldsymbol{\theta}_t$ may then change the neuronal activity of other regions, via the off-diagonal elements of $\mathbf{G}$ (or $\mathbf{G}_t$ if this matrix varies with time). These inputs directly influence the neuronal activity, and may be referred to as *driving* inputs. Additionally, it is also possible to specify *modulatory* inputs, which change the underlying neurodynamics (that is, the intrinsic and extrinsic connection strengths) (Penny et al., 2005). The coupling of brain regions in the presence of a modulatory input $x_t(j)$ may be described by an $n \times n$ coefficient matrix $\mathbf{C}_j$ (Ryali et al., 2011). The effects of these

two different types of input become clear if the state equation is extended as

$$\boldsymbol{\theta}_t = \left(\mathbf{G} + \sum_{j=1}^{J} x_t(j)\mathbf{C}_j\right)\boldsymbol{\theta}_{t-1} + \mathbf{D}\mathbf{u}_t + \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{W})$$

This is the state equation of the Multivariate Dynamical Systems (MDS) model of Ryali et al. (2011).

Both of these models operate in a discrete time framework. The LDS state equation may be expressed in continuous time, see Smith et al. (2013) for a detailed explanation of the parallels between the discrete and continuous time frameworks. Let $\dot{\boldsymbol{\theta}}$ be the first derivative of $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ be some set of time-invariant connectivity parameters (Razi and Friston, 2016). Define $\boldsymbol{\Theta} = \{\tilde{\mathbf{G}}, \tilde{\mathbf{C}}_1, \ldots, \tilde{\mathbf{C}}_J, \tilde{\mathbf{D}}, \boldsymbol{\Theta}^h\}$, where $\boldsymbol{\Theta}^h$ represents the parameters of the hemodynamic model and, following Smith et al. (2013), the ~ notation indicates continuous time variants of the matrices outlined above. The state equation of the widely-used, deterministic Dynamic Causal Model (DCM; Friston et al. (2003)) is

$$\dot{\boldsymbol{\theta}} = \left(\tilde{\mathbf{G}} + \sum_{j=1}^{J} u(j)\tilde{\mathbf{C}}_j\right)\boldsymbol{\theta} + \tilde{\mathbf{D}}\mathbf{u}.$$

Changes to the rate of change $\dot{\boldsymbol{\theta}}$ are known as second-order or bilinear effects (Kahan and Foltynie, 2013).

(a) Driving inputs **(b) Modulatory inputs**

Figure 1.5: Illustration of a state-space framework for models of effective connectivity. The neuronal activity of a region i is represented by an (unobservable) state variable $\theta(i)$ that varies in discrete or continuous time. **(a)** Driving inputs directly influence neuronal activity while modulatory inputs **(b)** affect neuronal activity indirectly by altering the connectivity strengths between nodes (the elements of $\mathbf{G}$). See Ryali et al. (2011) and Kahan and Foltynie (2013) for more specific illustrations of MDS and DCM respectively.

Within this general framework, illustrated in Figure 1.5, a stimulus $u_t(i)$ causes a change in some quasi-neural variable $\theta_t(i)$ which in turn causes a change in the measured response $Y_t(i)$, for some cortical region i at time t. The change in $\theta_t(i)$ may also cause

a change in some other region(s), e.g. $\theta_t(j)$ via the coefficient $\mathbf{G}^{(j,i)}$. Estimation of causal interactions is then equivalent to the estimation of the coefficient matrices ($\mathbf{G}$, $\mathbf{G}^{s_t}$ or $\tilde{\mathbf{G}}$, and $\mathbf{C}_j$ or $\tilde{\mathbf{C}}_j, j = 1, \ldots, J$). These types of models define causality in terms of the effect that one neural system exerts on another, in response to an input. To extend DCM to resting-state data, where there are no known external inputs, it becomes necessary to estimate not just the effective connectivity but also the hidden neuronal states that drive the endogenous activity of the system. This may be achieved with a stochastic DCM where the state equation becomes

$$\dot{\boldsymbol{\theta}} = \left(\tilde{\mathbf{G}} + \sum_{j=1}^{J} u_j \tilde{\mathbf{C}}_j\right)\boldsymbol{\theta} + \tilde{\mathbf{D}}\left(\mathbf{u} + \mathbf{w}^{(u)}\right) + \mathbf{w}^{(\theta)}$$

and $\mathbf{w}^{(\theta)}$ and $\mathbf{w}^{(u)}$ describe fluctuations in the states and hidden causes respectively (Li et al., 2011). However, inversion of stochastic models in the time domain is computationally-intensive. Alternatively, spectral DCM (spDCM) works in terms of the cross-spectra of the observed time series. Rather than estimating the (time-varying) hidden states, spDCM estimates their (time-invariant) covariance (Friston et al., 2014; Razi and Friston, 2016).

In order to more fully account for the hemodynamic response, these models replace the simple linear mapping in equation 1.3.2a by a more biophysically-informed relationship. For an individual region i, write

$$\mathbf{z}_t(i) = [\theta_t(i), \theta_{t-1}(i), \ldots, \theta_{t-L+1}(i)]^\top \tag{1.3.3a}$$

$$Y_t(i) = \mathbf{b}^\top(i)\boldsymbol{\Phi}\,\mathbf{z}_t(i) + v_t(i) \qquad v_t(i) \sim \mathcal{N}(0, V) \tag{1.3.3b}$$

where the hemodynamic response is represented by the product $\mathbf{b}^\top(i)\boldsymbol{\Phi}$, where $\mathbf{b}(i)$ provides region specific weights for the set of bases contained in $\boldsymbol{\Phi}$. These basis vectors span most of the variability in observed hemodynamic responses (Penny et al., 2005; Smith et al., 2010). The BOLD response is therefore modelled by the convolution of the hemodynamic response with the quasi-neural variable $\theta_t(i)$ L steps back into the past plus some zero-mean, uncorrelated, Gaussian observation error (Penny et al., 2005; Ryali et al., 2011). Observation models of this form are used in the SLDS and MDS models. DCMs, in comparison, use a more complex biophysical model (see Stephan et al. (2007) for a detailed description).

1.4 Multiregression Dynamic Models

1.4.1 The Multiregression Dynamic Model Equations

Imagine extending the linear dynamical state equations 1.3.2a and 1.3.2b so that the state transition matrix $\mathbf{G}$ and the observation matrix $\mathbf{F}$, as well as the state and observation covariances $\mathbf{W}$ and $\mathbf{V}$, may vary with time. As before, for a graph with n nodes, at some time t, there is an n-dimensional vector $\mathbf{Y}_t^\top = \{Y_t(1), Y_t(2), \ldots, Y_t(n)\}$ with observed values for each node $\mathbf{y}_t^\top = \{y_t(1), y_t(2), \ldots, y_t(n)\}$. Each observation

$Y_t(r)$ has a distribution determined by a $p_r \times 1$ state vector $\boldsymbol{\theta}_t(r)$, so for each node r, there is a linear dynamical system described by

$$\textit{Obs. equation} \qquad Y_t(r) = \mathbf{F}_t(r)^\top \boldsymbol{\theta}_t(r) + v_t(r) \qquad v_t(r) \sim \mathcal{N}[0, V_t(r)] \tag{1.4.1a}$$

$$\textit{state equation} \qquad \boldsymbol{\theta}_t(r) = \mathbf{G}_t(r)\boldsymbol{\theta}_{t-1}(r) + \mathbf{w}_t(r) \qquad \mathbf{w}_t(r) \sim \mathcal{N}[\mathbf{0}, \mathbf{W}_t(r)] \tag{1.4.1b}$$

where the observation matrix $\mathbf{F}$ has been replaced by a column vector with the same dimension as $\boldsymbol{\theta}_t(r)$. Denote the set of observations up to and including time t for region r by $\mathbf{Y}^t(r)^\top = \{Y_1(r), \ldots, Y_t(r)\}$, where the superscript indicates that we are considering *all* observations up to and including time t, rather than an individual time point. Similarly define $\mathbf{X}^t(r)^\top = \{\mathbf{X}_1(r)^\top, \ldots, \mathbf{X}_t(r)^\top\}$ and $\mathbf{Z}^t(r)^\top = \{\mathbf{Z}_1(r)^\top, \ldots, \mathbf{Z}_t(r)^\top\}$ with corresponding vectors of observations $\mathbf{x}^t(r)^\top = \{\mathbf{x}_1(r)^\top, \ldots, \mathbf{x}_t(r)^\top\}$ and $\mathbf{z}^t(r)^\top = \{\mathbf{z}_1(r)^\top, \ldots, \mathbf{z}_t(r)^\top\}$ such that

$$\mathbf{X}_t(r)^\top = \{Y_t(1), Y_t(2), \ldots, Y_t(r-1)\} \qquad 2 \le r \le n$$

$$\mathbf{Z}_t(r)^\top = \{Y_t(r+1), \ldots, Y_t(n)\} \qquad 2 \le r \le (n-1).$$

The column vector $\mathbf{F}_t(r)$ may then be defined as a known but arbitrary function of $\mathbf{x}^t(r)$ and $\mathbf{y}^{t-1}(r)$. It should not depend on $\mathbf{z}^t(r)$ or $y_t(r)$.

Denote a block diagonal matrix by $\mathsf{blockdiag}\{\}$. Define $\mathbf{G}_t = \mathsf{blockdiag}\{\mathbf{G}_t(1), \ldots, \mathbf{G}_t(n)\}$ and $\mathbf{W}_t = \mathsf{blockdiag}\{\mathbf{W}_t(1), \ldots, \mathbf{W}_t(n)\}$ where $\mathbf{G}_t(r)$ is the state matrix for node r and $\mathbf{W}_t(r)$ is the state variance for node r. These matrices may depend on past observations $\mathbf{x}^{t-1}(r)$ and $\mathbf{y}^{t-1}(r)$ but nothing else. The observation variance is denoted $V_t(r)$ such that $\mathbf{V}_t = \{V_t(1), \ldots, V_t(n)\}$. The observation and state error vectors, $\mathbf{v}_t = \{v_t(1), \ldots, v_t(n)\}$ and $\mathbf{w}_t^\top = \{\mathbf{w}_t(1)^\top, \ldots, \mathbf{w}_t(n)^\top\}$ respectively, are mutually independent with time, and the variables $v_t(1), \ldots, v_t(n)$ and $\mathbf{w}_t(1), \ldots, \mathbf{w}_t(n)$ are also mutually independent.

Finally, define some *initial information* to describe the system at time $t = 0$, given any information D_0 that is known *a priori*. Let $\boldsymbol{\theta}_0$ follow some distribution with moment parameters $\mathbf{m}_0$ and $\mathbf{C}_0$, where $\mathbf{m}_0$ is a vector $\mathbf{m}_0 = \{\mathbf{m}_0(1), \ldots, \mathbf{m}_0(n)\}$. Like $\mathbf{G}_t(r)$ and $\mathbf{W}_t(r)$, $\mathbf{C}_0$ is block diagonal and each $\mathbf{C}_0(r)$ is a $p_r \times p_r$ square matrix independent of everything except the past observations contained in $\mathbf{x}^{t-1}(r)$ and $\mathbf{y}^{t-1}(r)$.

Queen and Smith (1993) call $\{\mathbf{Y}_t\}_{t\ge1}$ a *Multiregression Dynamic Model* (MDM) if it is governed by n observation equations, a state equation[1] and initial information, defined

[1]Note that Queen and Smith (1993) refer to this as the *system* equation. For consistency with the state-space framework as outlined previously, we refer to the latent variables as the *state* variables when describing the MDM and DLM.

as

$$\begin{aligned}
&\textit{Obs. equations} && Y_t(r) = \mathbf{F}_t(r)^\top \boldsymbol{\theta}_t(r) + v_t(r) \quad v_t(r) \sim [0, V_t(r)] && (1.4.3a)\\
&\textit{State equation} && \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \mathbf{w}_t \qquad \mathbf{w}_t \sim (0, \mathbf{W}_t) && (1.4.3b)\\
&\textit{Initial information} && (\boldsymbol{\theta}_0 \,|\, D_0) \sim (\mathbf{m}_0, \mathbf{C}_0)\,. && (1.4.3c)
\end{aligned}$$

As $\mathbf{C}_0$ is block diagonal, the parameters for each variable are mutually independent at time $t = 0$ and the following conditional independence results hold:

Result 1
Given the observations up until time t, $\boldsymbol{\theta}_t(r)$ are mutually independent

$$\perp\!\!\!\perp_{r=1}^{n} \boldsymbol{\theta}_t(r) \mid \mathbf{y}^t.$$

Result 2
Given the observations up until time t for nodes 1,...,r, $\boldsymbol{\theta}_t(r)$ is independent of the rest of the past data

$$\boldsymbol{\theta}_t(r) \perp\!\!\!\perp \mathbf{z}^t(r) \mid \mathbf{x}^t(r), \mathbf{y}^t(r).$$

It follows from Result 1 that if the state variables $\boldsymbol{\theta}_0 = \{\boldsymbol{\theta}_0(1), \ldots, \boldsymbol{\theta}_0(n)\}$ are independent at time $t = 0$, the parameters associated with each variable remain independent over time and may be updated independently given data $\mathbf{y}^t$. Result 2 states that, given the current and previous observations from indexed series $1, \ldots, r$, the state vector $\boldsymbol{\theta}_t(r)$ is independent of the data from $(r+1), \ldots, n$.

The joint one-step-ahead forecast distribution for the MDM factors by node as

$$\begin{aligned}
p(\mathbf{y}_t \,|\, \mathbf{y}^{t-1}) &= \prod_{r=1}^{n} \int_{\boldsymbol{\theta}_t(r)} p[y_t(r) \,|\, \mathbf{x}^t(r), \mathbf{y}^{t-1}(r), \boldsymbol{\theta}_t(r)]\, p[\boldsymbol{\theta}_t(r) \,|\, \mathbf{y}^{t-1}]\, d\boldsymbol{\theta}_t(r)\\
&= \prod_{r=1}^{n} \int_{\boldsymbol{\theta}_t(r)} p[y_t(r) \,|\, \mathbf{x}^t(r), \mathbf{y}^{t-1}(r), \boldsymbol{\theta}_t(r)]\, p[\boldsymbol{\theta}_t(r) \,|\, \mathbf{x}^{t-1}(r), \mathbf{y}^{t-1}(r)]\, d\boldsymbol{\theta}_t(r).
\end{aligned}$$

Each observation follows a conditional, *univariate* Bayesian dynamic model.

The probability of the data over all time is

$$p[\mathbf{y}] = \prod_{t=1}^{T} p[\mathbf{y}_t \,|\, \mathbf{y}^{t-1}] = \prod_{t=1}^{T} \prod_{r=1}^{n} p[y_t(r) \,|\, \mathbf{x}^t(r), \mathbf{y}^{t-1}(r)]. \tag{1.4.4}$$

A full outline of MDM theory, including proofs for results 1 and 2, may be found in Queen and Smith (1993). The strength of the MDM is that it decomposes a complex multivariate system into n univariate ones (Queen and Smith, 1993). The individual probabilities $p[y_t(r) \,|\, \mathbf{x}^t(r), \mathbf{y}^{t-1}(r)]$ have a closed-form and may be calculated from the Dynamic Linear Model, as outlined in the next section.

1.4.2 The Dynamic Linear Model

We restrict our attention to *linear* Multiregression Dynamic Models, where the error distributions are Gaussian and the column vector $\mathbf{F}_t(r)$ is a linear function of $\mathbf{x}_t(r)$ with dimension $p_r \times 1$. Under these assumptions, the MDM equations 1.4.3a,1.4.3b and 1.4.3c as outlined by Queen and Smith (1993) may be simplified so that we may consider each individual node r in terms of a univariate Dynamic Linear Model (DLM), as described by West and Harrison (1997). If each state matrix $\mathbf{G}_t(r)$ is a $p_r \times p_r$ identity matrix and the observation variance is assumed to be constant over time, the DLM equations are

$$\begin{aligned}
&\textit{Obs. equation} && Y_t(r) = \mathbf{F}_t(r)^\top \boldsymbol{\theta}_t(r) + v_t(r) && v_t(r) \sim \mathcal{N}\left[0, \phi(r)^{-1}\right] \\
&\textit{State equation} && \boldsymbol{\theta}_t(r) = \boldsymbol{\theta}_{t-1}(r) + \mathbf{w}_t(r) && \mathbf{w}_t(r) \sim \mathcal{N}[\mathbf{0}, \mathbf{W}_t(r)] \\
&\textit{Initial information} && \boldsymbol{\theta}_0(r) \,|\, D_0 \sim \mathcal{N}[\mathbf{m}_0(r), \mathbf{C}_0(r)]\,.
\end{aligned}$$

At each time t, there is a $p_r \times 1$ state vector $\boldsymbol{\theta}_t(r)$. The $p_r \times 1$ state error vector is denoted by $\mathbf{w}_t(r)$ and follows a mean-zero multivariate normal distribution with $p_r \times p_r$ covariance matrix $\mathbf{W}_t(r)$. The observation variance is assumed to be normally- and independently-distributed with mean-zero and constant variance $\phi(r)^{-1}$. At time $t = 0$, any information known about the system may be represented in the initial information set D_0. This may include, for notational convenience, the (known) values of $\mathbf{F}_t(r)$ for all t. The $p_r \times 1$ prior mean vector $\mathbf{m}_0(r)$ and $p_r \times p_r$ covariance matrix $\mathbf{C}_0(r)$ must be specified *a priori*.

As the state variance $\mathbf{W}_t(r)$ is unknown, it is encoded through a scalar discount factor $\delta(r) \in [0.5, 1]$, such that

$$\mathbf{W}_t(r) = \frac{1 - \delta(r)}{\delta(r)} \mathbf{C}_{t-1}(r) \tag{1.4.6}$$

where $\mathbf{C}_{t-1}(r)$ is the posterior variance of the state variable $\boldsymbol{\theta}_t(r)$ at time $t - 1$. From equation 1.4.6, it is straightforward to see that if $\delta(r) = 1$, $\mathbf{W}_t(r) = 0$ for all time, and the corresponding model is static. Lower values of $\delta(r)$ treat the state variance as some fraction of the posterior variance at the previous time point; while this fraction is fixed, $\mathbf{C}_{t-1}(r)$ (and therefore $\mathbf{W}_t(r)$) may vary over time.

The posterior variance then becomes the 'prior' variance $\mathbf{R}_t(r)$ at time t, that is,

$$\mathbf{R}_t(r) = \mathbf{C}_{t-1}(r) + \mathbf{W}_t(r) = \frac{\mathbf{C}_{t-1}(r)}{\delta(r)}.$$

The posterior variance $\mathbf{C}_t(r)$ is updated at each time point t using the most recent observation $y_t(r)$.

The variances in the DLM that we need to estimate, the prior variance $\mathbf{R}_t$, the forecast variance Q_t and the posterior variance $\mathbf{C}_t$, may all be expressed as a product of

the observation variance (inverse precision) $\phi(r)^{-1}$ and a 'starred *scale-free*' variance parameter (West and Harrison, 1997, p.109), denoted by a *, i.e.

$$\mathbf{R}_t(r) = \phi(r)^{-1}\,\mathbf{R}_t^*(r) \qquad Q_t(r) = \phi(r)^{-1}\,Q_t^*(r) \qquad \mathbf{C}_t(r) = \phi(r)^{-1}\,\mathbf{C}_t^*(r).$$

Defining 'scale-free' variances in this way allows for these variance expressions to be updated via the DLM updating equations without any knowledge of $\phi(r)^{-1}$.

Define $D_t = \{D_0, y_1(r), \ldots, y_t(r)\}$, this is the initial information and the set of observations available up to and including time t. Denote the posterior mean for $\boldsymbol{\theta}_t(r)$ at time t as $\mathbf{m}_t(r)$, and the forecast mean at time t as $f_t(r)$. Then the system evolves according to

Posterior at time $t-1$	$p[\boldsymbol{\theta}_{t-1}(r)\,\vert\,\phi(r), D_{t-1}] \sim \mathcal{N}[\mathbf{m}_{t-1}(r), \phi(r)^{-1}\mathbf{C}_{t-1}^*(r)]$
Prior at time t	$p[\boldsymbol{\theta}_t(r)\,\vert\,\phi(r), D_{t-1}] \sim \mathcal{N}[\mathbf{m}_{t-1}(r), \phi(r)^{-1}\mathbf{R}_t^*(r)]$
One-step forecast	$p[Y_t(r)\,\vert\,\phi(r), D_{t-1}] \sim \mathcal{N}[f_t(r), \phi(r)^{-1}Q_t^*(r)]$
Posterior at time t	$p[\boldsymbol{\theta}_t(r)\,\vert\,\phi(r), D_t] \sim \mathcal{N}[\mathbf{m}_t(r), \phi(r)^{-1}\mathbf{C}_t^*(r)]$

with the parameters updated through

$$\begin{aligned}
f_t(r) &= \mathbf{F}_t(r)^\top \mathbf{m}_{t-1}(r) \\
Q_t^*(r) &= \mathbf{F}_t(r)^\top \mathbf{R}_t^*(r)\mathbf{F}_t(r) + 1 \\
\mathbf{m}_t(r) &= \mathbf{m}_{t-1}(r) + \frac{\mathbf{R}_t^*(r)\mathbf{F}_t(r)[Y_t(r) - f_t(r)]}{Q_t^*(r)} \\
\mathbf{C}_t^*(r) &= \mathbf{R}_t^*(r) - \frac{\mathbf{R}_t^*(r)\mathbf{F}_t(r)\mathbf{F}_t(r)^\top \mathbf{R}_t^*(r)}{Q_t^*(r)}.
\end{aligned}$$

At $t = t_0$, the prior on the precision is

$$p[\phi(r)\,\vert\,D_0] \sim \mathcal{G}\left(\frac{n_0(r)}{2}, \frac{d_0(r)}{2}\right) \tag{1.4.8}$$

where $\mathcal{G}(\cdot,\cdot)$ denotes the gamma distribution with shape and rate parameters. The prior hyperparameters $n_0(r)$ and $d_0(r)$ must be specified *a priori*. Specification of the hyperparameters will be discussed further in subsection 1.6.1. At any time t, the updated prior on the precision is

$$p[\phi(r)\,\vert\,D_t] \sim \mathcal{G}\left(\frac{n_t(r)}{2}, \frac{d_t(r)}{2}\right) \tag{1.4.9}$$

with the hyperparameters updated at each time point using

$$n_t(r) = n_{t-1}(r) + 1$$

$$d_t(r) = d_{t-1}(r) + \frac{[Y_t(r) - f_t(r)]^2}{Q_t^*(r)}.$$

At time t, the updated estimate for the observation variance is given by

$$S_t(r) = \frac{1}{\mathbb{E}[\phi(r) \,|\, D_t]} = \frac{d_t(r)}{n_t(r)}$$

Let $\mathcal{T}_{\cdot}(\cdot,\cdot)$ denote the t-distribution with degrees of freedom, and location and scale parameters. The final marginal distributions are then

Posterior at time $t-1$	$p[\boldsymbol{\theta}_{t-1}(r) \,|\, D_{t-1}] \sim \mathcal{T}_{n_{t-1}(r)}[\mathbf{m}_{t-1}(r), \mathbf{C}_{t-1}(r)]$	(1.4.10a)
Prior at time t	$p[\boldsymbol{\theta}_t(r) \,|\, D_{t-1}] \sim \mathcal{T}_{n_{t-1}(r)}[\mathbf{m}_{t-1}(r), \mathbf{R}_t(r)]$	(1.4.10b)
One-step forecast	$p[Y_t(r) \,|\, D_{t-1}] \sim \mathcal{T}_{n_{t-1}(r)}[f_t(r), Q_t(r)]$	(1.4.10c)
Posterior at time t	$p[\boldsymbol{\theta}_t(r) \,|\, D_t] \sim \mathcal{T}_{n_t(r)}[\mathbf{m}_t(r), \mathbf{C}_t(r)].$	(1.4.10d)

The estimates for the scale parameters are

$$\mathbf{R}_t(r) = S_{t-1}(r)\mathbf{R}_t^*(r) \qquad Q_t(r) = S_{t-1}(r)Q_t^*(r) \qquad \mathbf{C}_t(r) = S_t(r)\mathbf{C}_t^*(r).$$

Retrospective Distributions

Equations 1.4.10b and 1.4.10c give the one-step ahead forecast distributions for $\boldsymbol{\theta}_t(r)$ and $Y_t(r)$. The one-step forecast for $Y_t(r)$ provides a simple, closed-form formula for the likelihood stated in equation 1.6.1 while $\boldsymbol{\theta}_t(r)$ estimates the strength of the regressors (the parent nodes) at time t given data $y_1(r), \ldots, y_t(r)$. When examining the behaviour of $\boldsymbol{\theta}(r)$ over time, it is informative to consider not only the one-step estimates, but also *retrospective* estimates, $\{\boldsymbol{\theta}_T(r), \boldsymbol{\theta}_{T-1}(r), \ldots, \boldsymbol{\theta}_1(r)\}$ given all the data, $\mathbf{y}(r) = \{y_1(r), \ldots, y_T(r)\}$. These may be obtained in a similar, one-step manner via the recursive relations outlined below. In order to maintain the notation used by West and Harrison (1997), the (r) notation is dropped temporarily so that $\boldsymbol{\theta}_t(r) = \boldsymbol{\theta}_t$, $\phi(r) = \phi$ etc. Then the bracket notation denotes the parameters k steps back in time.

We have

$$p(\boldsymbol{\theta}_{t-k} \,|\, D_t) \sim \mathcal{T}_{n_t}\left[\mathbf{a}_t(-k), \frac{S_t}{S_{t-k}}\mathbf{R}_t^*(-k)\right] \qquad k \geq 0. \tag{1.4.11}$$

The parameters of this distribution may be obtained using the recursive relations

$$\mathbf{a}_t(-k) = \mathbf{m}_{t-k} + \mathbf{B}_{t-k}[\mathbf{a}_t(-k+1) - \mathbf{m}_{t-k}] \qquad \mathbf{a}_t(0) = \mathbf{m}_t$$

$$\mathbf{R}_t(-k) = \mathbf{C}_{t-k} + \mathbf{B}_{t-k}[\mathbf{R}_t(-k+1) - \mathbf{R}_{t-k+1}]\mathbf{B}_{t-k} \qquad \mathbf{R}_t(0) = \mathbf{C}_t \tag{1.4.12a}$$

where

$$\mathbf{B}_t = \mathbf{C}_t\,\mathbf{R}_{t+1}^{-1}.$$

Note that $\mathbf{C}_t\,\mathbf{R}_{t+1}^{-1} = \phi^{-1}\phi\,\mathbf{C}_t^*(\mathbf{R}_{t+1}^*)^{-1}$ and $\mathbf{R}_t^*(0) = \mathbf{C}_t^*$. For unknown variance ϕ^{-1}, we may write equation 1.4.12a in terms of S_t, its best estimate at time t:

$$\begin{aligned} S_t\mathbf{R}_t^*(-k) &= S_{t-k}\mathbf{C}_{t-k}^* + \mathbf{B}_{t-k}[S_t\mathbf{R}_t^*(-k+1) - S_{t-k}\mathbf{R}_{t-k+1}^*]\mathbf{B}_{t-k} \\ &= S_{t-k}\left[\mathbf{C}_{t-k}^* + \mathbf{B}_{t-k}\left[\frac{S_t}{S_{t-k}}\mathbf{R}_t^*(-k+1) - \mathbf{R}_{t-k+1}^*\right]\mathbf{B}_{t-k}\right]. \end{aligned}$$

Dynamic Linear Model theory is outlined in detail in West and Harrison (1997, Chapter 4).

Using these relations, it is possible to construct

$$p[\boldsymbol{\theta}_t(r)\,|\,\mathbf{y}(r)] \sim \mathcal{T}_{n_T(r)}[\boldsymbol{\mu}_t(r), \boldsymbol{\Sigma}_t(r)] \tag{1.4.13}$$

with

$$\boldsymbol{\mu}_t(r) = \mathbf{m}_t(r) + \mathbf{C}_t(r)\mathbf{R}_{t+1}(r)^{-1}[\boldsymbol{\mu}_{t+1}(r) - \mathbf{m}_t(r)] \tag{1.4.14a}$$

$$\boldsymbol{\Sigma}_t^*(r) = \mathbf{C}_t^*(r) + \mathbf{C}_t^*(r)\mathbf{R}_{t+1}^*(r)^{-1}[\boldsymbol{\Sigma}_{t+1}(r) - \mathbf{R}_{t+1}^*(r)]\,\mathbf{C}_t^*(r)\mathbf{R}_{t+1}^*(r)^{-1} \tag{1.4.14b}$$

$$\boldsymbol{\Sigma}_t(r) = S_T(r)\boldsymbol{\Sigma}_t^*(r). \tag{1.4.14c}$$

In this work, we use $\mathbf{m}_t(r)$ and $\mathbf{C}_t(r)$ to denote the parameters of equation 1.4.10d (that is, estimates for $\boldsymbol{\theta}_t(r)$ given the observations up until time t). We use $\boldsymbol{\mu}_t(r)$ and $\boldsymbol{\Sigma}_t(r)$ to denote the parameters of equation 1.4.11 (estimates for $\boldsymbol{\theta}_t(r)$ given all the data $\mathbf{y}(r)$).

1.4.3 MDM Interpretation

In this section, we describe the application of the MDM to fMRI data. We highlight some relevant features, which may be compared and contrasted to the models outlined above (in section 1.3).

MDMs describe contemporaneous, causal relationships between time series

We are interested in the dependence of node r on some set of parent nodes $Pa(r)$. Algorithms for discovering the parent set will be described in detail in section 1.6. For

now, assume for each r there is some known parent set at time t and this parent set is contained in the vector $\mathbf{X}_t(r)$. It follows that the column vector $\mathbf{F}_t(r)$ is a linear function of the parents of $Y_t(r)$.

We may write

$$Y_t(r) \perp\!\!\!\perp \left\{\mathbf{Y}^t(1), \mathbf{Y}^t(2), \ldots, \mathbf{Y}^t(r-1)\right\} \setminus Pa[\mathbf{Y}^t(r)] \,|\, Pa[\mathbf{Y}^t(r)], \mathbf{Y}^{t-1}(r)$$

which may be read as: given the values of the parent nodes up to and including time t, and the values of itself up to time $t-1$, node r at time t is independent of any nodes that are not in its parent set.

An MDM describes a dynamic Bayesian network. At each time t, there is a Bayesian network representing contemporaneous, causal relationships between the time series. For each variable in the parent set $Pa[Y_t(r)] \subseteq \{Y_t(1), \ldots, Y_t(n)\}$, there is a directed arc from the parent to $Y_t(r)$ (Queen and Albers, 2009). When applied to fMRI data, MDMs allow us not only to model connectivity that is directed, but also to distinguish between Markov equivalent graphs (Costa, 2014; Costa et al., 2015, 2017).

The state vector $\boldsymbol{\theta}_t(r)$ provides a measure of connectivity strength

Each $Y_t(r)$ is modelled by a regression Dynamic Linear Model where its parents are linear regressors (Queen and Albers, 2009). Using a Dynamic Linear Model, we may obtain estimates for the regression coefficients $\hat{\boldsymbol{\theta}}_t(r)$, which may be interpreted as instantaneous connectivity strengths. Note that if $Y_t(r)$ has no parents, it may be modelled by any appropriate DLM (Queen and Albers, 2009). The DLM parameter estimates are t-distributed (see equations 1.4.10d and 1.4.11) and can be quickly computed through one-step updating.

The Dynamic Linear Model estimates time-varying connectivity

As stated in equation 1.4.6, the dynamics are controlled by a single, scalar parameter, the discount factor $\delta(r)$. As an individual DLM is fitted to each node r, $\delta(r)$ may vary between nodes, hence the connectivity strengths are allowed to vary over time as much as is appropriate for the data. This includes the stationary model with $\delta(r) = 1$.

To fit a DLM, with some parent set $Pa(r)$, we need to specify the following parameters: the discount factor $\delta(r)$ and the prior hyperparameters $\mathbf{m}_0(r)$, $\mathbf{C}_0^*(r)$, $n_0(r)$ and $d_0(r)$. Because the DLM is specified in terms of one-step updating relations, with a new prior at each time t based on the distributions at time $t-1$ (see equations 1.4.10a to 1.4.10d), it is possible to choose weakly-informative values for the prior hyperparameters, such that, after a small number of initial time points, the effect of the prior hyperparameters on the updated parameter estimates $\boldsymbol{\theta}_t(r)$ is negligible. This was shown in Costa (2014).

The interpretation of the state variables and the nature of causality are therefore very different in the DLM/MDM framework than the linear dynamical systems models (e.g. SLDS, MDS) and Dynamic Causal Models described in section 1.3. While the MDM is a

state-space representation, we do not interpret the connectivity in terms of hidden neuronal states and the observation equation does not explicitly model the hemodynamic response. However, it provides a flexible and computationally-efficient framework to model dynamic, directed connectivity.

1.5 Network Discovery

In this section, we review methods for network discovery, some of which are based on the models outlined in section 1.3. In a now highly-cited paper, Smith et al. (2011) assessed the performance of a number of methods for network discovery, using a set of simulations designed to replicate fMRI data. Specifically, they assessed the ability of each method to detect the presence of edges and, where relevant, the ability to correctly identify directionality. The sensitivity of the methods to detect edge presence was quantified using the mean fractional rate of detecting true connections, a metric termed *c-sensitivity*. This metric uses the 95th percentile of the false positive distribution as a threshold, so that an edge with a higher connection strength than this threshold is considered a true positive (Smith et al., 2011). Additionally, *d-accuracy*, obtained by subtracting the connection strength in one direction from the connection strength in the opposite direction for each true connection and expressed as a mean fraction over subjects and edges, quantifies the effectiveness of a method of detecting directionality. (The definition of 'connection strength' varied across methods.) The c-sensitivity of these methods applied to human resting-state fMRI data was assessed by Dawson et al. (2013): in this study, the ground truth was based on detailed anatomical knowledge of the primate visual cortex, assuming that functional connectivity is reflective of the underlying anatomical connectivity.

1.5.1 Partial Correlation for Functional Connectivity

When inferring connectivity, we are interested in detecting the presence or absence of *direct* relationships between any two brain regions. If it is assumed there are no unmeasured regions that act as a common cause, the partial correlation between regions i and j, that is, the correlation when the influence of all other measured regions has been regressed out, may be interpreted as a *direct* influence between i and j. If data $\mathbf{Y}^\top = \{\mathbf{Y}(1), \ldots, \mathbf{Y}(n)\}$ are assumed to be drawn from a zero-mean, multivariate Gaussian with $n \times n$ covariance matrix $\mathbf{\Sigma}$ and precision (inverse covariance) matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, then zero elements in the precision matrix correspond to conditional independence relations, such that a matrix of partial correlations may be represented by an undirected graph. This graph is undirected because partial correlation networks are symmetric.

The partial correlation Π_{ij} between region i and region j is

$$\Pi_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}$$

(Marrelec et al., 2006). In practice, the precision matrix is unknown and must be estimated from the data. Adopting the notation of Friedman et al. (2008), let $\mathbf{S}$ denote

the sample covariance matrix. Then the maximum likelihood estimate (MLE) for the precision matrix is given by

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathbb{N}_n}{\arg\max} \left[\log |\boldsymbol{\Theta}| - \mathrm{Tr}(\mathbf{S}\boldsymbol{\Theta}) \right]$$

where $\mathbb{N}_n$ denotes the family of $n \times n$ positive-definite matrices. However, a unique solution (a unique precision matrix) only exists if $\boldsymbol{\Sigma}$ is positive-definite. This will not be the case if the number of observations T is smaller than the number of nodes n and, even in the case where $n < T$, the MLE may be ill-behaved (Pourahmadi, 2011; Hinne et al., 2015). For this reason, the graphical LASSO (Least Absolute Shrinkage and Selection Operator) method adds a penalty term to the MLE via a shrinkage parameter λ:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathbb{N}_n}{\arg\max} \left[\log |\boldsymbol{\Theta}| - \mathrm{Tr}(\boldsymbol{\Theta}\boldsymbol{\Sigma}) - \lambda \|\boldsymbol{\Theta}\|_1 \right]$$

(Friedman et al., 2008; Banerjee et al., 2008). One drawback of these methods is that both the maximum likelihood estimate and the penalised maximum likelihood estimate provide a point estimate so there is no quantification of the reliability of the estimate. Instead it may be advantageous to use a Bayesian framework which specifies a posterior distribution over $\boldsymbol{\Theta}$. For further discussion, and an application developed for fMRI functional connectivity inference, see Hinne et al. (2015).

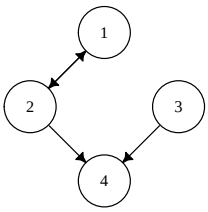
Partial correlation and regularised partial correlation (e.g. the graphical LASSO) only estimate functional connectivity and therefore only provide a *description* of the data (Smith, 2012). However, in both the Smith et al. (2011) and Dawson et al. (2013) studies, these methods proved to be some of the best-performing for correctly identifying edge presence: both partial correlation and regularised partial correlation achieved c-sensitivities of above 90 % on the simulated data (Smith et al., 2011), and c-sensitivities of 81 % and 84 % respectively on the human fMRI data (Dawson et al., 2013). These methods are also computationally-efficient and may readily be applied to larger-scale networks (i.e. networks with more than 20 brain regions). These reasons led Smith (2012) to recommend partial correlation, and ideally regularised partial correlation methods, to be some of the best approaches for functional connectivity estimation, as well as Bayesian network methods detailed in the following sections.

1.5.2 PC Algorithm

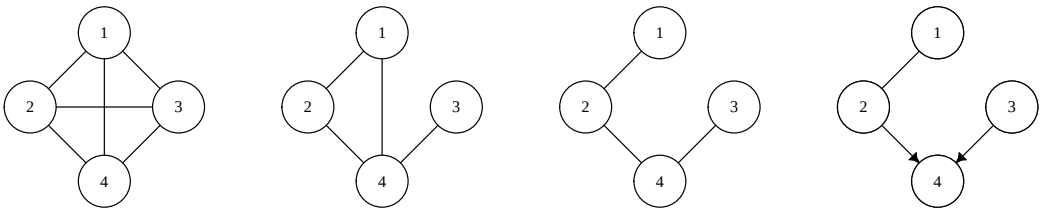
The Peter-Spirtes, Clark-Glymour (PC) algorithm is a method for Bayesian network discovery based on testing for conditional independence relations in the data. It consists of a two-step procedure, where the first step estimates edge presence and the second step orientates these edges to produce a directed or partially directed graph. The PC algorithm does not allow cycles, so an edge that cannot be oriented is returned as an undirected edge. An outline of the PC procedure is shown in Figure 1.6.

On both simulated and real data, the PC algorithm proved successful at identifying

edge presence, with c-sensitivity of above 90 % reported in the Smith et al. (2011) study and 77 % in the Dawson et al. (2013) study. However, its d-accuracy, a test of its ability to identify directionality (at the individual subject level), was found by the Smith et al. (2011) study to be no greater than chance.



Step	Instruction	Graph
0.	Start with complete undirected graph	**(a)**
1a.	Compute zero order independencies $Y(1) \perp\!\!\!\perp Y(3)$, $Y(2) \perp\!\!\!\perp Y(3)$	**(b)**
1b.	Compute first order independencies $Y(1) \perp\!\!\!\perp Y(4) \mid Y(2)$	**(c)**
2.	Orientate edges, return (partially) directed graph	**(d)**



(a) **(b)** **(c)** **(d)**

Figure 1.6: Illustration of the PC algorithm. The 'true' structure to be estimated is the directed cyclic graph on the top left. The PC algorithm begins with the complete, undirected graph in **(a)**, and uses conditional independence to remove edges (as in **(b)** and **(c)**). Edges may then be orientated, for example, $Y(4)$ is a *collider* because it is a common neighbour of $Y(2)$ and $Y(3)$ but is not in the conditioning set that rendered them independent. As the algorithm doesn't allow cycles, and both $Y(1) \rightarrow Y(2)$ and $Y(1) \leftarrow Y(2)$ could be true, this edge remains undirected and a partial DAG **(d)** is returned (Spirtes et al., 2000; Mumford and Ramsey, 2014).

1.5.3 GES and IMaGES

Another Bayesian network method tested by Smith et al. (2011) was the Greedy Equivalence Search (GES). Following Chickering and Meek (2002), two graphs $\mathcal{G}$ and $\mathcal{G}'$ are *equivalent* if they have the same probability distribution and the same independence constraints. Let ξ denote an equivalence class such that the equivalence of $\mathcal{G}$ and $\mathcal{G}'$ (written as $\mathcal{G} \approx \mathcal{G}'$) implies $\mathcal{G} \in \xi(\mathcal{G}')$ and $\mathcal{G}' \in \xi(\mathcal{G})$.

A graph $\mathcal{G}$ is *included* in a graph $\mathcal{H}$ if every probability distribution and independence constraint in $\mathcal{G}$ is also in $\mathcal{H}$, and this may be denoted by $\mathcal{G} \leq \mathcal{H}$. Then if $\mathcal{G} \leq \mathcal{G}'$ or $\mathcal{G}' \leq \mathcal{G}$, and the number of edges between the two graphs differs by one, the equivalence classes $\xi_1(\mathcal{G})$ and $\xi_2(\mathcal{G}')$ are said to be *adjacent*.

The Greedy Equivalence Search (GES) algorithm uses a *score equivalent* scoring criterion, where any DAGs within an equivalence class have the same score. It begins with an empty graph, scoring adjacent equivalent classes until it reaches a local maximum. This *forward equivalence search* (FES) proceeds such that if $\mathcal{G} \leq \mathcal{G}'$, the algorithm

moves from $\xi_1(\mathcal{G})$ to $\xi_2(\mathcal{G}')$. Once the local maximum is reached, a *backward equivalence search* moves between adjacent equivalence classes which have one less edge, until this single edge removal fails to improve the score (Chickering and Meek, 2002).

Imagine we have fMRI time series data with length T for S subjects. Denote the maximum likelihood estimate for subject s by ML_s and the number of free parameters (the number of directed edges plus the number of nodes) by k. A widely-used scoring criterion is the Bayesian information criterion (BIC)

$$\mathrm{BIC} = -2\log_e(ML_s) + k\log_e(T).$$

In the Smith et al. (2011) study, like the partial correlation methods and PC algorithm, GES achieved c-sensitivities of over 90 %. In the Dawson et al. (2013), it performed less well, with a c-sensitivity of ~60 %, although the authors note that Bayes net methods have the best performance out of all methods when the metric is based on the separation between the numbers of expected and unexpected connections, rather than the c-sensitivity (Dawson et al., 2013). Like the PC algorithm, the ability of GES to determine directionality was limited, with d-accuracy of less than 60 % (where the chance level is 50 %) reported by Smith et al. (2011).

An extension to GES is the Independent Multisample Greedy Equivalence Search (IMaGES) algorithm uses a BIC score combined over subjects

$$\mathrm{BIC} = -\frac{2}{S}\sum_{s=1}^{S}\log_e(ML_s) + c\,k\log_e(T)$$

where c is a penalty term to remove weaker (and potentially spurious) edges (Mumford and Ramsey, 2014). IMaGES was developed by Ramsey et al. (2010) and applied to the simulation data of Smith et al. (2011) in Ramsey et al. (2011). Using this method, edge identification could be as much as 100 %.

1.5.4 LiNGAM and LOFS

Table 1.1 shows the steps of the LiNG discovery algorithm. Consider an incorrect model represented by the path coefficient matrix $\mathbf{G}_*$, such that

$$\mathbf{Y} = \mathbf{G}_*\mathbf{Y} + \mathbf{r}$$

where $\mathbf{r}$ are the residuals. By writing

$$\mathbf{Y} = (\mathbb{I} - \mathbf{G}_0)^{-1}\boldsymbol{\epsilon} = (\mathbb{I} - \mathbf{G}_*)^{-1}\mathbf{r}$$

it is straightforward to see that the residuals of the incorrect model $\mathbf{G}_*$ are linear combinations of the residuals of the correct model $\mathbf{G}_0$

$$\mathbf{r} = (\mathbb{I} - \mathbf{G}_*)(\mathbb{I} - \mathbf{G}_0)^{-1}\boldsymbol{\epsilon}.$$

If $\mathcal{NG}()$ is some measure of 'non-Gaussianity', then for the correct model $\mathcal{NG}(\epsilon_i) = \mathcal{NG}(r_i)$. For any other model there will be, for some a, a score $\mathcal{NG}(a + \epsilon)$. As, by the Central Limit Theorem, the sum will be more Gaussian of any of its summands, it is possible to infer the correct model by maximising the non-Gaussianity of the residuals (Ramsey et al., 2011; Mumford and Ramsey, 2014). This is the basis of the LiNG Orientation Fixed Structure (LOFS) algorithms of Ramsey et al. (2011, 2014). Of particular relevance are LOFS-R1 and LOFS-R4 (Rule 1 and Rule 4), as these algorithms return graphs which may contain cycles. Given an undirected graph (which may be found with the PC algorithm, or GES or IMaGES), LOFS-R1 considers each node individually, choosing the set of parents (from the adjacent edges in the undirected graph) that maximises a score of non-Gaussianity, e.g. the Anderson-Darling statistic for normality. If an edge cannot be orientated, it can be returned undirected. As can be seen in Figure 1.7d, it is possible for this algorithm to identify cycles and 2-cycles (bidirectional edges). Ramsey et al. (2011) note that a 2-cycle may be due to actual feedback between two nodes, or an unrecorded, latent common cause (or both). LOFS-R4 is a simplified implementation of the LiNG algorithm (see Table 1.1), which does not permit self-loops. Edges which are absent in the undirected graph are replaced by zeros in the coefficient matrix, and the non-Gaussianity is maximised for each row of $\mathbf{W}_{ICA}$ (the matrix obtained by independent component analysis) (Lacerda et al., 2012; Ramsey et al., 2014).

In the Smith et al. (2011) and Dawson et al. (2013) studies, LiNGAM performed very poorly, in both cases achieving c-sensitivities of less than 20 % and direction accuracy only marginally above the level of chance. However, extensions based on the LiNG procedure have been more successful. Pairwise LiNGAM, for example, orientates each adjacent edge individually based on a log-likelihood ratio

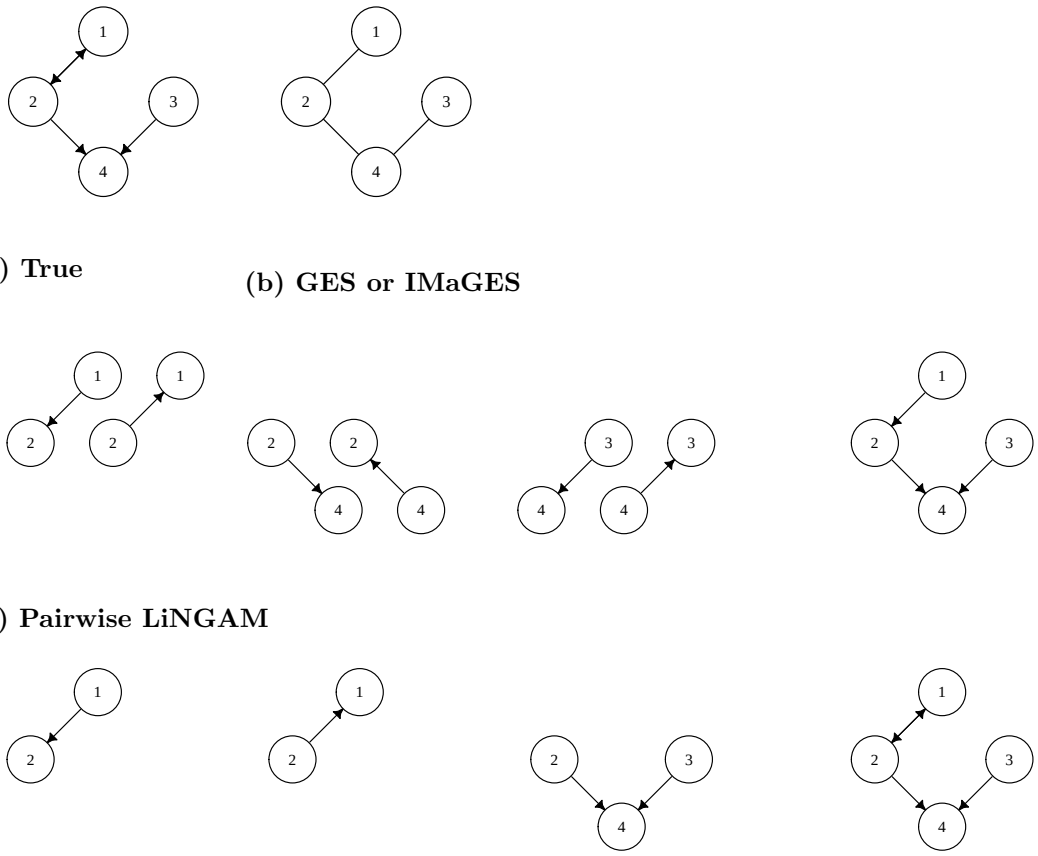
$$\frac{1}{T}\log L(X \to Y) - \frac{1}{T}\log L(Y \to X)$$

where positive values imply that $X \to Y$ is more likely and negative values imply the converse (Mumford and Ramsey, 2014). Using the Smith et al. (2011) data, Hyvärinen and Smith (2013) showed that pairwise LiNGAM could correctly orientate more than 75 % of edges correctly, or 100 % using a group analysis.

A comparison of procedures for edge orientation based on non-Gaussianity is provided in Ramsey et al. (2014).

Step	Instruction
0.	$\mathbf{Y} = (\mathbb{I} - \mathbf{G}_0)^{-1}\boldsymbol{\epsilon} = \mathbf{A}\boldsymbol{\epsilon}$
1.	Independent component analysis $\mathbf{W}_{ICA} = \mathbf{A}^{-1}$.
2.	Permute rows of $\mathbf{W}_{ICA}$ to get $\tilde{\mathbf{W}}$. A unique permutation gives a diagonal without zeros.
3.	Normalise $\tilde{\mathbf{W}}$ to get $\tilde{\mathbf{W}}'$. The diagonal of $\tilde{\mathbf{W}}'$ is all ones.
4.	Calculate $\hat{\mathbf{G}}_0 = \mathbb{I} - \tilde{\mathbf{W}}'$.
5.	Find $\tilde{\mathbf{G}}_\mathbf{0} = \mathbf{P}\hat{\mathbf{G}}_\mathbf{0}\mathbf{P}^\top$. $\tilde{\mathbf{G}}_\mathbf{0}$ is strictly lower triangular.

Table 1.1: The LiNG family of algorithms. The LiNGAM algorithm constrains $\mathbf{G}_0$ to be strictly lower triangular. The SEM is acyclic and there is a unique solution. Under the weaker constraint that the diagonal of $\mathbf{G}_0$ may not contain any ones, the more general LiNG algorithm returns many admissible models, which may contain cycles (Shimizu et al., 2006; Lacerda et al., 2012).



(a) True

(b) GES or IMaGES

(c) Pairwise LiNGAM

(d) LOFS (R1)

Figure 1.7: Procedures for edge orientation based on measures of non-Gaussianity. **(a)** The underlying graph to infer. **(b)** A search procedure (e.g. GES or IMaGES) finds the skeleton (undirected graph). **(c)** Pairwise LiNGAM infers directionality by calculating the log-likelihood ratio for each adjacent edge individually. **(d)** LOFS (R1) also considers each node individually, the parent set which gives the highest score of non-Gaussianity for the regression residuals is chosen.

1.5.5 Dynamic Causal Modelling

Model selection in a DCM framework involves calculating the model evidence (marginal likelihood) of the data $\mathbf{y}$ given some model $\mathcal{M}$ via

$$p(\mathbf{y} \,|\, \mathcal{M}, \mathbf{u}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\theta}} p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Theta} \,|\, \mathcal{M}, \mathbf{u}) \, d\boldsymbol{\theta} \, d\boldsymbol{\Theta}$$

where $\mathbf{u}$ are the external inputs, $\boldsymbol{\theta}$ are the hidden states and $\boldsymbol{\Theta}$ contains set of ordinary differential equations and parameters that model the hemodynamic response (Razi and Friston, 2016). In order to be able to score millions of candidate models, Friston et al. (2011) provide the following approximation to the model evidence. Consider the *full* model $\mathcal{M}_F$ where every connection exists and is reciprocal. The model evidence is denoted $p(\mathbf{y} \,|\, \mathcal{M}_F)$. All the other models are nested within this completely connected model. Let the parameters of the fully connected model be represented by $\boldsymbol{\Theta}_F$ and the parameters of a nested model $\mathcal{M}_j$ be represented by $\boldsymbol{\Theta}_j$. Then $\boldsymbol{\Theta}_j = 0$ denotes $\boldsymbol{\Theta}_F$ with some subset of the parameters set to zero. It is possible to write

$$p(\mathbf{y} \,|\, \mathcal{M}_j) = p(\mathbf{y} \,|\, \boldsymbol{\Theta}_j = 0, \mathcal{M}_F) = \frac{p(\boldsymbol{\Theta}_j = 0 \,|\, \mathbf{y}, \mathcal{M}_F) \, p(\mathbf{y} \,|\, \mathcal{M}_F)}{p(\boldsymbol{\Theta}_j = 0 \,|\, \mathcal{M}_F)}.$$

The log model evidence is then

$$\log_e[p(\mathbf{y} \,|\, \mathcal{M}_j)] = \log_e[p(\boldsymbol{\Theta}_j = 0 \,|\, \mathbf{y}, \mathcal{M}_F)] - \log_e[p(\boldsymbol{\Theta}_j = 0 \,|\, \mathcal{M}_F)] + \log_e[p(\mathbf{y} \,|\, \mathcal{M}_F)].$$

As the $\log_e[p(\mathbf{y} \,|\, \mathcal{M}_F)]$ term will be present for every model, it can be discarded. The first term may be approximated by a conditional Gaussian density $\log_e[q(\boldsymbol{\Theta}_j = 0 \,|\, \mathcal{M}_F)]$. The term $p(\boldsymbol{\Theta}_j = 0 \,|\, \mathcal{M}_F)$ is the prior probability that $\boldsymbol{\Theta}_j = 0$.

Unlike the other methods reviewed here, Friston et al. (2011) stipulate that all connections must be reciprocal as this, they argue, best represents the underlying biology. This has the advantage of significantly reducing the number of models scored by their method (Henry and Gates, 2017). See also Rosa et al. (2012).

1.6 MDM Directed Graph Model Search

One of the main strengths of the MDM is that the likelihood may be written in the form

$$p[\mathbf{y}] = \prod_{t=1}^{T} \prod_{r=1}^{n} p[y_t(r) \,|\, \mathbf{x}^t(r), \mathbf{y}^{t-1}(r)]$$

where $\mathbf{x}^t(r)$ contains the observations of some set of parent nodes up until time t and each $p[y_t(r) \,|\, \mathbf{x}^t(r), \mathbf{y}^{t-1}(r)]$ follows a univariate t-distribution (see equation 1.4.10c). As the observed values of the parent nodes at time t are contained in the observation vector $\mathbf{F}_t(r)$, which is assumed, for notational convenience, to be included in the initial information set D_0, we may easily convert each individual likelihood into DLM notation by writing $p[y_t(r) \,|\, D_{t-1}]$. As each observation vector contains the values of some set

of parent nodes, we are implicitly conditioning on some model with these parents. Throughout this work, some model i for node r with parent set $Pa_i(r)$ is denoted by $\mathcal{M}_i(r)$ and the model evidence is

$$p[\mathbf{y}(r) \,|\, \mathcal{M}_i(r)] = \prod_{t=1}^{T} p[y_t(r) \,|\, \mathcal{M}_i(r), \mathbf{y}^{t-1}(r)].$$

Within the MDM framework, $\mathbf{x}^t(r)$ is defined in such a way that for any MDM, the joint distribution (equation 1.4.4) will have an associated graph that is a DAG. However, in order to perform a search over the model space, we utilise the fact that the joint likelihood factors by node and find the highest scoring set of parents for each node individually. This approach is termed the MDM-DGM or Directed Graph Model search. We may then consider the highest-scoring model that is also a DAG. The MDM-IPA (Integer Programming Algorithm), outlined in the next section, provides one way to find this graph.

The number of candidate sets of parents *per node* is $N = 2^{n-1}$, so the total number of models to score is $n2^{n-1}$. The posterior model probability $p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)]$, again at the level of the individual node, is

$$p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)] = \frac{p[\mathbf{y}(r) \,|\, \mathcal{M}_i(r)]\, p[\mathcal{M}_i(r)]}{\sum_{i=1}^{N} p[\mathbf{y}(r) \,|\, \mathcal{M}_i(r)]\, p[\mathcal{M}_i(r)]}.$$

It follows that maximising the model probability is equivalent to maximising the likelihood, if the prior model probability $p[\mathcal{M}_i(r)]$ is assumed to be equal across all N potential models.

1.6.1 Implementation of the MDM-DGM Search

In practice, model selection is performed by maximising the Log Predictive Likelihood (LPL) over the set of models $\mathcal{M}(r) = \{\mathcal{M}_1(r), \ldots, \mathcal{M}_N(r)\}$ according to

$$\mathrm{LPL}[\mathcal{M}_i(r)] = \sum_{t=15}^{T} \log_e\{p[y_t(r) \,|\, Pa_i(r), D_{t-1}]\}. \tag{1.6.1}$$

The sum is from $t = 15$ (rather than $t = 1$) to minimise any effect of the choice of prior hyperparameters. Values for the hyperparameters are chosen so that the priors are weakly informative. The values were $\mathbf{m}_0(r) = \mathbf{0}$, $\mathbf{C}_0^*(r) = 3\,\mathbb{I}_{p_r}$ and $n_0(r) = d_0(r) = 0.001$, where p_r denotes the number of regressors in the candidate model. In the Dynamic Linear model, we condition on *initial information set* $D_{t-1} = \{D_0, y_1(r), \ldots, y_{t-1}(r)\}$ (see section 1.4.2) so, at any time t, we have the updated prior distributions

$$p[\boldsymbol{\theta}_t(r) \,|\, \phi(r), D_{t-1}] \sim \mathcal{N}\left(\mathbf{m}_{t-1}(r), \phi(r)^{-1} \frac{\mathbf{C}_{t-1}^*(r)}{\delta(r)}\right)$$

$$p[\phi(r) \,|\, D_{t-1}] \sim \mathcal{G}\left(\frac{n_{t-1}(r)}{2}, \frac{d_{t-1}(r)}{2}\right).$$

It follows that in practice the non-informative prior is replaced by a prior that is informed by the first few data points. We choose to disregard any influence of these initial data points on the prior model probability, assigning equal prior model probability $p[\mathcal{M}_i(r) \,|\, D_{t-1}]$ to all models (at time $t = 15$). Maximising the posterior model probability then corresponds to maximising the Log Predictive Likelihood, conditional on D_{t-1}. We prefer model $\mathcal{M}_i(r)$ over $\mathcal{M}_j(r)$ if

$$\mathrm{LPL}[\mathcal{M}_i(r)] > \mathrm{LPL}[\mathcal{M}_j(r)].$$

1.6.2 Log$_e$ Bayes Factors for Model Comparison

Throughout this work, we use log$_e$ Bayes factors to compare the evidence for different models. For two models i and j, we have

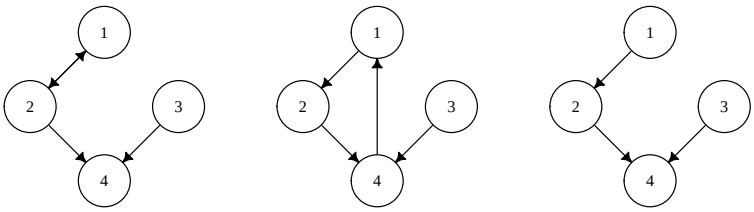
$$\begin{aligned}\log_e \mathrm{BF}_{ij} &= \log_e\{p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)]\} - \log_e\{p[\mathcal{M}_j(r) \,|\, \mathbf{y}(r)]\} \\ &= \mathrm{LPL}[\mathcal{M}_i(r)] - \mathrm{LPL}[\mathcal{M}_j(r)].\end{aligned}$$

The log$_e$ Bayes factor has been referred to as a 'weight of evidence' and may be interpreted as a relative measure of how well two models $\mathcal{M}_i(r)$ and $\mathcal{M}_j(r)$ predict the data $\mathbf{y}(r)$. To quantify the evidence for different models, we adopt the values proposed by Kass and Raftery (1995) for the natural logarithm of a Bayes factor and say that there is evidence for $\mathcal{M}_i(r)$ over $\mathcal{M}_j(r)$ if $\log_e \mathrm{BF}_{ij} > 1$. If $\log_e \mathrm{BF}_{ij} > 3$, the evidence is *strong*; if $\log_e \mathrm{BF}_{ij} > 5$, the evidence is *very strong*. Negative values indicate evidence to prefer model $\mathcal{M}_j(r)$ over $\mathcal{M}_i(r)$. If $\log_e \mathrm{BF}_{ij}$ is between -1 and 1, models $\mathcal{M}_i(r)$ and $\mathcal{M}_j(r)$ can be though of as being equivalent and we say that there is no evidence that one should be preferred over the other.

1.6.3 MDM Integer Programming Algorithm

The algorithm for finding the MDM which best fits the data involves calculating the Log Predictive Likelihood $\mathrm{LPL}[\mathcal{M}_i(r)]$ for all the candidate combinations of parents $Pa_i(r)$ for each node r individually, and then, if necessary, constraining the graph to be acyclic. These two steps comprise the MDM-IPA (Integer Programming Algorithm), described by Costa (2014), Costa et al. (2015) and Costa et al. (2017). The MDM-IPA is illustrated in Figure 1.8. The first step performs an MDM-DGM search. However, as can been seen in Figure 1.8 **(a)**, the resulting graph may not be a DAG. The algorithm then compares the LPL for each $\hat{\mathcal{M}}(r)$, and the lowest-scoring model is replaced by the next highest scoring model for that node. This process repeats until the resulting graph has no cycles or bidirectional edges. Applications of the MDM-IPA to resting-state fMRI data may be found in Costa (2014), Costa et al. (2015) and Costa et al. (2017).

Step	Instruction	Graph
1.	Find $\hat{Pa}(r)$ for $r = 1, \ldots, n$.	**(a)**
2.	Which LPL$[\hat{\mathcal{M}}(r)]$ is lowest? Replace lowest scoring parent set.	**(b)**
3.	If new graph is a DAG, terminate. If not, repeat step 2.	**(c)**



(a) **(b)** **(c)**

Figure 1.8: Illustration of the MDM-IPA. **(a)** The MDM-DGM search finds parents for each node individually so the resulting graph has the bidirectional edge $1 \leftrightarrow 2$. **(b)** If the model with $\hat{Pa}(1) = \{2\}$ has the lowest LPL, the MDM-IPA algorithm will replace $\hat{Pa}(1) = \{2\}$ with the next highest scoring parent set, in this example, $\hat{Pa}(1) = \{4\}$. **(c)** The graph in **(b)** now contains a cycle $1 \rightarrow 2 \rightarrow 4 \rightarrow 1$. If the model with $\hat{Pa}(1) = \{4\}$ still has the lowest LPL, the MDM-IPA algorithm will replace $\hat{Pa}(1) = \{4\}$ with the next highest scoring parent set, e.g. $\hat{Pa}(1) = \{\varnothing\}$. The resulting graph is now a DAG and the algorithm will terminate.

1.7 DAGs and Cyclic Graphs

When reviewing network discovery algorithms (Figures 1.6, 1.7 and 1.8), we have assumed that there is a single 'true' structure that we wish to infer, describing the relationship between 4 nodes. In this system, there is reciprocal connectivity between node 1 and node 2. If we assume that this bidirectional edge represents a genuine feedback loop between two nodes, rather than the presence of an unmeasured confounder, it is therefore interesting from the standpoint of understanding the underlying physiology. While the MDM-DGM algorithm is able to estimate graphs with cycles and bidirectional edges, doing so violates the principles of the MDM as outlined by Queen and Smith (1993): an MDM describes a dynamic *acyclic* Bayesian network. As previously discussed, the MDM-IPA algorithm is one way to impose a DAG constraint (see Figure 1.8) but in this work, we focus on the MDM-DGM because, as will be shown in the next chapter, permitting cycles and bidirectional edges enables us to capture behaviour that is likely to be more plausible physiologically. In doing so, we no longer have a strict definition of causality but nonetheless, we argue that the MDM-DGM is still a useful data-driven method for discovering correlations and predictive relationships between brain regions.

Chapter 2

Network Discovery with the MDM-DGM

2.1 Introduction

In this chapter, the MDM-DGM search is applied to two fMRI datasets with 15 brain regions, one a resting-state dataset (we also refer to this dataset as the 'safe' dataset) and the other a task condition where the participants were anticipating electric shock ('anticipation of shock'). Given these datasets, we first assess the ability of the MDM-DGM to detect edge presence, using partial correlation matrices for comparison. We construct an MDM-DGM network for each subject for both the 'safe' and 'anticipation of shock' data and compare the two experimental conditions based on the edges found to be consistently present or absent across subjects using a method based on the Binomial test. We also construct group networks by combining the LPL scores over subjects and use these group networks to show how we may obtain estimates for the time-varying connectivity strengths $\hat{\boldsymbol{\theta}}_t(r)$. Given these connectivity strengths, we test for differences between the two experimental conditions, 'safe' and 'anticipation of shock', over all subjects and based on group splits using measures of trait and induced anxiety. Taken together, these analyses will illustrate some of the key strengths, and potential weaknesses, of the MDM-DGM for the discovery and analysis of functional brain networks.

2.2 Datasets

We considered data consisting of BOLD time series from 17 brain regions, extracted from scans for 32 participants (14 male, all right-handed, 18-40 years, mean age 24.8 years). The voxel size was $2\,\mathrm{mm}^3$ and regions of interest (ROIs) were extracted either functionally or using the Harvard-Oxford template. The 17 extracted ROIs were the orbitofrontal cortex (OFC), dorsolateral prefrontal cortex (DLPFC), amygdala (Amyg), anterior insula (AntIns), posterior insula (PostIns), primary and secondary somatosensory cortices (SI, SII) (all from both the left and right hemisphere), the ventromedial prefrontal cortex (VMPFC), the anterior mid-cingulate cortex (aMCC) and the periaqueductal gray (PAG). As the primary and secondary somatosensory cortices were both highly correlated between the two hemispheres, these regions were replaced with their mean time series (SI-LR and SII-LR). This set of 15 ROIs is considered in this chapter.

Data were obtained under two experimental conditions: in both, participants were instructed to 'lie still, keep their eyes open, and stay awake'. In the second condition, participants were additionally told they would receive randomly timed electric shock stimuli *sometimes close together and sometimes with long gaps between stimuli* (Bijsterbosch et al., 2015, their emphasis). In practice, they received intermittent shocks for the first 5 and last 2 minutes of a 22 minute scan, leaving a 15 minute section where they were anticipating but not receiving shocks. The data analysed in this chapter therefore consist of two 15 minute scans for each participant, a resting-state scan, which we also refer to as the 'safe' scan, and an 'anticipation of shock' scan. The repetition time (TR) was 1140 milliseconds, giving time series with 790 time points.

Prior to the fMRI sessions, participants completed questionnaires to assess trait anxiety and depression. Using eight standardised measures of negative affect, the first component of a principle component factor analysis gave a *factor score*, which could be interpreted as a measure of trait anxiety. After each scan, participants were asked to rate how anxious they felt after each scan on a Visual Analogue Scale (VAS), where 1 was not at all and 7 was very much. The difference between the two scores provided a measure of induced anxiety. Responses were significantly higher after the scan where the participants received the electric shocks (paired t-test, $t = 2.84$, $p = 0.008$), indicating the 'shock' scan led to induced anxiety (Bijsterbosch et al., 2015).

More information about the data, including preprocessing steps and the trait and induced anxiety metrics, may be found in Bijsterbosch et al. (2015).

2.3 MDM-DGM Network Discovery

Given the datasets described above, we estimated an MDM-DGM network for each of the 32 participants, for both of the experimental paradigms, 'safe' and 'anticipation of shock'. As described in the previous chapter, section 1.6, we chose the set of parents $\hat{P}a_i(r)$ that maximised the Log Predictive Likelihood for each node r individually, $r = \{1, \ldots, 15\}$.

Models were scored using a C++ implementation of the Dynamic Linear Model available in the `multdyn` package for R[1,2]. We have developed this package for the application of the MDM-DGM to fMRI time series and it contains functions which allow the user to perform an exhaustive search over all candidate models for networks with up to 20 nodes. More details, including estimates of computation time and solutions for networks with more than 20 nodes, will be provided in the next chapter.

The chosen discount factor was the value of $\delta(r)$ in the range $\delta(r) \in [0.5, 1]$ which gave the highest LPL (grid search, step size 0.01). The discount factor for the winning model

[1] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`

[2] Schwab, S., Harbord, R., Costa, L., and Nichols, T. *multdyn: Multiregression Dynamic Models*, 2017b. URL `https://CRAN.R-project.org/package=multdyn`. R package version 1.5.1

$\hat{\mathcal{M}}_i(r)$ for each subject, node and experimental condition, is shown in Figure 2.1. The number of parents chosen for each subject, node and experimental condition, is plotted in Figure 2.2. Figure 2.1 shows that the optimal discount factor varies considerably between the nodes. For some nodes (e.g. the amygdala, the anterior and posterior insula and the PAG), the discount factor is consistently one or close to one, indicating the best model is a stationary (or near stationary) model, with static connection strengths. For other nodes, the chosen discount factor is consistently less than one, with greater variation over subjects, indicating that a dynamic model, with time-varying connection strengths, provides a better fit to the data.

The median values for both the discount factor and the number of parents appear similar for the two experimental conditions, with the notable exception of the VMPFC, where both the median discount factor and the median number of parents are visibly lower for the 'anticipation of shock' data.

2.4 Analysis Based on Partial Correlation

In addition to the MDM-DGM networks, a partial correlation matrix was also estimated for each participant for the resting-state ('safe') dataset[3]. As discussed in the previous chapter, partial correlation is a widely-used method for discovering functional brain networks and has been shown to have good sensitivity for detecting the presence of true connections on both simulated and real fMRI data (Smith et al., 2011; Dawson et al., 2013). For this reason, we use partial correlation networks as a tool to validate the MDM-DGM networks. We would expect that an undirected edge with a high (absolute) partial correlation would be replaced by a directed edge in the MDM-DGM network and this edge may or may not be bidirectional.

The lowest (absolute) partial correlation for which an edge was present in the corresponding MDM-DGM network was 5.2×10^{-4} (below the 5th percentile of all partial correlations), while the highest partial correlation for which an edge was absent was 0.49 (above the 95th percentile), so it may be concluded that, in itself, the absolute partial correlation between two regions is a poor predictor of whether an edge is present or absent in an MDM-DGM network. This is likely due in part to the fact that the MDM-DGM estimates directed edges, so a strong undirected partial correlation may be replaced by a directed edge in an MDM-DGM network. For example, the undirected edge VMPFC – OFC-L with partial correlation 0.49 becomes the directed edge VMPFC → OFC-L in the MDM-DGM network for this individual subject.

[3]Kim, S. *ppcor: Partial and Semi-Partial (Part) Correlation*, 2015. URL `https://CRAN.R-project.org/package=ppcor`. R package version 1.1

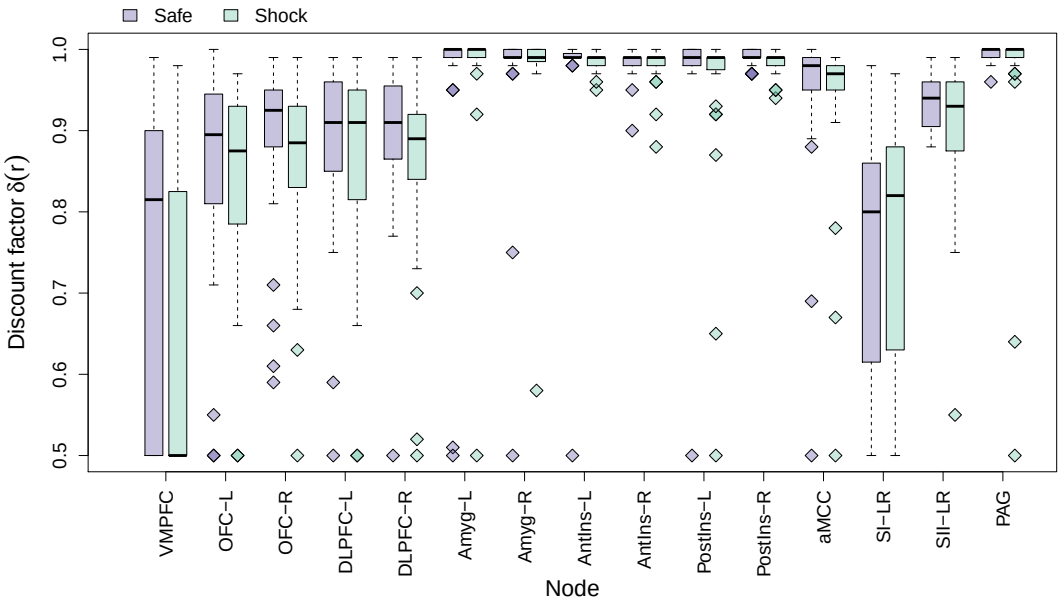


Figure 2.1: The optimal discount factor $\delta(r)$ for each subject across nodes and experimental conditions. For some nodes (e.g. the right posterior insula) $\delta(r)$ is consistently close (or equal) to one, suggesting the connectivity is static. For other nodes (e.g. the VMPFC), lower discount factors allow dynamic connectivity estimates. The Dynamic Linear Model captures information that would be lost in a static Bayesian regression model. The VMPFC is the only region where there is a noticeable difference between the two experimental conditions, with median values of 0.82 and 0.50 for the 'safe' and 'anticipation of shock' data respectively.

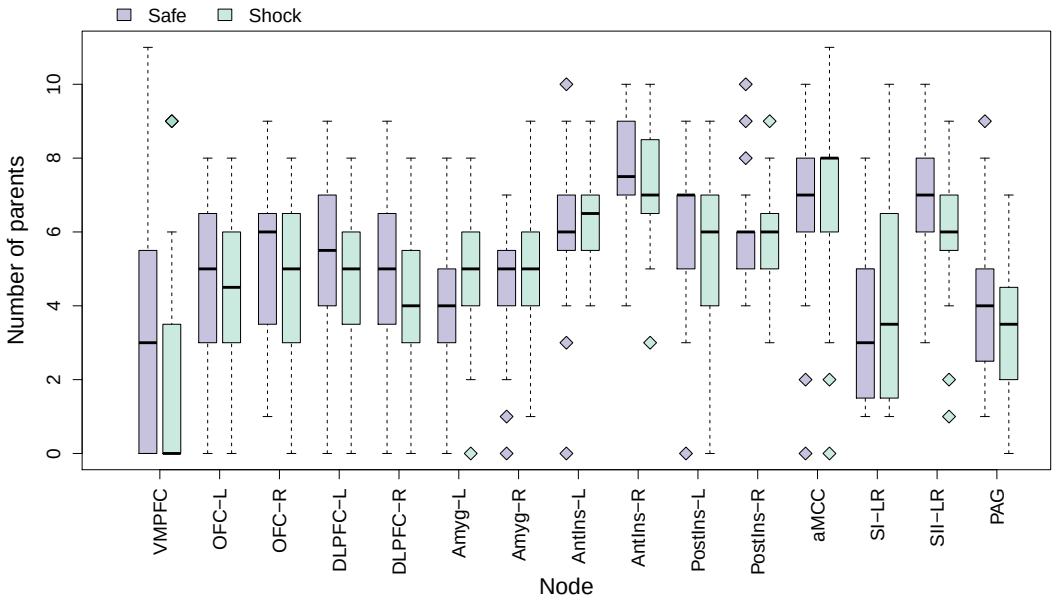


Figure 2.2: The number of parents for each subject across nodes and experimental conditions. The VMPFC displays the most noticeable difference in the number of parents with median values of 3 and 0 for the 'safe' and 'anticipation of shock' data respectively.

2.4.1 A Method to Quantify Consistency Over Subjects

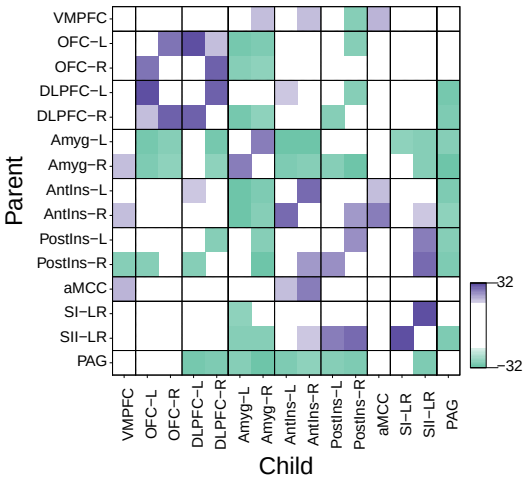
To quantitatively compare the partial correlation and MDM-DGM networks, we used a method for assessing whether edges were significantly present or absent (over subjects) based on one-sided Binomial tests. Let S denote the number of subjects and $\mathcal{E}_s$ denote the number of edges (excluding the diagonal) in the MDM-DGM network for subject s. The maximum number of possible edges is $n^2 - n$, where, as usual, n is the total number of nodes. The partial correlation matrices were thresholded by selecting the $\mathcal{E}_s$ edges with the highest absolute partial correlation. If $\mathcal{E}_s$ was an odd number, the $\mathcal{E}_s - 1$ edges with the highest absolute partial correlations were selected, to maintain the symmetry of the partial correlation matrices. To construct a group-level network, one-sided Binomial tests were used to assess whether each edge could be judged to be significantly present or absent. The probability of an edge occurring in a homogenous network was defined empirically, using the average proportion of subjects with an edge over the $(n^2 - n)$ possible edges

$$\hat{\pi}_{present} = \frac{1}{S} \sum_{s=1}^{S} \frac{\mathcal{E}_s}{(n^2 - n)}.$$

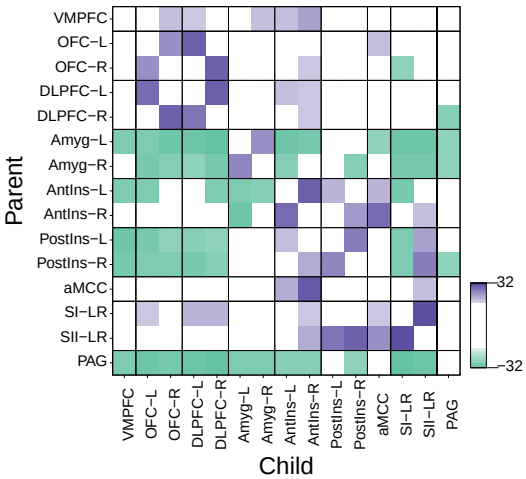
The probability of an edge being absent in a homogenous network was then $\hat{\pi}_{absent} = 1 - \hat{\pi}_{present}$. These values were $\hat{\pi}_{present} = 0.37$ and $\hat{\pi}_{absent} = 0.63$. Significant edges after false discovery rate (FDR) correction[4] ($\alpha = 0.05$) are shown in Figure 2.3. See also Costa (2014) and Costa et al. (2015, 2017) for similar analyses with a different dataset.

Strong, inter-hemispheric connectivity is present for both the partial correlation and MDM-DGM networks (Figures 2.3a and 2.3b), in particular OFC-L ↔ OFC-R, DLPFC-L ↔ DLPFC-R and Ant-Ins-L ↔ AntIns-R. Both networks also predict strong, bidirectional connectivity between the OFC and the DLPFC, the somatosensory cortices and the posterior insula and the secondary somatosensory cortex. This bidirectional connectivity would disappear if the networks were constrained to be acyclic. The networks agree that the periaqueductal gray has no parents or children. The differences between the two networks are shown in Figure 2.3c. One noticeable difference between the two networks is that the MDM-DGM estimates that the VMPFC has no parents. The parents found for the VMPFC by the partial correlation analysis, the aMCC, the Amyg-R and the AntIns-R, are not present in the MDM-DGM network, highlighting the potential of the MDM-DGM to capture asymmetries that are missed by an (undirected) partial correlation method. It should be noted here that this behaviour is consistent with the fact the VMPFC is known to play a top-down role in the regulation of negative emotion (see, for example, Motzkin et al. (2015)).

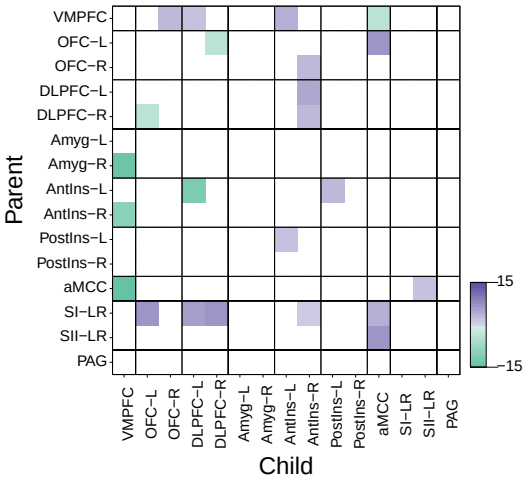
[4] All false discovery rate correction in this chapter used the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).



(a) Partial correlation, significant edges (safe)



(b) MDM-DGM, significant edges (safe)



(c) Difference, significant edge presence

Figure 2.3: Partial correlation and MDM-DGM estimate similar networks but MDM-DGM also infers directionality. Edges found to be significantly present or absent using a Binomial test and FDR correction with $\alpha = 0.05$ for **(a)** the partial correlation matrices and **(b)** the MDM-DGM matrices. Purple (positive values) indicates a high proportion of subjects (at least 19 out of 32, or 59 %) share an edge, while green (negative values) indicates an edge is absent in a high proportion of subjects (at least 27 out of 32, or 84 %). **(c)** Differences in the number of subjects when the partial correlation and MDM-DGM networks disagree on whether the presence of an edge is significant. Purple (positive values) indicates the MDM-DGM network has a higher number of subjects with a particular edge than the network based on partial correlations; green (negative values) indicates the reverse. The largest differences are aMCC → VMPFC (-15 subjects), Amyg-R → VMPFC (-14 subjects) and AntIns-R → VMPFC (-10 subjects) and AntIns-L → DLPFC-L (-12 subjects).

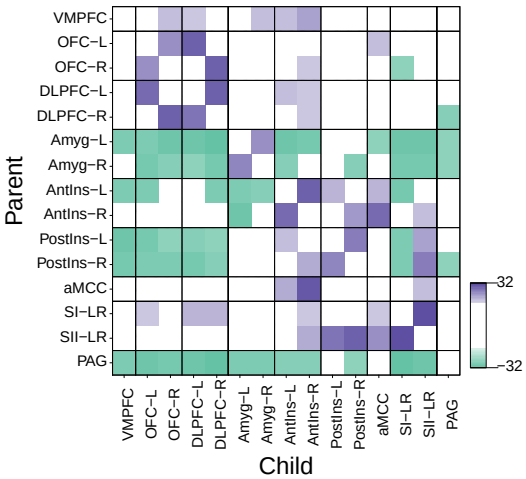
2.5 Safe vs. Anticipation of Shock: Comparing Networks

An identical analysis was performed to compare the networks for the 'safe' and 'anticipation of shock' data. (The value of $\hat{\pi}_{present}$ for the 'anticipation of shock' data was 0.36, so the mean over the two conditions gave a value of 0.37 as in the previous analysis). Figure 2.4 shows edges that are significantly present or absent for the 'safe' (Figure 2.4a, as Figure 2.3b) and 'anticipation of shock' (Figure 2.4b) datasets. Figure 2.4c shows the edges that are significantly present in one network but not the other. The edges with the largest differences were VMPFC → Amgy-L, VMPFC → aMCC, DLPFC-L → aMCC, and SI-LR → Amyg-R, which were significant in the 'anticipation of shock' network but not the 'safe' network, and OFC-L → aMCC and SI-LR → DLPFC-R, which were significant in the 'safe' network but not the 'anticipation of shock' network. However, it should be noted that none of the edges that were significantly present in one network were significantly absent in the other, suggesting the evidence for a difference between the networks is inconclusive.

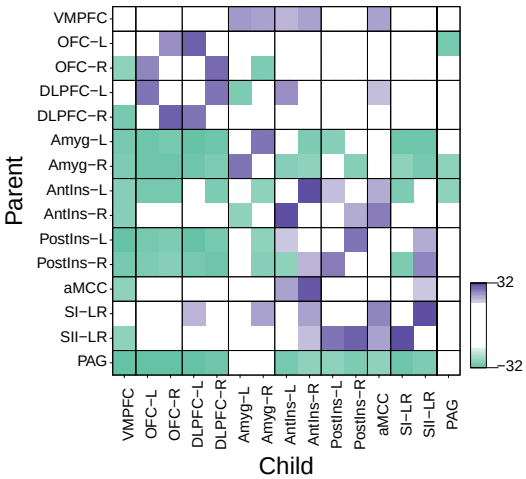
This method for significance testing allows the consistency over subjects to be quantified. For the partial correlation network, 45 % of the edges were found to be significantly present or absent, while for the MDM-DGM networks, it was 48 % for the 'safe' data and 50 % for the 'anticipation of shock' data. Because the Binomial method as described above uses $\hat{\pi}_{present} = 0.37$ as a threshold, only a relatively low percentage of subjects (59 %) need to share an edge for it to be found to be significant. Figures 2.5a and 2.5b show the edges which are present for at least 90 % of subjects (29 out of 32) for the 'safe' and 'anticipation of shock' networks. These two networks are identical, except for 3 edges missing from the 'anticipation of shock' network: DLPFC-L → OFC-L, DLPFC-L → DLPFC-R and Ant-Ins-R → aMCC. However, these edges all occur in at least 84 % of subjects, so it may be concluded that the strongest edges (when 'strongest' is defined as being shared by the highest number of subjects) are present in both the 'safe' and 'anticipation of shock' networks.

2.6 Log_e Bayes Factor Evidence for Model Differences

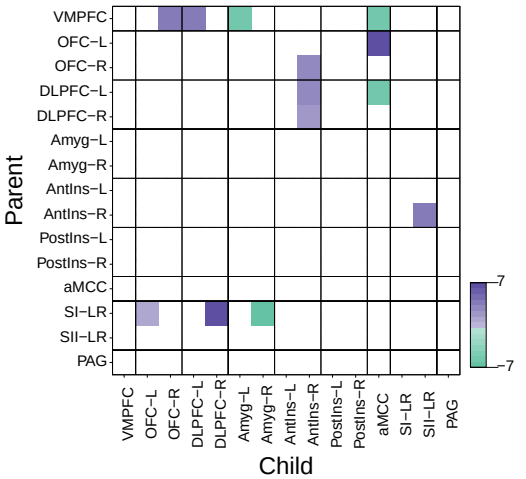
Another way to assess the significance of any differences between the networks is to fit the chosen set of parents for one dataset to the other dataset and compare the LPL scores. If, for a particular subject, the chosen parents for node r for the 'safe' data are $\hat{P}a(r)_{safe}$, we compare the difference between the score for this parent set and the chosen parents for the 'anticipation of shock' data $\hat{P}a(r)_{shock}$. As discussed in section 1.6, the $\log_e$ Bayes factor may be used to assess the strength of the evidence for one model over another, with a $\log_e$ Bayes factor of less than 1 interpreted as a lack of evidence for any difference between two models. Therefore, if the $\log_e$ Bayes factor between $\hat{P}a(r)_{safe}$ and $\hat{P}a(r)_{shock}$ on the same dataset ('safe' or 'anticipation of shock') is less than 1, we may conclude that both parent sets explain the data equally well. Conversely, higher values indicate that the chosen parent set has better explanatory power and a difference between the two parent sets may therefore be capturing some significant difference between the two experimental conditions.



(a) **MDM-DGM, significant edges (safe)**

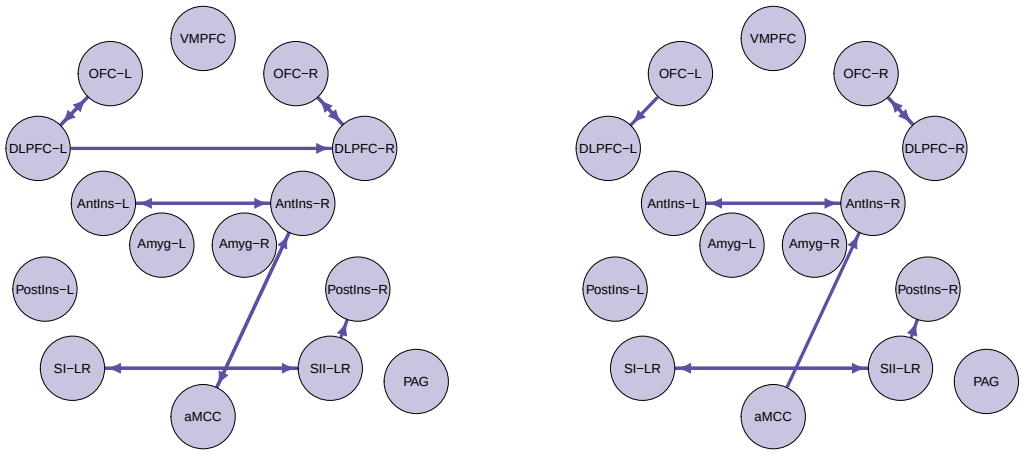


(b) **MDM-DGM, significant edges (shock)**



(c) **Difference, significant edge presence**

Figure 2.4: MDM-DGM networks are similar for the 'safe' and 'anticipation of shock' conditions. Edges found to be significantly present or absent using a Binomial test and FDR correction with $\alpha = 0.05$ for the two datasets **(a)** 'safe' and **(b)** 'anticipation of shock'. Purple (positive values) indicates a high proportion of subjects (at least 19 out of 32, or 59 %) share an edge, while green (negative values) indicates an edge is absent in a high proportion of subjects (at least 27 out of 32, or 84 %). **(c)** Differences in the number of subjects when the 'safe' and 'shock' MDM-DGM networks disagree on whether the presence of an edge is significant. Purple indicates the MDM-DGM safe network has a higher number of subjects with a particular edge than the MDM-DGM shock network; green values indicate the reverse. There is evidence for the edges SI-LR → DLPFC-R and OFC-L → aMCC in the 'safe' (resting-state) networks but not the 'anticipation of shock' networks. Conversely, there is evidence for the edges VMPFC → Amyg-L, VMPFC → aMCC, DLPFC-L → aMCC and SI-LR → Amyg-R in the 'shock' networks but not the 'safe' networks. Note that the maximum number of subjects by which these edges differ is 7.



(a) Safe **(b) Anticipation of shock**

Figure 2.5: Edges shared by > 90 % of subjects for (a) 'safe' and (b) 'anticipation of shock' datasets. The MDM-DGM networks are identical, except edges DLPFC-L → OFC-L, DLPFC-L → DLPFC-R and Ant-Ins-R → aMCC, which are found for 88 %, 88 % and 84 % of subjects respectively in the 'anticipation of shock' network.

Results of this analysis are shown in Figure 2.6a. Using the $\log_e$ Bayes factor criteria specified in Kass and Raftery (1995), only a small percentage of models may be thought of as being equivalent: when fitting the chosen parents for the 'safe' data to the 'anticipation of shock' data, 7.3 % of models have a $\log_e$ Bayes Factor of less than 1, while the number is 6.5 % when fitting the chosen parents for the 'anticipation of shock' data to the 'safe' data. In comparison, 89 % of models have a $\log_e$ Bayes Factor greater than 5, suggesting *very strong* evidence for a difference. This number is also 89 % when fitting the chosen parents for the 'anticipation of shock' data to the 'safe' data.

Looking at Figure 2.2, the median number of parents chosen for each node was comparable between the two experimental conditions. However, Figure 2.6b shows that the median number of parents that are different between the 'safe' and 'anticipation of shock' conditions was between 3 and 5. The fact that, at the individual subject level, the chosen parent sets tend to differ this much may explain the high $\log_e$ Bayes factor differences that are seen in Figure 2.6a.

In a similar analysis, we asked what percentage of parent sets were 'better' (had a higher LPL score) than the chosen parent set for the other dataset. This behaviour is explored in Figure 2.6c. When fitting the chosen parents for the 'safe' data to the 'anticipation of shock' data, 58 % of models were within the top 5 % of highest scoring models, 34 % were within the top 1 %. Fitting the chosen parents for the 'anticipation of shock' data to the 'safe' data, 57 % were within the top 5 % of highest scoring models and 32 % were within the top 1 %. This suggests that, while there is evidence of a significant difference between the chosen parent sets for the two experimental conditions, the parent set for the other condition may still provide one of the best models out of the large model space of possible candidates.

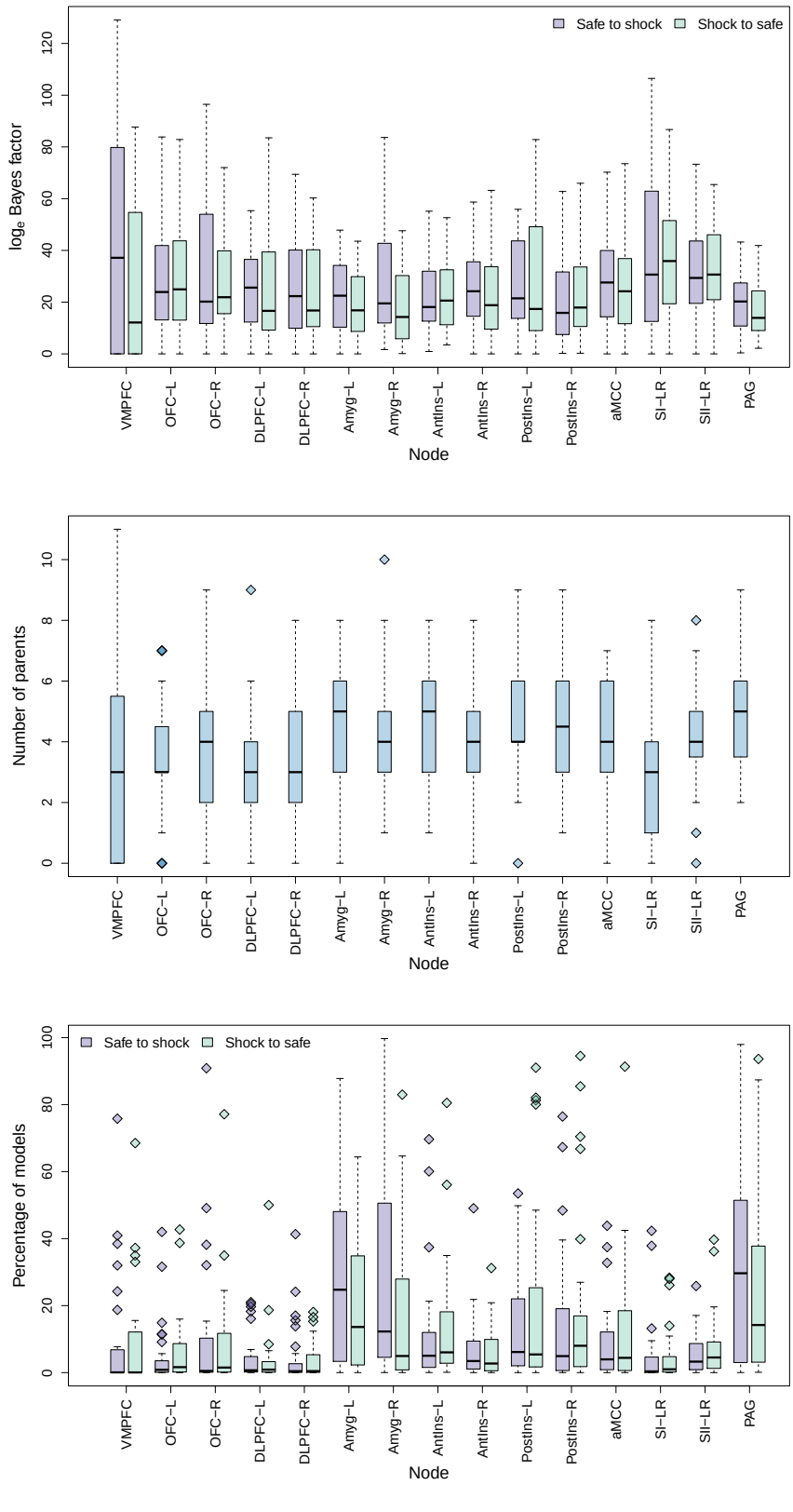


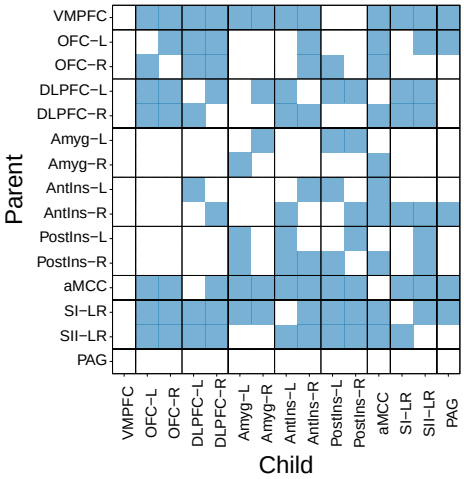
Figure 2.6: Evidence for a difference between the 'safe' and 'anticipation of shock' conditions. (a) For each subject and each node r, the $\log_e$ Bayes factor was calculated as the difference between the highest scoring parent set for the 'safe' data $\hat{Pa}(r)_{safe}$ and the highest scoring parent set for the 'anticipation of shock' data $\hat{Pa}(r)_{shock}$ fitted to the 'safe' data, and *vice versa*. For ease of visualisation, outliers are not shown. **(b)** The number of parents that are different between the two experimental conditions. **(c)** The percentage of models with a higher LPL than the chosen parent set for the other dataset.

2.7 Construction of an MDM-DGM Group Network

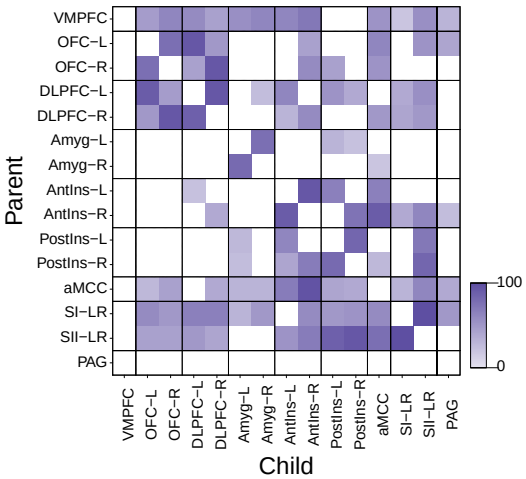
In the previous sections, the 'safe' and 'anticipation of shock' data, as well as the partial correlation matrices, were analysed at the group level by considering edge presence and absence. However, it is also informative to fit a 'common' or group network, which may be constructed by summing the Log Predictive Likelihood over subjects for each node, and then choosing the winning set of parents based on these summed scores. Denote this group network by $\hat{\mathcal{M}}_{\mathcal{G}} = \{\hat{\mathcal{M}}(1)_{\mathcal{G}}, \ldots, \hat{\mathcal{M}}(r)_{\mathcal{G}}, \ldots, \hat{\mathcal{M}}(n)_{\mathcal{G}}\}$, where each model represents some set of chosen parents $\hat{Pa}_{\mathcal{G}} = \{\hat{Pa}(1)_{\mathcal{G}}, \ldots, \hat{Pa}(r)_{\mathcal{G}}, \ldots, \hat{Pa}(n)_{\mathcal{G}}\}$. The networks based on the 'safe' and 'anticipation of shock' data may be denoted by $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ and $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$. Let $\hat{\mathcal{M}}_{\mathcal{G}_{safe+shock}}$ denote an overall group network constructed by summing the LPL scores over experimental conditions as well as subjects. A group network of this form may be thought of as a 'best' estimate of the network as it is based on all the subjects' data taken together.

The group networks $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ and $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$ are shown in Figures 2.7a and Figure 2.7c. Figures 2.7b and 2.7d show the percentage of subjects that have this edge using the individual MDM-DGM networks. The same strong, inter-hemisphere connections are present in the group networks, as well as the connections between the OFC and the DLPFC, and the insula (anterior and posterior) and the somatosensory cortices. The VMPFC is parentless is both networks, but is estimated to be the parent of most of the other ROIs, notably more than were found to be significant in the previous analysis (Figure 2.4). In fact, some of the edges in the group networks are only present in the individual networks for a few subjects, e.g. the connection OFC-L → PAG is present in the group network for the 'anticipation of shock' data but only occurs for 2 subjects in the individual networks. That the group networks are denser than the individual networks may be emphasised by the fact that the median number of edges in a individual network for the 'safe' data was 79.5 (the maximum was 103), while the number of edges in the group network was 102. For the 'anticipation of shock' data, the median number of edges was 77 (maximum 92), while the number of edges in the group network was 91. The number of edges in the combined network was 100.

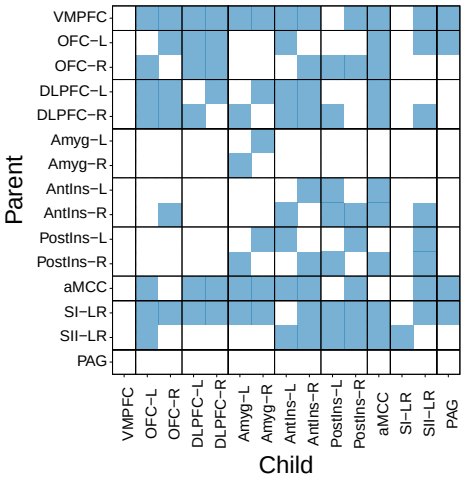
Figure 2.8 shows the difference between the two group networks. It can be seen that some edges were found in the group network for an experimental condition even if the same edge occurred for fewer subjects in the individual subject networks for this condition.



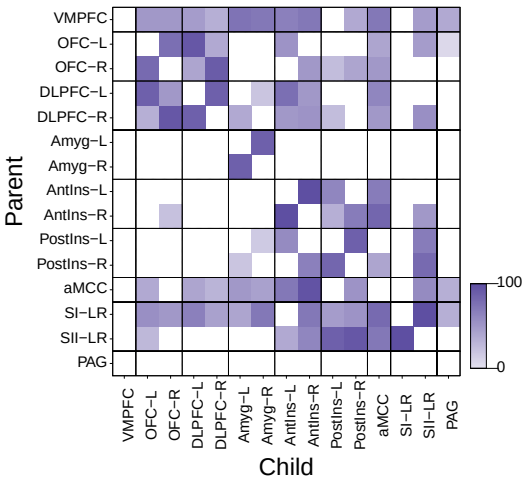
(a) **Group network,** $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$



(b)

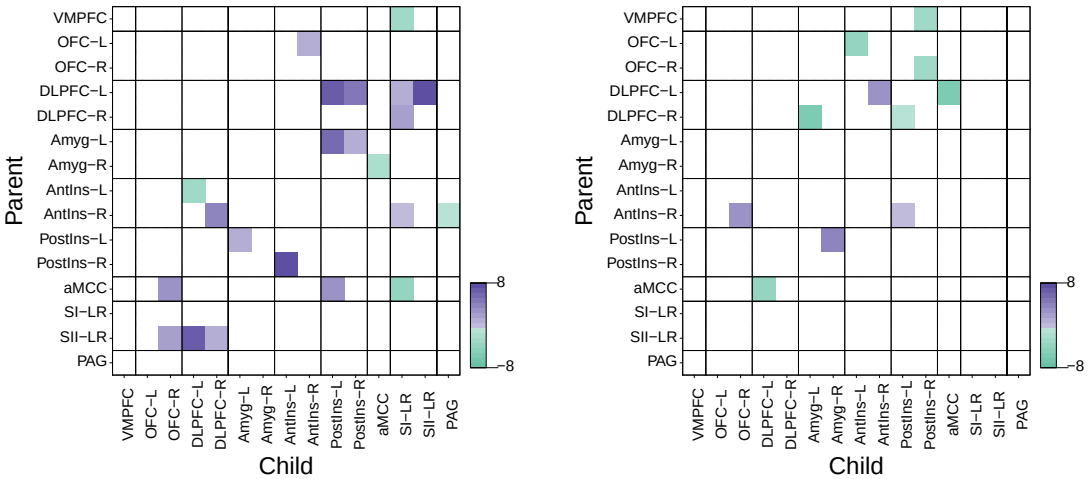


(c) **Group network,** $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$



(d)

Figure 2.7: MDM-DGM group networks for the 'safe' and 'anticipation of shock' datasets. **(a)** The group network $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ for the 'safe' data, obtained by maximising the sum of the LPL scores over subjects. **(b)** The percentage of subjects that have each edge in the group network, the minimum is 19 % (6 subjects). **(c)**, **(d)** are as **(a)** and **(b)** but for the 'anticipation of shock' data with group network $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$. The minimum number of subjects that have an edge in the group network is 6.2 % (2 subjects).



(a) Safe not shock **(b) Shock not safe**

Figure 2.8: MDM-DGM group networks differ between the 'safe' and 'anticipation of shock' datasets. Edges that occur only in the **(a)** 'safe' and **(b)** 'anticipation of shock' data, plotted as the difference in the number of subjects that have an edge in the individual networks. Purple indicates more subjects have an edge in the 'safe' data, and green that more subjects have an edge in the 'anticipation of shock' data.

Figure 2.9 shows the results of an identical analysis to that in Figure 2.6, except that this time the scores for winning parent sets $\hat{P}a(r)_{\mathcal{G}_{safe}}$ and $\hat{P}a(r)_{\mathcal{G}_{shock}}$ for each group model, $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ and $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$, were compared to the chosen parents in the individual networks. The $\log_e$ Bayes factors suggest there is strong evidence to prefer the parent sets of individual subject networks to the parent sets of the group model, with the median number of parents that are present in one network but absent in the other around 4 across the nodes. From Figure 2.9c, it is clear that for some nodes, such as the OFC and the DLPFC, the parent sets of the group models are consistently some of the highest scoring across subjects. In comparison, for the VMPFC and the PAG, there are subjects where the parent set of the group network is one of the *worst* performing models. We conclude that, for some nodes in particular, the parent sets of the group network may not be reflective of the parent sets of the individual networks.

2.8 Analysis of MDM-DGM Connectivity Strengths

Another advantageous feature of the MDM-DGM is that, alongside network discovery, we may obtain estimates for the time-varying regression coefficients $\hat{\boldsymbol{\theta}}_t(r)$, which may be interpreted as the instantaneous connectivity strength between two brain regions at time t. In this subsection, we assume a fixed network structure across subjects and analyse connectivity within and between subjects using $\hat{\boldsymbol{\theta}}_t(r)$ as a measure of connectivity. Rather than treating an edge as simply present or absent, each edge (in the network) instead has an associated connectivity strength that may be stronger or weaker between subjects or experimental conditions.

For each node r, given some set of parents $\hat{P}a(r)_{\mathcal{G}}$, a Dynamic Linear Model may be

fitted to obtain dynamic estimates for the parameter vector $\boldsymbol{\theta}(r) = \{\boldsymbol{\theta}_1(r), \ldots, \boldsymbol{\theta}_T(r)\}$ given data $\mathbf{y}(r)$. Let j denote an intercept term and p_r be one more than the number of parents in $\hat{P}a(r)_{\mathcal{G}}$, such that $\boldsymbol{\theta}_t(r) = \{\theta_{1t}(r), \ldots, \theta_{jt}(r), \ldots, \theta_{p_r t}(r)\}$. The index i then denotes a parent in $\hat{P}a(r)_{\mathcal{G}}$ so long as $i \neq j$. For each subject s and each time point t, there is a $p_r \times 1$ parameter vector $\hat{\boldsymbol{\theta}}_t(s,r)$ with $p_r \times 1$ location vector $\boldsymbol{\mu}_t(s,r)$ and $p_r \times p_r$ scale matrix $\boldsymbol{\Sigma}_t(s,r)$ (see equations 1.4.14a–1.4.14c). For each parent i, we define metrics $\bar{d}_{ir}$ and $\bar{\mu}_{ir}$ where

$$d_{it}(s,r) = \frac{\mu_{it}(s,r)}{\sqrt{\Sigma_{iit}(s,r)}}$$

so that

$$\bar{d}_{ir} = \frac{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{T}\sum_{t=1}^{T} d_{it}(s,r)\right)}{SD\left(\frac{1}{T}\sum_{t=1}^{T} d_{it}(s,r)\right)/\sqrt{S}}$$

and

$$\bar{\mu}_{ir} = \frac{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{T}\sum_{t=1}^{T} \mu_{it}(s,r)\right)}{SD\left(\frac{1}{T}\sum_{t=1}^{T} \mu_{it}(s,r)\right)/\sqrt{S}}.$$

These metrics provide a standardised and unstandardised t statistic for each edge, and so may be used to test whether, across subjects, the estimates for $\hat{\boldsymbol{\theta}}_t(r)$ are significantly different from zero. Applying false discovery rate correction ($\alpha = 0.05$) allows a threshold for $\bar{d}_{ir}$ to be determined. Results for the 'safe' and 'anticipation of shock' data, fitted to the group networks $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ and $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$ are shown in Figures 2.10c, 2.10d, 2.10e, 2.10f alongside the magnitude of the partial correlations for comparison (Figure 2.10a and Figure 2.10b). It can be seen that positive values of $\bar{d}_{ir}$ and $\bar{\mu}_{ir}$ correspond to positive partial correlations and negative values of $\bar{d}_{ir}$ and $\bar{\mu}_{ir}$ to negative partial correlation, providing further evidence that the MDM-DGM captures behaviour that would be captured using a partial correlation method, while also incorporating directionality. For example, the edges VMPFC → AntIns-L and VMPFC → AntIns-R have a negative mean partial correlation, as well as a negative $\bar{d}_{ir}$ and $\bar{\mu}_{ir}$ 'connectivity strength'. In agreement with the analysis based on the individual networks, the $\bar{d}_{ir}$ and $\bar{\mu}_{ir}$ metrics estimate strong connectivity between the orbitofrontal and dorsolateral prefrontal cortices, between the primary and secondary somatosensory cortices and between the hemispheres of the anterior insula. The presence of strong functional connectivity between brain regions that are anatomically close is widely supported in the literature, see, for example, Honey et al. (2009). Unlike the previous analyses based on edge presence, using this approach the PAG, while remaining childless, is estimated to have parents with relatively low but still significant connectivity strengths.

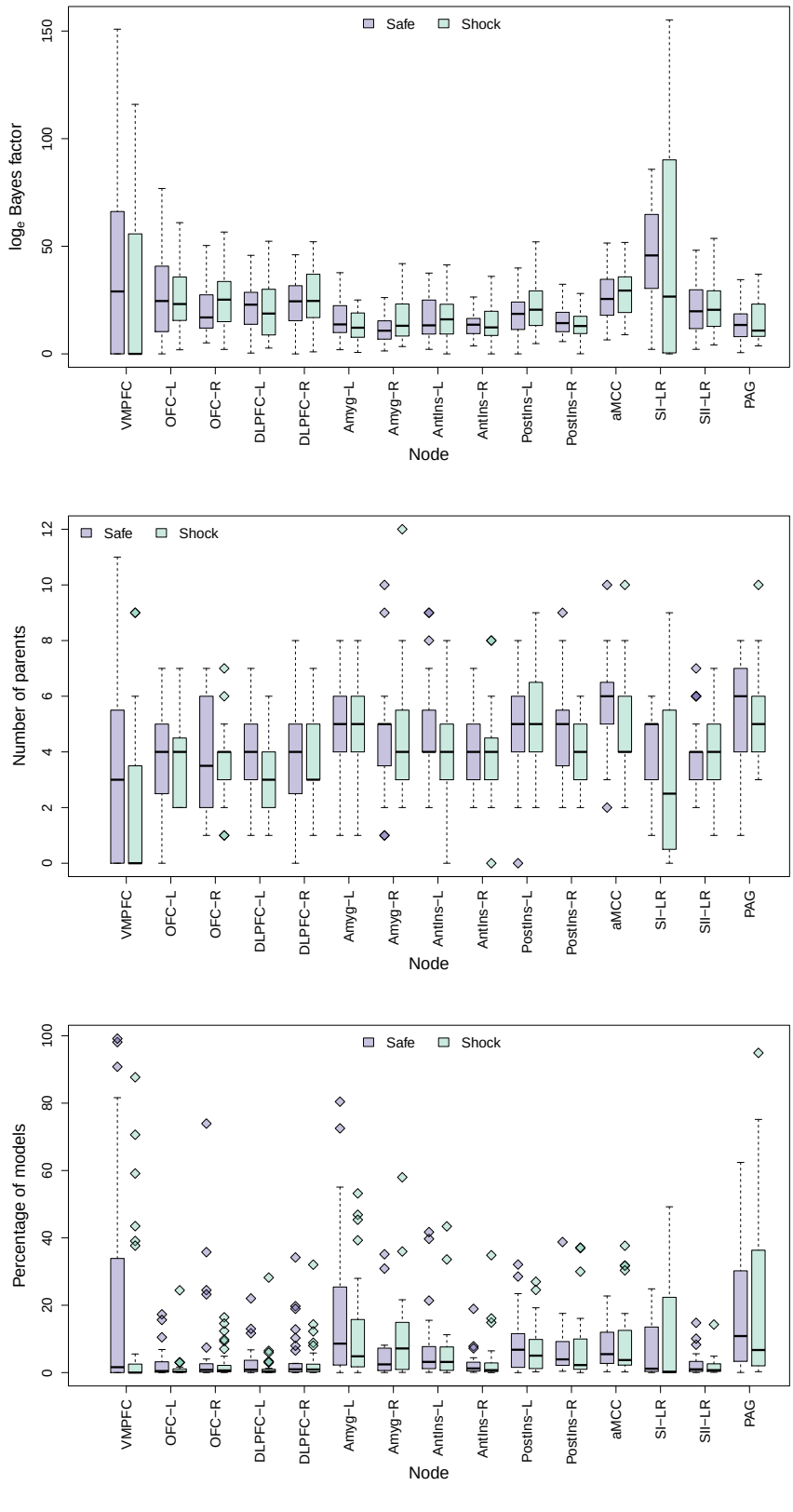
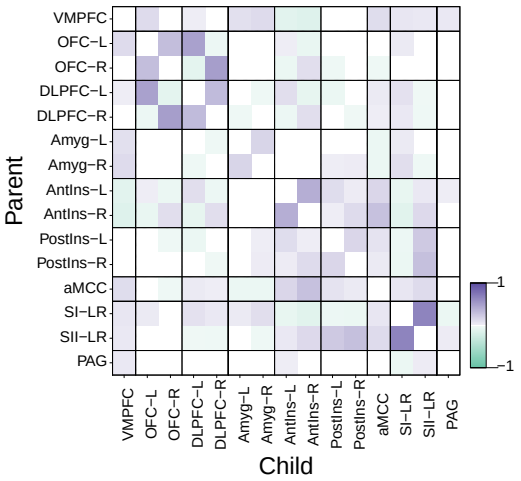
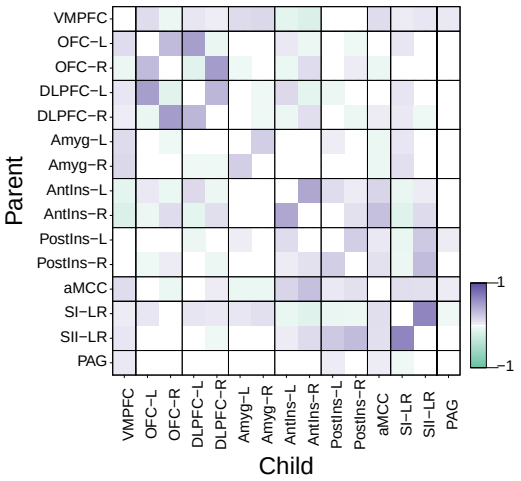


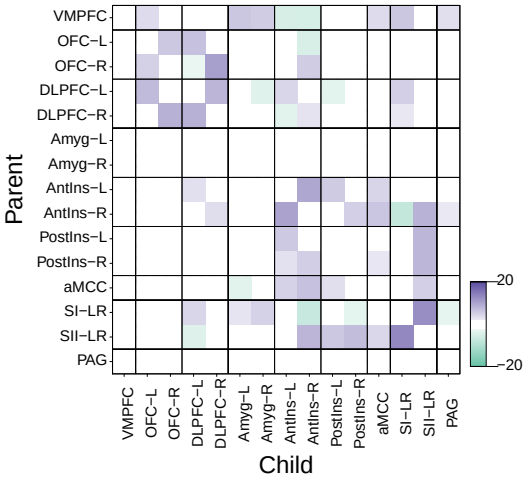
Figure 2.9: Evidence for a difference between the group and individual subject networks. (a) For each subject and each node r, the $\log_e$ Bayes factor was calculated as the difference between the highest scoring parent set for each subject and the highest scoring parent set for the group network for the 'safe' data (purple) and the 'anticipation of shock' data (green). For ease of visualisation, outliers are not shown. **(b)** Difference between the number of parents in the individual and group models for the 'safe' (purple) and 'anticipation of shock' (green) data. **(c)** The percentage of models with a higher LPL than the parent set in the group model for the 'safe' data (purple) and the 'anticipation of shock' data (green).



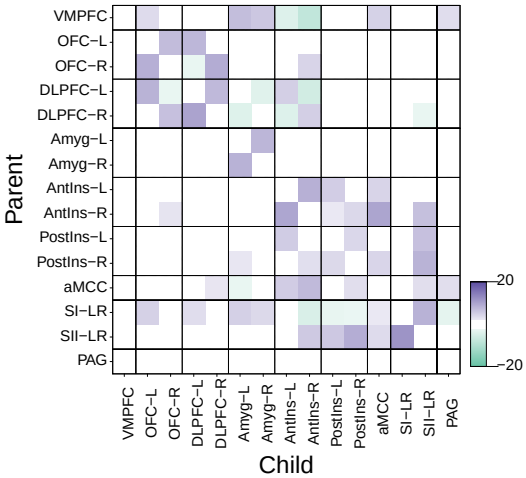
(a) **Mean partial correlation (safe)**



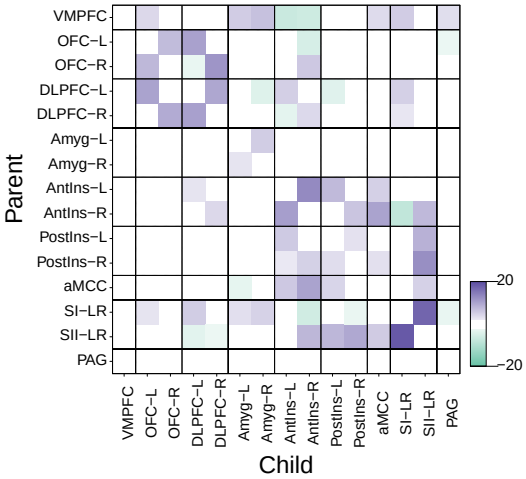
(b) **Mean partial correlation (shock)**



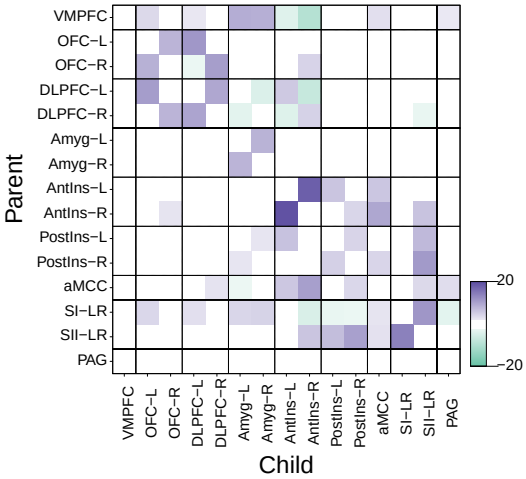
(c) **Safe**



(d) **Anticipation of shock**



(e) **Safe**



(f) **Anticipation of shock**

Figure 2.10: The state vector $\hat{\boldsymbol{\theta}}_t(r)$ provides a measure of connectivity strength. **(a)** The mean partial correlations over subjects, white indicates an (absolute) mean partial correlation less than 0.02. **(b)** The standardised t statistic $\bar{d}_{ir}$ for the 'safe' data after false discovery rate correction (α = 0.05) on the corresponding p-values. **(c)** As **(b)** but for the 'anticipation of shock' data. **(d)** and **(e)** As **(b)** and **(c)** but using the unstandardised t statistic $\bar{\mu}_{ir}$.

To compare the 'safe' and 'anticipation of shock' networks based on connectivity strength, we fitted the parent set of the model $\hat{\mathcal{M}}_{\mathcal{G}_{safe+shock}}$ and obtained the following paired t statistic for each edge:

$$d_{paired} = \frac{\frac{1}{S}\sum_{s=1}^{S}\left(\bar{d}_i(s,r)^{safe} - \bar{d}_i(s,r)^{shock}\right)}{SD\left(\bar{d}_i(s,r)^{safe} - \bar{d}_i(s,r)^{shock}\right)/\sqrt{S}}$$

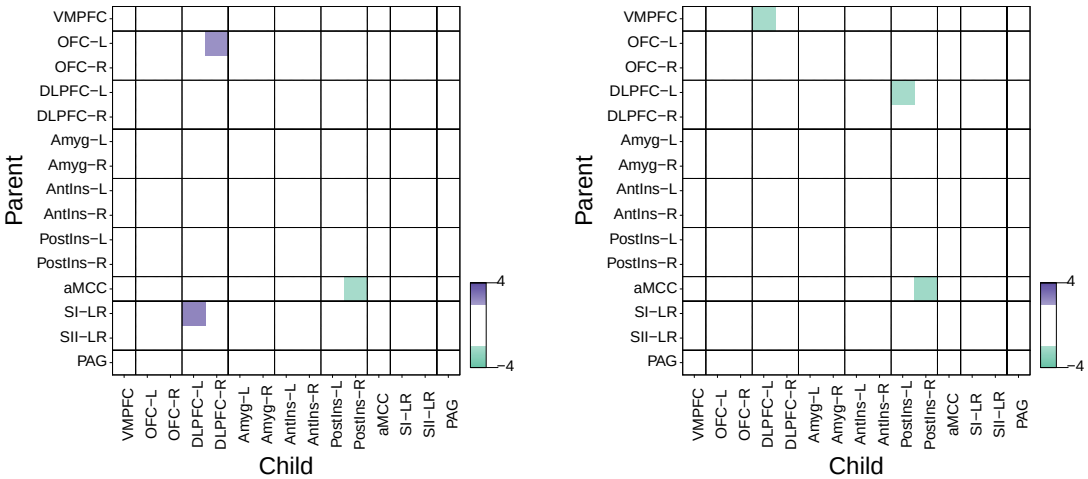
$$\mu_{paired} = \frac{\frac{1}{S}\sum_{s=1}^{S}\left(\bar{\mu}_i(s,r)^{safe} - \bar{\mu}_i(s,r)^{shock}\right)}{SD\left(\bar{\mu}_i(s,r)^{safe} - \bar{\mu}_i(s,r)^{shock}\right)/\sqrt{S}}$$

where

$$\bar{d}_i(s,r) = \frac{1}{T}\sum_{t=1}^{T} d_{it}(s,r) \qquad \bar{\mu}_i(s,r) = \frac{1}{T}\sum_{t=1}^{T} \mu_{it}(s,r)$$

Significant edges ($p < 0.05$) are shown in Figure 2.11. However, none of the edges found to be significant survived false discovery rate correction ($\alpha = 0.05$).

In Figure 2.6, we showed that the 'safe' and 'anticipation of shock' parent sets for each node differed by a median of 3 to 5 parents and had $\log_e$ Bayes factors that indicated strong evidence for a difference. This might seem to conflict with our failure to detect statistically significant differences between the two conditions. To further explore this behaviour, Figure 2.12 plots the absolute value of the unstandardised t statistic $\bar{\mu}_{ir}$ for each subject and each edge in the group networks $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ and $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$. It is straightforward to see that the strongest, most consistent edges also have the highest connectivity strengths. We showed in Figure 2.5 that the most consistent edges are shared across experimental conditions. From this, we may conclude that it is the weaker edges, in both presence and connectivity strength, that give rise to the observed differences between the networks. While it is possible that this is a reflection of real but subtle differences between the two conditions, it may be the case that the MDM-DGM tends to find overly-complex models. This is something we will come back to in Chapter 4.



(a) d_{paired} **(b)** μ_{paired}

Figure 2.11: A paired t-test does not find a significant difference between the connectivity strengths for the 'safe' and 'anticipation of shock' datasets. Purple indicates a higher value for the 'safe' data, green indicates a indicates a higher value for the 'anticipation of shock' data. None of these edges survive false discovery rate correction (α = 0.05).

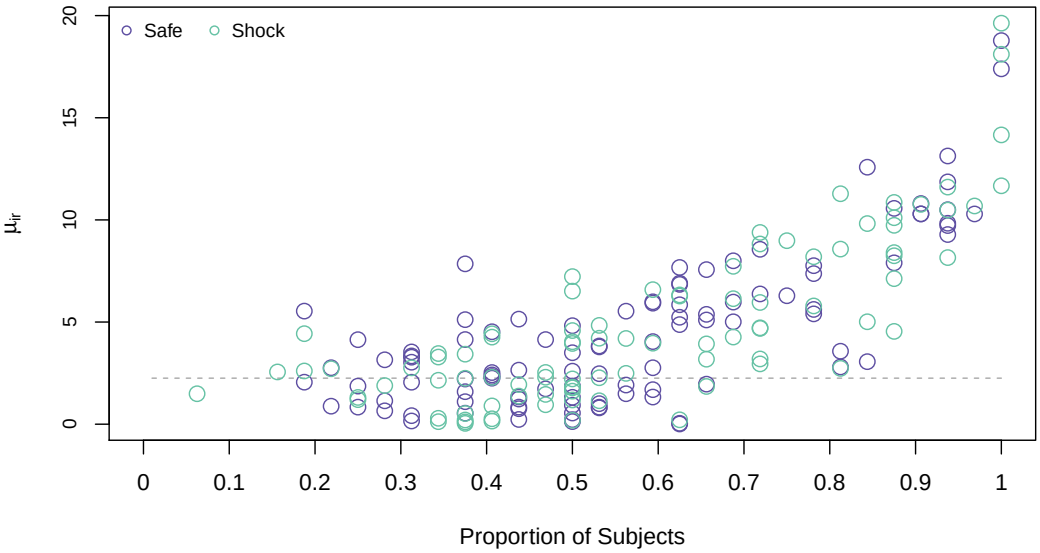
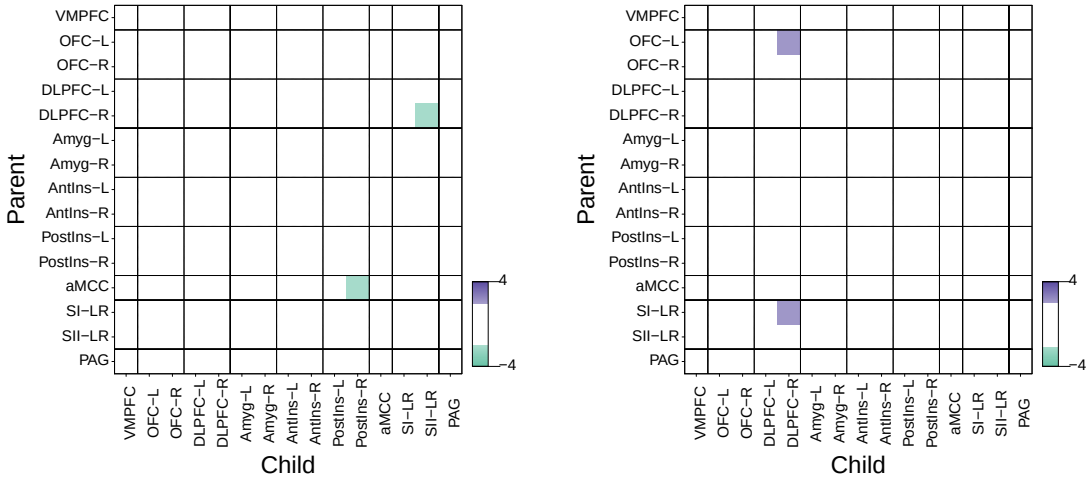


Figure 2.12: Edges shared by a high proportion of subjects have stronger connectivity strengths. The absolute value of $\bar{\mu}_{ir}$ for each edge in the networks $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ and $\hat{\mathcal{M}}_{\mathcal{G}_{shock}}$ against the proportion of subjects which have the edge in the individual networks. The dotted line indicates the significance threshold. It can be seen that all edges shared by more than 70 % of subjects have an (absolute) connectivity strength above the significance threshold.

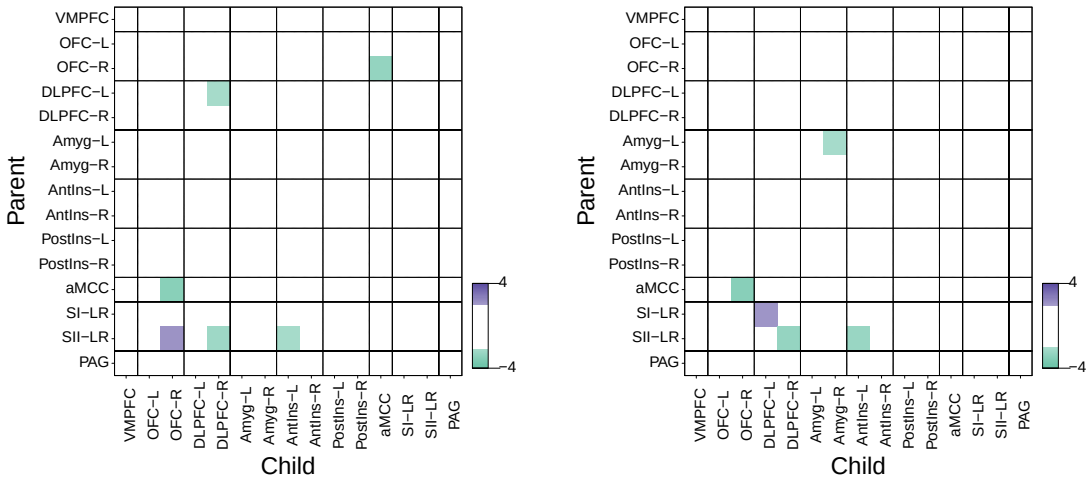
2.9 Detecting Differences Based on Trait and Induced Anxiety

We used the MDM-DGM to see if we could detect differences between the subjects based on measures of trait and induced anxiety. Based on a mean split, we grouped the subjects into no induced anxiety ($S = 14$) and induced anxiety ($S = 18$) using the VAS difference measure and low trait anxiety ($S = 19$) and high trait anxiety ($S = 13$) using the Factor Scores measure. These measures are detailed in Bijsterbosch et al. (2015) and described briefly in section 2.2. We then performed two-sample t-tests, using the $\bar{d}_{ir}$ and $\bar{\mu}_{ir}$ metrics for connectivity strength, to assess whether any of the connectivity strengths were significantly different between subgroups. The t statistics that were found to be significant ($p < 0.05$) are shown in Figure 2.13 for the 'safe' dataset and Figure 2.14 for the 'anticipation of shock' dataset. However, none of these edges survived false discovery rate correction ($\alpha = 0.05$).



(a) Induced anxiety $\bar{d}_{ir}$

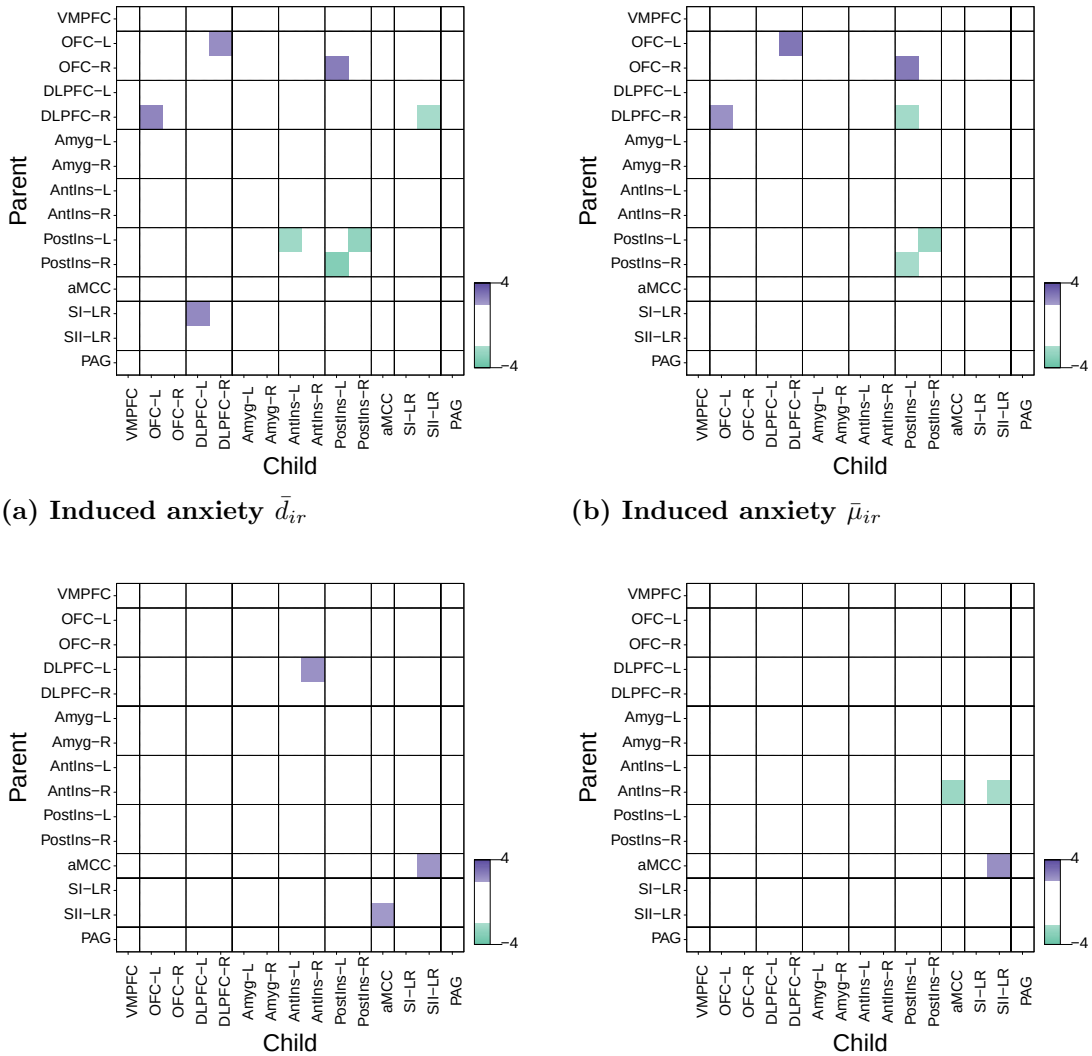
(b) Induced anxiety $\bar{\mu}_{ir}$



(c) Trait anxiety $\bar{d}_{ir}$

(d) Trait anxiety $\bar{\mu}_{ir}$

Figure 2.13: Differences between the 'safe' networks when the subjects are split based on induced and trait anxiety. Significant edges ($p < 0.05$, before FDR adjustment) are shown in terms of the t statistic for a two-sample t-test, purple indicates a higher value for the 'safe' data and green a higher value for the 'anticipation of shock' data. None of these edges survive false discovery rate correction ($\alpha = 0.05$).



(a) **Induced anxiety** $\bar{d}_{ir}$

(b) **Induced anxiety** $\bar{\mu}_{ir}$

(c) **Trait anxiety** $\bar{d}_{ir}$

(d) **Trait anxiety** $\bar{\mu}_{ir}$

Figure 2.14: Differences between the 'anticipation of shock' networks when the subjects are split based on induced and trait anxiety. Differences between the 'anticipation of shock' networks when the subjects are split based on induced and trait anxiety. Significant edges ($p < 0.05$, before FDR adjustment) are shown in terms of the t statistic for a two-sample t-test, purple indicates a higher value for the 'safe' data and green a higher value for the 'anticipation of shock' data. None of these edges survive false discovery rate correction ($\alpha = 0.05$).

2.10 Discussion

In this chapter, we have used the two datasets of Bijsterbosch et al. (2015) to discover and analyse directed, functional connectivity networks using the MDM-DGM search. In this section, we review some of the features of the MDM-DGM search outlined in this chapter and the previous one, highlighting some key strengths and weaknesses.

MDM-DGM estimates dynamic, directed connectivity

As partial correlation networks have been shown to have good sensitivity for detecting edge presence, we used partial correlation networks to assess the ability of the MDM-

DGM to infer edge presence. The MDM-DGM recovers very similar networks to the partial correlation networks. This is noticeable because the MDM-DGM networks tend to be symmetric, with strong connectivity between the hemispheres. We can model this bidirectional connectivity because we do not constrain the MDM-DGM networks to be acyclic. The VMPFC showed reasonably consistent asymmetric connectivity and using the group-level analyses it was was found to have no parents but multiple children. This is consistent with its known top-down and regulatory role in the processing of negative emotion (Bijsterbosch et al., 2015; Motzkin et al., 2015).

Given a set of parents discovered by the MDM-DGM search, we can fit a Dynamic Linear Model and obtain estimates for the strength of the connectivity. Like partial correlation, these estimates may be positive or negative, but unlike partial correlation, for the case where the discount factor $\delta(r) < 1$, these estimates may vary over time. As $\delta(r)$ is chosen for each node by maximising the Log Predictive Likelihood, the amount of dynamics in the regression coefficients is *data driven* and the case where $\delta(r) = 1$ allows for a static model if this provides the best fit to the data.

MDM-DGM can estimate both subject and group level networks

The MDM-DGM search consists of calculating a (log) likelihood that factors by subject and by node. Not only is this hugely advantageous from a computational perspective, as the model search may readily be parallelised, but it also allows us to construct and analyse individual subject networks. It is straightforward to sum the scores to obtain a group-level network. We constructed an MDM-DGM network for each subject for both the 'safe' and 'anticipation of shock' data, finding that around half of the edges are consistently present or absent across subjects and experimental conditions. Some differences between the 'safe' and 'anticipation of shock' conditions were detected between the individual networks, and between group networks, but these differences tended to be limited to a small number of subjects.

Given the group networks, we obtained estimates for the time-varying connectivity strengths $\hat{\boldsymbol{\theta}}_t(r)$. These showed that connectivity (e.g. connectivity between the orbitofrontal and dorsolateral prefrontal cortices) that was consistent over subjects in terms of edge presence was associated with high connectivity strengths.

Further analysis comparing the $\log_e$ Bayes factor to assess model fit suggested strong evidence for a difference between the chosen parent sets for each experimental condition, but due to the lack of a statistically significant difference between the networks, we hypothesise that this is being driven by the presence of less consistent edges with smaller connection strengths. This may be because there are real but subtle differences between the networks, although if this were the case, we might expect these differences to become more pronounced in the splits based on anxiety metrics. Alternatively, we may hypothesise that these inconsistent edges with low connectivity weights are spurious, or at least, not as physiologically interesting as the stronger edges, and therefore compromise the robustness of the estimated networks. We will explore methods to

minimise the impact of these 'spurious' edges in Chapter 4.

MDM-DGM networks are consistent between 'safe' and 'anticipation of shock'

When comparing the 'safe' and 'anticipation of shock' networks, the MDM-DGM did not find any significant differences that survived false discovery rate correction, even when the subjects were split into low and high anxiety subgroups. However, the consistency between the estimated networks lends support to the validity of the MDM-DGM approach. As discussed in section 1.2.1, it has been shown that resting-state networks strongly correlate with their associated task-based networks (Smith et al., 2009; Tavor et al., 2016) and this behaviour is apparent in the results presented here.

Chapter 3

Scaling-up the MDM-DGM with Stepwise Regression

3.1 Introduction

While the MDM-DGM search allows us to construct biophysically-plausible networks and test a number of hypotheses, it would be desirable to be able to work with networks with many more nodes, encompassing many more brain regions. The more brain regions that may be included, the more it is possible to interpret functional connectivity in terms of the underlying architecture and to be able to say that the *direct* connectivity that we estimate represents a true direct causal influence, rather than the influence of other, unmeasured regions. However, for many of the widely-used methods for directed connectivity, networks with more than a handful of nodes have not been computationally feasible. As the number of nodes increases, so does the number of combinations of directed edges, and methods such as IMaGES and DCM, as well as the MDM-IPA and MDM-DGM, become severely limited by the size of the model space (Henry and Gates, 2017), although recently methods for larger networks have emerged. For example, Razi et al. (2017) demonstrate a spectral DCM approach using an empirical dataset with 36 brain regions. In this chapter, we show how the size of the MDM-DGM model space grows exponentially, making a search impossible for network with more than 20 nodes. In order to construct larger MDM-DGM networks in reasonable computational time, we consider stepwise methods which, as we will show, dramatically reduce the number of models it is necessary to score.

3.1.1 MDM-DGM Computation Time

As previously described, the Log Predictive Likelihood has closed-form and factors by node and by model (candidate set of parents). The search over the model space may therefore be readily parallelised and can be performed very quickly. Using a MacBook Pro, 2.7 GHz Intel Core i5, 8GB RAM running R version 3.4.0 and the C++ implementation available in the `multdyn` package (Schwab et al., 2017b), the run time for the all-parent model for an individual node (number of nodes $n = 15$) for a single value of the discount factor was 0.004 seconds. A grid search over a range of discount factors ($\delta(r) \in [0.5, 1]$, step size 0.01) is therefore achievable in around 0.2 seconds and a search over all $N = 2^{n-1}$ (16,384) candidate sets of parents may be performed in less than an hour. We use the term *exhaustive* to refer to an MDM-DGM model search

which scores every candidate model. Estimates for the run times of the MDM-DGM exhaustive search are provided in Table 3.1. It is immediately clear that while a 15 node network may be scored in a reasonable amount of time, networks with only a few additional nodes require significantly more computation time, such that it is not feasible to perform an exhaustive search on networks with more than 20 nodes.

No. of nodes	**No. of models**	**Approx. run time**
5	16	3.2 seconds
10	512	1.7 minutes
15	1.6×10^4	55 minutes
20	5.2×10^5	29 hours
50	5.6×10^{14}	3.6×10^6 years

Table 3.1: Estimated run time of the MDM-DGM, per subject, per node, for increasing numbers of nodes. The approximate computation time to score an individual model (with 790 time points) was 0.2 seconds, using a C++ implementation available in the `multdyn` package for R. A parallelised model search over 15 nodes may be performed in less than a hour but, as the size of the model space increases exponentially, an exhaustive model search is not feasible for more than 20 nodes.

3.1.2 MDM-DGM Computational Complexity

Table 3.1 is based on the assumption that the run time of an individual model is approximately the same regardless of the number of parents in the candidate model. In this subsection, we provide a more rigorous assessment of the computational complexity of the MDM-DGM search. As part of this, we examine the Dynamic Linear Model as implemented in the `dlm.lpl` function in the `multdyn` package for R (Schwab et al., 2017b). This function uses one-step updating to calculate the Log Predictive Likelihood for a specified set of parents and discount factor. It takes as inputs the time series of a particular node r and the time series of the parents in the parent set, a scalar value of the discount factor and values for the prior hyperparameters at time $t = 0$. The dimensions of the hyperparameters $\mathbf{m}_0(r)$ and $\mathbf{C}_0^*(r)$ depend on the number of parents. The computational complexity of scoring an individual model (candidate set of parents) using this function is determined by the number of parents and the number of time points T; complexity increases linearly with the number of time points. To find the parents for an individual node, for each candidate parent set, this function is called for a fixed number of potential values of $\delta(r)$ (usually 51, $\delta(r) \in [0.5, 1]$, step size 0.01), that is, the function is called $51 \times N$ times.

Table 3.2 shows how the complexity of a single iteration in the Dynamic Linear Model depends on the number of parents. The updating relations are those described in section 1.4.2. For compactness, the r notation is dropped for the rest of this subsection. At time t, the vectors $\mathbf{F}_t$, $\mathbf{m}_t$ and $\mathbf{A}_t$ have length p_r where p_r is the number of parents plus one to account for an intercept, so the all-parent model has $p_r = n$. The matrices $\mathbf{R}_t^*$,

$\mathbf{R}_t$, $\mathbf{C}_t^*$ and $\mathbf{C}_t$ have dimension $p_r \times p_r$. It follows that the complexity (and subsequent computation time) depends quadratically on the number of parents in the model being scored. However, as previously stated, the number of models increases exponentially with the number of nodes n, so this increase in computational complexity will be dwarfed by the increase in the size of the model space.

Calculation	Dimension	Complexity
$\mathbf{R}_t^* = \mathbf{C}_{t-1}^* / \delta$	$p_r \times p_r$	$O(n^2)$
$\mathbf{R}_t = \mathbf{R}_{t-1}^* S_{t-1}$	$p_r \times p_r$	$O(n^2)$
$f_t = \mathbf{F}_t^\top \mathbf{m}_{t-1}$	–	$O(n)$
$Q_t^* = 1 + \mathbf{F}_t^\top \mathbf{R}_t^* \mathbf{F}_t$	–	$O(n^2)$
$Q_t = Q_t^* S_{t-1}$	–	$O(1)$
$e_t = Y_t - f_t$	–	$O(1)$
$\mathbf{A}_t = \mathbf{R}_t^* \mathbf{F}_t / Q_t^*$	$p_r \times 1$	$O(n^2)$
$\mathbf{m}_t = \mathbf{m}_{t-1} + \mathbf{A}_t e_t$	$p_r \times 1$	$O(n)$
$n_t = n_{t-1} + 1$	–	$O(1)$
$d_t = d_{t-1} + e_t^2 / Q_t^*$	–	$O(1)$
$S_t = d_t / n_t$	–	$O(1)$
$\mathbf{C}_t^* = \mathbf{R}_t^* - \mathbf{A}_t \mathbf{A}_t^\top Q_t^*$	$p_r \times p_r$	$O(n^2)$
LPL	–	$O(1)$

Table 3.2: Computational complexity of the Dynamic Linear Model. The first column shows the updating steps of the Dynamic Linear Model needed to calculate the Log Predictive Likelihood as implemented in the `multdyn` package for R. The second column shows the dimension of the output, '–' indicates a scalar and p_r is the number of parents plus one to include an intercept. The third column is the computational complexity of the calculation. The complexity of a single iteration is therefore quadratic.

3.2 Forward Selection and Backward Elimination

We explore stepwise methods for model selection: our goal is to recover the same sets of parents without scoring all 2^{n-1} combinations. We exploit the fact that the LPL factors by node so the highest-scoring set of parents may be found for each node individually. We consider two complementary stepwise methods: forward selection (FS) and backward elimination (BE) (see, for example, Davison (2003), Chapter 8).

Forward selection starts by scoring the zero parent (intercept-only) model and all the one parent models. The parent (if any) which gives the biggest increase in the LPL is selected for inclusion. The algorithm then scores all the two-parent models which include this parent and so on, until the inclusion of additional parents does not increase the LPL. Figure 3.1b illustrates how this process may reproduce the graph in Figure 3.1a. The algorithm first scores the model with $Pa(4) = \{\varnothing\}$ (Step 1) and then $Pa(4) = \{1\}$, $Pa(4) = \{2\}$, $Pa(4) = \{3\}$ (Step 2). Suppose after these first two steps $\hat{Pa}(4) = \{3\}$.

The third step is to score $Pa(4) = \{1,3\}$ and $Pa(4) = \{2,3\}$, replacing $\hat{P}a(4) = \{3\}$ with $\hat{P}a(4) = \{1,3\}$. The final model to score is $Pa(4) = \{1,2,3\}$, which will fail to increase the LPL, so the algorithm successfully returns the same parent set as the MDM-DGM.

Backward elimination (BE) obeys similar principles. It begins by scoring the model that includes all $n-1$ candidate parents, removing parents one at a time until doing so fails to increase the LPL. This procedure is shown in Figure 3.1c, where Step 1 scores the model with parents $Pa(4) = \{1,2,3\}$ and Step 2 scores $Pa(4) = \{1,2\}$, $Pa(4) = \{1,3\}$ and $Pa(4) = \{2,3\}$, choosing the 'correct' model $\hat{P}a(4) = \{1,3\}$. Step 3 then scores $Pa(4) = \{1\}$ and $Pa(4) = \{3\}$ but neither improve the LPL so the algorithm terminates.

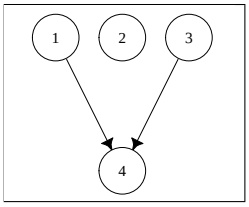
The maximum number of models that can be scored (per node) using these methods is $N_{step} = 1 + \sum_{k=1}^{(n-1)}(n-k)$. While this is 88 % of the total for a 4 node network (i.e. these methods may end up scoring 7 out of the 8 candidate sets of parents), for a 15 node network, the maximum number of models that can be scored is 106 out of 16,384 (or 0.65 % of the total). The dramatic reduction in the size of the model space for models with increasing numbers of nodes is shown in Table 3.3.

No. of nodes	No. of models	% of total	Approx. run time
5	11 (16)	69	2.2 seconds
10	46 (512)	9	9.2 seconds
15	106 (1.6×10^4)	0.65	21 seconds
20	191 (5.2×10^5)	0.036	38 seconds
50	1226 (5.6×10^{14})	2.2×10^{-10}	4.1 minutes

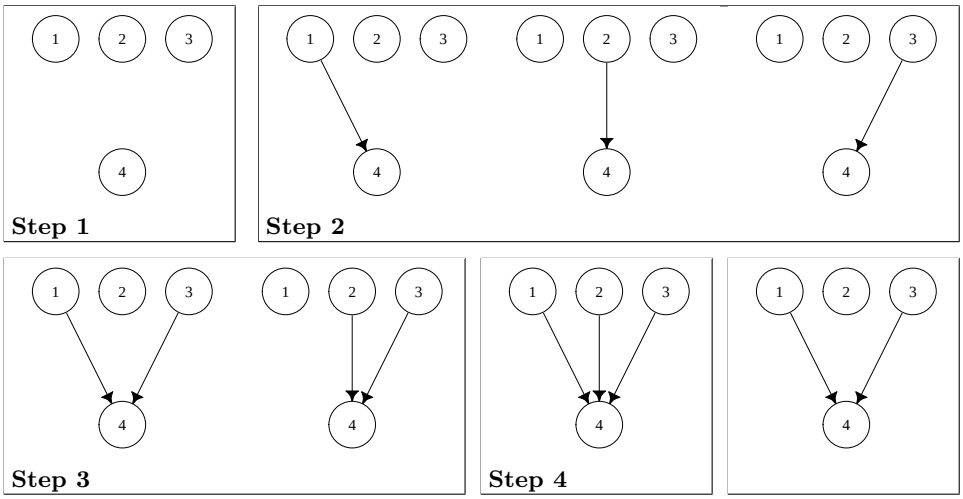
Table 3.3: Stepwise methods dramatically reduce the number of models to score. As the number of nodes increases, the maximum number of models that may be scored by either forward selection or backward elimination, as a percentage of the total number of candidate models, significantly decreases.

The example in Figure 3.1b assumes that the LPL for the zero-parent model is lower than for the selected one-parent model with $Pa(4) = \{3\}$. If this is not the case, the algorithm will terminate, returning the $Pa = \{\varnothing\}$ as the winning model and the stepwise solution will have two missing parents. Similarly, consider the backward elimination example if the winning model in the exhaustive search is $Pa(4) = \{3\}$, rather than $Pa(4) = \{1,3\}$ and that $Pa(4) = \{1,2,3\}$ has an higher score than $Pa(4) = \{1,2\}$, $Pa(4) = \{1,3\}$ or $Pa(4) = \{2,3\}$. The algorithm will terminate, returning $Pa(4) = \{1,2,3\}$ as the winning model and the stepwise solution will have two extra parents. To avoid these errors, we simply need to remove the instruction to terminate if the LPL cannot be improved. As can be seen in Table 3.3, the stepwise algorithms score such a tiny fraction of the model space that any computational speed-up that may be obtained from instructing the algorithm to terminate will be negligible.

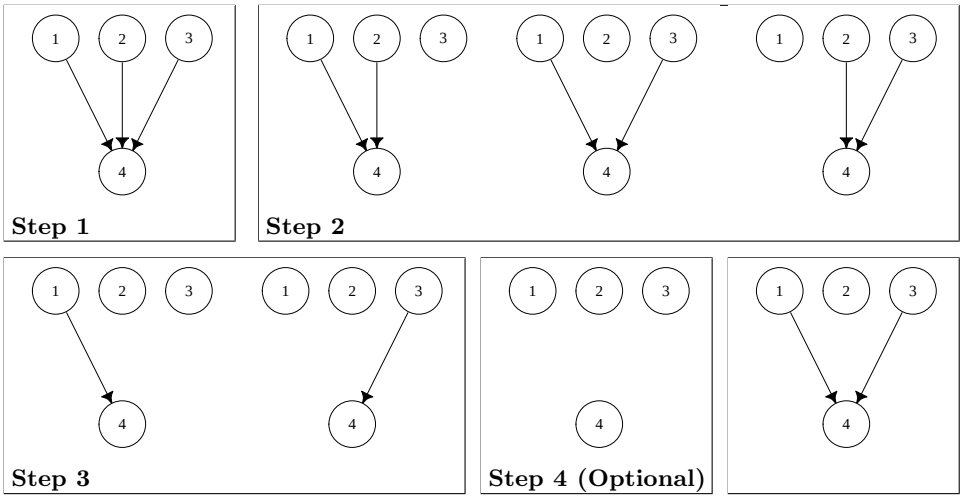
These two methods were run on the 15 node resting-state ('safe') data, using code we



(a) MDM-DGM exhaustive search



(b) Forward selection



(c) Backward elimination

Figure 3.1: Illustration of stepwise algorithms for the MDM-DGM. To goal is to discover the same parent set as an exhaustive MDM-DGM search e.g. the parent set shown in **(a)**. The forward selection (FS) algorithm in **(b)** begins with the zero-parent (intercept-only) model, adding parents one at a time. The backward elimination (BE) algorithm in **(c)** begins with the all-parent model and removes parents one at a time.

have made available[1]. As our aim is to recover the same networks as those found by the exhaustive search, we quantify performance in terms of the number of edges (out of a possible 210, for each subject) that are the same (present or absent) as in the exhaustive networks. It is also informative to distinguish missing edges (edges present in the exhaustive network but absent in the stepwise network) and extra edges (edges absent in the exhaustive network but present in the stepwise network). Note that we use the terms 'correct' and 'incorrect' (as well as 'missing' and 'extra') only the context of comparison of the stepwise to the exhaustive networks, not with any reference to 'correctness' of the network as a reflection of the underlying physiology.

The algorithms described here, like the MDM-DGM exhaustive search, assume that there is a single maximum LPL and a single winning set of parents. Therefore, the choice of parent to include or exclude may be based on an LPL that is only fractionally higher and there may be other parents that may be included or excluded giving a model with equivalent evidence in terms of the $\log_e$ Bayes factor. For this reason, it is also informative to assess the performance of the stepwise algorithms in terms of $\log_e$ Bayes factors. If the winning parents for node r in an exhaustive search are $\hat{P}a(r)$, as before, and the winning parents in a forward selection or backward elimination search are $\hat{P}a(r)_{step}$, then we may argue that the stepwise algorithm has discovered an equally correct model if $\mathrm{LPL}[\hat{P}a(r)] - \mathrm{LPL}[\hat{P}a(r)_{step}] < 1$. In line with Kass and Raftery (1995), a $\log_e$ Bayes factor of less than 1 indicates that there is no evidence to prefer one model over another.

3.2.1 Performance of Forward Selection and Backward Elimination

Results of forward selection and backward elimination algorithms applied to the 15 node resting-state data are presented in Figures 3.2 and 3.3. The performance of these algorithms is reasonably good. In the case where the algorithm terminated if at any step no improvement in the LPL was obtained, forward selection correctly identified 91.2 % of edges on average (minimum 82.9 %, maximum 97.1 %, s.d. 3.7 %). Of the incorrectly specified edges, across subjects, 79.2 % were missing. This is clear from Figure 3.2b: the plot on the left hand side shows the number of missing edges and the plot on the right hand side the number of extra edges, compared to the exhaustive search. Figures 3.2a and 3.2c show the improved performance obtainable when the algorithm iterates until the all-parent model. In this case, 95.3 % of edges are correctly identified on average (minimum 89.5 %, maximum 99.5 %, s.d. 2.5 %). Looking at Figure 3.2c, it is clear that this increase in performance occurs because the number of missing edges is significantly reduced.

Backward Elimination performed comparably, correctly identifying 90.9 % of edges (minimum 79 %, maximum 98.1 %, s.d. 4.1 %) when the algorithm terminated if it reached a local maximum. Of the incorrectly specified edges, across subjects, 76.6 %

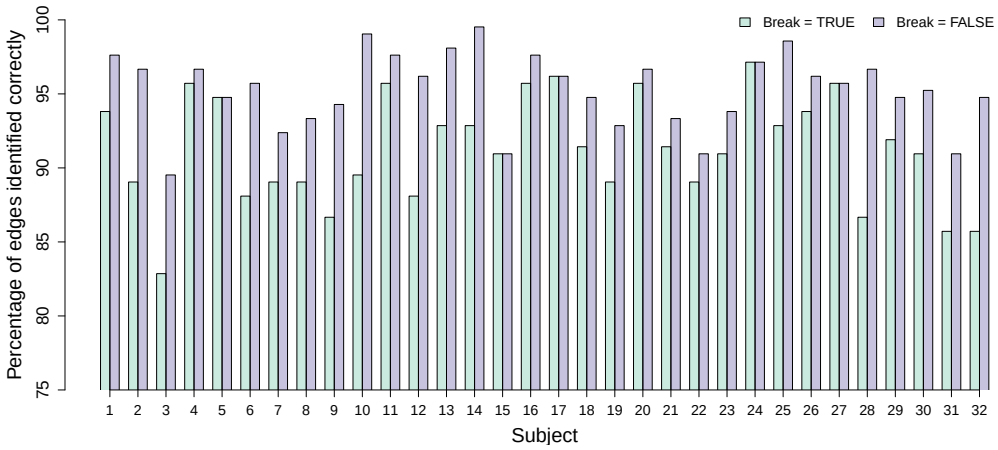
[1]Schwab, S., Harbord, R., Costa, L., and Nichols, T. *multdyn: Multiregression Dynamic Models*, 2017a. URL `https://CRAN.R-project.org/package=multdyn`. R package version 1.6

were extra. The numbers of missing and extra edges over subjects are shown in Figure 3.3b. In the case where the algorithm continued to the zero-parent model, backward elimination correctly identified 95.5 % of edges on average (minimum 88.1 %, maximum 100 %, s.d. 2.8 %).

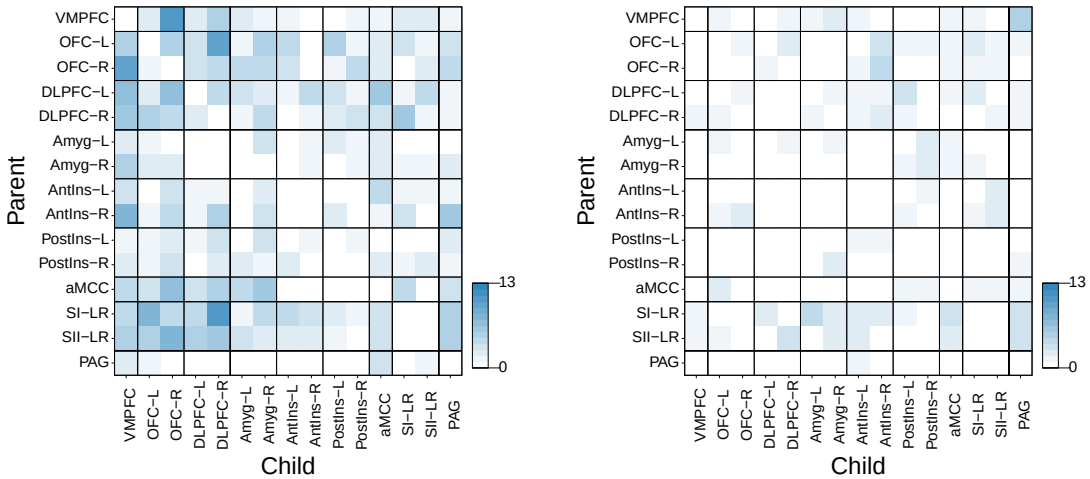
It is clear (and not unexpected) that forward selection tends to miss edges whereas the converse is true for backward elimination. From Figures 3.2b and 3.3b, it can be seen that these errors occur relatively uniformly across the 15 nodes, rather than there being particular nodes where the stepwise method is more or less successful at identifying the correct parents, although the number of extra parents for the VMPFC (consistently predicted to be parentless in the group analyses) is noticeable when using backward elimination.

When the algorithm runs until the all-parent model, the number of edges missed using forward selection is reduced from 6.9 % (of the total number of edges) to 2.5 %, meaning the total numbers of missing and extra edges are now roughly equivalent (53.8 % missing, 46.2 % extra). For backward elimination, the number of extra edges is reduced from 6.9 % (of the total number of edges) to 2.2 %, and again the total numbers of missing and extra edges are now roughly equivalent (51.5 % missing, 48.5 % extra). Another test of the effectiveness of these methods is to calculate the $\log_e$ Bayes factor between the parent sets found by the exhaustive and stepwise searches. Results are shown in Figure 3.4. For forward selection, 76.7 % of models (out of $S \times n = 480$) had a $\log_e$ Bayes factor of zero. For a further 5.8 %, the $\log_e$ Bayes factor was less than 1, while 9.4 % had a $\log_e$ Bayes factor greater than 1 (but less than 3), indicating evidence to prefer the model discovered by the exhaustive search. Only 8.1 % had a $\log_e$ Bayes factor greater than 3, indicating strong evidence for a difference. For backward elimination, the results were almost identical, with a $\log_e$ Bayes factor of zero for 76.2 % of models and a $\log_e$ Bayes factor of less than 1 for 10 %, while 8.1 % had a $\log_e$ Bayes factor greater than 1 (but less than 3) and only 5.6 % had a $\log_e$ Bayes factor greater than 3.

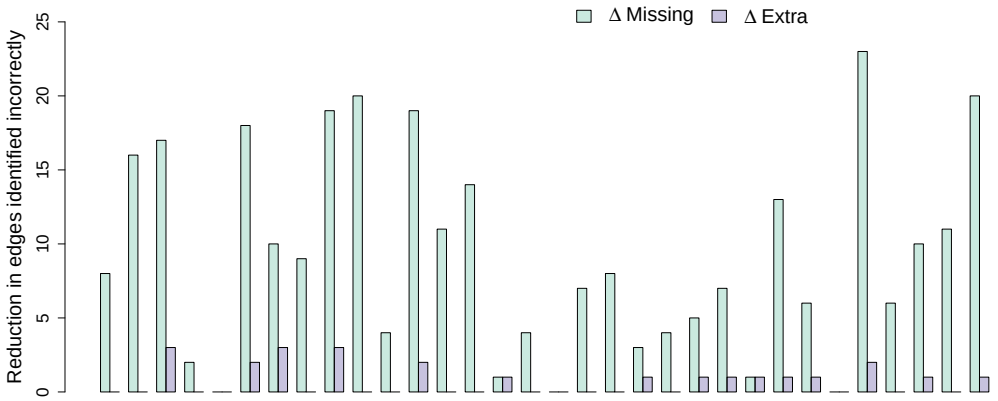
It may be concluded that these methods can successfully reproduce the MDM-DGM exhaustive networks when the number of nodes $n = 15$. For this dataset, the performance of forward selection and backward elimination is equivalent, and the accuracy is comparable over subjects.



(a)

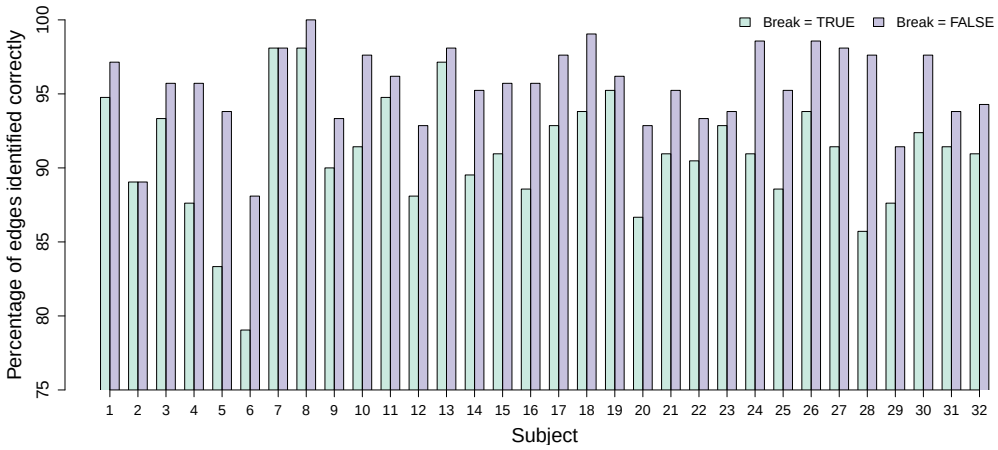


(b)

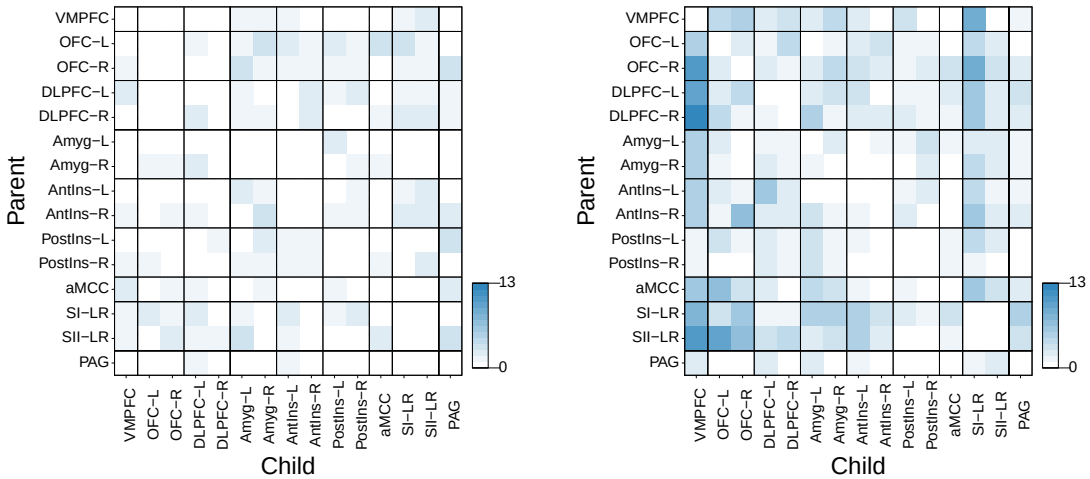


(c)

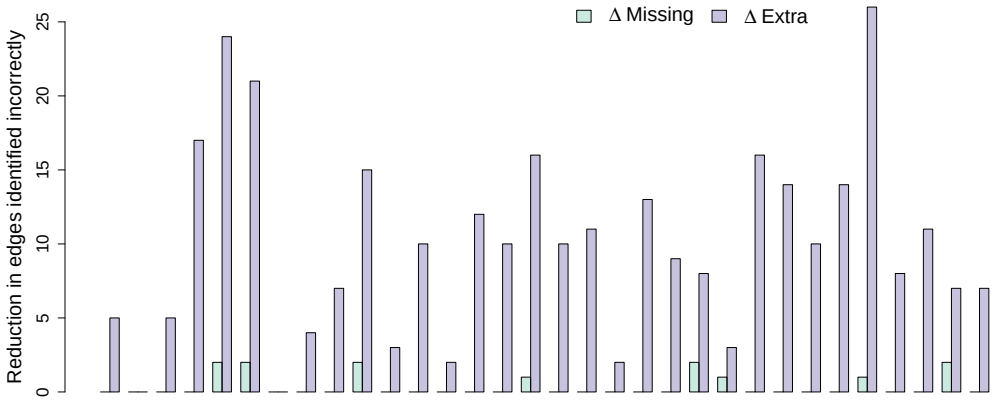
Figure 3.2: The performance of forward selection. (a) For each subject, the percentage of edges correctly identified by the forward selection algorithm when the algorithm terminates if a local maximum is found (green) and when the algorithm terminates at the all-parent model (purple). **(b)** The number of missing (left) and extra (right) edges across all subjects using the algorithm that terminates if a local maximum is found. **(c)** The reduction in the number of 'missing' and 'extra' edges when the algorithm continues until the all-parent model.



(a)



(b)



(c)

Figure 3.3: The performance of backward elimination. (a) For each subject, the percentage of edges correctly identified by the backward elimination algorithm when the algorithm terminates if a local maximum is found (green) and when the algorithm terminates at the zero-parent (intercept-only) model (purple). **(b)** The number of missing (left) and extra (right) edges across all subjects using the algorithm that terminates when a local maximum is found. **(c)** The reduction in the number of 'missing' and 'extra' edges when the algorithm continues until the zero-parent model.

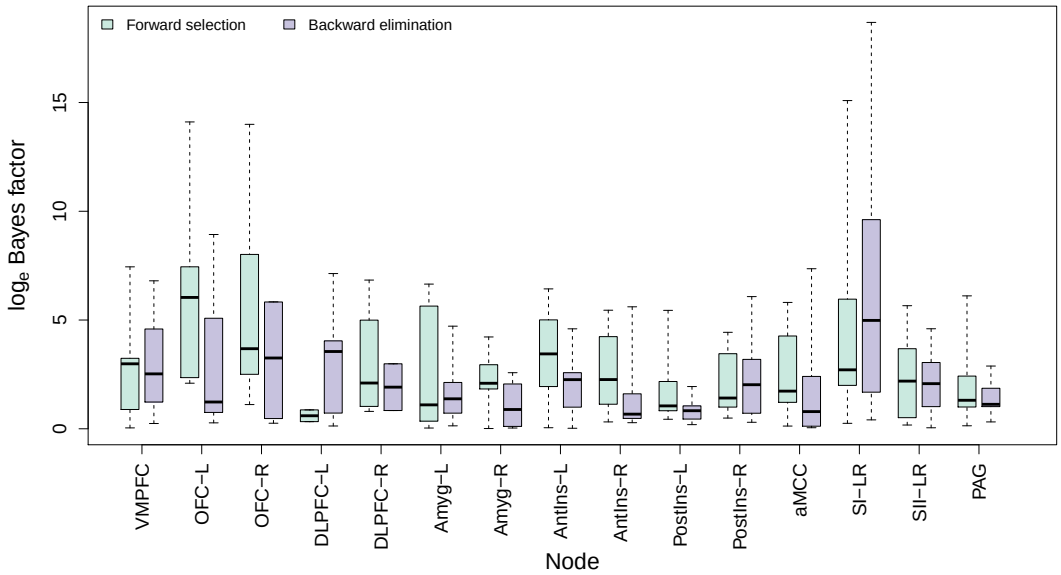
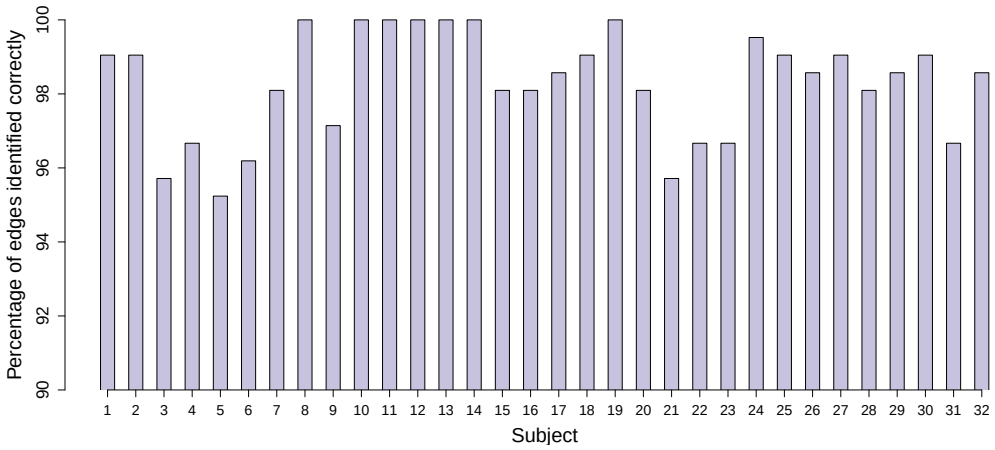


Figure 3.4: Log$_e$ Bayes factor comparison between the parent sets discovered by exhaustive and forward selection and exhaustive and backward elimination searches. Plotted values are for the case when the stepwise algorithm fails to find the set of parents identified by an exhaustive search. Whiskers show the minimum and maximum values.

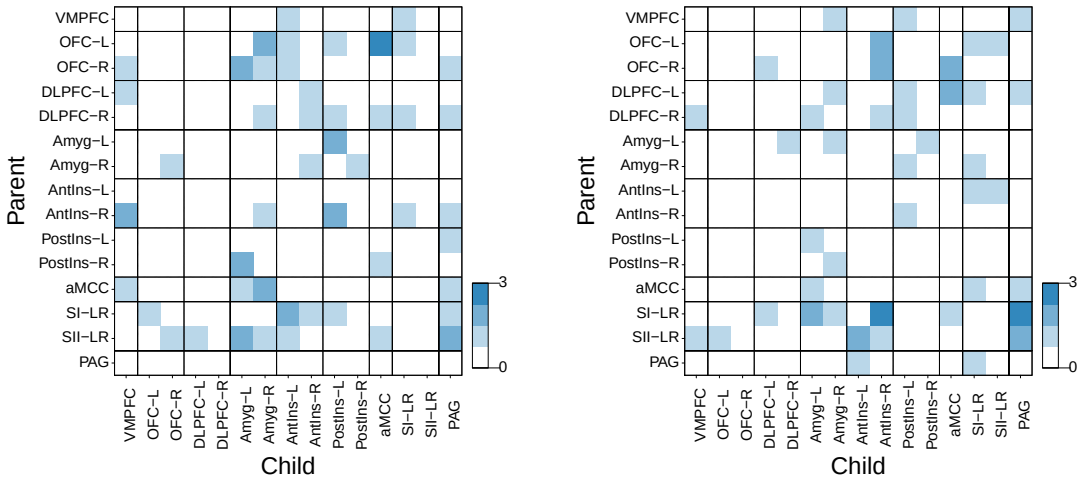
3.3 Combining Forward Selection and Backward Elimination

Due to the significant reduction in the size of the model space using stepwise methods, it is feasible computationally to run both forward selection and backward elimination: for a 15 node network, this means scoring 212 models per subject, per node (assuming the termination constraint is removed). We may then compare the scores for the winning models using each method and, if they are different, select the set of parents with the higher LPL. Results are shown in Figure 3.5a. Mean accuracy for edges correctly identified was 98.3 % (minimum 95.2 %, maximum 100%, s.d. 1.4 %) and for 7 out of the 32 subjects the exhaustive MDM-DGM networks were reproduced in full (i.e. 100 % of edges were correctly identified). The *worst* performance still identified 95.2 % of edges correctly. Looking at Figure 3.5b, again we can see that the small proportion of edges that were incorrectly specified were not associated with a particular node (or nodes) and that the ratio of missing to extra edges was almost equal (52.2 % missing (0.9 % of the total number of edges); 47.8 % extra (0.8 % of the total number of edges)). This is further illustrated in Figure 3.5c, which shows the number of missing and extra edges per subject.

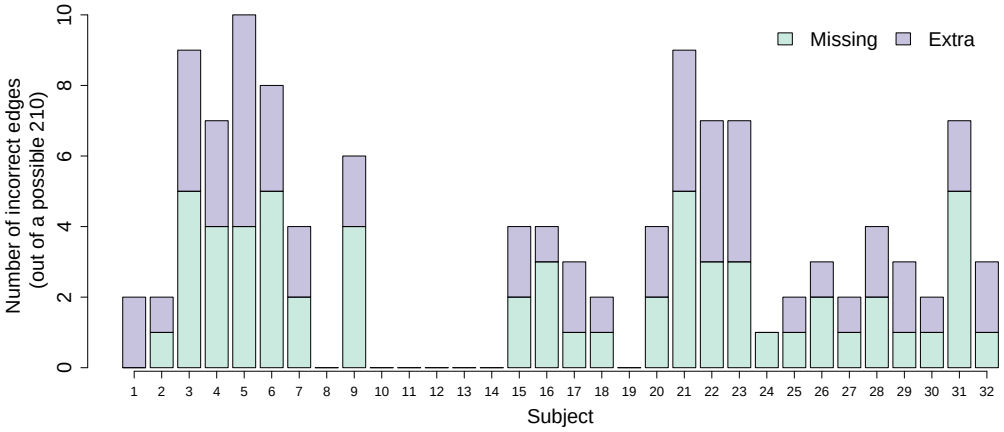
Finally, we look at the log$_e$ Bayes factors for the misspecified models (see Figure 3.6). Across all subjects and all nodes, the stepwise algorithm returned a different parent set for 40 out of 480 (8.3 %). However, out of these, 22 had a log$_e$ Bayes factor difference of less than 1. This meant that, using the combined forward selection and backward elimination algorithm, for only 3.7 % of cases the stepwise algorithm returned a parent set where it could be argued that the parent set identified by the exhaustive search should be preferred.



(a)



(b)



(c)

Figure 3.5: Combining forward selection and backward elimination. **(a)** The percentage of edges correctly identified by the combined FS and BE algorithm. **(b)** The number of missing (left) and extra (right) edges across all subjects. **(c)** The number of missing and extra edges across nodes for each subject.

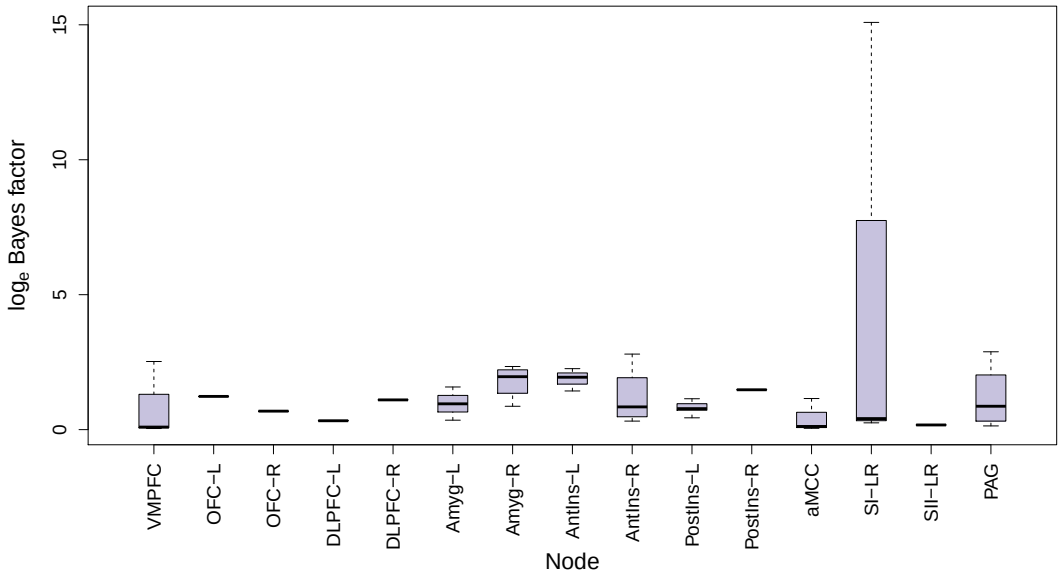


Figure 3.6: Log$_e$ Bayes factor comparison between the parent sets discovered by exhaustive and combined forward selection and backward elimination searches. Plotted values are for the case when the stepwise algorithm fails to find the set of parents identified by an exhaustive search. Whiskers show the minimum and maximum values. Interestingly, the noticeable difference for the SI-LR results from a single region (the AntIns-L) being present in the stepwise parent set but absent in the exhaustive parent set.

3.4 Accuracy of Stepwise Methods for Increasing Numbers of Nodes

In this section, we explore whether we might expect the accuracy of forward selection and backward elimination to be maintained, or compromised, as the number of nodes increases. As it is not feasible to run an exhaustive search on large networks, we consider smaller networks, specifically subnetworks of the 15 node 'safe' dataset with 6, 8, 10 and 12 nodes (as detailed in Table 3.4). We compared the performance of the stepwise approaches (without the termination instruction) on these smaller networks with the performance on the 15 node networks. Results are shown in Figures 3.7 - 3.9.

Subnetwork	Nodes
6	VMPFC, OFC-L, OFC-R, aMCC, DLPFC-L, DLPFC-R
8	Subnetwork 6, Amyg-L, Amyg-R
10	Subnetwork 8, AntIns-L, AntIns-R
12	Subnetwork 10, Post-Ins-L, Post-Ins-R

Table 3.4: Brain regions included in the subnetworks of the 15 node 'safe' dataset.

Figure 3.7 shows the accuracy for forward selection and backward elimination, both individually and combined, where, as in Figures 3.2a, 3.3a and 3.5a, accuracy is defined as the number of edges correctly identified as present or absent when the stepwise

networks are compared with the networks found in an exhaustive search. There is a small but noticeable reduction in accuracy as the number of nodes increases: in the 8 node subnetwork, the combined forward selection and backward elimination method reproduced the exhaustive networks with 100 % accuracy for all 32 subjects whereas for the 10 node subnetworks, there were 3 subjects with 2 or 3 incorrect edges. For 12 nodes, the median is still 100 %, falling to 98.6 % for the 15 node network. While this reduction in accuracy is small, it emphasises that we should not assume the accuracies observed with small numbers of nodes will necessarily be maintained for larger networks. As the size of the model space increases exponentially, the number of 'incorrect' models also increases exponentially. However, as previously discussed, assessing 'correctness' in terms of a single model (the model with the highest Log Predictive Likelihood) may not be the most appropriate method. As we showed in Figures 3.4 and 3.6, the reduction in accuracy can be mitigated by relaxing our definition of correctness to allow models where there is insufficient evidence for a difference i.e. where $\mathrm{LPL}[\hat{P}a(r)] - \mathrm{LPL}[\hat{P}a(r)_{step}] < 1$.

Figure 3.8 shows that for increasing numbers of nodes, the $\log_e$ Bayes factor when comparing the 'best' model to the second best (i.e. the parent set with the second highest LPL) tends towards lower values. As the size of the model space increases, we might expect the number of models with equivalent evidence to increase also. This is confirmed in Figure 3.9a, which shows the number of 'equivalent' models using $\log_e\mathrm{BF} < 1$ and $\log_e\mathrm{BF} < 3$. It is interesting to compare Figure 3.9a with Figure 3.9b, which shows the number of models with equivalent evidence as a percentage of the total number of models (the size of the model space). As the number of nodes increases, the number of equivalent models increases but *decreases* as a percentage of the size of the model space. This suggests that, within a very large model space, there will be a relatively small number of models which provide a good fit to the data. Further discussion will be provided in Chapter 5.

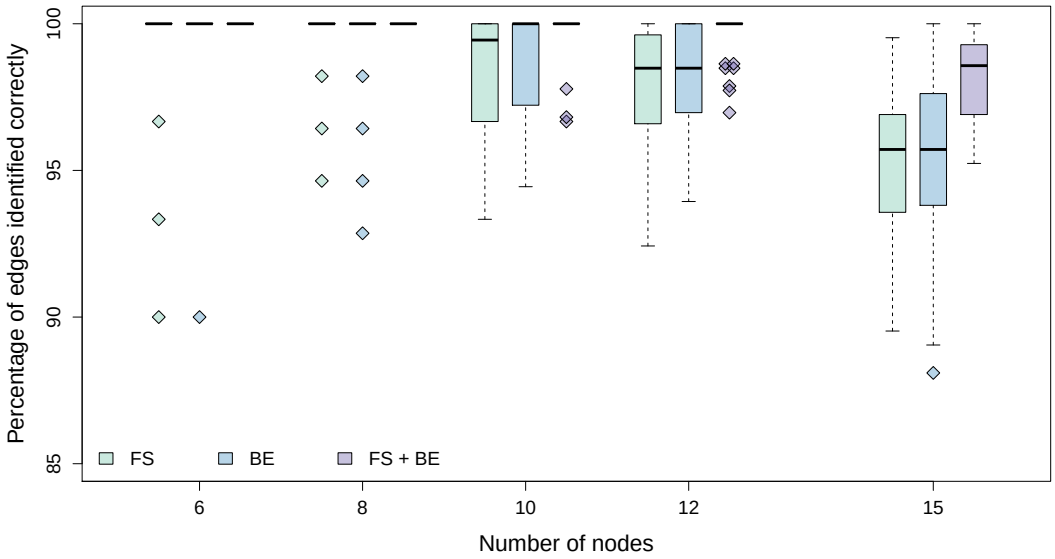


Figure 3.7: The accuracy of the stepwise approaches decreases as the number of nodes increases. Boxes show the accuracy of forward selection, backward elimination and combined forward selection, backward elimination for subnetworks of the 15 node resting-state ('safe') data. Accuracy is expressed as the percentage of edges correctly identified over the whole network for each subject.

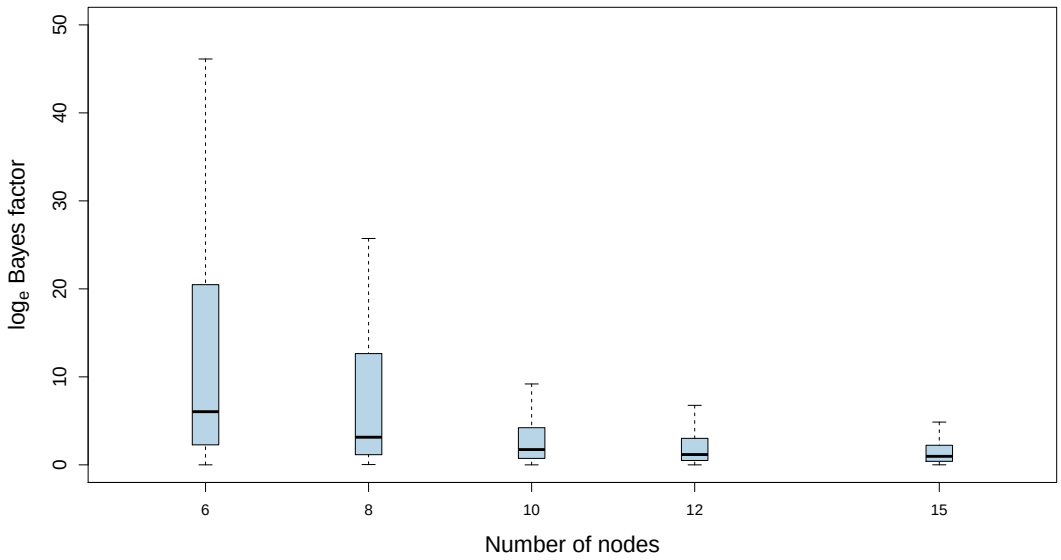
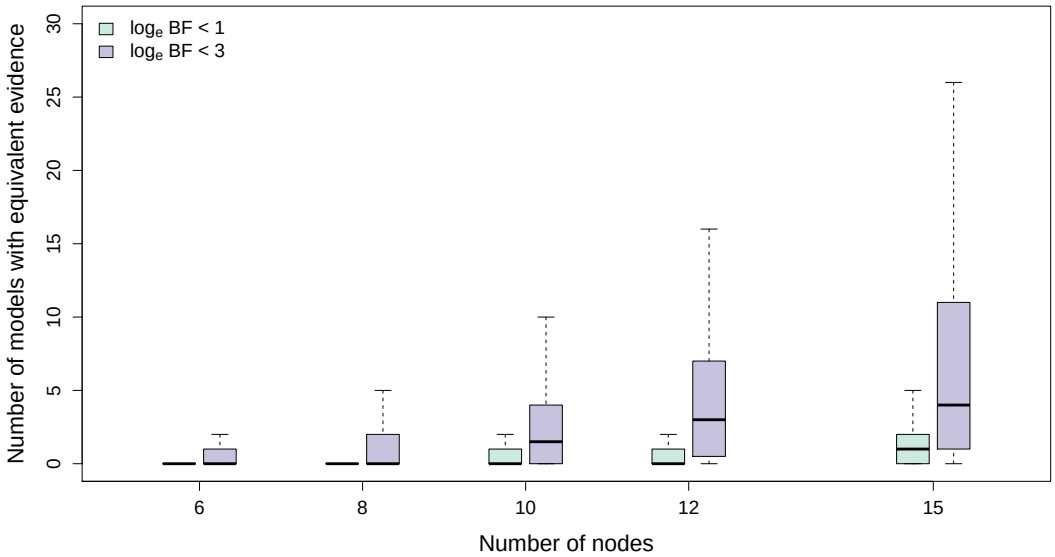
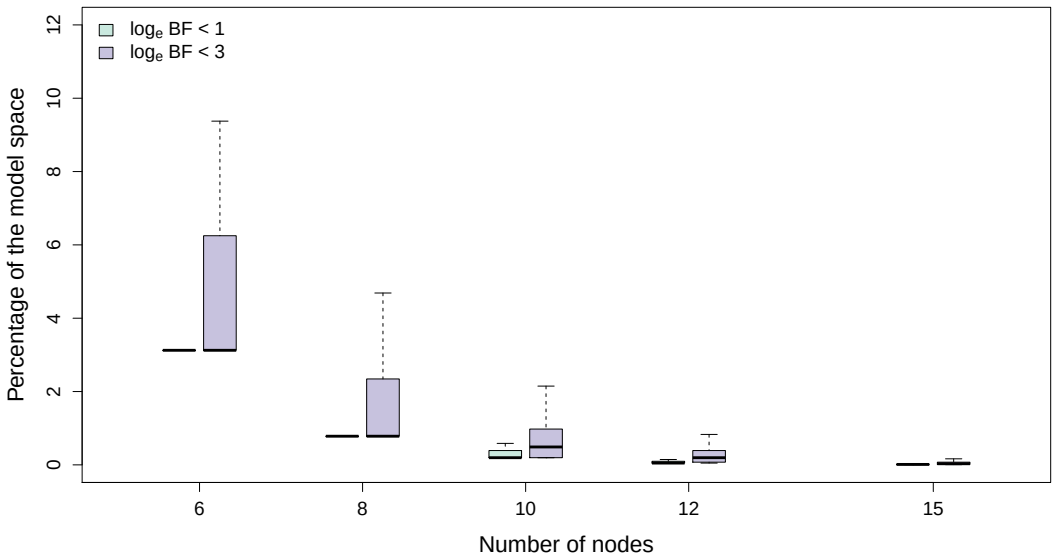


Figure 3.8: The $\log_e$ Bayes factor for the highest scoring vs. the next highest scoring model decreases as the number of nodes increases. Boxes show the $\log_e$ Bayes factor comparing the highest scoring with the second highest across all subjects and nodes for subnetworks of the 15 node resting-state ('safe') data. For easier visualisation, outliers are not shown.



(a)



(b)

Figure 3.9: The number of models with equivalent evidence increases as the number of nodes increases, but decreases as a percentage of the model space. (a) For subnetworks of the 15 node resting-state ('safe') data, we found the number of models with a $\log_e$ Bayes factor of less than 1 (green) and less than 3 (purple) compared to the highest scoring model, across all subjects and nodes. **(b)** As **(a)** but expressed as a percentage of the size of the model space. For easier visualisation, outliers are not shown.

3.5 Discussion

In this chapter, we have exploited the fact that the Log Predictive Likelihood of the MDM-DGM factors by node in order to replace an exhaustive model search with a stepwise one. For a resting-state network with 15 nodes, stepwise regression methods can successfully reproduce the networks estimated by an exhaustive search over the

model space. We have shown that forward selection and backward elimination may be used in isolation, but combining the results of both allows for greater accuracy, as much as 100 % in some cases. The reduction in the size of the model space is so dramatic that these methods may readily be combined without compromising the massive reduction in computation time that they offer.

While these methods may readily be applied to networks of with 50 or even hundreds of nodes, it should be noted that there is no guarantee the performance of the stepwise algorithms would be replicated on larger networks. As mentioned in Chapter 1, section 1.5, for networks with numbers of nodes close to the number of time points require a regularisation term to ensure the stability of the partial correlation matrix. As we showed in the previous chapter, the MDM-DGM tends to detect multiple edges with low connectivity strengths which may be potentially spurious. The following chapter considers how we might introduce a penalty term to ensure more robust edge detection, while potential extensions to the model selection algorithms will be discussed in Chapter 5. For the moment, we conclude that stepwise methods allow fast reconstruction of small networks, and, if interpreted with due caution, may allow the MDM-DGM to be applied to much larger networks than have previously been feasible.

Chapter 4

Dynamic Linear Models with Non-Local Priors

4.1 Motivation

As shown in Chapter 2 (see Figure 2.12), when fitting the MDM-DGM to individual subject data, there is a strong correspondence between the consistency of an edge over subjects and the magnitude of $\hat{\boldsymbol{\theta}}_t(r)$. Returning to the 15 node resting-state ('safe') dataset, this behaviour is illustrated in Figure 4.1, which shows the location parameter $\boldsymbol{\mu}_t(r)$ (see equation 1.4.14a) over all subjects and all time for the parents of the anterior mid-cingulate cortex (aMCC). The numbers at the bottom show the proportion of subjects that had each parent. The parent set of the group model $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$ contained 10 parents but in the individual networks only half of these parents were shared by more than 59 % of subjects (the threshold for significance defined by the Binomial test method, see Chapter 2, section 2.4.1). From Figure 4.1, the secondary somatosensory cortex and the anterior insula, which occur in a higher proportion of subjects, have noticeably higher values of $\mu_{it}(r)$. As previously discussed, we might expect that the consistent edges with high connectivity strengths are more likely to represent genuine, physiologically-interesting relationships, whereas the more inconsistent edges may potentially be spurious. Therefore in order to improve the robustness of the MDM-DGM networks, it may be desirable to impose a penalty on the Log Predictive Likelihood to reduce the model evidence for unnecessarily complex models. We use non-local priors for penalised model selection within a Bayesian framework. As we will show, non-local priors penalise models with regression coefficients that are near zero at individual time points, or consistently near zero over all time. We focus on extending the work of Johnson and Rossell (2012) and Rossell and Telesca (2017), developed for Bayesian linear models, to the Dynamic Linear Model. One major advantage of non-local priors (in the form we introduce here) is that we retain closed-form expressions for the model evidence. As we will show in this chapter, the 'dynamics' of the Dynamic Linear Model present some methodological challenges. However, we argue that non-local priors provide a novel and flexible extension to the DLM with the potential to improve the stability of the MDM-DGM networks.

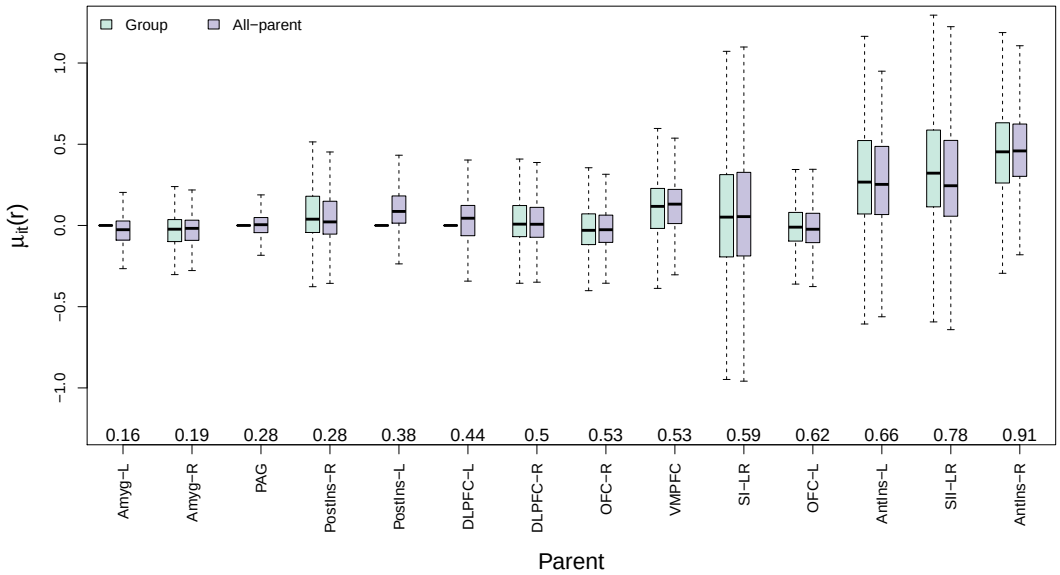


Figure 4.1: Correspondence between the consistency of edge presence and connectivity strength for the aMCC. Boxes show $\mu_{it}(r)$ over all subjects and all time for the parents of the aMCC when the group model (green) and the all-parent model (purple) were fitted (flat lines represent parents absent in the group network $\hat{\mathcal{M}}_{\mathcal{G}_{safe}}$. For easier visualisation, outliers are not shown). Numbers at the bottom show the proportion of subjects that were found to have this parent in the individual networks.

4.2 Introduction to Non-Local Priors

Consider a simple Bayesian linear model with a single regression coefficient θ. There are two candidate models

$$\mathcal{M}_0 : \theta = 0$$

$$\mathcal{M}_1 : \theta \neq 0.$$

Let the prior on θ follow a normal distribution with mean $\mu = 0$ and variance σ^2 so that we may write

$$p(\theta \,|\, \mathcal{M}_1) \sim \mathcal{N}(0, \sigma^2).$$

As the normal distribution is symmetric and centred around zero, it is straightforward to see that this prior assigns the highest probability to the values of θ that are closest to zero. Imagine replacing this prior with a prior distributed as

$$p(\theta) \sim Z\,\mathcal{N}(0, \sigma^2)\,\theta^{2h}$$

where Z is a normalisation constant and h is an integer known as the *order* of the density (see section 4.3). A *non-local* probability density of this form is shown in Figure 4.2 (solid blue line, $h = 1$). This density is close to zero when θ is close to zero and assigns higher probability to values of θ that are not to close but not too far from

zero.

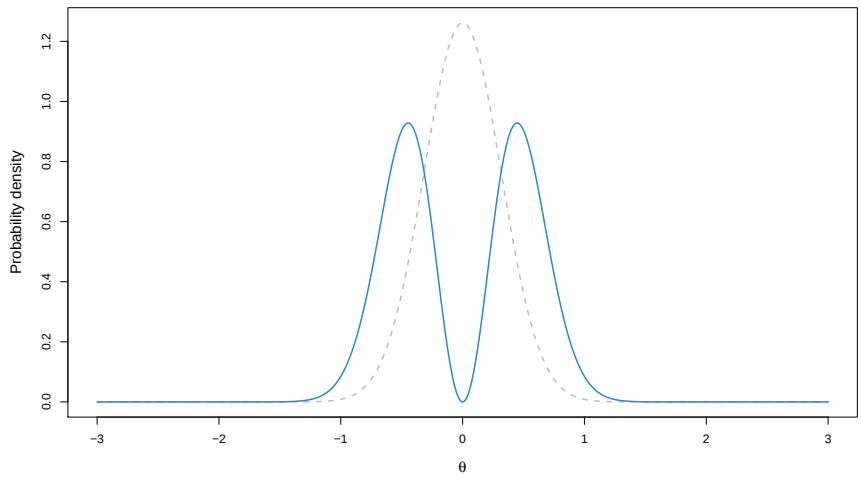


Figure 4.2: A univariate normally-distributed prior and its non-local equivalent.

This simple example illustrates the basic principle behind a non-local prior. A full discussion of the theoretical basis of non-local priors is provided in Rossell and Telesca (2017). Some of the key ideas, which we draw upon in order to incorporate a non-local prior into the Dynamic Linear Model, are described below.

Let $\mathcal{M}_j$ denote some candidate model with parameters of interest $\boldsymbol{\theta}^{(j)} \in \Theta^{(j)} \subseteq \Theta$ and let $\phi^{(j)}$ be a fixed-dimension nuisance parameter. Define $\mathcal{M}_0$ to be a submodel of $\mathcal{M}_j$ with parameters $\boldsymbol{\theta}^{(0)}$. If we can specify a function $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]$ in such a way that $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}] \to 0$ as $\boldsymbol{\theta}^{(j)} \to \boldsymbol{\theta}^{(0)}$, then any non-local prior is proportional to a local prior $p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j]$ multiplied by this function $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]$. We use the superscript NL to indicate a non-local prior, writing

$$p^{\mathrm{NL}}[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j] \propto f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]\, p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j].$$

This representation is always possible because

$$\begin{aligned} p^{\mathrm{NL}}[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j] &= \frac{p^{\mathrm{NL}}[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j]}{p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j]}\, p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j] \\ &= f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]\, p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j]. \end{aligned}$$

Assume that $p^{\mathrm{NL}}[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathcal{M}_j]$ is proper (we will discuss normalisation in the following sections). Denote the model evidence under a local prior by $m_j(\mathbf{y})$, i.e. $m_j(\mathbf{y}) = p(\mathbf{y} \,|\, \mathcal{M}_j)$. The model evidence under a non-local prior is

$$m_j^{\mathrm{NL}}(\mathbf{y}) = m_j(\mathbf{y}) \int_{\phi} \int_{\boldsymbol{\theta}} f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]\, p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathbf{y}]\, d\boldsymbol{\theta}\, d\phi \tag{4.2.1}$$

that is, it is equal to the model evidence under a local prior multiplied by the expec-

tation of the function $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]$ with respect to the posterior distribution of the model parameters (Rossell and Telesca, 2017).

4.2.1 The Bayes Factor under a Non-Local Prior

Following Rossell and Telesca (2017), write

$$g_j(\mathbf{y}) = \int_{\phi} \int_{\boldsymbol{\theta}} f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]\, p[\boldsymbol{\theta}^{(j)}, \phi^{(j)} \,|\, \mathbf{y}]\, d\boldsymbol{\theta}\, d\phi.$$

Assuming a uniform prior on the model probabilities $p(\mathcal{M}_i)$ and $p(\mathcal{M}_j)$, the Bayes factor for models $\mathcal{M}_i$ and $\mathcal{M}_j$ under a non-local prior is

$$\mathrm{BF}_{ij} = \frac{m_i(\mathbf{y})\, g_i(\mathbf{y})}{m_j(\mathbf{y})\, g_j(\mathbf{y})}$$

so we may write

$$\log_e \mathrm{BF}_{ij} = \log_e[m_i(\mathbf{y})] - \log_e[m_j(\mathbf{y})] + \log_e[g_i(\mathbf{y})] - \log_e[g_j(\mathbf{y})].$$

Imagine a simple example with two regressors and a single observation. Suppose we have

$$Y = \theta^{(1)} + F_2\, \theta^{(2)}$$

where $\theta^{(1)}$ represents an intercept (which should not be penalised under a non-local prior) and $\theta^{(2)}$ is the regression coefficient of interest. Our two candidate models are therefore

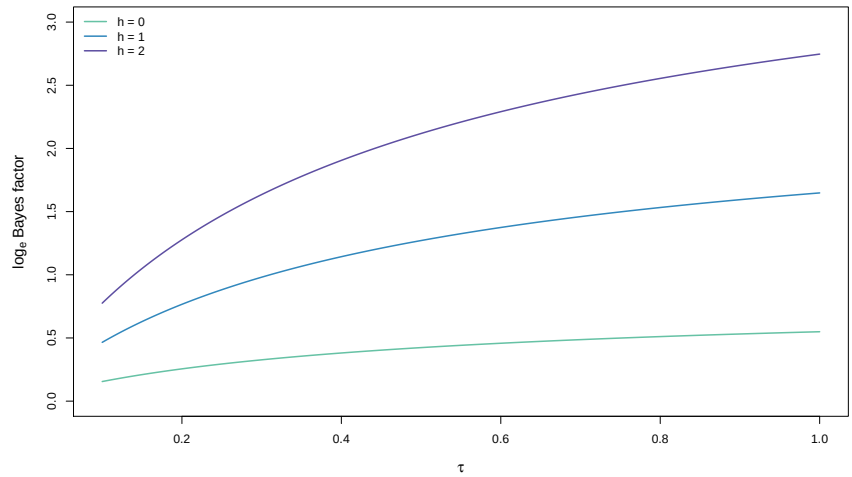
$$\mathcal{M}_0 : \theta^{(2)} = 0$$

$$\mathcal{M}_1 : \theta^{(2)} \neq 0.$$

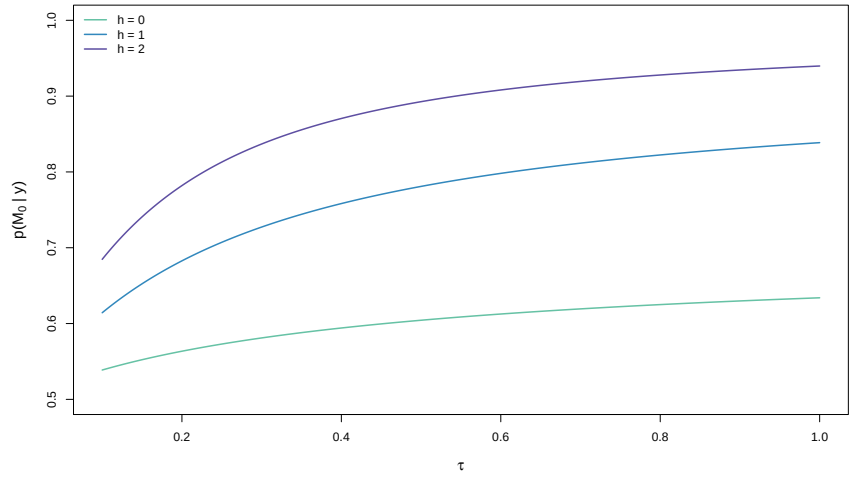
Let the model evidence for the intercept-only model be the same under the local and non-local priors. The $\log_e$ Bayes factor is then

$$\log_e \mathrm{BF}_{01} = \log_e[m_0(\mathbf{y})] - \log_e[m_1(\mathbf{y})] - \log_e[g_1(\mathbf{y})]$$

and it follows that low values of $g_1(\mathbf{y})$ $(0 < g_1(\mathbf{y}) < 1)$ will reduce the evidence for $\mathcal{M}_1$. This behaviour is illustrated in Figure 4.3, which shows the how the evidence for the sparser model increases under a non-local prior.



(a)



(b)

Figure 4.3: Illustration of the influence of a non-local prior on the log$_e$ Bayes factor and posterior model probabilities. We performed a Bayesian regression with a single time point using $Y = 0$, $\mathbf{F}^\top = \{1, 2\}$, $\phi^{-1} = 1$ and $\mathbf{R}^* = \tau \mathbb{I}_2$ for local priors ($h = 0$) and non-local priors ($h = 1$, $h = 2$) with varying τ. **(a)** The log$_e$ Bayes factor is the evidence for model $\mathcal{M}_0$ over model $\mathcal{M}_1$. **(b)** The posterior model probabilities for $\mathcal{M}_0$. Higher values of the dispersion parameter τ and the order of the density h increase the evidence for the sparser model.

4.3 Candidate Non-Local Priors

In this section, we consider potential forms of the function $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]$. We start by detailing the *product moment* (pMOM) non-local prior, as applied to a Bayesian linear model by Johnson and Rossell (2012). Given this framework, we show how it may be extended and modified so that we may apply similar non-local priors to the Dynamic Linear Model. While we focus on product moment non-local priors, it should be noted that Rossell and Telesca (2017) have also developed product inverse moment and product exponential moment non-local prior formulations; the application of product inverse moment non-local priors for the Bayesian linear model is described in Johnson and Rossell (2012).

4.3.1 Product Moment Non-Local Priors

Consider a Bayesian linear regression model with vector of regression coefficients $\boldsymbol{\theta}$ with dimension $p \times 1$. Let $\mathbf{Y}^\top = \{Y_1, \ldots, Y_T\}$ represent a random vector corresponding to a set of T observations $\mathbf{y}^\top = \{y_1, \ldots, y_T\}$. Define a matrix of real numbers $\mathbf{F}$ with dimension $T \times p$. In multiple linear regression, it is usual to denote $\mathbf{F}$ and $\boldsymbol{\theta}$ by $\mathbf{X}$ and $\boldsymbol{\beta}$ respectively. We use our notation for consistency with the notation of West and Harrison (1997) for Dynamic Linear Models and by Queen and Smith (1993) for Multiregression Dynamic Models. By adopting this notation, the parallels with the Dynamic Linear Model are clearer and the formulae for DLMs with non-local priors follow more naturally from the those presented in Johnson and Rossell (2012).

Bayesian linear regression is described in full in Chapter 9 of O'Hagan (2004). Observations are modelled using the linear relation

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \mathbf{v} \qquad v_t \sim \mathcal{N}\left(0, \phi^{-1}\right).$$

Conditioning on unknown, constant observation variance ϕ^{-1}, the observations are distributed as

$$p\left(\mathbf{Y} \,|\, \boldsymbol{\theta}, \phi\right) \sim \mathcal{N}_{\mathbf{Y}}\left(\mathbf{F}\boldsymbol{\theta}, \phi^{-1}\mathbb{I}_T\right).$$

The coefficient vector $\boldsymbol{\theta}$ is normally-distributed with $p \times 1$ mean vector $\mathbf{a}$ and $p \times p$ covariance matrix $\mathbf{R} = \phi^{-1}\mathbf{R}^*$ so that

$$p(\boldsymbol{\theta} \,|\, \phi) \sim \mathcal{N}_{\boldsymbol{\theta}}\left(\mathbf{a}, \phi^{-1}\mathbf{R}^*\right). \tag{4.3.1}$$

As with the Dynamic Linear Model, in a Bayesian linear regression model there is a gamma-distributed prior on the precision with shape $\frac{n_0}{2}$ and rate $\frac{d_0}{2}$:

$$p(\phi) \sim \mathcal{G}_\phi\left(\frac{n_0}{2}, \frac{d_0}{2}\right).$$

Returning to the non-local prior framework, let the vector of regression coefficients $\boldsymbol{\theta}$ be the parameters of interest and the observation variance ϕ^{-1} be a nuisance parameter. Suppose there are two candidate models $\mathcal{M}_j$ and $\mathcal{M}_k$ where $\mathcal{M}_k$ is nested in $\mathcal{M}_j$ and contains one less regressor θ_i such that $\boldsymbol{\theta}^{(j)} = \{\theta_1, \ldots, \theta_i, \ldots, \theta_p\}$ and $\boldsymbol{\theta}^{(k)} = \{\theta_1, \ldots, \theta_{(i-1)}, \theta_i = 0, \theta_{(i+1)} \ldots, \theta_p\}$. Model selection may therefore be framed as testing the following hypotheses

$$\mathcal{M}_j : \theta_i \neq 0$$

$$\mathcal{M}_k : \theta_i = 0.$$

Define a function

$$f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}] = \prod_{i=1}^{p} \frac{\theta_i^{2h}}{\phi^{-h}}.$$

where h is an integer. It is straightforward to see that $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}] \to 0$ as $\boldsymbol{\theta}^{(j)} \to \boldsymbol{\theta}^{(k)}$. As a non-local prior may be expressed as the product of a local density and the function $f^{(j)}[\boldsymbol{\theta}^{(j)}, \phi^{(j)}]$, we may write

$$p^{\mathrm{NL}}(\boldsymbol{\theta} \,|\, \phi, h) \propto p(\boldsymbol{\theta} \,|\, \phi) \prod_{i=1}^{p} \frac{\theta_i^{2h}}{\phi^{-h}}$$

If we introduce some normalisation constant Z and dispersion parameter $\tau > 0$, this is the product moment non-local prior described in Johnson and Rossell (2012) and Rossell and Telesca (2017):

$$p^{\mathrm{NL}}(\boldsymbol{\theta} \,|\, \phi, \tau, h) \sim Z\, \mathcal{N}_{\boldsymbol{\theta}}\big(\mathbf{0}, \phi^{-1}\tau\, \mathbf{R}^*\big) \prod_{i=1}^{p} \frac{\theta_i^{2h}}{(\phi^{-1}\tau)^h}.$$

The behaviour of this prior (in the univariate case) is illustrated in Figure 4.4. Figures 4.4a and 4.4b show how the strength of the penalty is influenced by the two parameters, dispersion parameter τ and the order of the density h. The effect of these parameters is to increase or decrease the width of the window around zero. The $(\phi^{-1}\tau)^h$ term in the denominator allows the normalisation constant Z to be calculated independent of ϕ and τ. As shown in Appendix 4.A, if $\boldsymbol{\gamma} = \phi^{\frac{1}{2}}\tau^{-\frac{1}{2}}\boldsymbol{\theta}$, the normalisation constant is

$$\frac{1}{Z} = \int_{\boldsymbol{\gamma}} \mathcal{N}_{\boldsymbol{\gamma}}(\mathbf{0}, \mathbf{R}^*) \prod_{i=1}^{p} \gamma_i^{2h}\, d\boldsymbol{\gamma}.$$

4.3.2 DLM-pMOM Non-Local Priors

In the Dynamic Linear Model, the vector of regression coefficients $\boldsymbol{\theta}$ is replaced by $p_r \times 1$ the state vector $\boldsymbol{\theta}_t(r)$ for node r at time t. Imagine a Bayesian linear model as outlined above with $T = 1$. Replace equation 4.3.1 with the prior of the Dynamic Linear Model at some time t

$$p[\boldsymbol{\theta}_t(r) \,|\, \phi(r), D_{t-1}] \sim \mathcal{N}_{\boldsymbol{\theta}_t(r)}\big[\mathbf{a}_t(r), \phi(r)^{-1}\mathbf{R}_t^*(r)\big]. \tag{4.3.2}$$

It should be noted that this 'prior' may depend on the previous observations, i.e. it is conditioned on D_{t-1}.

Consider replacing equation 4.3.2 with a pMOM prior of the form

$$p^{\mathrm{NL}}[\boldsymbol{\theta}_t(r) \,|\, \phi(r), h_r, D_{t-1}] = Z_t(r)\, \mathcal{N}_{\boldsymbol{\theta}_t(r)}\big[\mathbf{a}_t(r), \phi(r)^{-1}\mathbf{R}_t^*(r)\big] \prod_{i \neq j}^{p_r} \theta_i(r)^{2h_r} \tag{4.3.3}$$

where h_r is the order of the density. There are a number of subtle but important differences between this non-local prior and the one described by Johnson and Rossell

(2012) for the Bayesian linear model. We make the distinction clear by referring to our application as a *Dynamic Linear Model Product Moment* (DLM-pMOM) prior. As we wish to apply this prior at each individual time point, we cannot assume that the mean $\mathbf{a}_t(r)$ is zero for all t. The effect this has on the pMOM-like prior is illustrated in Figure 4.4c. The (time-dependent) normalisation constant $Z_t(r)$ will therefore depend on $\phi(r)$. For this reason, we drop the scaling by $\phi(r)^{-h_r}$ in the denominator of the DLM-pMOM prior. However, it is still possible to obtain a closed-form expression in terms of the DLM one-step distributions (see Appendix 4.A). We obtain

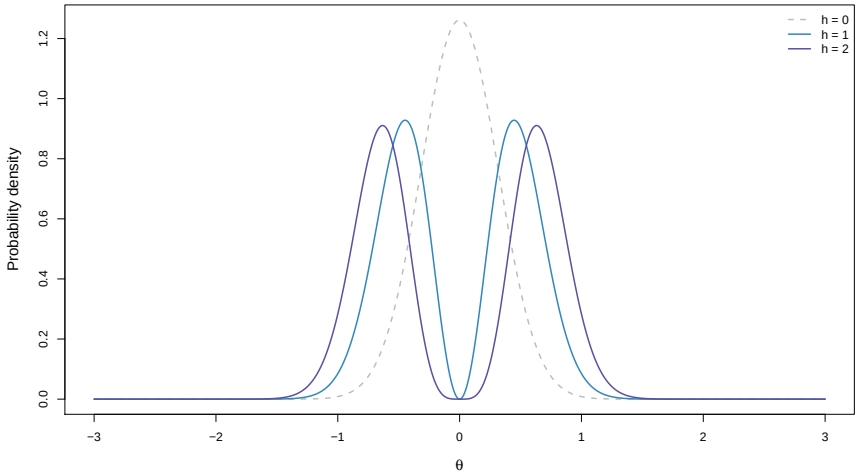
$$\frac{1}{Z_t(r)} = \int_{\boldsymbol{\theta}_t(r)} \mathcal{T}_{n_{t-1}(r)}[\mathbf{a}_t(r), \mathbf{R}_t(r)] \prod_{i \neq j}^{p_r} \theta_i(r)^{2h_r} \, d\boldsymbol{\theta}_t(r).$$

As previously discussed, Johnson and Rossell (2012) define a dispersion parameter τ, which controls the strength of the penalty by scaling the covariance of the regression coefficients. However, recall that in the Dynamic Linear Model, the prior covariance is defined via

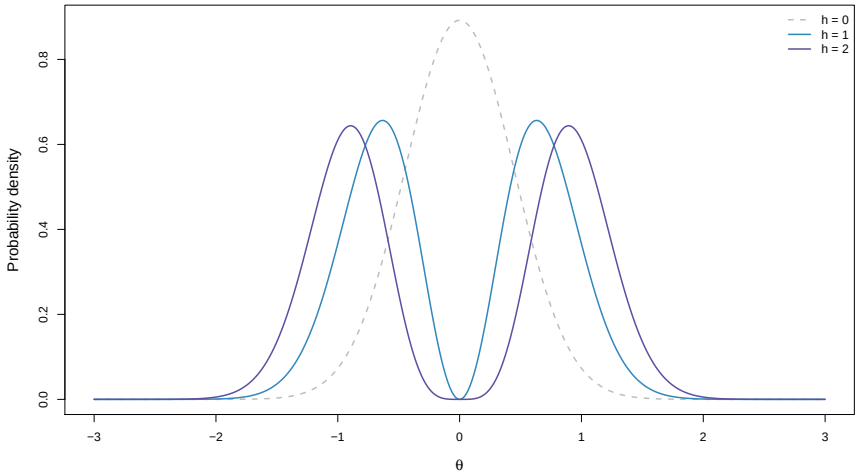
$$\mathbf{R}_t^*(r) = \frac{\mathbf{C}_{t-1}^*(r)}{\delta(r)}.$$

A dispersion parameter will therefore influence the system in the same way as the discount factor $\delta(r)$, affecting the width of the distribution of $\boldsymbol{\theta}_t(r)$ around its mean value. For this reason, we drop τ and note that lower values of $\delta(r)$ will increase the variance and consequently the strength of the penalty. If we fix $\delta(r)$, the only control we have over the strength of the penalty is the order of the density h_r. This behaviour will be explored further in section 4.5.

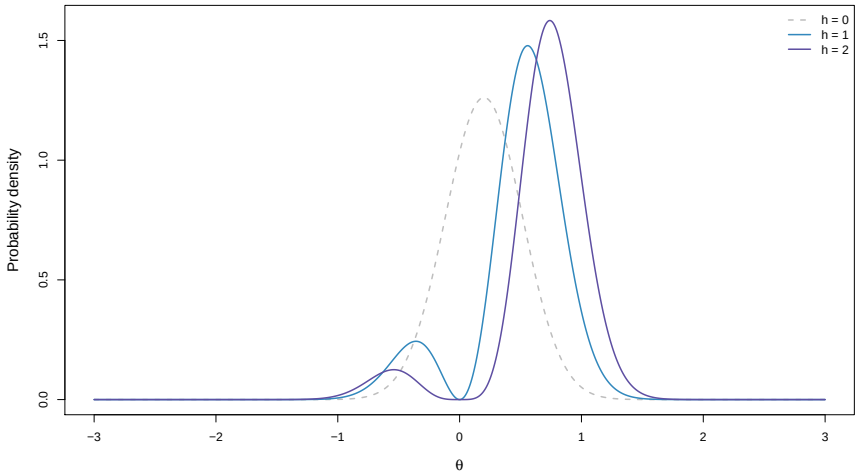
A final difference is that, when applying the Dynamic Linear Model to fMRI time series, it is desirable to include an intercept term. We do not wish there to be any penalty associated with this intercept term. Denote the intercept by the subscript j such that $\boldsymbol{\theta}_t(r) = \{\theta_{1t}(r), \ldots, \theta_{jt}(r), \ldots, \theta_{pt}(r)\}$ where $\theta_{it}(r)$ corresponds to a parent of node r at time t for all $i \neq j$. It is then straightforward to exclude the intercept from the product term. Calculation of the normalisation constant $Z_t(r)$ is then with respect to the marginal distribution $p[\boldsymbol{\theta}_{i \neq j,t}(r) \,|\, \phi(r), D_{t-1}]$.



(a) $\mu = 0$, $\sigma^2 = 0.1$, $\tau = 1$



(b) $\mu = 0$, $\sigma^2 = 0.1$, $\tau = 2$



(c) $\mu = 0.2$, $\sigma^2 = 0.1$, $\tau = 1$

Figure 4.4: Examples of a univariate product moment non-local prior. Each non-local prior density is proportional to a product a normal density with mean μ and variance σ^2 (grey, dashed line) and $f[\theta] = \theta^{2h}$. The parameter h is the order of the density (solid blue line $h = 1$, solid purple line $h = 2$). Both h and dispersion parameter τ influence the strength of the penalty via the width of the window around zero, as shown in **(a)** and **(b)**. If μ is non-zero, the non-local prior distribution loses its symmetry, as shown in **(c)**.

Again consider two models $\mathcal{M}_j(r)$ and $\mathcal{M}_k(r)$ where $\mathcal{M}_k(r)$ is nested in $\mathcal{M}_j(r)$. Because the regression coefficients now vary over time, model selection may be interpreted as a hypothesis test such that

$$\mathcal{M}_j(r) : \theta_{it}(r) \neq 0 \quad \text{for some } t$$

$$\mathcal{M}_k(r) : \theta_{it}(r) = 0 \quad \forall t.$$

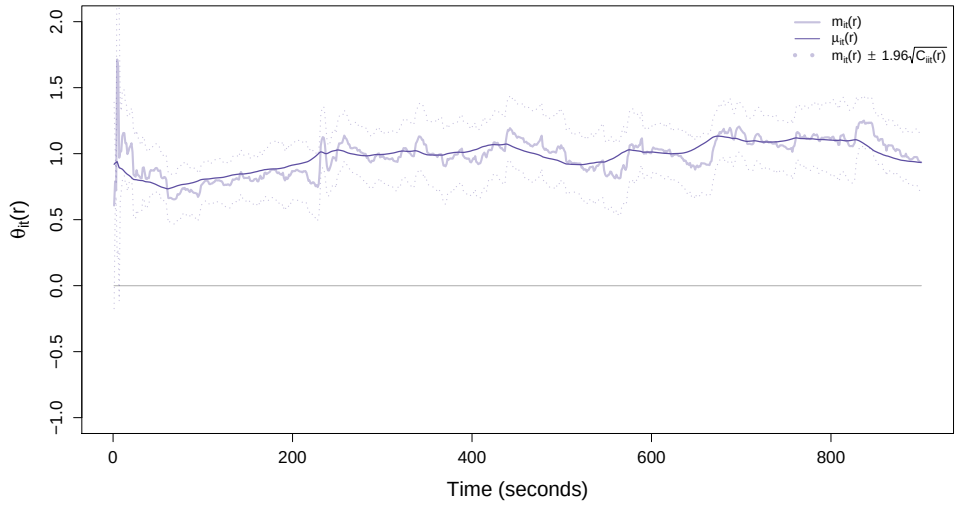
We are interested in models where a particular regressor (the connectivity from a particular brain region) is non-zero at least one time point. However, a DLM-pMOM non-local prior will penalise the case $\theta_{it}(r) = 0$ for any t. The hypothesis test corresponding to model selection will become

$$\mathcal{M}_j(r) : \theta_{it}(r) \neq 0 \quad \forall t$$

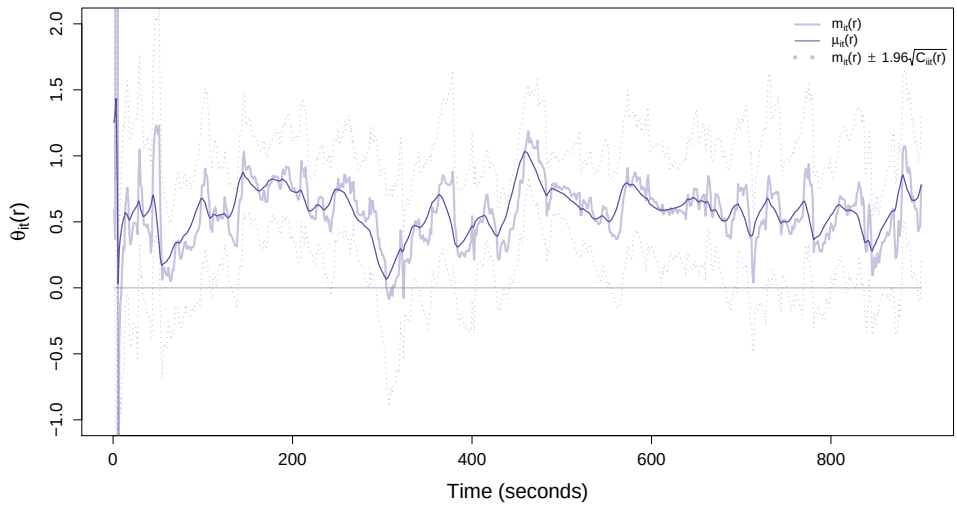
$$\mathcal{M}_k(r) : \theta_{it}(r) = 0 \quad \forall t.$$

This is potentially problematic in the Dynamic Linear Model, where $\theta_{it}(r)$ may be far from zero at some time points and zero, or close to zero, at other time points, particularly for lower values of the discount factor $\delta(r)$. Imagine a time series where some brain region i influences another region r for some period of time after which its influence becomes close to zero. Alternatively, imagine a change point (or multiple change points) where the connectivity switches from high positive to high negative values. Both of these cases may represent underlying physiological changes that it would be desirable to detect and these models may be unduly penalised using a DLM-pMOM prior.

This is illustrated in Figure 4.5, which shows the one-step and retrospective estimates ($m_{it}(r)$ and $\mu_{it}(r)$ respectively) for the edge DLPFC-L → OFC-L in a model with parents VMPFC, DLPFC-L and Amyg-L. For both subjects, the estimates are consistently above zero. However, for the subject in Figure 4.5b, with $\delta(r) = 0.9$, the estimate falls through zero at around 300 seconds.



(a) $\delta(r) = 0.97$



(b) $\delta(r) = 0.9$

Figure 4.5: The time-varying regression coefficient for the edge DLPFC-L → OFC-L in the model with parent set VMPFC, DLFPC-L and Amyg-L. The one-step and retrospective estimates are denoted by $m_{it}(r)$ and $\mu_{it}(r)$ respectively. **(a)** For a subject with a discount factor $\delta(r)$ that is close to one, the dynamic regression coefficient is consistently above zero. **(b)** For a subject with a lower value of the discount factor $\delta(r)$, the dynamic regression coefficient takes both high positive and near zero values.

4.3.3 DLM-Quadratic Form Non-Local Priors

To address the potential limitations of the DLM-pMOM non-local prior, consider a penalty function

$$f[\boldsymbol{\theta}^{(j)}(r), \phi^{(j)}(r)] = \prod_{i \neq j}^{p_r} \frac{\sum_{t=1}^{T} \theta_{it}^{2h_r}}{[\phi(r)^{-1}]^{h_r}}.$$

It follows that this function will only be equal to zero if $\theta_{it}(r) = 0$ for all time t. Restricting our attention to the case where $h_r = 1$, we can specify a non-local prior of the form

$$p^{\mathrm{NL}}[\boldsymbol{\theta}(r) \,|\, \phi(r), D_0] \propto p[\boldsymbol{\theta}(r) \,|\, \phi(r), D_0] \prod_{i \neq j}^{p_r} \frac{\boldsymbol{\theta}_i(r)^{\top} \boldsymbol{\theta}_i(r)}{[\phi(r)^{-1}]^{h_r}}.$$

where $\boldsymbol{\theta}_i(r)^{\top} = \{\theta_{i1}(r), \ldots, \theta_{it}(r), \ldots, \theta_{iT}(r)\}$. Rather than being defined in terms of the one-step prior distributions of the Dynamic Linear Model, this non-local prior is defined in terms of the joint distribution of the state variables over all time $\boldsymbol{\theta}(r) = \{\boldsymbol{\theta}_1(r), \ldots, \boldsymbol{\theta}_t(r), \ldots, \boldsymbol{\theta}_T(r)\}$. This is a vector with dimension $p_r T \times 1$. As will be shown in section 4.4.2, the local distribution is multivariate normal with $p_r T \times 1$ mean vector $\underline{\mathbf{a}}(r)$ and $p_r T \times p_r T$ covariance matrix $\underline{\mathbf{R}}(r) = \phi(r)^{-1} \underline{\mathbf{R}}^{*}(r)$. We introduce the underscore notation to denote parameters of the Dynamic Linear Model joint distributions. We have

$$p^{\mathrm{NL}}[\boldsymbol{\theta}(r) \,|\, \phi(r), D_0] \propto \mathcal{N}_{\boldsymbol{\theta}(r)}\big[\mathbf{0}, \phi(r)^{-1} \underline{\mathbf{R}}^{*}(r)\big] \prod_{i \neq j}^{p_r} \frac{\boldsymbol{\theta}_i(r)^{\top} \boldsymbol{\theta}_i(r)}{\phi(r)^{-1}}.$$

If this prior has mean $\underline{\mathbf{a}}(r) = \mathbf{0}$, the normalisation constant $Z_{QF}(r)$ may be obtained without knowledge of $\phi(r)$ via

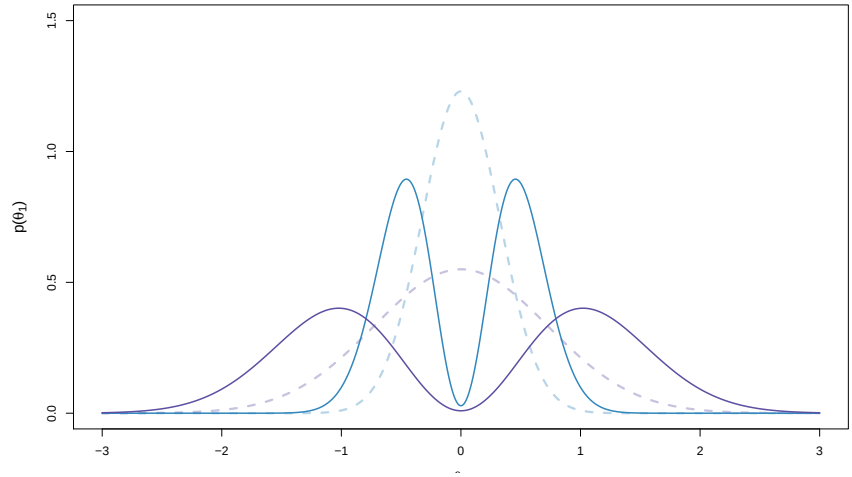
$$\frac{1}{Z_{QF}(r)} = \int_{\boldsymbol{\gamma}(r)} \mathcal{N}_{\boldsymbol{\gamma}(r)}[\mathbf{0}, \underline{\mathbf{R}}^{*}(r)] \prod_{i \neq j}^{p_r} \boldsymbol{\gamma}_i(r)^{\top} \boldsymbol{\gamma}_i(r) \, d\boldsymbol{\gamma}(r)$$

as shown in Appendix 4.A.

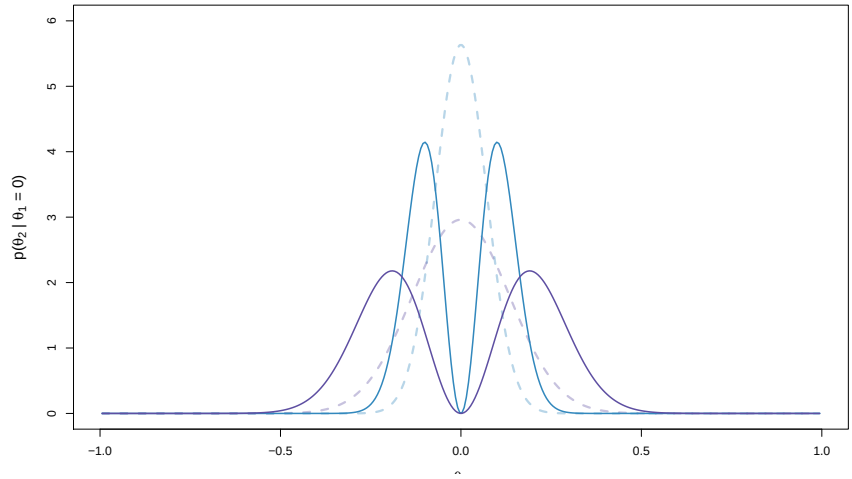
The final form of this non-local prior is therefore

$$p^{\mathrm{NL}}[\boldsymbol{\theta}(r) \,|\, \phi(r), D_0] = Z_{QF}(r) \, \mathcal{N}_{\boldsymbol{\theta}(r)}\big[\mathbf{0}, \phi(r)^{-1} \underline{\mathbf{R}}^{*}(r)\big] \prod_{i \neq j}^{p_r} \frac{\boldsymbol{\theta}_i(r)^{\top} \boldsymbol{\theta}_i(r)}{\phi(r)^{-1}}. \tag{4.3.4}$$

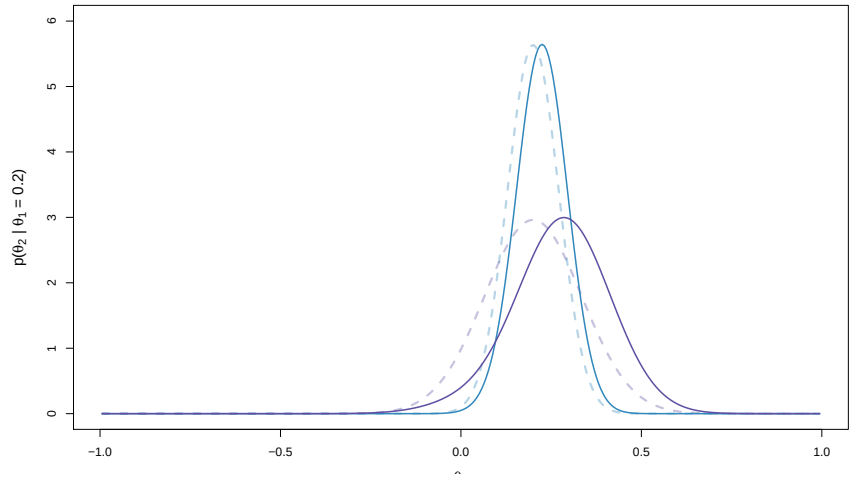
We call the density described by equation 4.3.4 a *Dynamic Linear Model Quadratic Form* (DLM-QF) non-local prior. The behaviour of this type of prior in the simplest case is illustrated in Figure 4.6.



(a)



(b)



(c)

Figure 4.6: Example of a DLM-QF non-local prior. Imagine a simple model with one regressor and two time points, θ_1 and θ_2. The prior covariance matrix was constructed using $F_1 = F_2 = 1$, $\delta = 0.95$ and $\phi = 1$. The marginal $p(\theta_1)$ is shown in **(a)** for $C_0^* = 0.1$ (blue) and $C_0^* = 0.5$ (purple). Dotted and solid lines indicate local and non-local priors respectively. **(b)** and **(c)** The conditional distributions $p(\theta_2 \,|\, \theta_1 = 0)$ and $p(\theta_2 \,|\, \theta_1 = 0.2)$ (using the same parameter values as in **(a)**). Note that if $\theta_1 = 0$, the non-local prior distribution is that of a pMOM prior.

4.3.4 The Dynamic Linear Model Joint Distributions

The additive non-local prior outlined in this section is defined in terms of the joint prior distribution, i.e. the distribution over all time before any observation is made. We have the state vector $\boldsymbol{\theta}(r)^\top = \{\boldsymbol{\theta}_1(r)^\top, \ldots, \boldsymbol{\theta}_T(r)^\top\}$ where each $\boldsymbol{\theta}_t(r)$ is a $p_r \times 1$ vector, such that $\boldsymbol{\theta}(r)$ has dimension $p_r T \times 1$. Using the Dynamic Linear Model equations, the joint distributions $p[\boldsymbol{\theta}(r) \,|\, \phi(r), D_0]$, $p[\mathbf{y}(r), \boldsymbol{\theta}(r) \,|\, \phi(r), D_0]$ and $p[\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r), \phi(r), D_0]$ may be constructed. Full derivations are provided in Appendix 4.B. The key results are summarised here.

The joint prior distribution is multivariate normal

$$p[\boldsymbol{\theta}(r) \,|\, \phi(r), D_0] \sim \mathcal{N}\big[\underline{\mathbf{a}}(r), \phi^{-1}(r)\underline{\mathbf{R}}^*(r)\big]$$

where $\underline{\mathbf{a}}(r)$ is the $p_r \times 1$ vector

$$\underline{\mathbf{a}}(r) = \mathbb{E}\left[\boldsymbol{\theta}(r)\right] = \begin{pmatrix} \mathbf{m}_0(r) \\ \vdots \\ \mathbf{m}_0(r) \end{pmatrix}$$

and $\underline{\mathbf{R}}^*(r)$ is the $p_r T \times p_r T$ matrix

$$\underline{\mathbf{R}}^*(r) = \begin{pmatrix} \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \cdots & \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) \\ \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \sum_{t=1}^2 \mathbf{W}_t^*(r) & \cdots & \mathbf{C}_0^*(r) + \sum_{t=1}^2 \mathbf{W}_t^*(r) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \sum_{t=1}^2 \mathbf{W}_t^*(r) & \cdots & \mathbf{C}_0^*(r) + \sum_{t=1}^T \mathbf{W}_t^*(r) \end{pmatrix}.$$

Recall that $\mathbf{m}_0(r)$ is $p_r \times 1$ vector and $\mathbf{C}_0^*(r)$ is a $p_r \times p_r$ matrix, and that these hyper-parameters must be specified *a priori*.

Note that $\mathbf{W}_t^*(r)$ implicitly depends on $\mathbf{F}_t(r)$, the discount factor $\delta(r)$ and the (scale-free) prior variance $\mathbf{C}_0^*(r)$ because of the recursive relation

$$\begin{aligned} \mathbf{W}_t^*(r) &= \frac{1-\delta(r)}{\delta(r)} \mathbf{C}_{t-1}^*(r) \\ &= \mathbf{R}_{t-1}^*(r) - \mathbf{R}_{t-1}^*(r)\mathbf{F}_{t-1}(r)[1 + \mathbf{F}_{t-1}(r)^\top \mathbf{R}_{t-1}^*(r)\mathbf{F}_{t-1}(r)]^{-1}\mathbf{F}_{t-1}(r)^\top \mathbf{R}_{t-1}^*(r) \end{aligned}$$

where

$$\mathbf{R}_{t-1}^*(r) = \frac{\mathbf{C}_{t-2}^*(r)}{\delta(r)}.$$

The joint distribution of $\mathbf{Y}(r)$ and $\boldsymbol{\theta}(r)$, conditional on $\phi(r)$ and D_0, is

$$\left(\begin{matrix} \mathbf{Y}(r) \\ \boldsymbol{\theta}(r) \end{matrix} \middle| \phi(r), D_0\right) \sim \mathcal{N}\left(\begin{pmatrix} \underline{\mathbf{F}}(r)^\top \underline{\mathbf{a}}(r) \\ \underline{\mathbf{a}}(r) \end{pmatrix}, \phi(r)^{-1} \begin{pmatrix} \underline{\mathbf{Q}}^*(r) & \underline{\mathbf{F}}(r)^\top \underline{\mathbf{R}}^*(r) \\ \underline{\mathbf{R}}^*(r)\underline{\mathbf{F}}(r) & \underline{\mathbf{R}}^*(r) \end{pmatrix}\right)$$

where $\underline{\mathbf{F}}(r)$ is a $p_r T \times T$ block diagonal matrix

$$\underline{\mathbf{F}}(r) = \begin{pmatrix} \mathbf{F}_1(r) & & & \\ & \mathbf{F}_2(r) & & \\ & & \ddots & \\ & & & \mathbf{F}_T(r) \end{pmatrix}$$

and $\underline{\mathbf{Q}}^*(r)$ is a $T \times T$ matrix

$$\underline{\mathbf{Q}}^*(r) = \underline{\mathbf{F}}(r)^\top \underline{\mathbf{R}}^*(r) \underline{\mathbf{F}}(r) + \mathbb{I}_T.$$

The joint posterior distribution is then multivariate Gaussian

$$p[\boldsymbol{\theta}(r) \,|\, \mathbf{Y}(r), \phi(r), D_0] \sim \mathcal{N}\big[\underline{\mathbf{m}}(r), \phi(r)^{-1} \underline{\mathbf{C}}^*(r)\big]$$

where $\underline{\mathbf{m}}(r)$ is $p_r T \times 1$ vector

$$\underline{\mathbf{m}}(r) = \underline{\mathbf{a}}(r) + \underline{\mathbf{R}}^*(r) \underline{\mathbf{F}}(r) \underline{\mathbf{Q}}^*(r)^{-1} \big[\mathbf{y}(r) - \underline{\mathbf{F}}(r)^\top \underline{\mathbf{a}}(r)\big]$$

and $\underline{\mathbf{C}}^*(r)$ is the $p_r T \times p_r T$ matrix

$$\underline{\mathbf{C}}^*(r) = \underline{\mathbf{R}}^*(r) - \underline{\mathbf{R}}^*(r) \underline{\mathbf{F}}(r) \underline{\mathbf{Q}}^*(r)^{-1} \underline{\mathbf{F}}(r)^\top \underline{\mathbf{R}}^*(r).$$

4.4 The Model Evidence under a Non-Local Prior

Recall equation 4.2.1, which states that the model evidence under a non-local prior will have the form

$$m_j^{\mathrm{NL}}[\mathbf{y}(r)] = m_j[\mathbf{y}(r)] \int_{\phi(r)} \int_{\boldsymbol{\theta}(r)} f[\boldsymbol{\theta}(r), \phi(r)] \, p[\boldsymbol{\theta}(r), \phi(r) \,|\, \mathbf{y}(r)] \, d\boldsymbol{\theta}(r) \, d\phi(r).$$

For any of the non-local priors described in the previous section, we can express the penalty function as $f[\boldsymbol{\theta}(r), \phi(r)] = g[\boldsymbol{\theta}(r)] \, H[\phi(r)]$ so that, assuming that we are implicitly conditioning on some model $\mathcal{M}_j(r)$, we may write

$$p^{\mathrm{NL}}[\mathbf{y}(r)] = Z_*(r) \int_{\phi(r)} p[\phi(r)] H[\phi(r)] \left(\int_{\boldsymbol{\theta}(r)} p[\mathbf{y}(r) \,|\, \boldsymbol{\theta}(r), \phi(r)] \, p[\boldsymbol{\theta}(r) \,|\, \phi(r)] \, g[\boldsymbol{\theta}(r)] \, d\boldsymbol{\theta}(r) \right) d\phi(r)$$

where $Z_*(r)$ is the appropriate normalisation constant. With some rearrangement,

$$p^{\mathrm{NL}}[\mathbf{y}(r)] = Z_*(r) \, p[\mathbf{y}(r)] \int_{\phi(r)} p[\phi(r) \,|\, \mathbf{y}(r)] \, H[\phi(r)] \left(\int_{\boldsymbol{\theta}(r)} p[\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r), \phi(r)] \, g[\boldsymbol{\theta}(r)] \, d\boldsymbol{\theta}(r) \right) d\phi(r). \tag{4.4.1}$$

We therefore need to evaluate

$$\int_{\phi(r)} p[\phi(r)\,|\,\mathbf{y}(r)]\,H[\phi(r)] \int_{\boldsymbol{\theta}(r)} p[\boldsymbol{\theta}(r)\,|\,\mathbf{y}(r),\phi(r)]\,g[\boldsymbol{\theta}(r)]\,d\boldsymbol{\theta}(r)\,d\phi(r).$$

4.4.1 The Model Evidence under a DLM-pMOM Non-Local Prior

Under a DLM-pMOM non-local prior, equation 4.4.1 is evaluated at each time point

$$\begin{aligned} p^{\mathrm{NL}}[y_t(r)\,|\,D_{t-1}] =& Z_t(r)\,p[y_t(r)\,|\,D_{t-1}] \\ & \int_{\phi(r)} p[\phi(r)\,|\,D_{t-1}] \int_{\boldsymbol{\theta}_t(r)} p[\boldsymbol{\theta}_t(r)\,|\,\phi(r),D_{t-1}]\,g[\boldsymbol{\theta}_t(r)]\,d\boldsymbol{\theta}_t(r)\,d\phi(r) \\ =& Z_t(r)\,p[y_t(r)\,|\,D_{t-1}]\,\mathbb{E}_{\mathcal{T}_{n_t(r)}[\mathbf{m}_t(r),\mathbf{C}_t(r)]}\left[\prod_{i\neq j}^{p_r} \theta_{it}^{2h_r}\right]. \end{aligned} \tag{4.4.2}$$

4.4.2 The Model Evidence under a DLM-QF Non-Local Prior

Given the joint distributions (outlined in subsection 4.3.4), we need to evaluate

$$\begin{aligned} p^{\mathrm{NL}}[\mathbf{y}(r)\,|\,D_0] = Z_{QF}(r)\,p[\mathbf{y}(r)\,|\,D_0] \int_{\phi(r)} p[\phi(r)\,|\,\mathbf{y}(r),D_0]\,H[\phi(r)] \\ \left(\int_{\boldsymbol{\theta}(r)} p[\boldsymbol{\theta}(r)\,|\,\mathbf{y}(r),\phi(r),D_0]\,g[\boldsymbol{\theta}(r)]\,d\boldsymbol{\theta}(r)\right) d\phi(r). \end{aligned}$$

The integral term is

$$\int_{\phi(r)} p[\phi(r)\,|\,\mathbf{y}(r),D_0]\,\phi(r)^{(p_r-1)} \left(\int_{\boldsymbol{\theta}(r)} p[\boldsymbol{\theta}(r)\,|\,\mathbf{y}(r),\phi(r),D_0] \prod_{i\neq j}^{p_r} \boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\,d\boldsymbol{\theta}(r)\right) d\phi(r).$$

The prior on the precision is proportional to a gamma distribution with probability density

$$\begin{aligned} & p[\phi(r)\,|\,\mathbf{y}(r),D_0]\,\phi(r)^{(p_r-1)} \\ &= \frac{1}{\Gamma\left[\frac{T+n_0(r)}{2}\right]} \left[\frac{d_0(r)+\tilde{d}(r)}{2}\right]^{\frac{T+n_0(r)}{2}} \phi(r)^{\frac{2(p_r-1)+T+n_0(r)}{2}} \exp\left\{-\phi(r)\left[\frac{d_0(r)+\tilde{d}(r)}{2}\right]\right\} \end{aligned}$$

where $\tilde{d}(r) = \mathbf{y}(r)^\top \mathbf{Q}^*(r)^{-1}\mathbf{y}(r)$.

It follows that we need to evaluate

$$\begin{aligned} & \int_{\phi(r)} p[\phi(r)\,|\,\mathbf{y}(r),D_0]\,\phi(r)^{(p_r-1)}\,\mathbb{E}_{\mathcal{N}[\underline{\mathbf{m}}(r),\phi(r)^{-1}\underline{\mathbf{C}}^*(r)]}\left[\prod_{i\neq j}^{p_r} \boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\right] d\phi(r) \\ &= \frac{\Gamma\left[\frac{\nu(r)}{2}\right]}{\Gamma\left[\frac{T+n_0(r)}{2}\right]} \left[\frac{d_0(r)+\tilde{d}(r)}{2}\right]^{-(p_r-1)} \mathbb{E}_{\mathcal{T}_{\nu(r)}[\underline{\mathbf{m}}(r),\underline{\mathbf{C}}(r)]}\left[\prod_{i\neq j}^{p_r} \boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\right] \end{aligned}$$

where $\nu(r) = 2(p_r - 1) + T + n_0(r)$ is the adjusted degrees of freedom of the posterior distribution. The adjusted degrees of freedom influence the estimated variance $S_T(r)$ because

$$S_T(r) = \mathbb{E}[\phi(r) \,|\, \mathbf{y}(r), D_0] = \frac{d_0(r) + \tilde{d}(r)}{\nu(r)}.$$

The final form of the model evidence under a DLM-QF non-local prior is therefore

$$p^{\mathrm{NL}}[\mathbf{y}(r) \,|\, D_0] = Z_{QF}(r)\, p[\mathbf{y}(r) \,|\, D_0]$$
$$\frac{\Gamma\left[\frac{\nu(r)}{2}\right]}{\Gamma\left[\frac{T+n_0(r)}{2}\right]} \left[\frac{d_0(r) + \tilde{d}(r)}{2}\right]^{-(p_r-1)} \mathbb{E}_{\mathcal{T}_{\nu(r)}[\underline{\mathbf{m}}(r), \underline{\mathbf{C}}(r)]} \left[\prod_{i \neq j}^{p_r} \boldsymbol{\theta}_i(r)^{\top} \boldsymbol{\theta}_i(r)\right]. \quad (4.4.3)$$

4.5 Application of a DLM-pMOM Non-Local Prior

In this section and the next, we illustrate the behaviour of the the DLM-pMOM and DLM-QF non-local priors when applied to real data. Returning to the 15 node, resting-state dataset, we consider a subnetwork with 4 nodes: the orbitofrontal cortex, the dorsolateral prefrontal cortex, the amygdala (all left hemisphere) and the ventromedial prefrontal cortex. As an example, we focus on discovering the parent set (out of the 8 possible candidates) for the OFC-L in this subnetwork. As can be seen from Figure 4.7, the DLPFC-L, found to be a parent in the group model, is associated with higher values of $\mu_{it}(r)$ than the VMPFC or the amygdala. The amygdala was only found to be a parent in the (sub)network of one subject and has consistently low values for $\mu_{it}(r)$. Therefore, we would expect the non-local prior to strongly penalise any model which contains the amygdala.

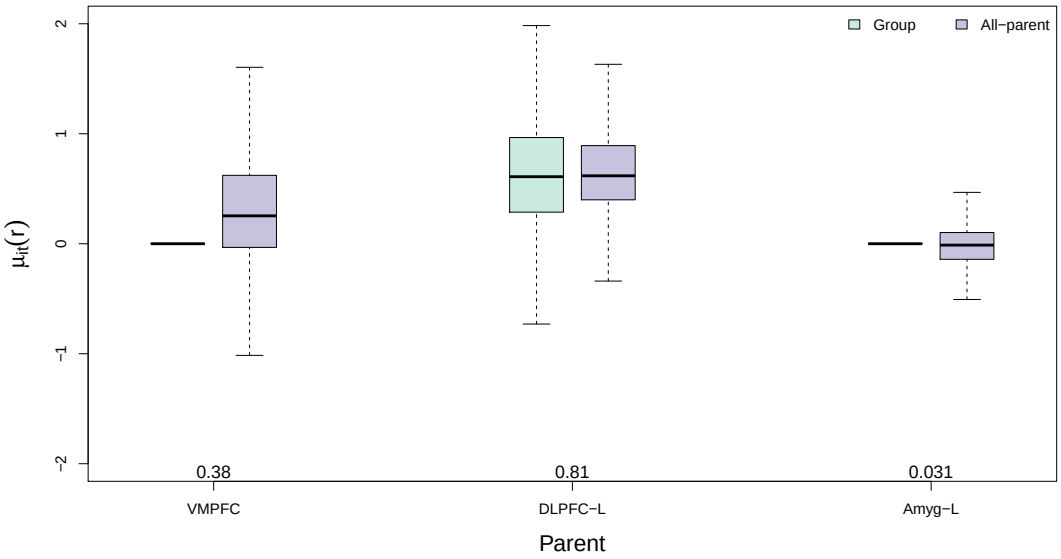


Figure 4.7: Correspondence between the consistency of edge presence and connectivity strength for the OFC-L subnetwork. Boxes show $\mu_{it}(r)$ over all subjects and all time for the parents of the OFC-L when the group model (green) and the all-parent model (purple) were fitted (flat lines represent parents absent in the group network; for easier visualisation, outliers are not shown). Values at the bottom show the proportion of subjects that were found to have this parent in the individual subject networks.

4.5.1 Implementation of a DLM-pMOM Non-Local Prior

From the form of equation 4.4.2, it is clear that calculating the penalised model evidence under a DLM-pMOM non-local prior is simply a case of calculating the normalisation term

$$\frac{1}{Z_t(r)} = \mathbb{E}_{\mathcal{T}_{n_{t-1}(r)}[\mathbf{a}_t(r),\mathbf{R}_t(r)]}\left[\prod_{i\neq j}^{p_r}\theta_{it}^{2h_r}\right]$$

and the posterior expectation term

$$\mathbb{E}_{\mathcal{T}_{n_t(r)}[\mathbf{m}_t(r),\mathbf{C}_t(r)]}\left[\prod_{i\neq j}^{p_r}\theta_{it}^{2h_r}\right]$$

at every time t. We know the location and scale parameters, and the degrees of freedom, at each time point from the DLM updating relations. Closed-form, computationally-efficient formulae for expectation of the product of normally- or t-distributed random variables raised to some power are available in Kan (2008) and may be implemented using the `eprod` function in the `mombf` package for R[1,2]. For small numbers of parents,

[1] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`

[2] Rossell, D., Cook, J.D., Telesca, D., and Roebuck, P. *mombf: Moment and Inverse Moment Bayes Factors*, 2017. URL `https://CRAN.R-project.org/package=mombf`. R package version 1.9.5

these calculations are very fast and calling the `eprod` function at each time point is feasible. However, the computational effort increases exponentially with the model size (Johnson and Rossell, 2012) so for larger networks it may be necessary to approximate these expectation terms. One straightforward approach would be assume $\theta_{it}(r)$ to be independent of the other $\boldsymbol{\theta}_t(r)$ so that

$$\mathbb{E}_{\mathcal{T}_{n_t(r)}[\mathbf{m}_t(r),\mathbf{C}_t(r)]}\left[\prod_{i\neq j}^{p_r}\theta_{it}^{2h_r}\right] \approx \prod_{i\neq j}^{p_r}\mathbb{E}_{\mathcal{T}_{n_t(r)}[\mathbf{m}_t(r),\mathbf{C}_t(r)]}\left[\theta_{it}^{2h_r}\right].$$

As this non-local prior is specified individually at each time point, as with the local case, we can discard the first few time points and only implement the penalty term once the model is influenced by the data and not the choice of prior hyperparameters[3].

4.5.2 The Effect of $\delta(r)$ on the Penalty Strength

Figure 4.8 shows the relationship between the number of parents in a model and the chosen discount factor $\delta(r)$.

As previously discussed, under a DLM-pMOM prior the discount factor will influence the strength of the penalty via its influence on the 'prior' variance $\mathbf{R}_t^*(r)$. Lower values of $\delta(r)$ increase the prior variance and in turn the width of the window around zero. This effect can be seen clearly in Figure 4.9a, which shows the relationship between the discount factor and the Log Predictive Likelihood or the penalised Log Predictive Likelihood with $h_r = 1$ and $h_r = 2$ for an individual subject.

It follows that we would expect the chosen value of the discount factor (the $\delta(r)$ with the highest LPL score) to increase (tend towards the static model, where $\delta(r) = 1$) under a DLM-pMOM prior. This is indeed what is found, shown across all 32 subjects, in Figure 4.9b.

To fully understand this behaviour, Figure 4.10 shows the penalty strengths for the 8 candidate parents sets when the discount factor is optimised for each parent set and each h_r (Figure 4.10a), when the discount factor is optimised for each parent set for the local model (Figure 4.10b) and when the discount factor is fixed across all parent sets (Figure 4.10c). From Figure 4.10a, we can see that, somewhat counterintuitively, the strength of the penalty decreases as the number of parents increases. The all-parent model (VMPFC, DLPFC-L, Amyg-L) has higher values of $\delta(r)$ and this results in a smaller penalty than for the model which has the Amyg-L as a single parent. As discussed above, our aim is to incorporate a penalty term that reduces the likelihood of models which contain regressors with consistently low coefficients, in this case the amygdala. Figure 4.10b illustrates that we still see this behaviour when we fix $\delta(r)$ at its chosen value under the local model. This behaviour may be avoided by fixing

[3]Note that, in Figures 4.9, 4.10 and 4.13, we calculated the model evidence and optimised $\delta(r)$ by summing from $t = 50$ (i.e. discarding the first 49 time points), rather than $t = 15$ as in previous chapters.

$\delta(r)$ across *all* the candidate parent sets. As is clear from Figure 4.10c, the penalty strength now increases with the number of parents. It is not desirable to fix the discount factor in this way, as doing so would mean compromising one the key strengths of the Dynamic Linear Model, its ability to choose the amount of variance that best fits the data. However, looking at Figure 4.8, it is clear that for models with higher number of parents, the range of chosen discount factors is much narrower (between 0.9 and 1), suggesting the undue severity of the penalty would mostly be of concern for the sparser models.

4.6 Application of a DLM-QF Non-Local Prior

4.6.1 Implementation of a DLM-QF Non-Local Prior

Given the limitations of the DLM-pMOM prior, we also developed the DLM-QF prior. From equation 4.4.3, calculating the penalised model evidence involves calculating the normalisation constant $Z_{QF}(r)$, the posterior expectation term and additional terms which arise from the scaling by $\phi(r)$ in the penalty term. Recall that the matrices $\underline{\mathbf{R}}^*(r)$ and $\underline{\mathbf{C}}(r)$ have dimension $p_r T \times p_r T$, so for the 15 node dataset with $T = 790$, these matrices would have dimension 11850×11850 for the all-parent model. We make the approximations

$$\mathbb{E}_{\mathcal{N}[\mathbf{0},\underline{\mathbf{R}}^*(r)]}\left[\prod_{i\neq j}^{p_r} \boldsymbol{\gamma}_i(r)^\top \boldsymbol{\gamma}_i(r)\right] \approx \prod_{i\neq j}^{p_r} \mathbb{E}_{\mathcal{N}[\mathbf{0},\underline{\mathbf{R}}^*(r)]}\left[\boldsymbol{\gamma}_i(r)^\top \boldsymbol{\gamma}_i(r)\right]$$

$$\mathbb{E}_{\mathcal{T}_{\nu(r)}[\underline{\mathbf{m}}(r),\underline{\mathbf{C}}(r)]}\left[\prod_{i\neq j}^{p_r} \boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\right] \approx \prod_{i\neq j}^{p_r} \mathbb{E}_{\mathcal{T}_{\nu(r)}[\underline{\mathbf{m}}(r),\underline{\mathbf{C}}(r)]}\left[\boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\right]$$

These allow fast, straightforward computation of the normalisation and posterior expectation terms. The adjustment term

$$\frac{\Gamma\left[\frac{\nu(r)}{2}\right]}{\Gamma\left[\frac{T+n_0(r)}{2}\right]}\left[\frac{d_0(r)+\tilde{d}(r)}{2}\right]^{-(p_r-1)}$$

is easily calculated from the one-step distributions because $d_0(r) + \tilde{d}(r) = d_T(r)$.

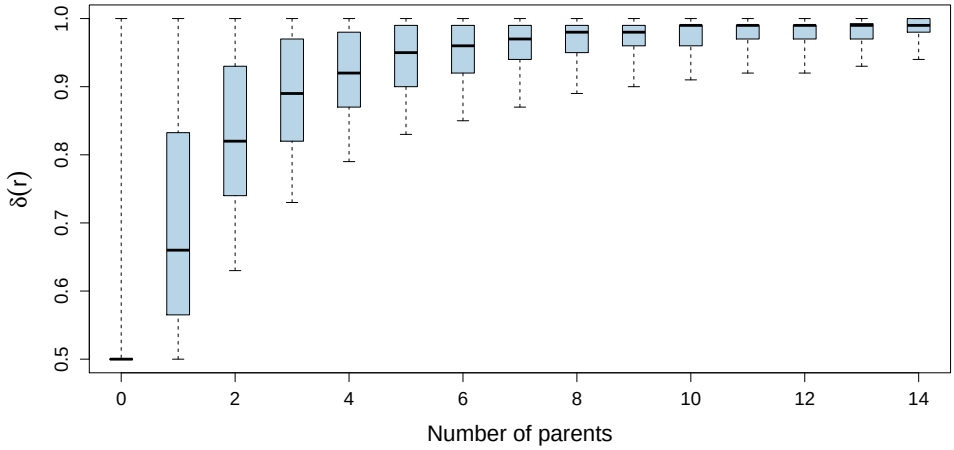
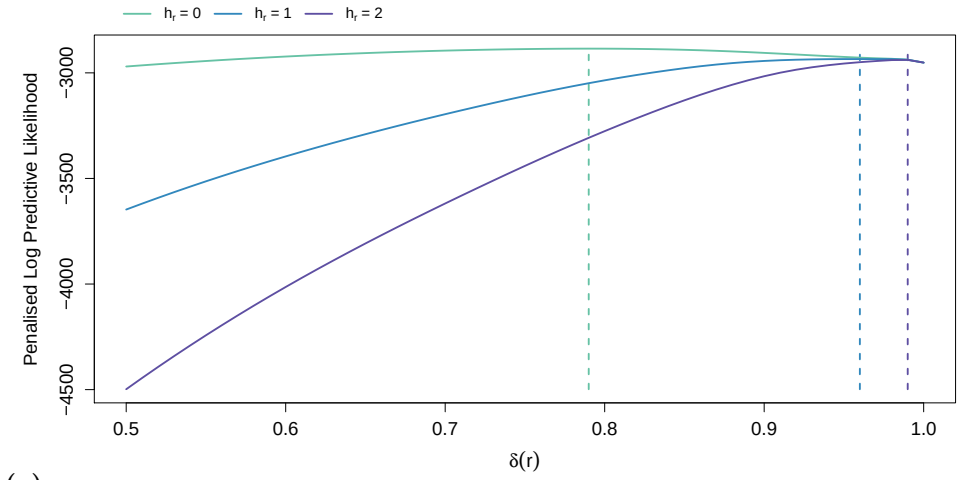
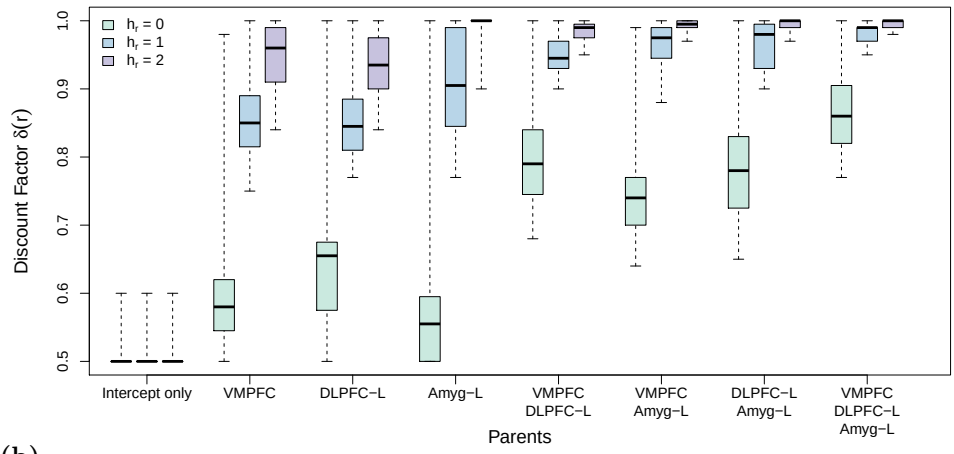


Figure 4.8: As the number of parents increases, the discount factor $\delta(r)$ tends towards higher values. Using the 15 node resting-state data, for each subject and each node, we calculated the median $\delta(r)$ for models with each number of parents. Whiskers show minimum and maximum values.

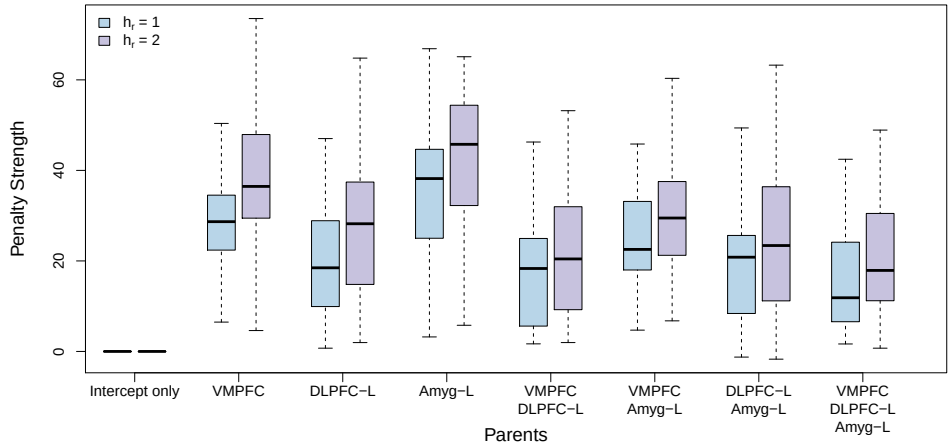


(a)

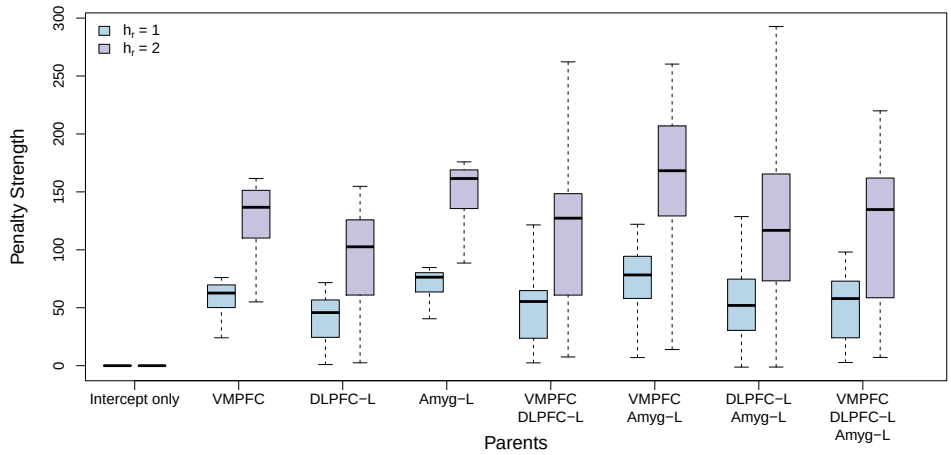


(b)

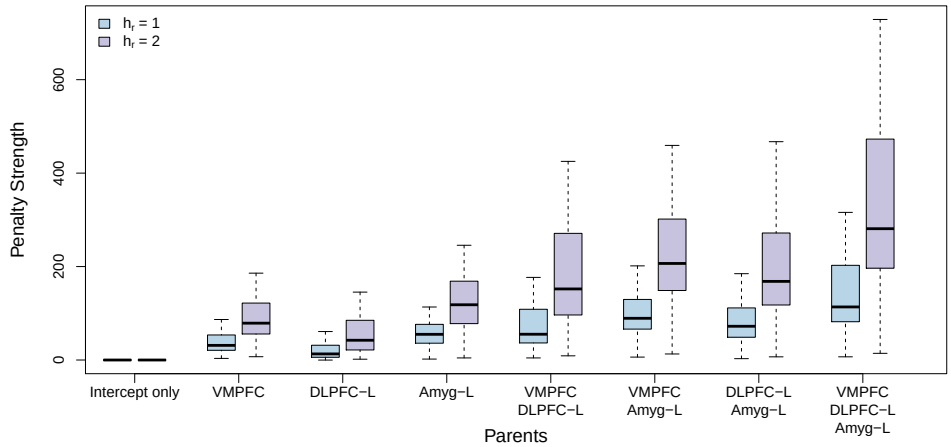
Figure 4.9: Under a DLM-pMOM prior, the optimum $\delta(r)$ is higher than in under a local prior. (a) The (penalised) Log Predictive Likelihood against the discount factor $\delta(r)$ for an individual subject, for the parents VMPFC and DLPFC-L. **(b)** This behaviour is consistent across subjects (whiskers show minimum and maximum values). As the order of the density and the number of parents increase, the optimal discount factor is pushed towards 1 (the static model).



(a)



(b)



(c)

Figure 4.10: The effect of the discount factor $\delta(r)$ on the strength of the penalty. The 'penalty strength' is the difference between the Log Predictive Likelihood under a local prior and under a DLM-pMOM prior with $h_r = 1$ (blue) and $h_r = 2$ (purple), each box is across subjects and outliers are not shown. In **(a)** the discount factor was optimised for each h_r, in **(b)** the discount factor was the optimal discount factor for the local prior ($h_r = 0$) and **(c)** the discount factor was fixed (for each subject) at its median value across all parent sets and all orders $h_r = 0$, $h_r = 1$ and $h_r = 2$.

4.6.2 Sensitivity to $\mathbf{C}_0^*(r)$

Recall the form of the joint (local) prior:

$$\begin{pmatrix} \boldsymbol{\theta}_1(r) \\ \boldsymbol{\theta}_2(r) \\ \vdots \\ \boldsymbol{\theta}_T(r) \end{pmatrix} \Bigg| \phi(r), D_0 \Bigg) \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \phi(r)^{-1} \begin{pmatrix} \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \cdots & \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) \\ \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \sum_{t=1}^{2} \mathbf{W}_t^*(r) & \cdots & \mathbf{C}_0^*(r) + \sum_{t=1}^{2} \mathbf{W}_t^*(r) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \sum_{t=1}^{2} \mathbf{W}_t^*(r) & \cdots & \mathbf{C}_0^*(r) + \sum_{t=1}^{T} \mathbf{W}_t^*(r) \end{pmatrix} \right).$$

As previously discussed, one advantage of the one-step relations is that we use the first few time points to obtain empirical values for the hyperparameters $\mathbf{m}_0(r)$, $\mathbf{C}_0^*(r)$, $n_0(r)$ and $d_0(r)$. However, using the DLM-QF prior, we obtain a form of the model evidence in terms of *all* the data and calculate the normalisation constant $Z_{QF}(r)$ independent of any observation $y_1(r), y_2(r)$ etc. This means the prior hyperparameter $\mathbf{C}_0^*(r)$ will influence both the likelihood (in the local model) and the strength of the penalty. While this may be advantageous in that we may potentially use $\mathbf{C}_0^*(r)$ to control the amount of sparsity we wish to impose, it also means it is necessary to give careful consideration when specifying a value. For this reason, the rest of this section explores the effect of the choice of $\mathbf{C}_0^*(r)$ on the local and non-local models. We fitted the all-parent model to the OFC-L, using $\mathbf{C}_0^*(r) = 3\,\mathbb{I}_{p_r}$ (a diffuse, non-informative value) and $\mathbf{C}_0^*(r) = \mathbf{C}_T^*(r)$, which may be thought of as the 'best' estimate as it has been informed by all the data $\mathbf{y}(r)$. Note that the values $\mathbf{C}_T^*(r)$ are typically much smaller than $3\,\mathbb{I}_{p_r}$, see Figure 4.11.

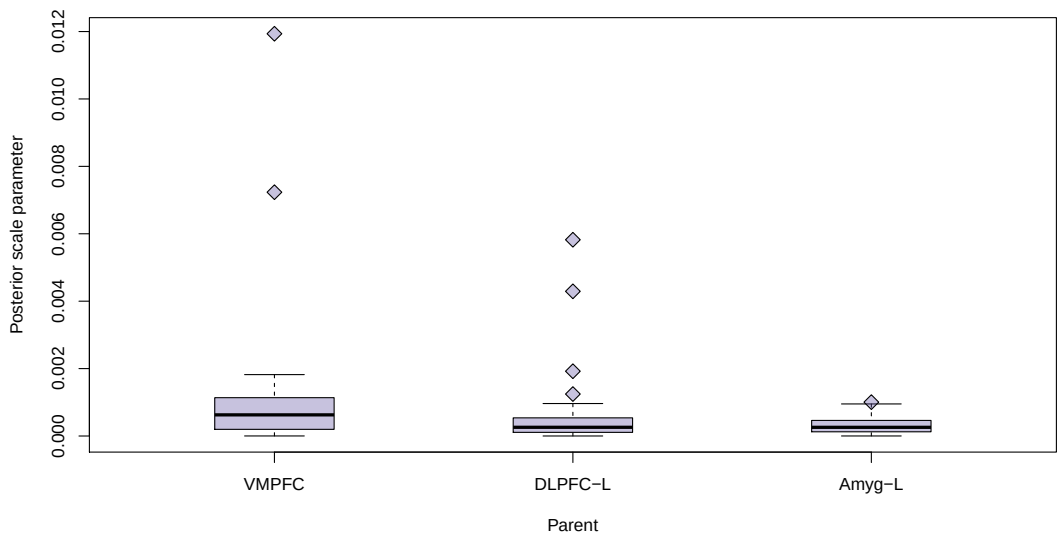
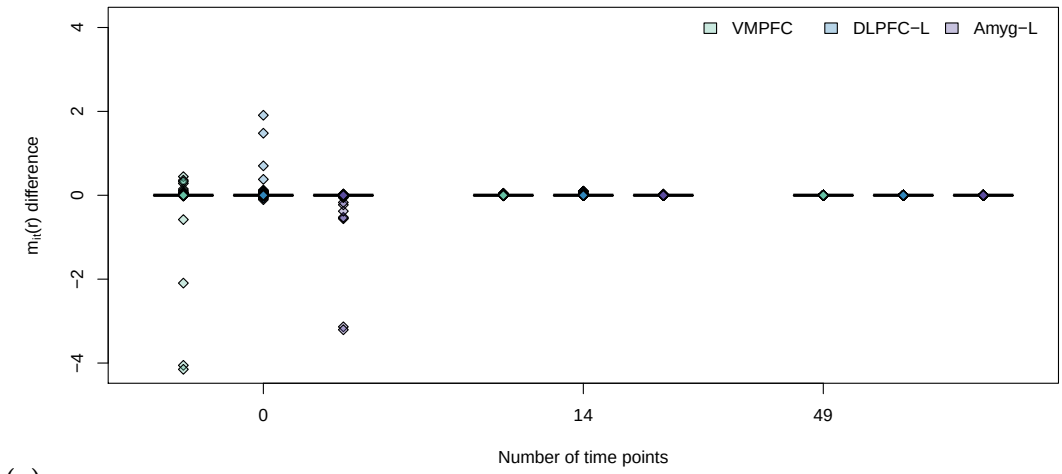


Figure 4.11: The posterior scale parameter $\mathbf{C}_T^*(r)$ across subjects for each parent in the subnetwork. Values are for the optimum discount factor $\delta(r) \in [0.5, 1]$, step size 0.02.

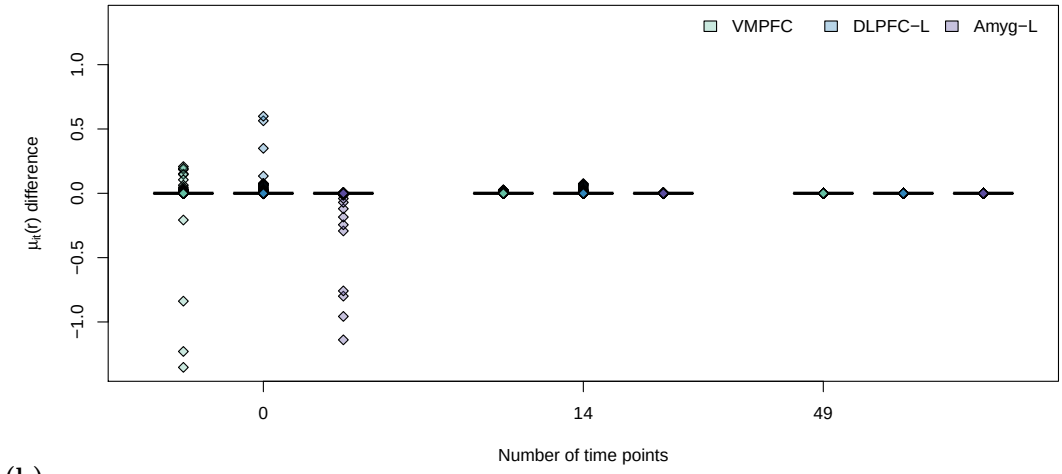
We compared the estimates for the time-varying regression coefficients under the two values of $\mathbf{C}_0^*(r)$, that is, the one-step posterior estimates $\mathbf{m}_t(r)$ and the retrospective estimates $\boldsymbol{\mu}_t(r)$. Results are shown in Figure 4.12. Figure 4.12a shows the difference

(for a single subject) between each $m_{it}(r)$ for each parent node for $t = 1, \ldots, T$ (left), $t = 15, \ldots, T$ (centre) and $t = 50, \ldots, T$ (right). It is clear that the largest discrepancies occur for the initial time points as the largest outliers disappear when these first few time points are discarded. Figure 4.12b shows that comparable behaviour is observed for the retrospective posterior mean $\mu_{it}(r)$, although (as would be expected), the magnitude of the outliers is smaller.

Figure 4.13 shows the normalisation term, the posterior expectation term and the DLM-QF penalty (across subjects). As would be expected from Figure 4.12, the posterior expectation term is much less strongly affected by the choice of $\mathbf{C}_0^*(r)$ than the normalisation term.



(a)



(b)

Figure 4.12: The influence of the prior hyperparameter $\mathbf{C}_0^*(r)$ on the estimates for the regression coefficients. **(a)** The difference in posterior mean of the one-step distribution $\mathbf{m}_t(r)$ for $\mathbf{C}_0^*(r) = 3\,\mathbb{I}_{p_r}$ and $\mathbf{C}_0^*(r) = \mathbf{C}_T^*(r)$ for an individual subject, over all time (left), with the first 14 time points removed (centre) and with the first 49 time points removed (right). It is clear that by removing the initial time points, the difference is strongly reduced and the effect of the choice of $\mathbf{C}_0^*(r)$ is minimised. **(b)** As **(a)** but for the retrospective posterior mean $\boldsymbol{\mu}_t(r)$.

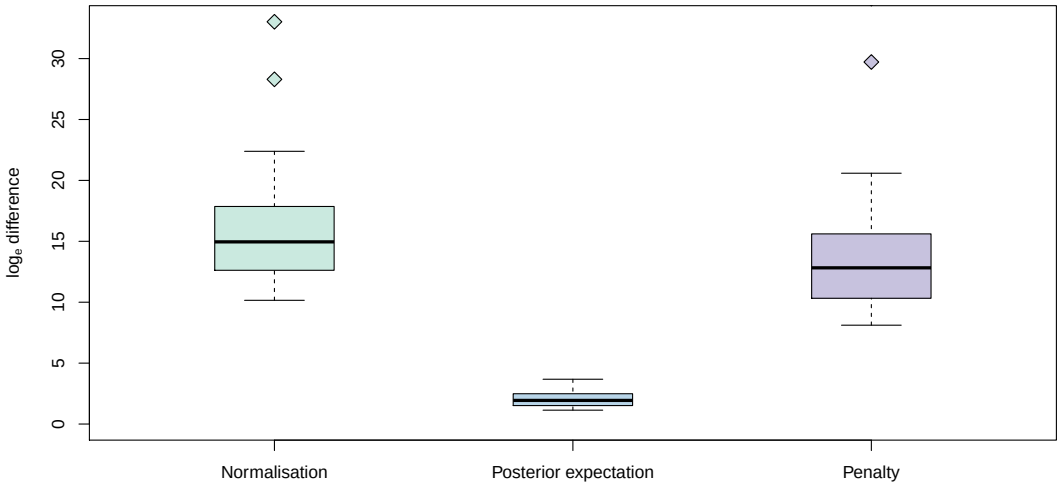


Figure 4.13: The influence of the prior hyperparameter $\mathbf{C}_0^*(r)$ on the strength of the penalty. When the penalty term of the DLM-QF prior (right) is divided into its component terms, it can be seen that the normalisation term is strongly influenced by the choice of $\mathbf{C}_0^*(r)$. We first calculated the penalty with $\mathbf{C}_0^*(r) = 3\mathbb{I}_{p_r}$ and the optimal discount factor $\delta(r)$ for the local model. We then used $\mathbf{C}_0^*(r) = \mathbf{C}_T^*(r)$ as the prior hyperparameter and re-optimised $\delta(r)$. Boxes show the (absolute) difference between the $\log_e$ normalisation term (green), the $\log_e$ posterior expectation (blue) and the overall $\log_e$ penalty (purple).

4.7 Discussion

In this chapter, we have derived two closed-form expressions for the model evidence which incorporate a penalty on weaker, and therefore potentially spurious, edges. The DLM-pMOM prior extends the pMOM prior of Johnson and Rossell (2012) to the Dynamic Linear Model. One clear advantage of this type of prior is that it may be implemented at each time point, thereby maintaining some of the computational-efficiency of the one-step distributions, although exact calculation of the expectation of the product of random variables will become prohibitively slow as the number of parents increases, making some kind of approximation necessary. The main drawback of the DLM-pMOM prior is that the discount factor $\delta(r)$ influences the strength of the penalty in such a way that low values of $\delta(r)$ (e.g. $\delta(r) = 0.5$) will introduce a severe penalty. Therefore, we may inadvertently end up penalising parent nodes with physiologically-interesting, time-varying connectivity, in favour of a more static model. As the optimum discount factor, at least on the data considered here, tends to be higher (suggesting stationary or near stationary connectivity strengths) for models with larger number of parents, the DLM-pMOM prior, with some fixed $\delta(r)$, may be appropriate.

However, in this work, we attempted to overcome the limitations of the DLM-pMOM non-local prior by constructing the joint prior distribution of the Dynamic Linear Model. While this allows us to specify a form for the penalty that may be considered more appropriate for dynamic, biological data, we must now be much more cautious in our specification of the prior hyperparameter $\mathbf{C}_0^*(r)$ and further work is necessary to

fully quantify the effect of the choice of $\mathbf{C}_0^*(r)$ on both the local and non-local models.

4.A The Normalisation Constant under a Non-Local Prior

Consider a local prior density that follows a multivariate normal distribution

$$p(\boldsymbol{\theta} \,|\, \phi) \sim \mathcal{N}_{\boldsymbol{\theta}}\big(\mathbf{0}, \phi^{-1}\mathbf{R}^*\big)\,.$$

with a non-local prior term $f(\boldsymbol{\theta}, \phi)$ that may be expressed as $f(\boldsymbol{\theta}, \phi) = g(\boldsymbol{\theta})H(\phi)$. We require that

$$\int_\phi \int_{\boldsymbol{\theta}} p^{\mathrm{NL}}(\boldsymbol{\theta}, \phi \,|\, \tau, h_r)\, d\boldsymbol{\theta}\, d\phi = 1$$

or equivalently that

$$\int_\phi p(\phi) H(\phi) \left(\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \phi)\, g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right) d\phi = 1.$$

4.A.1 The Normalisation Constant under a pMOM Non-Local Prior

Under the pMOM non-local prior of Johnson and Rossell (2012), $g(\boldsymbol{\theta}) = \prod_{i=1}^p \theta_i^{2h}$ and $H(\phi) = \phi^{hp}\tau^{-hp}$.

The inner integral is

$$\int_{\boldsymbol{\theta}} \mathcal{N}_{\boldsymbol{\theta}}\big(\mathbf{0}, \phi^{-1}\tau\mathbf{R}^*\big) \prod_{i=1}^p \theta_i^{2h} d\boldsymbol{\theta}.$$

Write $\boldsymbol{\theta} = \phi^{-\frac{1}{2}}\tau^{\frac{1}{2}}\boldsymbol{\gamma}$ and $d\boldsymbol{\theta} = \phi^{-\frac{p}{2}}\tau^{\frac{p}{2}}\, d\boldsymbol{\gamma}$ (because $d\boldsymbol{\theta} = |\mathbf{J}|\, d\boldsymbol{\gamma}$ where $\mathbf{J}$ is the Jacobian). Then, by changing the variable of integration, the inner integral is

$$\begin{aligned}
&\int_{\boldsymbol{\theta}} \mathcal{N}_{\boldsymbol{\theta}}\big(\mathbf{0}, \phi^{-1}\tau\mathbf{R}^*\big) \prod_{i=1}^p \theta_i^{2h} d\boldsymbol{\theta} \\
&\quad = (2\pi)^{-\frac{p}{2}} \phi^{\frac{p}{2}}\, \tau^{-\frac{p}{2}}\, |\mathbf{R}^*|^{-\frac{1}{2}} \int_{\boldsymbol{\theta}} \exp\left\{ -\frac{\phi}{2} \left[\boldsymbol{\theta}^\top (\tau\mathbf{R}^*)^{-1} \boldsymbol{\theta} \right] \right\} \prod_{i=1}^p \theta_i^{2h}\, d\boldsymbol{\theta} \\
&\quad = (2\pi)^{-\frac{p}{2}}\, |\mathbf{R}^*|^{-\frac{1}{2}} \int_{\boldsymbol{\gamma}} \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\gamma}^\top (\mathbf{R}^*)^{-1} \boldsymbol{\gamma} \right] \right\} \prod_{i=1}^p \left[\phi^{-\frac{1}{2}} \tau^{\frac{1}{2}} \gamma_i \right]^{2h} d\boldsymbol{\gamma} \\
&\quad = \phi^{-hp}\, \tau^{hp} \int_{\boldsymbol{\gamma}} \mathcal{N}_{\boldsymbol{\gamma}}(\mathbf{0}, \mathbf{R}^*) \prod_{i=1}^p \gamma_i^{2h}\, d\boldsymbol{\gamma}.
\end{aligned}$$

Then, because $H(\phi) = \phi^{hp}\,\tau^{-hp}$, the normalisation constant is

$$\begin{aligned}
\frac{1}{Z} &= \int_{\phi} p(\phi)\, H(\phi) \left(\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \phi)\, g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right) d\phi \\
&= \int_{\phi} p(\phi) \left(\int_{\boldsymbol{\gamma}} \mathcal{N}_{\boldsymbol{\gamma}}(\mathbf{0}, \mathbf{R}^{*}) \prod_{i=1}^{p} \gamma_i^{2h}\, d\boldsymbol{\gamma} \right) d\phi \\
&= \int_{\boldsymbol{\gamma}} \mathcal{N}_{\boldsymbol{\gamma}}(\mathbf{0}, \mathbf{R}^{*}) \prod_{i=1}^{p} \gamma_i^{2h}\, d\boldsymbol{\gamma}.
\end{aligned}$$

When the non-local prior is mean-zero, the integral over $\boldsymbol{\theta}$ may be calculated without knowledge of ϕ or τ.

4.A.2 The Normalisation Constant under a DLM-pMOM Non-Local Prior

We have $g[\boldsymbol{\theta}_t(r)] = \prod_{i \neq j}^{p_r} \theta_{it}^{2h_r}$. We need to find, for each time t,

$$\begin{aligned}
\frac{1}{Z_t(r)} &= \int_{\phi(r)} \int_{\boldsymbol{\theta}(r)} p[\boldsymbol{\theta}_t(r) \,|\, \phi(r), D_{t-1}]\, p[\phi(r) \,|\, D_{t-1}]\, g[\boldsymbol{\theta}_t(r)]\, d\boldsymbol{\theta}_t(r)\, d\phi(r) \\
&= \int_{\phi(r)} p[\phi(r) \,|\, D_{t-1}] \int_{\boldsymbol{\theta}_t(r)} p[\boldsymbol{\theta}_t(r) \,|\, \phi(r), D_{t-1}]\, g[\boldsymbol{\theta}_t(r)]\, d\boldsymbol{\theta}_t(r)\, d\phi(r) \\
&= \int_{\phi(r)} p[\phi(r) \,|\, D_{t-1}]\, \mathbb{E}_{\boldsymbol{\theta}_t(r)}\{g[\boldsymbol{\theta}_t(r)] \,|\, \phi(r), D_{t-1}\}\, d\phi(r) \\
&= \mathbb{E}_{\phi(r)}\{\mathbb{E}_{\boldsymbol{\theta}_t(r)}\{g[\boldsymbol{\theta}_t(r)] \,|\, \phi(r), D_{t-1}\}\} = \mathbb{E}_{\boldsymbol{\theta}_t(r)}\{g[\boldsymbol{\theta}_t(r)] \,|\, D_{t-1}\}
\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{\theta}_t(r)}\{\cdot\}$ and $\mathbb{E}_{\phi(r)}\{\cdot\}$ denote expectations with respect to the distribution of $\boldsymbol{\theta}_t(r)$ and $\phi(r)$ respectively.

We know the marginal distribution of $\boldsymbol{\theta}_t(r)$ is

$$p[\boldsymbol{\theta}_t(r) \,|\, D_{t-1}] \sim \mathcal{T}_{n_{t-1}(r)}[\mathbf{a}_t(r), \mathbf{R}_t(r)]$$

where $\mathbf{R}_t(r) = S_{t-1}\mathbf{R}_t^{*}(r)$.

4.A.3 The Normalisation Constant under a DLM-Quadratic-Form Non-Local Prior

Under a DLM-QF non-local prior, we have $g[\boldsymbol{\theta}(r)] = \prod_{i \neq j}^{p_r} \boldsymbol{\theta}_i(r)^{\top}\boldsymbol{\theta}_i(r)$ and $H[\phi(r)] = \phi(r)^{(p_r - 1)}$.

Write $\boldsymbol{\theta}(r) = \phi(r)^{-\frac{1}{2}}\boldsymbol{\gamma}(r)$ so that $\boldsymbol{\theta}_i(r)^{\top}\boldsymbol{\theta}_i(r) = \phi(r)^{-1}\boldsymbol{\gamma}_i(r)^{\top}\boldsymbol{\gamma}_i(r)$ and $d\boldsymbol{\theta}(r) = \phi(r)^{-\frac{p_r T}{2}}\, d\boldsymbol{\gamma}(r)$.

Then, by changing the variable of integration, the inner integral is

$$\int_{\boldsymbol{\theta}(r)} \mathcal{N}_{\boldsymbol{\theta}(r)}\big[\mathbf{0}, \phi(r)^{-1}\underline{\mathbf{R}}^*(r)\big] \prod_{i\neq j}^{p_r} \boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\, d\boldsymbol{\theta}(r)$$

$$= (2\pi)^{-\frac{p_r T}{2}} \phi^{\frac{p_r T}{2}} |\underline{\mathbf{R}}^*|^{-\frac{1}{2}} \int_{\boldsymbol{\theta}(r)} \exp\left\{-\frac{\phi(r)}{2}\left[\boldsymbol{\theta}(r)^\top [\underline{\mathbf{R}}^*(r)]^{-1} \boldsymbol{\theta}(r)\right]\right\} \prod_{i\neq j}^{p_r} \boldsymbol{\theta}_i(r)^\top \boldsymbol{\theta}_i(r)\, d\boldsymbol{\theta}(r)$$

$$= (2\pi)^{-\frac{p_r T}{2}} |\underline{\mathbf{R}}^*|^{-\frac{1}{2}} \int_{\boldsymbol{\gamma}(r)} \exp\left\{-\frac{1}{2}\left[\boldsymbol{\gamma}(r)^\top [\underline{\mathbf{R}}^*(r)]^{-1} \boldsymbol{\gamma}(r)\right]\right\} \prod_{i\neq j}^{p_r} \left[\phi(r)^{-1} \boldsymbol{\gamma}_i(r)^\top \boldsymbol{\gamma}_i(r)\right] d\boldsymbol{\gamma}(r)$$

$$= \phi(r)^{-(p_r-1)} \int_{\boldsymbol{\gamma}(r)} \mathcal{N}_{\boldsymbol{\gamma}(r)}[\mathbf{0}, \underline{\mathbf{R}}^*(r)]\, d\boldsymbol{\gamma}(r).$$

Then, because $H[\phi(r)] = \phi(r)^{(p_r-1)}$, the normalisation constant is

$$\frac{1}{Z_{QF}(r)} = \int_{\phi(r)} p[\phi(r) \,|\, D_0] \left(\int_{\boldsymbol{\gamma}(r)} \mathcal{N}_{\boldsymbol{\gamma}(r)}[\mathbf{0}, \underline{\mathbf{R}}^*(r)]\, d\boldsymbol{\gamma}(r) \right) d\phi(r)$$

$$= \int_{\boldsymbol{\gamma}(r)} \mathcal{N}_{\boldsymbol{\gamma}(r)}[\mathbf{0}, \underline{\mathbf{R}}^*(r)] \prod_{i\neq j}^{p_r} \boldsymbol{\gamma}_i(r)^\top \boldsymbol{\gamma}_i(r)\, d\boldsymbol{\gamma}(r).$$

4.B Derivation of the DLM Joint Distributions

Here we provide a more detailed derivation of the results presented in section 4.3.4.

The Dynamic Linear Model equations are

Obs. equation	$Y_t(r) = \mathbf{F}_t(r)^\top \boldsymbol{\theta}_t(r) + v_t(r)$	$v_t(r) \sim \mathcal{N}\big[0, \phi(r)^{-1}\big]$	(4.B.1)
State equation	$\boldsymbol{\theta}_t(r) = \boldsymbol{\theta}_{t-1}(r) + \mathbf{w}_t(r)$	$\mathbf{w}_t(r) \sim \mathcal{N}[\mathbf{0}, \mathbf{W}_t(r)]$	(4.B.2)
Initial information	$\boldsymbol{\theta}_0(r) \,\|\, D_0 \sim \mathcal{N}[\mathbf{m}_0(r), \mathbf{C}_0(r)]$		(4.B.3)

From equation 4.B.2, it is straightforward to show that at any time t, prior to any observations, the expectation of $\boldsymbol{\theta}_t(r)$ is simply the expectation of $\boldsymbol{\theta}_0(r)$ because

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\theta}_t(r)\right] &= \mathbb{E}\left[\boldsymbol{\theta}_{t-1}(r)\right] + \mathbb{E}\left[\mathbf{w}_t(r)\right] \\ &= \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right] + \mathbb{E}\left[\mathbf{w}_1(r)\right] + \cdots + \mathbb{E}\left[\mathbf{w}_t(r)\right]\end{aligned}$$

and $\mathbb{E}\left[\mathbf{w}_t(r)\right] = \mathbf{0}$ for all t.

Employing the single underline notation to denote mean and variance (or location and scale) parameters defined over all time, rather than at an individual time point, we

may define the prior expectation $\underline{\mathbf{a}}(r)$, with dimension $p_r T \times 1$ as

$$\underline{\mathbf{a}}(r) = \mathbb{E}\left[\boldsymbol{\theta}(r)\right] = \begin{pmatrix} \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right] \\ \vdots \\ \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right] \end{pmatrix} = \begin{pmatrix} \mathbf{m}_0(r) \\ \vdots \\ \mathbf{m}_0(r) \end{pmatrix}.$$

By the same logic, the variance of $\boldsymbol{\theta}_t(r)$ may be expressed as

$$\begin{aligned} \mathrm{Var}[\boldsymbol{\theta}_t(r)] &= \mathrm{Var}[\boldsymbol{\theta}_{t-1}(r)] + \mathrm{Var}[\mathbf{w}_t(r)] \\ &= \mathrm{Var}[\boldsymbol{\theta}_0(r)] + \mathrm{Var}[\mathbf{w}_1(r)] + \cdots + \mathrm{Var}[\mathbf{w}_t(r)] \\ &= \mathbf{C}_0(r) + \mathbf{W}_1(r) + \cdots + \mathbf{W}_t(r). \end{aligned}$$

The covariance of $\boldsymbol{\theta}_t(r)$ with some past value of itself $\boldsymbol{\theta}_{t-k}(r)$ is

$$\mathrm{Cov}[\boldsymbol{\theta}_{t-k}(r), \boldsymbol{\theta}_t(r)] = \mathrm{Cov}[\boldsymbol{\theta}_{t-k}(r), \boldsymbol{\theta}_{t-k}(r) + \mathbf{w}_{t-k+1}(r) +, \ldots, + \mathbf{w}_t(r)] = \mathrm{Var}[\boldsymbol{\theta}_{t-k}(r)].$$

The joint covariance structure, denoted $\underline{\mathbf{R}}(r)$, is then

$$\underline{\mathbf{R}}(r) = \begin{pmatrix} \mathbf{C}_0(r) + \mathbf{W}_1(r) & \mathbf{C}_0(r) + \mathbf{W}_1(r) & \cdots & \mathbf{C}_0(r) + \mathbf{W}_1(r) \\ \mathbf{C}_0(r) + \mathbf{W}_1(r) & \mathbf{C}_0(r) + \mathbf{W}_1(r) + \mathbf{W}_2(r) & \cdots & \mathbf{C}_0(r) + \mathbf{W}_1(r) + \mathbf{W}_2(r) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_0(r) + \mathbf{W}_1(r) & \mathbf{C}_0(r) + \mathbf{W}_1(r) + \mathbf{W}_2(r) & \cdots & \mathbf{C}_0(r) + \mathbf{W}_1(r) + \mathbf{W}_2(r) +, \ldots, + \mathbf{W}_T(r) \end{pmatrix}.$$

As with the one-step distributions, it is possible to write $\underline{\mathbf{R}}(r) = \phi(r)^{-1}\,\underline{\mathbf{R}}^*(r)$, so that we may express the joint prior as

$$\begin{pmatrix} \boldsymbol{\theta}_1(r) \\ \boldsymbol{\theta}_2(r) \\ \vdots \\ \boldsymbol{\theta}_T(r) \end{pmatrix} \Bigg| \phi(r), D_0 \Bigg) \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}_0(r) \\ \mathbf{m}_0(r) \\ \vdots \\ \mathbf{m}_0(r) \end{pmatrix}, \phi(r)^{-1} \begin{pmatrix} \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \cdots & \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) \\ \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \sum_{t=1}^{2} \mathbf{W}_t^*(r) & \cdots & \mathbf{C}_0^*(r) + \sum_{t=1}^{2} \mathbf{W}_t^*(r) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_0^*(r) + \mathbf{W}_1^*(r) & \mathbf{C}_0^*(r) + \sum_{t=1}^{2} \mathbf{W}_t^*(r) & \cdots & \mathbf{C}_0^*(r) + \sum_{t=1}^{T} \mathbf{W}_t^*(r) \end{pmatrix} \right).$$

We now use a similar approach to find the joint distribution $p[\mathbf{y}(r), \boldsymbol{\theta}(r) \,|\, \phi(r), D_0]$. Combining equations 4.B.1 and 4.B.2, we may write

$$\begin{aligned} Y_t(r) &= \mathbf{F}_t(r)^\top [\boldsymbol{\theta}_{t-1}(r) + \mathbf{w}_t(r)] + v_t(r) \\ &= \mathbf{F}_t(r)^\top [\boldsymbol{\theta}_0(r) + \mathbf{w}_1(r) + \cdots + \mathbf{w}_t(r)] + v_t(r). \end{aligned}$$

In the MDM, $\mathbf{F}_t(r)$ is a known linear function, independent of everything except $\mathbf{x}^t(r)$ (observations of the parent nodes) and $\mathbf{y}^{t-1}(r)$ (Queen and Smith, 1993). It follows that

$$\mathbb{E}\left[Y_t(r)\right] = \mathbf{F}_t(r)^\top \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right]$$

(see also West and Harrison (1997) pp. 638-9). We define the joint prior expectation

for $\mathbf{Y}(r)$ as a vector with length T

$$\mathbb{E}\left[\mathbf{Y}(r)\right] = \begin{pmatrix} \mathbf{F}_1(r)^\top \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right] \\ \mathbf{F}_2(r)^\top \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right] \\ \vdots \\ \mathbf{F}_T(r)^\top \mathbb{E}\left[\boldsymbol{\theta}_0(r)\right] \end{pmatrix} = \underline{\mathbf{F}}(r)^\top \underline{\mathbf{a}}(r)$$

The $p_r T \times T$ matrix $\underline{\mathbf{F}}(r)$ is a block diagonal matrix defined so that

$$\underline{\mathbf{F}}(r) = \begin{pmatrix} \mathbf{F}_1(r) & & & \\ & \mathbf{F}_2(r) & & \\ & & \ddots & \\ & & & \mathbf{F}_T(r) \end{pmatrix} \qquad \underline{\mathbf{F}}(r)^\top \boldsymbol{\theta}(r) = \begin{pmatrix} \mathbf{F}_1(r)^\top \boldsymbol{\theta}_1(r) \\ \mathbf{F}_2(r)^\top \boldsymbol{\theta}_2(r) \\ \vdots \\ \mathbf{F}_T(r)^\top \boldsymbol{\theta}_T(r) \end{pmatrix}.$$

With this definition of $\underline{\mathbf{F}}(r)$, and the vector for the observation variance $\mathbf{v}(r) = \{v_1(r), \ldots, v_T(r)\}$ corresponding to the variance matrix $\phi(r)^{-1}\,\mathbb{I}_T$, it follows that the variance of $\mathbf{Y}(r)$ is

$$\begin{aligned} \mathrm{Cov}[\mathbf{Y}(r), \mathbf{Y}(r)] &= \mathrm{Cov}\big[\underline{\mathbf{F}}(r)^\top \boldsymbol{\theta}(r) + \mathbf{v}(r), \underline{\mathbf{F}}(r)^\top \boldsymbol{\theta}(r) + \mathbf{v}(r)\big] \\ &= \underline{\mathbf{F}}(r)^\top \mathrm{Var}[\boldsymbol{\theta}(r)]\, \underline{\mathbf{F}}(r) + \mathrm{Var}[\mathbf{v}(r)] \\ &= \phi(r)^{-1}\big[\underline{\mathbf{F}}(r)^\top \underline{\mathbf{R}}^*(r) \underline{\mathbf{F}}(r) + \mathbb{I}_T\big] = \phi(r)^{-1} \underline{\mathbf{Q}}^*(r) \end{aligned}$$

The covariance structure of $\mathbf{Y}(r)$ and $\boldsymbol{\theta}(r)$ is

$$\mathrm{Cov}[\mathbf{Y}(r), \boldsymbol{\theta}(r)] = \mathrm{Cov}\big[\underline{\mathbf{F}}(r)^\top \boldsymbol{\theta}(r) + \mathbf{v}(r), \boldsymbol{\theta}(r)\big] = \underline{\mathbf{F}}(r)^\top \mathrm{Cov}[\boldsymbol{\theta}(r), \boldsymbol{\theta}(r)] = \underline{\mathbf{F}}(r)^\top \mathrm{Var}[\boldsymbol{\theta}(r)]\,.$$

Now we construct the joint distribution of $\mathbf{Y}(r)$ and $\boldsymbol{\theta}(r)$ (conditional on $\phi(r)$ and D_0)

$$\begin{pmatrix} \mathbf{Y}(r) \\ \boldsymbol{\theta}(r) \end{pmatrix} \Bigg| \phi(r), D_0 \Bigg) \sim \mathcal{N}\left(\begin{pmatrix} \underline{\mathbf{F}}(r)^\top \underline{\mathbf{a}}(r) \\ \underline{\mathbf{a}}(r) \end{pmatrix}, \phi(r)^{-1} \begin{pmatrix} \underline{\mathbf{Q}}^*(r) & \underline{\mathbf{F}}(r)^\top \underline{\mathbf{R}}^*(r) \\ \underline{\mathbf{R}}^*(r) \underline{\mathbf{F}}(r) & \underline{\mathbf{R}}^*(r) \end{pmatrix} \right).$$

West and Harrison (1997) (pp. 638-9) provide a simpler example based on constant matrices.

It follows from the properties of the multivariate Gaussian that

$$\begin{aligned} p[\mathbf{Y}(r) \,|\, \boldsymbol{\theta}(r), \phi(r), D_0] &\sim \mathcal{N}\big[\underline{\mathbf{F}}(r)^\top \boldsymbol{\theta}(r), \phi(r)^{-1} \mathbb{I}_T\big] \\ p[\boldsymbol{\theta}(r) \,|\, \mathbf{Y}(r), \phi(r), D_0] &\sim \mathcal{N}\big[\underline{\mathbf{m}}(r), \phi(r)^{-1} \underline{\mathbf{C}}^*(r)\big] \end{aligned}$$

with

$$\underline{\mathbf{C}}^*(r) = \underline{\mathbf{R}}^*(r) - \underline{\mathbf{R}}^*(r)\underline{\mathbf{F}}(r)\underline{\mathbf{Q}}^*(r)^{-1}\underline{\mathbf{F}}(r)^\top\underline{\mathbf{R}}^*(r)$$

$$\underline{\mathbf{m}}(r) = \underline{\mathbf{a}}(r) + \underline{\mathbf{R}}^*(r)\underline{\mathbf{F}}(r)\underline{\mathbf{Q}}^*(r)^{-1}\left[\mathbf{y}(r) - \underline{\mathbf{F}}(r)^\top\underline{\mathbf{a}}(r)\right].$$

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we have presented a number of extensions to the work of Costa (2014), Costa et al. (2015) and Costa et al. (2017), focusing on the Multiregression Dynamic Model Directed Graph Model (MDM-DGM) search. Using this algorithm, we may infer directed fMRI networks and obtain dynamic estimates for the connectivity weights. As discussed in Chapter 1, a number of competing methods exist and these differ from the MDM-DGM in a number of regards. For example, methods such as Multivariate Dynamical Systems (MDS; Ryali et al. (2011)) and Switching Linear Dynamic Systems (SLDS; Smith et al. (2010)) use a state-space framework, representing neuronal activity by a set of latent (state) variables, which are then mapped to the observed BOLD response. By comparison, in the MDM-DGM, the state variables represent time-varying connectivity strengths. Network discovery involves finding the set of parent nodes that maximise the model evidence. As the model evidence is closed-form and the parents are found for each node individually, the model search of the MDM-DGM is computationally-efficient and may readily be parallelised. Unlike some other fMRI Bayesian network discovery methods, MDM-DGM networks are not necessarily acyclic. Without acyclicity, we lose the definition of contemporaneous causality described by Queen and Albers (2009). The MDM-Integer Programming Algorithm (MDM-IPA) algorithm of Costa (2014) and Costa et al. (2015, 2017) provides a method by which the MDM-DGM networks may be constrained to be DAGs. However, we argue that the ability to infer cycles and bidirectional edges is advantageous as it allows networks with a more biophysical interpretation. For example, we showed in Chapter 2 that the MDM-DGM networks estimate strong inter-hemispheric connectivity, a feature that would be missed if an acyclicity constraint was imposed.

Unlike Dynamic Causal Modelling (DCM), the MDM-DGM has no generative model relating the observed BOLD response to the underlying biophysical processes (Friston et al., 2003, 2011). However, the ability of the MDM-DGM to infer directed and dynamic connectivity allows us deeper insights than a method such as partial correlation (an established method for inferring edge presence). In Chapter 2, we showed that the MDM-DGM search could estimate directed, physiologically-interpretable networks for

a system with 15 brain regions. We constructed both individual and group networks for two experimental conditions, 'safe' and 'anticipation of shock' (Bijsterbosch et al., 2015). We used partial correlation to validate the MDM-DGM networks, showing that the networks estimated by both methods were very similar. Interestingly, using the MDM-DGM we found that the ventromedial prefrontal cortex (VMPFC) had multiple children but few or no parents, an insight not possible using an undirected method such as partial correlation. We also performed an alternative analysis using the time-varying estimates for the regression coefficients, although we were unable to detect statistically significant differences between the two experimental conditions, or between subgroups using splits based on measures of trait and induced anxiety. Possible extensions to these analyses are discussed in section 5.2.

During our analysis of these networks, we identified two limitations of the MDM-DGM, which became the focus of the following two chapters. Firstly, the size of the model space increases exponentially, limiting an exhaustive search over all the candidate parent sets to networks with no more than 20 nodes. In Chapter 3, we showed that stepwise algorithms could reproduce the 15 node networks with as much as 100 % accuracy. As these algorithms score only a tiny fraction of the model space, a dramatic reduction in computation time is possible. Potential improvements to the model selection procedure are discussed in section 5.3.

As we touched upon in Chapter 3, the performance of the MDM-DGM on larger networks is yet to be established. As other connectivity methods are applied to larger number of nodes (see, for example, Razi et al. (2017)), these will provide a benchmark for comparison.

The second limitation was that while the MDM-DGM search identified some edges that were highly consistent across subjects and had high values of the regression coefficients, it also tended to identify weaker, less consistent edges that compromised the robustness of the estimated group networks. In Chapter 4, we considered non-local priors as a potential method to incorporate a penalty on the model evidence. We developed two non-local prior formulations, the DLM-pMOM non-local prior, which is implemented at the level of the one-step distributions, and the DLM-QF non-local prior, which is defined using the joint (over time) distributions of the Dynamic Linear Model. Advantageously, under these non-local priors the model evidence retains its closed-form. However, as discussed in Chapter 4, a number of theoretical and computational challenges must be addressed before these non-local priors may be applied appropriately to larger networks. Some of these challenges are discussed briefly in section 5.4.

5.2 Point Estimate vs. Bayesian Model Averaging

Throughout this work, we have assumed that there is a single 'best' model (set of parents for each node) which can be found by maximising the Log Predictive Likelihood. When we obtain estimates for the dynamic regression coefficients, we are conditioning

on a single model $\hat{\mathcal{M}}(r)$, such that

$$\boldsymbol{\mu}_t(r) = \mathbb{E}\{\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r), \hat{\mathcal{M}}(r)\}. \tag{5.2.1}$$

However, as discussed in Madigan and Raftery (1994), conditioning on a single model fails to account for model uncertainty. As we showed in Chapter 3, there is typically a small subset of models that may be thought of as having equivalent evidence ($\log_e$ Bayes factor $\pm$ 1, Kass and Raftery (1995)). It would be straightforward to replace equation 5.2.1 with the Bayesian Model Average (BMA)

$$\mathbb{E}\{\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r)\} = \sum_{i=1}^{N} \mathbb{E}\{\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r), \mathcal{M}_i(r)\}\, p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)] \tag{5.2.2}$$

where the posterior model probabilities are

$$p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)] = \frac{p[\mathbf{y}(r) \,|\, \mathcal{M}_i(r)]\, p[\mathcal{M}_i(r)]}{\sum_{i=1}^{N} p[\mathbf{y}(r) \,|\, \mathcal{M}_i(r)]\, p[\mathcal{M}_i(r)]}$$

as outlined in Chapter 2, section 1.6.

However, as argued by Madigan and Raftery (1994), it makes sense to discard the large number of models which predict the data much less well (when using Bayes factors as a measure of model fit). They define some subset of models, relative to the model with the maximum posterior model probability $\max\{p[\mathcal{M}_j(r) \,|\, \mathbf{y}(r)]\}$, for which the Bayes factor evidence for a difference is inconclusive, based on some threshold c

$$\mathcal{A}' = \left\{ \mathcal{M}_i(r) : \frac{\max\{p[\mathcal{M}_j(r) \,|\, \mathbf{y}(r)]\}}{p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)]} \leq c \right\}.$$

For example, $c \leq 20$ would correspond to a $\log_e$ Bayes factor of ≤ 3. To exclude unnecessarily complex models, any model $\mathcal{M}_j(r)$ is rejected if there is a model $\mathcal{M}_i(r)$ nested in $\mathcal{M}_j(r)$ and $p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)] > p[\mathcal{M}_j(r) \,|\, \mathbf{y}(r)]$. More formally, there is a subset of models

$$\mathcal{B} = \left\{ \mathcal{M}_j(r) : \exists\, \mathcal{M}_i(r) \in \mathcal{A}', \mathcal{M}_i(r) \subset \mathcal{M}_j(r), \frac{p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r)]}{p[\mathcal{M}_j(r) \,|\, \mathbf{y}(r)]} > 1 \right\}.$$

Madigan and Raftery (1994) propose replacing equation 5.2.2 with equation 5.2.3, which only includes models in the subset $\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}$

$$\mathbb{E}\{\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r), \mathcal{A}\} = \sum_{i=1}^{N} \mathbb{E}\{\boldsymbol{\theta}(r) \,|\, \mathbf{y}(r), \hat{\mathcal{M}}(r)\}\, p[\mathcal{M}_i(r) \,|\, \mathbf{y}(r), \mathcal{A}]. \tag{5.2.3}$$

As we showed in Chapter 3, section 3.4, for MDM-DGM networks with 12-15 nodes, there is a large model space containing a small number of parent sets with comparable $\log_e$ Bayes factor evidence. This type of model averaging may be an effective method

for analysing the dynamic regression coefficients.

5.3 Alternative Model Selection Strategies

In Chapter 3, we considered forward selection and backward elimination individually. However, the term *stepwise regression* often refers to algorithms which use forward selection and backward elimination principles, where at each stage a regressor (parent) may be added (as in forward selection) but at the same step a previously included parent (or parents) may be removed (Hocking, 1976; Davison, 2003). Figure 5.1 illustrates the basic principle behind this type of search. When used in isolation, forward selection cannot recover the parent set $Pa(4) = \{1,3\}$ if the parent set $Pa(4) = \{2\}$ has a higher Log Predictive Likelihood than either $Pa(4) = \{1\}$ or $Pa(4) = \{3\}$. However, if after scoring $Pa(4) = \{1,2,3\}$, the algorithm may also score $Pa(4) = \{1,3\}$ (i.e. test the removal of parent 2), it can now return the parent set found using an exhaustive search.

Implementing a search of this kind may be a natural extension to the approaches explored in Chapter 3, although the number of models scored using this type of method would have to be determined empirically, as in principle this algorithm could score much larger regions of the model space. One advantage of the combined method presented in Chapter 3 is that the number of models in the reduced model space is fixed. Another limitation of this method is that it is still a search that assumes there is a single local maximum. In order to implement equation 5.2.3, we require an algorithm that returns the subset of models with equivalent evidence. The Occam's window algorithm, a variant of the greedy search based on the principles outlined in the previous section, provides a possible alternative to the stepwise search. Consider two models $\mathcal{M}_0(r)$ and $\mathcal{M}_1(r)$ where $\mathcal{M}_0(r)$ is nested in $\mathcal{M}_1(r)$ and let O_R be a positive constant (O_R may be zero). Using this algorithm, there are three options

1. There is conclusive evidence for the simpler model. Reject $\mathcal{M}_1(r)$.

$$\log\left\{\frac{p[\mathcal{M}_0(r)\,|\,\mathbf{y}(r)]}{p[\mathcal{M}_1(r)\,|\,\mathbf{y}(r)]}\right\} > O_R$$

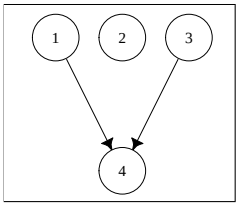
2. The evidence is inconclusive. Consider $\mathcal{M}_0(r)$ and $\mathcal{M}_1(r)$.

$$-\log(c) \leq \log\left\{\frac{p[\mathcal{M}_0(r)\,|\,\mathbf{y}(r)]}{p[\mathcal{M}_1(r)\,|\,\mathbf{y}(r)]}\right\} \leq O_R$$

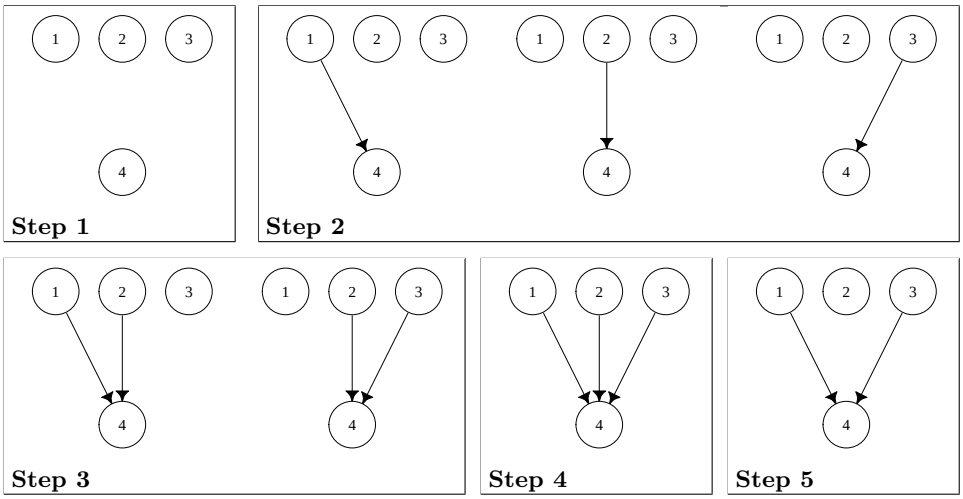
3. There is conclusive evidence for the more complex model. Reject $\mathcal{M}_0(r)$.

$$\log\left\{\frac{p[\mathcal{M}_0(r)\,|\,\mathbf{y}(r)]}{p[\mathcal{M}_1(r)\,|\,\mathbf{y}(r)]}\right\} < -\log(c)$$

For more details see Madigan and Raftery (1994) and Raftery et al. (1997).



(a) MDM-DGM exhaustive search



(b) Stepwise regression

Figure 5.1: A stepwise regression algorithm has the potential to improve accuracy. If at step 2, $\hat{Pa}(r) = \{2\}$, forward selection alone will not be able to identify the model in **(a)** found by an exhaustive search. However, the 'correct' parent set can be found if an additional step (step 5) removes parent 2 after the inclusion of parents 1 and 3.

5.4 Development of Non-Local Priors

In Chapter 4, we considered two candidate non-local priors for the Dynamic Linear Model. However, theoretical and computational considerations at this stage currently prohibit a penalised model search over a network of with enough nodes to be physiologically-interesting. In order to implement a DLM-pMOM prior, we might consider whether it is possible to de-couple the discount factor $\delta(r)$ from the penalty term. We showed empirically in Chapter 4 that the optimal $\delta(r)$ tends to be higher (closer to 1) as the number of parents increases, suggesting that the highly dynamic regression coefficients that will be unduly penalised with a DLM-pMOM prior are likely to be a feature of models with small numbers of parents. In this work, we did not fully explore this behaviour. One possible extension might be to place a prior on the discount factor, so that we may quantify the variance associated with our estimate.

Using the joint distributions of the Dynamic Linear Model, we introduced the DLM-QF non-local prior, focusing on the necessity of specifying an appropriate value for the prior hyperparameter $\mathbf{C}_0^*(r)$. There are a number of ways which we could go about this: the first would be to consider a penalty term proportional to $\prod_{i \neq j}^{p_r} \boldsymbol{\theta}_{i,t>t'}(r)^\top \boldsymbol{\theta}_{i,t>t'}(r)$ and use the first $t' - 1$ time points to train the model. Another option would be to put a hyperprior on $\mathbf{C}_0^*(r)$ in order to formally incorporate our uncertainty about our choice of value, although we note that implementation of such a prior may not be trivial.

References

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Bielza, C. and Larrañaga, P. Bayesian networks in neuroscience: a survey. *Frontiers in Computational Neuroscience*, 8:131, 2014.

Bijsterbosch, J., Smith, S., and Bishop, S.J. Functional connectivity under anticipation of shock: Correlates of trait anxious affect versus induced anxiety. *Journal of Cognitive Neuroscience*, 27(9):1840–1853, 2015.

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995.

Chang, C. and Glover, G. H. Time–frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage*, 50(1):81–98, 2010.

Chen, G., Glen, D. R., Saad, Z. S., Hamilton, J. P., Thomason, M. E., Gotlib, I. H., and Cox, R. W. Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis. *Computers in Biology and Medicine*, 41 (12):1142–1155, 2011.

Chickering, D. M. and Meek, C. Finding optimal Bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 94–102. Morgan Kaufmann Publishers Inc., 2002.

Costa, L. *Studying effective brain connectivity using multiregression dynamic models.* PhD thesis, The University of Warwick, 2014.

Costa, L., Smith, J., Nichols, T., Cussens, J., Duff, E. P., and Makin, T. R. Searching multiregression dynamic models of resting-state fMRI networks using integer programming. *Bayesian Analysis*, 10(2):441–478, 2015.

Costa, L., Nichols, T., and Smith, J. Q. Studying the effective brain connectivity using

multiregression dynamic models. *Brazilian Journal of Probability and Statistics*, 31 (4):765–800, 2017.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

Davison, A. C. *Statistical models*, volume 11. Cambridge University Press, 2003.

Dawson, D. A., Cha, K., Lewis, L. B., Mendola, J. D., and Shmuel, A. Evaluation and calibration of functional network modeling methods based on known anatomical connections. *NeuroImage*, 67:331–343, 2013.

Foster, B. L., Rangarajan, V., Shirer, W. R., and Parvizi, J. Intrinsic and task-dependent coupling of neuronal population activity in human parietal cortex. *Neuron*, 86(2):578–590, 2015.

Fox, M. D. and Raichle, M. E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9): 700, 2007.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Friston, K., Moran, R., and Seth, A. K. Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology*, 23(2):172–178, 2013.

Friston, K. J., Harrison, L., and Penny, W. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

Friston, K. J., Li, B., Daunizeau, J., and Stephan, K. E. Network discovery with DCM. *NeuroImage*, 56(3):1202–1221, 2011.

Friston, K. J., Kahan, J., Biswal, B., and Razi, A. A DCM for resting state fMRI. *NeuroImage*, 94:396–407, 2014.

Friston, K.J., Frith, C.D., Liddle, P.F., and Frackowiak, R.S.J. Functional connectivity: the principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14, 1993.

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Handwerker, D. A., Gonzalez-Castillo, J., D'Esposito, M., and Bandettini, P. A. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *NeuroImage*, 62(2):1017–23, 2012.

Henry, T. and Gates, K. Causal search procedures for fMRI: review and suggestions. *Behaviormetrika*, 44(1):193–225, 2017.

Hindriks, R., Adhikari, M. H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N. K., and Deco, G. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *NeuroImage*, 127:242–256, 2016.

Hinne, M., Janssen, R. J., Heskes, T., and van Gerven, M. A. J. Bayesian estimation of conditional independence graphs improves functional connectivity estimates. *PLoS Computational Biology*, 11(11):e1004534, 2015.

Hocking, R. R. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

Honey, C.J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., and Hagmann, P. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., et al. Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, 80:360–378, 2013.

Hyvärinen, A. and Smith, S. M. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14 (Jan):111–152, 2013.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11(May):1709–1731, 2010.

Johnson, V. E. and Rossell, D. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

Jurasinski, G., Koebsch, F., Guenther, A., and Beetz, S. *flux: Flux Rate Calculation from Dynamic Closed Chamber Measurements*, 2014. URL `https://CRAN.R-project.org/package=flux`. R package version 0.3-0.

Kahan, J. and Foltynie, T. Understanding DCM: ten simple rules for the clinician. *NeuroImage*, 83:542–549, 2013.

Kan, R. From moments of sum to moments of product. *Journal of Multivariate Analysis*, 99(3):542–554, 2008.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Keilholz, S. D. The neural basis of time-varying resting-state functional connectivity. *Brain Connectivity*, 4(10):769–779, 2014.

Keilholz, S. D., Pan, W.-J., Billings, J., Nezafati, M., and Shakil, S. Noise and non-

neuronal contributions to the BOLD signal: applications to and insights from animal studies. *NeuroImage*, 2017.

Kim, S. *ppcor: Partial and Semi-Partial (Part) Correlation*, 2015. URL `https://CRAN.R-project.org/package=ppcor`. R package version 1.1.

Lacerda, G., Spirtes, P. L., Ramsey, J., and Hoyer, P. O. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

Laumann, T. O., Snyder, A. Z., Mitra, A., Gordon, E. M., Gratton, C., Adeyemo, B., Gilmore, A. W., Nelson, S. M., Berg, J. J., Greene, D. J., et al. On the stability of BOLD fMRI correlations. *Cerebral Cortex*, 2016.

Li, B., Daunizeau, J., Stephan, K. E., Penny, W., Hu, D., and Friston, K. Generalised filtering and stochastic DCM for fMRI. *NeuroImage*, 58(2):442–457, 2011.

Madigan, D. and Raftery, A. E. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

Mannino, M. and Bressler, S. L. Foundational perspectives on causality in large-scale brain networks. *Physics of Life Reviews*, 15:107–123, 2015.

Mantini, D., Perrucci, M. G., Del Gratta, C., Romani, G. L., and Corbetta, M. Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104(32):13170–13175, 2007.

Mark, C. I., Mazerolle, E. L., and Chen, J. J. Metabolic and vascular origins of the BOLD effect: Implications for imaging pathology and resting-state brain function. *Journal of Magnetic Resonance Imaging*, 42(2):231–46, 2015.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., and Benali, H. Partial correlation for functional brain interactivity investigation in functional MRI. *NeuroImage*, 32(1):228–237, 2006.

Motzkin, J. C., Philippi, C. L., Wolf, R. C., Baskaya, M. K., and Koenigs, M. Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. *Biological Psychiatry*, 77(3):276–284, 2015.

Mumford, J. A. and Ramsey, J. D. Bayesian networks for fMRI: a primer. *NeuroImage*, 86:573–582, 2014.

Neuwirth, E. *RColorBrewer: ColorBrewer Palettes*, 2014. URL `https://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-2.

O'Hagan, A. *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian inference*, volume 2B. Arnold, 2004.

Penny, W., Ghahramani, Z., and Friston, K. Bilinear dynamical systems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1457):983–993, 2005.

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. Modelling functional integration: a comparison of structural equation and dynamic causal models. *NeuroImage*, 23:S264–S274, 2004.

Pourahmadi, M. Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

Queen, C. M. and Albers, C. J. Intervention and causality: forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association*, 104(486):669–681, 2009.

Queen, C. M. and Smith, J. Q. Multiregression dynamic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 849–870, 1993.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

Raftery, A. E., Madigan, D., and Hoeting, J. A. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., and Glymour, C. Six problems for causal inference from fMRI. *NeuroImage*, 49(2): 1545–1558, 2010.

Ramsey, J. D., Hanson, S. J., and Glymour, C. Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simulation study. *NeuroImage*, 58(3):838–848, 2011.

Ramsey, J. D., Sanchez-Romero, R., and Glymour, C. Non-Gaussian methods and high-pass filters in the estimation of effective connections. *NeuroImage*, 84:986–1006, 2014.

Razi, A. and Friston, K. J. The connected brain: causality, models, and intrinsic dynamics. *IEEE Signal Processing Magazine*, 33(3):14–35, 2016.

Razi, A., Seghier, M. L., Zhou, Y., McColgan, P., Zeidman, P., Park, H.-J., Sporns, O., Rees, G., and Friston, K. J. Large-scale DCMs for resting state fMRI. *Network Neuroscience*, 2017.

Rosa, M. J., Friston, K., and Penny, W. Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, 208(1):66–78, 2012.

Rossell, D. and Telesca, D. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017.

Rossell, D., Cook, J.D., Telesca, D., and Roebuck, P. *mombf: Moment and Inverse Moment Bayes Factors*, 2017. URL `https://CRAN.R-project.org/package=mombf`. R package version 1.9.5.

Roweis, S. and Ghahramani, Z. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

Ryali, S., Supekar, K., Chen, T., and Menon, V. Multivariate dynamical systems models for estimating causal interactions in fMRI. *NeuroImage*, 54(2):807–823, 2011.

Schwab, S., Harbord, R., Costa, L., and Nichols, T. *multdyn: Multiregression Dynamic Models*, 2017a. URL `https://CRAN.R-project.org/package=multdyn`. R package version 1.6.

Schwab, S., Harbord, R., Costa, L., and Nichols, T. *multdyn: Multiregression Dynamic Models*, 2017b. URL `https://CRAN.R-project.org/package=multdyn`. R package version 1.5.1.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct): 2003–2030, 2006.

Shmuel, A. and Maier, A. Locally measured neuronal correlates of functional MRI signals. In *fMRI: From Nuclear Spins to Brain Functions*, pages 105–128. Springer, 2015.

Smith, J. F., Pillai, A., Chen, K., and Horwitz, B. Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. *NeuroImage*, 52(3):1027–1040, 2010.

Smith, J. F., Pillai, A. S., Chen, K., and Horwitz, B. Effective connectivity modeling for fMRI: six issues and possible solutions using linear dynamic systems. *Frontiers in Systems Neuroscience*, 5:104, 2012.

Smith, J. F., Chen, K., Pillai, A. S., and Horwitz, B. Identifying effective connectivity parameters in simulated fMRI: a direct comparison of switching linear dynamic system, stochastic dynamic causal, and multivariate autoregressive models. *Frontiers in Neuroscience*, 7, 2013.

Smith, S. M. The future of FMRI connectivity. *NeuroImage*, 62(2):1257–1266, 2012.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. Network modelling methods for fMRI. *NeuroImage*, 54(2):875–891, 2011.

Soetaert, K. *shape: Functions for Plotting Graphical Shapes, Colors*, 2014. URL `https://CRAN.R-project.org/package=shape`. R package version 1.4.2.

Spirtes, P. Directed cyclic graphical representations of feedback models. In *Proceedings*

of the Eleventh Conference on Uncertainty in Artificial Intelligence, pages 491–498. Morgan Kaufmann Publishers Inc., 1995.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search.* MIT press, 2000.

Sporns, O. Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience*, 17(5):652, 2014.

Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J. Comparing hemodynamic models with DCM. *NeuroImage*, 38(3):387–401, 2007.

Tavor, I., Jones, O. P., Mars, R. B., Smith, S. M., Behrens, T. E., and Jbabdi, S. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

Ugurbil, K. What is feasible with imaging human brain function and connectivity using functional magnetic resonance imaging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1705), 2016.

West, M. and Harrison, P. J. *Bayesian Forecasting & Dynamic Models.* Springer, 2nd edition, 1997.

Xie, Y. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2017. URL `https://yihui.name/knitr/`. R package version 1.17.

Zeki, S. and Shipp, S. The functional logic of cortical connections. *Nature*, 335(6188): 311–317, 1988.

Zilles, K. and Amunts, K. Anatomical basis for functional specialization. In *fMRI: From Nuclear Spins to Brain Functions*, pages 27–66. Springer, 2015.

R Packages

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`

Kim, S. *ppcor: Partial and Semi-Partial (Part) Correlation*, 2015. URL `https://CRAN.R-project.org/package=ppcor`. R package version 1.1

Jurasinski, G., Koebsch, F., Guenther, A., and Beetz, S. *flux: Flux Rate Calculation from Dynamic Closed Chamber Measurements*, 2014. URL `https://CRAN.R-project.org/package=flux`. R package version 0.3-0

Neuwirth, E. *RColorBrewer: ColorBrewer Palettes*, 2014. URL `https://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-2

Rossell, D., Cook, J.D., Telesca, D., and Roebuck, P. *mombf: Moment and Inverse Moment Bayes Factors*, 2017. URL `https://CRAN.R-project.org/package=mombf`. R package version 1.9.5

Schwab, S., Harbord, R., Costa, L., and Nichols, T. *multdyn: Multiregression Dynamic Models*, 2017b. URL `https://CRAN.R-project.org/package=multdyn`. R package version 1.5.1

Soetaert, K. *shape: Functions for Plotting Graphical Shapes, Colors*, 2014. URL `https://CRAN.R-project.org/package=shape`. R package version 1.4.2

Xie, Y. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2017. URL `https://yihui.name/knitr/`. R package version 1.17