Modelling Dynamic Functional Brain Connectivity

Søren Føns Vind Nielsen

Supervisors:
Morten Mørup, DTU Compute
Mikkel N. Schmidt, DTU Compute
Rasmus Røge, DTU Compute
Kristoffer H. Madsen, Danish Research Centre for Magnetic Resonance

M.Sc. Thesis, Master of Science in Engineering
Kongens Lyngby 2015

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Building 303B, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45253031
www.compute.dtu.dk

Abstract

Functional brain connectivity, the statistical dependence between activity in segregated brain regions, has been studied extensively over the last two decades. Most models that describe functional connectivity have parameters in the model that do not change over time, implicitly assuming that functional connectivity is temporally static. Recent research shows that a wealth of information can be gained by modeling functional connectivity in a dynamic setting, i.e. that the brain can be in different states throughout an experiment. We investigated two different non-parametric Bayesian models of dynamic functional connectivity, one simple model with relatively few parameters, and another more complex model with relatively many parameters. Both were based on the infinite hidden Markov model to model transitions between brain states. We investigated how model complexity can affect the number of states extracted from data, and how that affects our interpretation of dynamic functional connectivity. We first conducted several synthetic experiments, generating data from the two models considered and afterwards ran inference by Markov chain Monte Carlo on the same data. The aim of this was to study the behaviour of the models in a setting where there was a clear model-mismatch. Furthermore, we investigated whether the models were able to characterize task and resting state functional magnetic resonance imaging (fMRI) data from the Danish Research Center for Magnetic Resonance (DRCMR) and from the Human Connectome Project (HCP). On synthetic data we showed that the simple model found many states on data generated from the complex model, but that the complex model was able to find the true number of states in data from the simple model. We found that the more complex model with only one state could characterize real-world data better than the simple model that found evidence for multiple states. The fact that the complex model only found one state in real-world data contradicts our intuition that multiple brain states should be present, but this could be explained by the dimensionality reduction carried out in this project. The results of this thesis indicate that one must always interpret dynamics in functional connectivity in terms of the model used and especially its limitations. We suspect that preprocessing and dimensionality reduction has a huge impact on the conclusions that can be drawn. This should be investigated further.

Keywords: Dynamic Functional Connectivity, Functional Magnetic Resonance Imaging, Bayesian Non-parametric Modeling, Human Connectome Project

Preface

The thesis was written as part of obtaining a masters degree in Mathematical Modeling and Computation from the department of Applied Mathematics and Computer Science at the Technical University of Denmark.

Kgs. Lyngby, June 26th, 2015

Søren Føns Vind Nielsen
(s103226)

Acknowledgements

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

I thank my supervisors, Rasmus Røge, Mikkel N. Schmidt, Kristoffer H. Madsen and Morten Mørup for their support, inspiration and catching enthusiasm for science. They have all been an enormous help throughout the project.

Contents

Chapter 1

Introduction

The human brain has been studied in centuries due to its magnificence and complexity, and we are by no means done with this investigation. As our technology has become more advanced we have come up with advanced, both invasive and non-invasive, tools to measure how the brain looks and works. Historically, functional magnetic resonance imaging (fMRI) has been used as a technique to investigate the activity of different brain regions in a non-invasive manner (Ogawa et al. [1992], Kwong et al. [1992]). A large portion of the studies in the last decade that analyse fMRI data have been focused on the synchronous activity of spatially different regions in the brain, this field being termed *functional connectivity*. It has for instance been shown that functional brain networks of connectivity extracted from fMRI can be used as biomarkers for diseases such as schizophrenia or Alzheimers (cf. Calhoun et al. [2009], Buckner et al. [2009]). But a problem with most models used in the past is that they most likely represent an oversimplification of true underlying physical phenomena, since they explicitly or implicitly assume that parameters governing the functional interaction between brain regions do not change over time (cf. Hutchison et al. [2013]). We can imagine that this is not the case - most brains regions most likely do not interact in the same way during sleep as they do during a stressful examination in advanced statistics. We thus need statistical models that can account for the variability and change of parameters over time, i.e. *dynamic* models. This thesis will consider the problem of modeling dynamic functional brain connectivity using Bayesian non-parametric statistics.

1.1 Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging technique that started seeing use in the 1990's. fMRI indirectly locates areas of the brain that are active by identifying the oxygen-needs of groups of neurons, i.e. the rationale here is that neurons that need oxygen are also active (cf. Stippich et al. [2007]). The measured signal is often referred to as the blood oxygen level dependent (BOLD) signal (cf. Ogawa et al. [1992] and Kwong et al. [1992]). After a stimulus is applied to a neuron, oxygenated blood runs to the area around it, but there is a delay from the onset of the stimuli and until the BOLD signal reaches its peak. The phenomenon is called the *hemodynamic response*, and the function describing the BOLD signal as a function of time after an onset of a stimuli is called the *hemodynamic response function* (HRF). The HRF is modelled often as a sum of two gamma functions and assumed to be the same over the entire brain but that is a simplification. A lot of research is focused on modeling the HRF differently over subjects and areas in the brain (cf. for instance Gössl et al. [2001]).

The key thing that makes these measurements possible is the magnetic properties of haemoglobin. Oxygenated haemoglobin is paramagnetic (and thus attracted by magnetic forces) whereas de-oxygenated haemoglobin is diamagnetic (and thus repelled by magnetic forces). Using a strong magnetic field (like the magnet inside an MRI machine) can align the magnetic moments of all molecules forming the basis for a measurable signal. In an fMRI scan the brain is partitioned into *voxels* (small cubes) where the BOLD signal is extracted from each voxel. The spatial resolution is a term used to define how many voxels that have been used, whereas the temporal resolution defines how often we can capture the measurements. Compared to many other neuroimaging methods fMRI has a good spatial resolution (in some cases the voxel size is 1 mm^3) but relatively worse temporal resolution (especially compared to electroencephalography (EEG)). Ogawa et al. [1992] and Kwong et al. [1992] were the first teams to use fMRI to show appropriate brain activity patterns when subjects were asked to clench their hands and look into flashing lights respectively. This was a huge victory for this non-invasive method which paved the way for new neuroimaging research.

1.2 Functional Connectivity and the Default Mode Network

As mentioned, fMRI saw its birth in the early 1990's and was in the beginning used exclusively for identifying areas of activity associated with a specific brain function, called *functional segregation*. The interest here is localizing brain function, whereas *functional integration* is the concept of how spatially segregated brain regions interact in a given mental state. But as Friston [2011] points out *"functional segregation is only meaningful in the context of functional integration and vice versa"*. What this means is that we cannot talk about localized brain function without also analyzing the relation to other segregated regions, and we cannot describe an interaction between brain regions without defining the regions of functional segregation. This brings us to the term *functional connectivity*, which by Friston [2011] is defined as the *"statistical dependencies among remote neurophysiological events"*. Thus if we with some statistical power can establish activity in two (or more) spatially separated brain regions we say that they are functionally connected given the circumstances of the experiment. Over the last decade functional connectivity (FC) studies have grown in number, and the scientific questions we can ask increase in complexity (cf. Smith [2012]). Commonly, the correlation coefficient between the BOLD time series of two regions has been used as a measure of statistical dependence, from which we can extract weighted networks of FC. One of the most studied brain networks is the default mode network (DMN) associated with and found in resting-state brain data. Raichle et al. [2001] searched for a baseline brain-state and found, using fMRI and positron emission tomography (PET), that a number of areas consistently decreased in activity during a variety of task experiments, thus defining the DMN. The DMN has from that point on been studied at length, for instance by Greicius et al. [2004] in the context of diagnosing Alzheimers disease (AD). They showed in a motor-task experiment with 26 subjects, 13 with AD and 13 healty, that the AD group displayed decreased connectivity in parts of the DMN compared to the control group, thus yielding a potential non-invasive biomarker for AD.

1.3 Effective Connectivity

Friston [2011] points out that neurophysiological events and activity as measured in fMRI can arise from a number of factors not directly linked to a functional meaning. Therefore he argues that *effective connectivity*, defined as *"the influence that one neuronal system exerts over another"*, is a more appropriate concept to analyse. Harrison et al. [2003] investigated the use of a multivariate autoregressive (VAR) process to describe fMRI data and to model effective connectivity. A directed network was extracted to show what connections between brain regions existed and in what direction the connectivity was present. The whole framework was adopted from a fully Bayesian approach, allowing model order selection from Bayesian evidence. Assuming that the data was generated from a VAR(p) process, i.e. a VAR process dependent on the previous p time points, each connection's VAR-coefficients were tested if they were significantly non-zero. The p-values from these tests were used as strengths in the directed network extracted. Further research of effective connectivity resulted in Friston et al. [2003] publishing the famous dynamic causal model (DCM), a differential equation model embedded in the Bayesian framework. In the DCM, neuronal activity is modelled as a continuous latent variable and the observed signal as a non-linear transformation of the neuronal activity. A very desirable property of the DCM is that the hemodynamic response function (HRF), is directly accounted for in the non-linear transformation. Inference in the model is carried out using variational Bayes, specifically by minimizing the free energy, and model comparison can easily be carried out in this Bayesian setting by the evidence of the models in consideration. Shortly after Harrison et al. [2003] presented their VAR framework, Goebel et al. [2003] presented a method for analyzing the Granger causality (sometimes called G-causality) of two multivariate time series, i.e. whether the prediction on one time series can be improved by including the other in the model. The method compares the covariance estimates from three VAR models; two models trained on the two time series at hand and a third model trained on the stacked time series. This gives a measure of how much covariance that can be "gained" by modelling the two time series together. The DCM and G-causality model have been the two major models describing effective connectivity throughout the 2000's. Both methods have been criticized in the literature, DCM for the lack of robustness of the variational inference and G-causality for the lack of hemodynamic response modelling (cf. Stephan and Roebroeck [2012]).

1.4 Dynamic Functional Connectivity

Previously, almost all studies have either implicitly or explicitly assumed temporal stationary integration between segregated brain areas during the scan period. But as pointed out by many, this might be a simplification of the true underlying process, and intuitively it would make sense that brain regions interact in different ways at different times. Hutchison et al. [2013] present in their review paper of recent research that multiple authors find it of increasing importance to model functional brain connections dynamically. One very popular way to model temporal dynamics of the BOLD signal is by a sliding window approach. Each time-series is windowed and functional connectivity (FC) measures and models are calculated on each of the windows extracted. An example of this can be seen in figure 1.1, where the correlation is used as a measure of FC yielding correlation matrices.

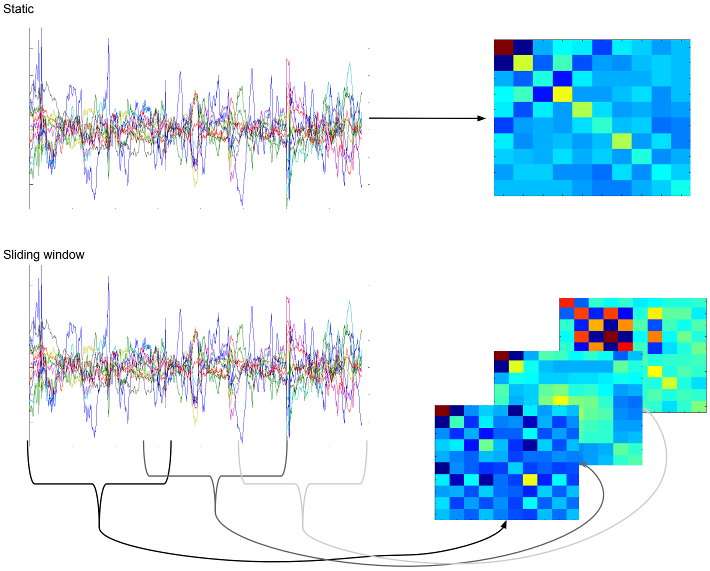


Figure 1.1: An illustration of the sliding window approach to extract correlation matrices from subsequences of a mulitvariate signal. In the static approach, the correlation matrix is calculated based on the entire time series. In the sliding window approach (applied to the same time series), the time series is divided into subsequences (in this case 3) of a fixed length (called the window length), and the correlation matrix is extracted from each of the subsequences.

Allen et al. [2012] used the group independent components (IC) (cf. Calhoun et al. [2001]) from 405 subject's resting state data to create correlation matrices from each subsequence extracted by windowing the IC time-series. The upper triangular part of the covariance matrix was stacked into a vector, and k-means clustering was performed on all the correlation matrices extracted from all windows and subjects. The conclusion was that some of the cluster centroids, i.e. archetypal brain networks, were identified with the traditional DMN and some with previously unanticipated functional connectivity patterns. Previous studies analyzing the DMN by Kiviniemi et al. [2011] suggested partly the same conclusion, namely that the DMN exhibits spatial variability over time.

Another dynamic sliding window approach was investigated by Yu et al. [2015] to distinguish between schizophrenic (SZ) patients and a healthy control group. They extracted windowed correlation matrices from spatial group ICA using only IC's pertaining to physiological meaning. Looking at the correlation matrices as a time evolving graph, a number of network statistics (such as the clustering coefficient) were calculated. These quantities were then used to statistically test the dynamics of the connections. They found that the network statistics considered had a significantly lower variance in the SZ group compared to the heatly control group, which could be useful for characterizing and diagnosing the disease.

Zalesky et al. [2014] used windowed correlation matrices to test the pairwise functional stationarity of all regions in the Automated Anatomical Labeling (AAL) atlas. They used a stable two-dimensional VAR model for each connection, fitted it to the true data, and generated a new dataset from the VAR model, called the null data, i.e. a data set that satisfied the null hypothesis of stationarity. The true data was tested against the null data for each of the 6670 connections repeated over 10 subjects, and the null hypothesis was rejected on average for around 300 connections. The top-100 dynamic connections were analyzed further and were found to be fairly consistent over subjects, suggesting a modular functional structure in the brain were the large scale organization is static and that a few connections are dynamic.

A drawback of using the sliding window approach is influence of window length. One way to overcome this is by using more advanced frequency analysis methods. Chang and Glover [2010] used a wavelet transform analysis to generate two-dimensional maps of the BOLD-signal correlation between two regions of interest in both time and frequency. They showed that the posterior cingulate cortex (PPC), normally associated with the DMN, varied in correlation with other regions outside the DMN in both time and frequency, suggesting a dynamic behaviour of the PPC.

Stemming from the viewpoint that a significant dynamic connectivity pattern is one that is repeated during recording, Majeed et al. [2011] investigated a novel method to detect recurring functional configurations in rat and human brains. Starting from a random initial time point, a subsequence with a user defined window length is extracted as a template for the recurring pattern. This template is then alternatingly updated in the following two steps; 1) Sliding window correlation between the template and the original sequence (across all regions) is calculated and timepoints above a certain threshold are identified. 2) The template is updated by averaging the identified timepoint's spatial maps. The authors found that the recurring patterns were identifiable in the data analysed and that the maps found were robust toward choice of window length.

In work by Tagliazucchi et al. [2012] it was pointed out that the networks extracted from previous research, i.e. DMN or task positive network (TPN), are dominated by a few time-point measurements. Liu and Duyn [2013] developed a framework to utilize this and find single time instances of spontaneous activity resembling these networks, rather than blurring out these configurations by averaging over time. In their approach a single-volume is considered as a seed region, and by thresholding the activity in that particular volume, time points of interest are identified. From these time-points the activity from all voxels was collected and stacked into vectors, and afterwards a k-means clustering was performed with the correlation distance. All instances from the same cluster were averaged yielding k so-called co-activation patterns (CAP's). Notice here that the threshold level can be used to go from very fine-grained spontaneous activity (high threshold) to a more averaging based approach (low threshold). They confirmed that only a few time-points dominate the networks known from literature (DMN and TPN).

1.4.1 Validation of Found Dynamics

In the preceding section we have described multiple ways of finding dynamic functional connectivity and connectivity states. But now the question is how do we validate that they have a physiological meaning? Hutchison et al. [2013] reviews some of the efforts that have been made in this direction and two of the frameworks will be highlighted here. The first framework is based on having a simultaneous measurement in another modality, such as EEG (cf. Duyn [2012]) or local field potential (LFP) (cf. Schölvinck et al. [2010]). If the connectivity networks extracted from fMRI are somehow consistent with time-evolving networks extracted from the other modality, we can with higher certainty conclude that dynamics are present. The second framework to validate dynamics comprises correlation of the functional connectivity patterns with a behavioural response from a task-experiment for instance. Thompson et al. [2013] showed that a high anti-correlation between the DMN and a task network a few seconds before the task stimuli was predictive of faster reaction time by the subject. This means that the BOLD dynamics can be validated if a 'ground truth' is available, i.e. some human behaviour that can be explained by the networks extracted.

1.5 Project Outline

In this master thesis we will investigate different models for modelling dynamic functional connectivity. The models considered will be extensions of already existing state-of-the-art frameworks for this type of analysis (i.e. Allen et al. [2012], Zalesky et al. [2014], Korzen et al. [2014]). The extensions will be based on Bayesian non-parametric methods to avoid choosing certain parameters in the existing models, especially the number of states. We will in particular consider using VAR models to model a filtering process of the brain signal, and covariance matrices to model functional connectivity patterns. The dynamics, i.e. switching from one state to another, will be modelled as a non-parametric hidden Markov model as described in Van Gael [2012].

Two models will be analysed, one where only the covariance of the signal is dynamic and another where both the covariance and an accompanying VAR process can change over time. A study of synthetic data from both models, will answer how the models behave on data generated from a different model. We will thereby analyse what the consequences are of choosing a simple model (in terms of parameters) for a complex problem and vice versa. It is suspected that the ′simple′ model will find too many states, and therefore is relatively worse to characterize the ′dynamics′ of the data compared to a more ′complex′ model.

Finally, the models will be applied to real world data. Data from the Danish Research Center for Magnetic Resonance (DRCMR) and from the Human Connectome Project (HCP)(cf. Van Essen et al. [2012]) is available throughout the project. We wish to validate the use of dynamic models on real-world data by quantifying with the predictive likelihood how well the models capture dynamics in previously unseen data. We have multiple task-experiment data sets from multiple subjects and expect the dynamics to be different from task to task, which should be reflected in the predictive likelihood.

Korzen et al. [2014] showed results that indicated functional connectivity dynamics not being shared over subjects, i.e. that each subject displayed its own brain states not found in other subjects. In this project, we will therefore mainly focus on modelling single-subject brain dynamics. We expect that a model fitted to one task should perform well in terms of predictive likelihood on unseen data from the same task carried out by the same subject.

The main research questions can thus be formulated as follows,

- How can functional brain dynamics be modelled in terms of non-parametric Bayesian statistics?
- How does the choice of model influence the interpretation of dynamic functional connectivity?
- Can the models be used to characterize brain states in real-world data from single-subject simple task-based fMRI studies?

The thesis will have the following structure. In chapter 2 we will present all the necessary methods for modeling dynamic functional connectivity, including a detailed description of the models we will use. In chapter 3 we will present the real-world data analysed and some of the preprocessing that was carried out. The main results of the thesis will be presented in chapter 4, both from synthetic and real-world experiments. In chapter 5 the research questions will answered and discussed. Finally, the thesis will be summarized briefly in chapter 6.

Chapter 2

Theory

In this chapter we present the methods and models used to analyse functional connectivity in a dynamic setting. In section 2.1 the vector autoregressive (VAR) model will be described, followed by a description in section 2.2 of an extension into a mixture of VAR models. In section 2.3 we briefly introduce the hidden Markov model, a general framework for sequential data, that will be necessary to understand its non-parametric extension described in detail in section 2.4. In this section we will delve into two observed data models, namely an inverse Wishart mixture and a mixture of VAR's. In section 2.5 we briefly describe the switching vector autoregressive model proposed by Willsky et al. [2009]. In section 2.6 we describe how we estimate the parameters in the generative models by Markov chain Monte Carlo sampling. Finally, in section 2.7 we will describe a general framework for predicting on new data given the models earlier described.

2.1 Vector Autoregressive Model

The vector autoregressive (VAR) process is a model for multidimensional signals that depend directly on their past values (cf. Kirchgässner et al. [2012] for an introduction to AR and VAR models). It has seen use in many applications in economics and neuroscience, and in the latter the VAR model has been used both for modeling effective connectivity (cf. Goebel et al. [2003]) and for modeling the noise process in fMRI data (cf. Lund et al. [2006]). Formally we write that the P-dimensional signal at time $t = 0..T-1$, $\mathbf{x}_t$, follows the VAR-equation,

$$\mathbf{x}_t = \left(\sum_{\tau=1}^{M} \mathbf{A}_\tau \mathbf{x}_{t-\tau} \right) + \boldsymbol{\epsilon}_t, \tag{2.1}$$

in which M is the model order (i.e. how many past values we use to regress on the present), $\mathbf{A}_\tau$ is a matrix of size $P \times P$ containing the lag specific coefficients, and $\boldsymbol{\epsilon}_t$ is the P-dimensional noise (sometimes called the *innovation*). Often statistical assumptions are made on the expectation of both the signal and the covariance of the noise when estimating the coefficients in the model, for instance a mean zero signal and no covariance between two successive innovation terms, i.e. white noise. Collecting all model parameters in one large matrix can greatly simplify the later least-squares estimation of aforementioned parameters, i.e. rewriting (2.1) and disregarding the noise yields,

$$\mathbf{X} = \mathbf{A}\bar{\mathbf{X}}, \tag{2.2}$$

where $\mathbf{X}$ is a $p \times (T - M)$ matrix, $\mathbf{A}$ is a $p \times (p \cdot M)$ matrix and $\bar{\mathbf{X}}$ is a $(p \cdot M) \times (T - M)$ matrix.

These are defined as follows,

$$\begin{aligned}
\mathbf{X} &= [\mathbf{x}_M \; \mathbf{x}_{M+1} \; \cdots \; \mathbf{x}_T] \\
\mathbf{A} &= [\; A_1 \; A_2 \; \cdots \; A_M] \\
\bar{\mathbf{X}} &= \begin{bmatrix} \mathbf{x}_{M-1} & \mathbf{x}_M & \cdots & \mathbf{x}_{T-1} \\ \mathbf{x}_{M-2} & \mathbf{x}_{M-1} & \cdots & \mathbf{x}_{T-2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{T-M} \end{bmatrix}.
\end{aligned}$$

The model parameters can for instance be found using the Moore-Penrose inverse (cf. Penrose [1955]) in (2.2). If we denote the Moore-Penrose inverse of a matrix $\mathbf{Y}$ as $\mathbf{Y}^\dagger$, the solution to (2.2) becomes,

$$\mathbf{A} = \mathbf{X}\bar{\mathbf{X}}^\dagger = \mathbf{X}\bar{\mathbf{X}}^T \left(\bar{\mathbf{X}}\bar{\mathbf{X}}^T \right)^{-1}. \tag{2.3}$$

2.2 Mixture of Vector Autoregressive Models

One simple way to model time dynamics in fMRI data is to use a mixture of k vector autoregressive (VAR) models. This following framework is considered only as a stepping stone to the more advanced non-parametric models that are the cornerstone of this project. Each VAR process can be formally written as,

$$\mathbf{f}_k(t) = \left(\sum_{\tau=1}^{M} \mathbf{A}_\tau^{(k)} \mathbf{x}_{t-\tau} \right), \quad M \leq t \leq T \tag{2.4}$$

in which the signal $\mathbf{x}$ is assumed to have zero mean in expectation and $\mathbf{A}_\tau^{(k)}$ is the model coefficient matrix for the k'th process at time lag τ. Viewing this as a discrete latent variable model, a simple generative model can be written as follows,

$$\mathbf{z} \sim \mathrm{Cat}(\boldsymbol{\pi}, K), \tag{2.5}$$

$$vec(\mathbf{A}^k) \sim \mathcal{N}(0, \mathbf{R}), \tag{2.6}$$

$$\sigma_t^2 \sim \mathcal{G}^{-1}(\beta_1, \beta_2), \tag{2.7}$$

$$\mathbf{x_t} \sim \mathcal{N}(\mathbf{A}^{(z_t)} \bar{\mathbf{x}}_t, \sigma_t^2 \mathbf{I}), \tag{2.8}$$

in which Cat(...,K) is the K-categorical distribution, $\boldsymbol{\pi}$ is a vector of mixing coefficients, $vec(\cdot)$ is the vectorization operator (i.e. stacking values in a vector), $\mathcal{N}$ is the multivariate normal distribution, $\mathbf{R}$ is a diagonal $PPM \times PPM$-matrix defined as the kroenecker product $\mathbf{R} = \mathbf{R}_\tau \otimes \mathbf{I}_{PP}$, where $\mathbf{R}_\tau$ is a vector of length M containing the prior variances on the time lags of $A^{(k)}$, $\mathcal{G}^{-1}$ is the Inverse-Gamma distribution and $\bar{\mathbf{x}}_t$ the corresponding past to time t stacked into a vector. A detailed description of the expectation-maximization inference procedure implemented in this project is described in appendix A.1. This model has the downside that heuristics must be enforced to choose the number of components K and it does not take into account the obvious sequential structure of the data in the clustering. Thus the next section will briefly describe a general class of models that incorporate time-dependencies.

2.3 Hidden Markov Models

A hidden Markov model (HMM) is a special case of the latent state-space model, in which the observed data is assumed to have a latent representation with discrete values, called the *state-sequence* $\mathbf{z} = \{z_1, z_2, ..., z_N\}$ (cf. [Bishop et al., 2006, chapter 13]). The model is analogous to a mixture model (or a clustering), with the exception that the cluster assignments or state values are dependent on the previous observation (in time). Formally, we say the state-sequence is 1st order Markovian, i.e. that $p(z_n|z_1, z_2, ..., z_N, \theta) = p(z_n|z_{n-1}, \theta)$, in which θ represents all relevant model parameters. This means that the conditional distribution over the state assignment of one datum is only conditioned on the previous datum and its state assignment. A sketch of the model can be seen in figure 2.1. As with the mixture model described in the previous section, an EM-approach, called the *forward-backward* algorithm (cf. Rabiner [1989]), is most commonly used for the inference. But this approach again has the downside that the number of potential states in the model must be chosen by some heuristic. In the next section we describe a non-parametric extension of the HMM that can learn the number of hidden states from the data.

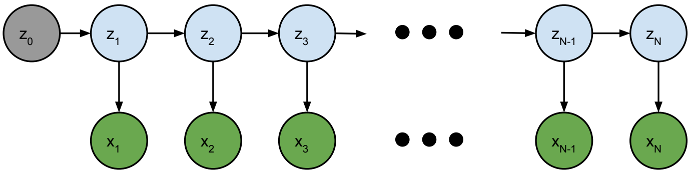


Figure 2.1: A schematic of the general hidden Markov model (HMM). Each observed datum x_n is assumed to be emitted from a latent state space value given by z_n, which is discrete.

2.4 The Infinite Hidden Markov Model

The infinite hidden Markov model (IHMM), first proposed by Beal et al. [2001], can be summarized by its generative representation,

$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma), \tag{2.9}$$

$$\pi^{(\mathbf{k})}|\boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \tag{2.10}$$

$$z_t|z_{t-1} \sim \mathrm{Multinomial}(\pi^{(\mathbf{z_{t-1}})}), \tag{2.11}$$

$$\theta_k \sim H \tag{2.12}$$

$$x_t \sim F(\theta_{z_t}) \tag{2.13}$$

in which GEM is the stick-breaking construction (cf. Sethuraman [1994], Pitman [2002]), DP is the Dirichlet process, γ and α are positive hyperparameters controlling the state sequence (cf. section 2.4.1), H is a distribution over the state specific parameters and F is the distribution of the observed sequence. The graphical model corresponding to this is given in Figure 2.2.

The Dirichlet process defines a probability distribution over a random probability measure, and has been used extensively in non-parametric Bayesian mixture models as a prior over the mixture components. Blackwell and MacQueen

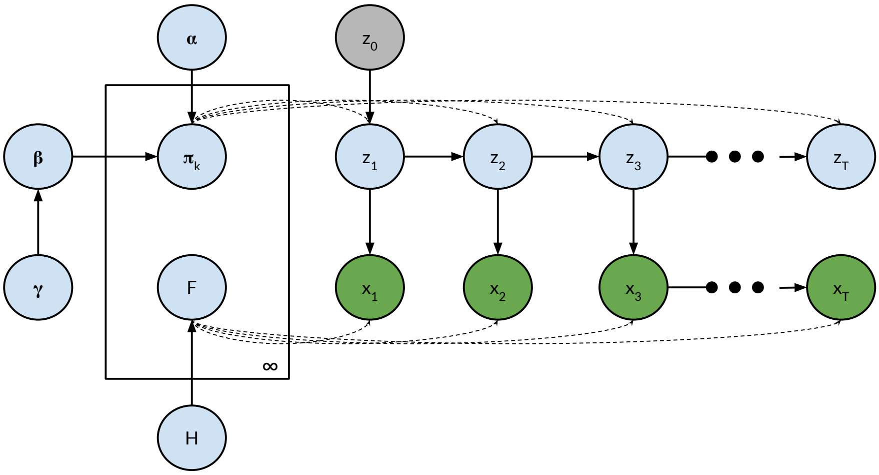


Figure 2.2: The Infinite Hidden Markov Model represented as a graphical model. Note here an abuse of notation - H and F are distributions, not parameters.

[1973] described this process using a Pólya urn scheme. In a clustering setting, we could represent each data point's assignment as a coloured ball, the ball being the data point and the color its clustering assignment. All balls are kept in an urn, and when drawing from the urn (assigning a new data point to a cluster) we draw a color proportional to the number of balls with that color. Afterwards, we place the ball back in the urn together with a new ball of same color. Furthermore, we draw a ball with a previously unseen color with probability proportionally to the positive parameter α, and a ball of the new color is added to the urn. This discrete clustering, with a potentially infinite number of clusters, is also know as the Chinese restaurant process (cf. Aldous [1985]). The GEM or the stick-breaking construction distrbution has been shown by Sethuraman [1994] to be equivalent to a DP, but we will in accordance to Van Gael [2012] keep using the GEM-notation as the distribution over $\boldsymbol{\beta}$. In short, imagine that we have a stick of length 1 that we break into smaller pieces from the end of the stick, where each piece corresponds to a probability mass for a certain cluster. The end of the stick is the mass associated with generating a new cluster. Thus the stick represent a probability measure over the integers.

However, the IHMM introduces a prior over the parameters in the DP, namely a stick breaking construction (GEM). Since the GEM is equal to a DP, we can interpret this as a hierarchy of DP's, the GEM as the root node and a potential for infinitely many DP's in the layer below. As we can see the elements of the state sequence, z_t, can be generated independently from the observed data, x_t, and in the following section it will be described how one can sample a state sequence from a hierarchy of DP's.

2.4.1 Hierarchical Polya Urn Scheme

Teh et al. [2006] showed that the generative process described by (2.9), (2.10) and (2.11), is equivalent to a hierarchical Polya urn scheme. Since this is a more intuitive description of a state sequence with a potentially infinite number of distinct state values, this section will be dedicated to its details and relation to the generative model described above (cf. Van Gael [2012] for more details on this subject). In the hierarchical Polya urn scheme we imagine that we have an infinite number of urns with colored balls in them, where each color represents a state. Each urn represents the transitions from each state, i.e. each urn also has a color and the balls in the urn represent the transition counts out of that state. Furthermore, we introduce an *oracle* urn that controls the average distribution of states. In each time step we sample a color, z_t, based on the previous ball color, z_{t-1}, by querying either the urn of color z_{t-1} or the oracle urn. If we query the z_{t-1}-urn, the new ball is dropped in that urn, and if we query the oracle, a ball of the new color (z_t) is dropped in both the oracle urn and the z_t-urn.

The transition probability is given as,

$$p(z_t = j | z_{t-1} = i, \alpha) = \frac{n_{ij}}{\sum_{j'} n_{ij'} + \alpha} \tag{2.14}$$

in which n_{ij} denotes the number of balls with color j in the i'th urn, and α is a positive concentration parameter that controls how often we query the oracle urn. The probability of querying the oracle urn becomes

$$p(oracle | z_{t-1} = i, \alpha) = \frac{\alpha}{\sum_{j'} n_{ij'} + \alpha}. \tag{2.15}$$

Given that we have queried the oracle, the transition probability becomes,

$$p(z_t = j | z_{t-1} = i, oracle, \gamma) = \begin{cases} \frac{c_j}{\sum_{j'} c_{j'} + \gamma}, & j = \text{existing color} \\ \frac{\gamma}{\sum_{j'} c_{j'} + \gamma}, & j = \text{new color} \end{cases} \tag{2.16}$$

in which c_j is the number of balls with color j in the oracle urn. Note here that in the IHMM this fraction $\frac{c_j}{\sum_{j'} c_{j'} + \gamma}$ is replaced by the stick-breaking construction (i.e. some β-value between 0 and 1). This means that if α is high we will tend to query the oracle urn more often and therefore arrive at a sequence which is distributed according to the stick. In contrast if α tends to zero one state will gain all the mass. The parameter γ controls how often we add a new color. A schematic of the sampling process can be seen in Figure 2.3.

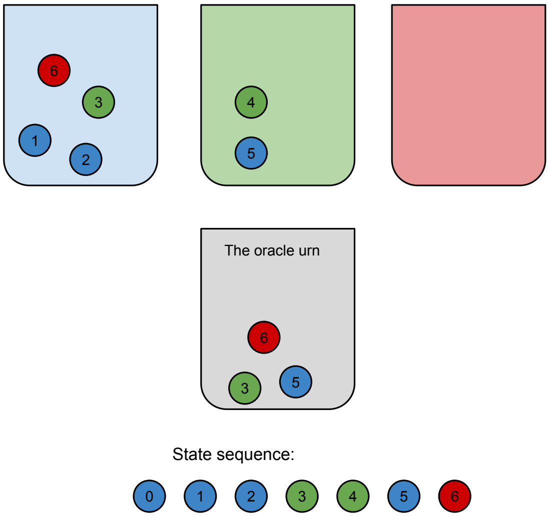


Figure 2.3: A schematic of sampling a state sequence from a hierarchical Pólya urn scheme. The state sequence $\mathbf{z} = \{z_1, z_2, z_3, ...\}$ transformed into colors can be written as $\{blue, blue, green_o, green, blue_o, red_o\}$, where $color_o$ represents a ball drawn after querying the oracle urn. The first ball z_0 is not counted in the state sequence but is a neccesary starting point, and the color can be arbitrarily chosen.

2.4.2 Normal-Inverse-Wishart Model

To complete the IHMM model we need to specify the observed data likelihood, F, and distribution over relevant latent parameters H. To use the framework presented in section 2.4, we must choose F and H to be conjugate to each other so that we can marginalize over the cluster specific parameters from H (more on this in section 2.6.5). Korzen et al. [2014] proposed a model for fMRI, where each time point is drawn from a normal distribution with a certain covariance structure drawn from an inverse Wishart distribution. Each time-point belongs to a cluster and each cluster has its own covariance structure. A CRP was used as a prior over the clustering and there was therefore no time dependencies is then clustering.

We extend this model to accommodate a latent time dependency, by embedding the model in the IHMM framework. In the context of fMRI, we can interpret this as a dynamic functional connectivity model, where functional connectivity patterns are given by corresponding covariance matrices that change over time according to the latent state sequence. The generative model for the observed data, extending (2.9), (2.10) and (2.11), can be written as

$$\Sigma^{(k)} \sim \mathcal{W}^{-1}(\eta\Sigma_0, v_0), \tag{2.17}$$

$$x_t \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \Sigma^{(z_t)}), \tag{2.18}$$

in which $\mathcal{W}^{-1}(\Sigma_0, v_0)$ is the inverse Wishart distribution with probability density function,

$$p(\Sigma|\Sigma_0, v_0) = \frac{|\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})}|\Sigma|^{\frac{-(v_0+p+1)}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\Sigma_0(\Sigma)^{-1}\right)\right).$$

Here Γ_p is the p-variate gamma function, v_0 is the degrees of freedom which in this project is fixed at $v_0 = p$ and $\operatorname{tr}(\cdot)$ is the trace of a matrix, i.e $\operatorname{tr}(A) = \sum_i A_{ii}$. The role of Σ_0 in the context of functional connectivity, can be thought of as the default connectivity present in the data. The parameter η is a positive scaling parameter which we intend to learn by Metropolis-Hastings sampling (cf. section 2.6.2). Also we allow for a time specific scaling of the covariance structure, σ_t^2, which can be interpreted as a noise parameter, to overcome inferring the same structure on different scales in two states. In the context of fMRI we could imagine that there are non-physiological noise artefacts such as spikes or drift that can corrupt the signal, and we hope to better model this with a time-dependent noise parameter.

We will show momentarily how we can marginalize the noise covariance $\Sigma^{(k)}$ out of the joint likelihood. This means that we can obtain collapsed sampling of the state sequence in the later inference (cf. 2.6.1 and further), i.e. we augment sampling any other cluster specific parameters. Collecting all hyperparameters in $\boldsymbol{\theta} = \{\sigma_t^2, v_0, \eta, \Sigma_0\}$, the joint likelihood of the model can be written as,

$$\begin{aligned} p(\mathbf{X}, \boldsymbol{\Sigma}|\mathbf{z}, \boldsymbol{\theta}) = & \prod_k \frac{|\Sigma_0|^{\frac{v}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})}|\Sigma^{(k)}|^{\frac{-(v_0+p+1)}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\Sigma_0(\Sigma^{(k)})^{-1}\right)\right) \\ & \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}}|\Sigma^{(z_t)}|^{\frac{-1}{2}} \exp\left(-\frac{1}{2}x_t^T(\sigma_t^2\Sigma^{(z_t)})^{-1}x_t\right) \\ = & \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k \frac{|\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})}|\Sigma^{(k)}|^{\frac{-(v_0+p+1+n_k)}{2}} \\ & \exp\left(-\frac{1}{2}\left(\operatorname{tr}\left(\Sigma_0(\Sigma^{(k)})^{-1}\right) + \sum_{t:z_t=k} x_t^T(\Sigma^{(k)})^{-1}x_t\right)\right), \end{aligned} \tag{2.19}$$

in which n_k is the number of time points assigned to state k, and the time specific noise parameters σ_t^2 have been multiplied onto each $\mathbf{x}_t$.

Utilizing that

$$x_t^T(\Sigma^{(k)})^{-1}x_t = \sum_{i,j}(xx^t)_{ij}(\Sigma^{(k)})^{-1}_{ij} = \text{tr}\left((xx^t)(\Sigma^{(k)})^{-1}\right),$$

we can rewrite (2.19) to be

$$p(\mathbf{X}, \mathbf{\Sigma}|\mathbf{z}, \boldsymbol{\theta}) = \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k \frac{|\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})}|\Sigma^{(k)}|^{\frac{-(v_0+p+1+n_k)}{2}}$$
$$\exp\left(-\frac{1}{2}\,\text{tr}\left(\left(\Sigma_0 + \sum_{t:z_t=k} x_t x_t^T\right)(\Sigma^{(k)})^{-1}\right)\right) \quad (2.20)$$

Marginalizing out $\mathbf{\Sigma}$ we can arrive at

$$p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{\Sigma}|\mathbf{z}, \boldsymbol{\theta})d\mathbf{\Sigma}$$
$$= \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k \frac{|\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})} \frac{2^{\frac{(v_0+n_k)p}{2}}\Gamma_p(\frac{v_0+n_k}{2})}{|\Sigma_0 + \mathbf{X}^{(k)}|^{\frac{v_0+n_k}{2}}}, \quad (2.21)$$

in which $\mathbf{X}^{(k)} = \sum_{t:z_t=k} x_t x_t^T$.

2.4.3 Multiple-State Vector Autoregressive Model

Willsky et al. [2009] and Fox [2009] proposed a switching vector autoregressive model (described later in section 2.5), where the signal we model is assumed to be generated by a number of VAR's that switch on and off in different time points (one at a time). Following this, another way of completing the IHMM is by assuming that each state can be represented by a VAR-process with an accompanying noise covariance. In an fMRI context, we imagine that each state in the dynamic signal is characterized by instantaneous activity patterns, modelled by the 'noise' covariance $\Sigma^{(k)}$. Afterwards the signal is filtered and distributed over time to other regions, modelled by the VAR process $\mathbf{A}^{(k)}$. We will in this section show how we again can use conjugacy to allow for collapsed sampling of the state sequence. The generative model for the observed data can be written as,

$$\Sigma^{(k)} \sim \mathcal{W}^{-1}(\eta\Sigma_0, v_0) \tag{2.22}$$

$$\mathbf{A}^{(k)} \sim \mathcal{MN}(\mathbf{0}, \Sigma^{(k)}, \mathbf{R}), \tag{2.23}$$

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t, \sigma_t^2\Sigma^{(k)}), \tag{2.24}$$

in which $\bar{\mathbf{x}}_t$ is the vector containing the appropriate past of $\mathbf{x}_t$, $\mathbf{R}$ is a diagonal $PM \times PM$-matrix containing the prior variances ($\sigma^2_{\tau_m}$) of the time lags of $A^{(k)}$ defined as the Kroenecker product $\mathbf{R} = \mathbf{R}_\tau \otimes \mathbf{I}_P$, where

$$\mathbf{R}_\tau = \begin{bmatrix} \sigma^2_{\tau_1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{\tau_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma^2_{\tau_M} \end{bmatrix},$$

and $\mathcal{MN}(\mathbf{M}, \mathbf{V}, \mathbf{U})$ is the matrix normal distribution with probability density function for the $P \times N$-matrix $\mathbf{X}$, with mean $\mathbf{M}$, row-variance $\mathbf{V}$ and column variance $\mathbf{U}$,

$$\begin{aligned} p(\mathbf{X}|\mathbf{M}, \mathbf{V}, \mathbf{U}) = {} & (2\pi)^{-PN/2}|\mathbf{V}|^{-P/2}|\mathbf{U}|^{-N/2} \\ & \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})^T\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M}))\right). \end{aligned}$$

Collecting all hyperparameters in $\boldsymbol{\theta} = \{\sigma_t^2, v_0, \eta, \Sigma_0, \sigma^2_{\tau_m}\}$, the joint likelihood of the observed data and the coefficients of the VAR-processes can be written as,

$$\begin{aligned} p(\mathbf{A}, \mathbf{X}, \boldsymbol{\Sigma}|\mathbf{z}, \boldsymbol{\theta}) = {} & \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)^T(\sigma_t^2\Sigma^{(z_t)})^{-1}(\mathbf{x}_t - \mathbf{A}^{(z_k)}\bar{\mathbf{x}}_t)\right) \\ & \prod_k (2\pi)^{\frac{-ppM}{2}}|\mathbf{R}|^{\frac{-p}{2}}|\Sigma^{(k)}|^{\frac{-pM}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{R}^{-1}\mathbf{A}^{(k)T}\Sigma^{-(k)}\mathbf{A}^{(k)})\right) \\ & \prod_k \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})}|\Sigma^{(k)}|^{\frac{-v_0+p+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\eta\Sigma_0\Sigma^{-(k)}\right)\right) \\ = {} & \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k (2\pi)^{\frac{-ppM}{2}}|\mathbf{R}|^{\frac{-p}{2}} \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})}|\Sigma^{(k)}|^{\frac{-(v_0+n_k+pM)+p+1}{2}} \\ & \exp\left(-\frac{1}{2}\operatorname{tr}((\mathbf{X}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{X}}^{(k)})^T\Sigma^{-(k)}(\mathbf{X}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{X}}^{(k)})\right. \\ & \left. +\mathbf{R}^{-1}\mathbf{A}^{(k)T}\Sigma^{-(k)}\mathbf{A}^{(k)} + \eta\Sigma_0\Sigma^{-(k)}\right) \end{aligned}$$

in which $\mathbf{X}^{(k)}$ is the collection of all data points belonging to process k, $\bar{\mathbf{X}}^{(k)}$ is the appropriate past corresponding to $\mathbf{X}^{(k)}$, the time dependent noise variances σ_t^2 have been multiplied onto the corresponding columns of $\mathbf{X}^{(k)}$ and $\bar{\mathbf{X}}^{(k)}$, n_k is the number of time points belonging to process k, and $\Sigma^{-(k)}$ is the inverse of $\Sigma^{(k)}$.

Utilizing conjugacy we can collapse both the VAR-coefficients and $\Sigma^{(k)}$ yielding the likelihood,

$$\begin{aligned} p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) = \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k |\mathbf{R}|^{\frac{-p}{2}} \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})} \\ |\mathbf{S}_{\bar{x}\bar{x}}|^{-\frac{p}{2}} \frac{2^{\frac{(v_0+n_k)p}{2}}\Gamma_p(\frac{v_0+n_k}{2})}{|\hat{\mathbf{S}}|^{\frac{v_0+n_k}{2}}}, \end{aligned} \tag{2.25}$$

in which,

$$\begin{aligned} \mathbf{S}_{\bar{x}\bar{x}} &= \bar{\mathbf{X}}^{(k)}\bar{\mathbf{X}}^{(k)T} + \mathbf{R}^{-1} \\ \mathbf{S}_{x\bar{x}} &= \mathbf{X}^{(k)}\bar{\mathbf{X}}^{(k)T} \\ \mathbf{S}_{xx} &= \mathbf{X}^{(k)}\mathbf{X}^{(k)T} + \eta\Sigma_0 \\ \hat{\mathbf{S}} &= \mathbf{S}_{xx} - \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1}\mathbf{S}_{x\bar{x}}^T. \end{aligned}$$

For a detailed derivation see section A.2 of the appendix.

2.5 Switching Vector Autoregressive Model

Willsky et al. [2009] described a switching vector autoregressive model, i.e. a framework that models multiple VAR-proccesses that switch on and off at different times in time series data (only one process at a time). In real-world data we expect persistent states over longer time scales, i.e. more self-transitions in the Markov chain, and for that reason a sticky hierarchical Dirichlet process hidden Markov model is used (HDP-HMM) (cf. Fox et al. [2008]). For improved mixing properties of the model, i.e. converging to the true number of states faster, Fox et al. [2008] claimed that using a truncated version of the infinite HMM with an upper bound on the number of states helps. The generative model can be written as,

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma), \tag{2.26}$$

$$\pi^{(\mathbf{k})}|\boldsymbol{\beta} \sim \text{DP}(\alpha+\kappa, \frac{\alpha\boldsymbol{\beta}+\kappa\delta_k}{\alpha+\kappa}), \tag{2.27}$$

$$z_t|z_{t-1} \sim \text{Multinomial}(\pi^{(\mathbf{z_{t-1}})}), \tag{2.28}$$

$$\Sigma^{(k)} \sim \mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, n_0), \tag{2.29}$$

$$\mathbf{A}^{(k)} \sim \mathcal{MN}(\mathbf{M}, \Sigma^{(k)}, \mathbf{K}), \tag{2.30}$$

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t, \Sigma^{(z_t)}), \tag{2.31}$$

in which δ_k is a vector of zeros with a 1 in the k'th entry, $\mathcal{MN}(\mathbf{M}, \mathbf{V}, \mathbf{K})$ is the matrix-normal distribution with mean $\mathbf{M}$ and row- and column-covariance $\mathbf{V}$ and $\mathbf{K}$ respectively, and $\bar{\mathbf{x}}_t$ is the past observations needed for the autoregression. The VAR process parameters, $\mathbf{A}^{(k)}$, have been stacked in the same way as in (2.2).

2.6 Inference

Up until now we have described a generative model; a way to generate data given hyperparameters in the model. But what we are really interested in is finding the most likely parameters, θ, given the data, X. More generally we want to compute the posterior distribution $p(\theta|X)$, which can be done using Bayes theorem,

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta')p(\theta')d\theta'}. \tag{2.32}$$

The denominator of (2.32), called the *evidence*, is in most cases not analytically possible to calculate, so we must turn to approximate methods. The following sections will be dedicated to introduce Markov chain Monte Carlo (MCMC) methods for sampling the posterior, and how we have implemented an MCMC-inference procedure for the IHMM.

2.6.1 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a class of algorithms that can be used to iteratively sample from a desired probability distribution (cf. [Bishop et al., 2006, Chapter 11]). The idea is to create a Markov chain of samples, i.e. each sample is dependent on the previous, where in the limit the samples created come from the desired distribution. In a general Markov chain each new element, $\mathbf{x}^{(t)}$, is generated from a transition distribution, $T(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$, and if we run the chain long enough (and the chain satisfies certain conditions), then we will arrive at the *equilibrium distribution*, $p^*(\mathbf{x})$. This means that another step in the chain leaves the distribution unchanged, a property called *invariance*. The equilibrium distribution thus satisfies,

$$p^*(\mathbf{x}) = \sum_{\mathbf{x}'} T(\mathbf{x}|\mathbf{x}')p^*(\mathbf{x}') \tag{2.33}$$

Invariance of the equilibrium distribution can be proved by showing that the transition distribution satisfies the *detailed balance* condition given by,

$$p^*(\mathbf{x})T(\mathbf{x}'|\mathbf{x}) = p^*(\mathbf{x}')T(\mathbf{x}|\mathbf{x}'). \tag{2.34}$$

This can be shown by looking at the right hand side of (2.33) and using detailed balance (2.34), i.e.

$$\sum_{\mathbf{x}'} T(\mathbf{x}|\mathbf{x}')p^*(\mathbf{x}') = \sum_{\mathbf{x}'} T(\mathbf{x}'|\mathbf{x})p^*(\mathbf{x}) = p^*(\mathbf{x}) \sum_{\mathbf{x}'} T(\mathbf{x}'|\mathbf{x}) = p^*(\mathbf{x}),$$

since T is a probability distribution and therefore $\sum_{\mathbf{x}'} T(\mathbf{x}'|\mathbf{x}) = 1$.

To ensure that the chain converges to the equilibrium distribution, we must require that this convergence is not dependent on initialization of the chain. This property is called *ergodicity*, and also implies the uniqueness of the equilibrium. For MCMC only mild restrictions on the equilibrium and transition distributions yield an ergodic chain (cf. Neal [1993], Bishop et al. [2006]). Now, we will describe an MCMC sampling scheme, and how we choose the transition distribution such that the samples generated come from a desired distribution.

2.6.2 Metropolis-Hastings Sampling

Hastings [1970] was the first to describe the *Metropolis-Hastings* algorithm, an extension of the *Metropolis* algorithm first described by Metropolis and Ulam [1949]. As in other MCMC algorithms, we generate samples forming a Markov chain from a transition distribution, $T(\mathbf{x}'|\mathbf{x})$, to approximate the distribution $p(\mathbf{x})$. In the Metropolis-Hastings algorithm T is split into a *proposal* distribution, $q(\mathbf{x}'|\mathbf{x})$, and an *acceptance* probability, $\alpha_{\mathbf{x}',\mathbf{x}}$, yielding $T(\mathbf{x}'|\mathbf{x}) = \alpha_{\mathbf{x}',\mathbf{x}}q(\mathbf{x}'|\mathbf{x})$, where

$$\alpha_{\mathbf{x}',\mathbf{x}} = \min\left(1, \frac{p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right). \tag{2.35}$$

Note here that the evaluation of the acceptance probability does not require the full distribution $p(\mathbf{x})$, since the normalization constant for p cancels out in the ratio. The choice of q has a large impact on the performance of the algorithm, especially how long it takes to arrive at the equilibrium distribution.

As described in the previous sections, the detailed balance condition (2.34) is a property of our chain that we need, in order to guarantee that the algorithm, if run long enough, converges to the desired distribution. Using the transition probability and (2.35), we can write,

$$\begin{aligned}
T(\mathbf{x}'|\mathbf{x})p(\mathbf{x}) &= q(\mathbf{x}'|\mathbf{x})\alpha_{\mathbf{x}',\mathbf{x}}p(\mathbf{x}) \\
&= \min\left(p(\mathbf{x})q(\mathbf{x}'|\mathbf{x}), \frac{p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right) \\
&= \min\left(p(\mathbf{x})q(\mathbf{x}'|\mathbf{x}), p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')\right) \\
&= \min\left(\frac{p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}{p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}, 1\right) q(\mathbf{x}|\mathbf{x}')p(\mathbf{x}') \\
&= \alpha_{\mathbf{x},\mathbf{x}'}q(\mathbf{x}|\mathbf{x}')p(\mathbf{x}') \\
&= T(\mathbf{x}|\mathbf{x}')p(\mathbf{x}'),
\end{aligned}$$

and this proves detailed balance for the Metropolis-Hastings algorithm.

2.6.3 Gibbs Sampling

A special case of the Metropolis-Hastings sampling is Gibbs sampling, which relies on sampling the conditional distribution. For multidimensional variables, $\mathbf{x} = (x_1, x_2, ..., x_N)$, we can propose a new element in the chain by only changing one of the variable's dimensions, for instance $\mathbf{x}' = (x_1', x_2, ..., x_N)$. In Gibbs sampling this new variable is sampled from the distribution over one variable-dimension conditioned on all the others, i.e.

$$x_n' \sim p(x_n|x_1, x_2, ..., x_{n-1}, x_{n+1}, ..., x_N) \equiv p(x_n|\mathbf{x}_{\setminus n}) \tag{2.36}$$

Each variable is then sampled in turn fixing the others at their current value, thus in each step only changing one variable at a time. Looking at this in terms of a Metropolis-Hastings algorithm, we see that the proposal distribution (2.36) yields the following acceptance probability

$$\begin{aligned}\alpha_{\mathbf{x}',\mathbf{x}} &= \min\left(1, \frac{p(\mathbf{x}')p(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x})p(\mathbf{x}'|\mathbf{x})}\right)\\ &= \min\left(1, \frac{p(\mathbf{x}'_{\setminus n})p(x'_n|p(\mathbf{x}'_{\setminus n})p(\mathbf{x}_n|\mathbf{x}'_{\setminus n})}{p(\mathbf{x}_{\setminus n})p(x_n|p(\mathbf{x}_{\setminus n})p(\mathbf{x}'_n|\mathbf{x}_{\setminus n})}\right)\\ &= \min\left(1, \frac{p(\mathbf{x}'_{\setminus n})p(x'_n|p(\mathbf{x}'_{\setminus n})p(x_n|\mathbf{x}'_{\setminus n})}{p(\mathbf{x}_{\setminus n})p(x_n|p(\mathbf{x}_{\setminus n})p(x'_n|\mathbf{x}_{\setminus n})}\right)\\ &= 1,\end{aligned}$$

due to the fact that $\mathbf{x}_{\setminus n} = \mathbf{x}'_{\setminus n}$. This means that we always accept the samples generated by a Gibbs sampler, but also that these small changes can yield very correlated samples. Often a principle of *thinning* is applied where we only save every T'th sample. Detailed balance of this algorithm is ensured because it is a special case of the Metropolis-Hastings sampler.

2.6.4 Split-Merge Sampling

Split-merge sampling was proposed by Jain and Neal [2004] to overcome mixing issues with Gibbs sampling for Dirichlet process mixture models, and was first applied to hidden Markov models by Hughes et al. [2012]. The procedure revolves around randomly splitting and merging clusters (or in this case states) and accepting or rejecting the new configuration by the Metropolis-Hastings ratio. It has been shown that this procedure can help the Gibbs sampler escape local maxima, where it would be stuck otherwise (cf. Phillips and Smith [1996]). In a merge-move, we merge the two states chosen and compare this to the old configuration (equivalent to splitting the merged state). In a split-move, rather than randomly assigning data points to each of the two new clusters generated, Jain and Neal [2004] proposed using a restricted Gibbs-scan. In this procedure the state assignment of the data points is sampled from the conditional distribution over states yielding that the partitioning is consistent with the data.

The procedure starts by picking two distinct random observations, denoted z_{τ_1} and z_{τ_2}, uniformly over all observations from an initial state sequence, $\mathbf{z}^{(old)}$. If the two observations are in the same state ($z_{\tau_1} = z_{\tau_2}$), then a split move is proposed, and if the two observations are in distinct states ($z_{\tau_1} \neq z_{\tau_2}$) a merge move is proposed. In a split move, the two chosen observations are each placed in a separate state, one of them in the old state ($z_{\tau_1}^{(new)} = z_{\tau_1}$) and the other in a new state ($z_{\tau_2}^{(new)} = z^*$) . Then for each observation from the originally chosen state, z_{τ_1}, we Gibbs sample new state assignments restricted to choose between $z_{\tau_1}^{(new)}$ and $z_{\tau_2}^{(new)}$. The new configuration, $\mathbf{z}^{(split)}$, is evaluated by the Metropolis-Hastings acceptance ratio $\alpha_{\mathbf{z}^{(split)},\mathbf{z}^{(old)}}$. In the Metropolis-Hastings ratio we must evaluate the probability of making the opposite move of what we are proposing, and in this case the opposite of splitting a component in two is exactly merging the two newly proposed states yielding $\mathbf{z}^{(old)}$.

In a merge move (where $z_{\tau_1} = z_{\tau_2}$), the procedure is straightforward; we simply merge the two components chosen such that $z_t^{(new)} = z_{\tau_1}, \forall t : z_t = z_{\tau_1}$. The new configuration, $\mathbf{z}^{(merge)}$, is then compared to the split-move that generates the old clustering. The probability of that split-move can be calculated by a similar procedure to the restricted Gibbs-scan.

Typically, split-merge sampling is run after a normal Gibbs sampling sweep. A number of split-merge iterations is run for each normal Gibbs sweep, and the restricted Gibbs sweep in the split-procedure is also run a couple of times (2-3), to get a good estimation of the best split move possible according to the data. Split-merge sampling has the potential of being computationally expensive, if we for instance consistently try to split a large state. The restricted Gibbs-scan will be accordingly long and scale in the number of data points that needs to be re-sampled.

2.6.5 Infinite Hidden Markov Model Revisited

Van Gael [2012] described a Gibbs sampler for the IHMM (cf. section 2.4) that alternatingly re-samples the state sequence and the stick-parameter β. Van Gael [2012] showed that re-sampling the state sequence, given β, requires sampling the conditional,

$$p(z_t|z_{-t}, x_{1:T}, \alpha, \beta, \gamma, H, F) \propto p(z_t|z_{-t}, \alpha, \beta, \gamma)p(x_t|z_t, z_{-t}, H, F), \tag{2.37}$$

in which z_{-t} denotes the state sequence without the t-th time point. Computing the distribution $p(z_t|z_{-t}, \alpha, \beta, \gamma)$ provides the probability of transitioning from the state z_{t-1} to any state times the probability of transitioning from any state to the state z_{t+1}. Using the Polya urn scheme representation of the state sequence with the transition probabilities (2.14), (2.15) and (2.16) we can write this probability as,

$$\begin{aligned}(p(z_t = k|z_{t-1}, \alpha) + p(oracle|z_{t-1}, \alpha) \cdot \beta_k) \cdot \\ \left(p(z_{t+1}|z_t = k, \alpha) + p(oracle|z_t = k, \alpha) \cdot \beta_{z_{t+1}}\right)\end{aligned} \tag{2.38}$$

We have replaced the oracle-urn term by the stick-breaking probability β. (2.38) disregards any kind of special case (e.g. end-points of the sequence), so the true conditional becomes the following (from Van Gael [2012]),

$$p(z_t = k | z_{-t}, \alpha, \beta, \gamma) \propto \begin{cases} (n^{-t}_{z_{t-1},k} + \alpha\beta_k) \frac{n^{-t}_{k,z_{t+1}} + \alpha\beta_{z_{t+1}}}{(\sum_{j'} n^{-t}_{k,j'}) + \alpha} & \text{if } k \leq K, \\ (n^{-t}_{z_{t-1},k} + \alpha\beta_k) \frac{n^{-t}_{k,z_{t+1}} + 1 + \alpha\beta_{z_{t+1}}}{(\sum_{j'} n^{-t}_{k,j'}) + 1 + \alpha} & \text{if } k = z_{t-1} = z_{t+1}, \\ (n^{-t}_{z_{t-1},k} + \alpha\beta_k) \frac{n^{-t}_{k,z_{t+1}} + \alpha\beta_{z_{t+1}}}{(\sum_{j'} n^{-t}_{k,j'}) + 1 + \alpha} & \text{if } k = z_{t-1} \neq z_{t+1}, \\ \alpha\beta_k\beta_{z_{t+1}} & \text{if } k = K + 1 \end{cases}, \tag{2.39}$$

in which K is the current number of states and n^{-t}_{ij} is the count of the number of transitions out of state i into state j without the t-th timepoint. For completeness we must start the chain somewhere, in a z_0, and we will follow the convention used by Van Gael in his iHMM-Toolbox (Van Gael [2010]) that $z_0 = 1$. Details on how to sample α, β and γ can be found in Van Gael [2012].

The other half of the Gibbs sampling equation (2.37) concerned with the observed data can be calculated based on the collapsed likelihood $p(x|z)$ as shown in (2.21) for the IHMM-Wish and (2.25) for the IHMM-MVAR. In each of the collapsed likelihood equations we have a product over the K states, and thus evaluating (2.37) boils down to evaluating the gain in likelihood of placing the data point in each of the K states or a new state. We want to recalculate as little as possible in the implementation of this, so it can be relevant to identify what variables that change when changing the state sequence. The elements in the collapsed likelihood that change if we add a data point to a new state, denoted the *sufficient statistics*, are for the IHMM-Wish n_k and $\mathbf{X}^{(k)}$ and for the IHMM-MVAR n_k, $\mathbf{S}_{xx}$, $\mathbf{S}_{x\bar{x}}$ and $\mathbf{S}_{\bar{x}\bar{x}}$. Apart from n_k, each of the sufficient statistics can be up- and downdated by adding or subtracting outer-prodcuts. For example if we were to remove the t'th data point x_t from the k'th state in the IHMM-Wish model we would make the following down-dates

$$n_k \leftarrow n_k - 1$$
$$\mathbf{X}^{(k)} \leftarrow \mathbf{X}^{(k)} - x_t x_t^T$$

Similar rules can be derived for up-dates and for the IHMM-MVAR sufficient statistics.

2.6.6 Implementation of Inference Procedure

Our implementation of the models described above follow the way Jurgen Van Gael has implemented his IHMM in the iHMM-Toolbox (Van Gael [2010]). The IHMM-Wish was implemented (but not tested) by supervisor Morten Mørup, and building on that the implementation the IHMM-MVAR was written in MATLAB as part of this thesis. Both implementations underwent testing and validation of their correctness also as part of this project (cf. section 2.6.7). A sketch of the full inference procedure for the IHMM model can be seen in Algorithm 1. Other than the inference already described in section 2.6.5 and 2.6.4, we add a random-walk Metropolis-Hastings for hyperparameters of the prior distributions associated with the observed data, i.e. in case of the Normal-Inverse-Wishart model the covariance scale η and the time specific noise σ_t^2, and in addition the lag specific variance $\sigma_{\tau_m}^2$ for the multiple-state vector autoregressive model. We add an improper $1/\mathcal{X}$ prior on both σ_t^2 and $\sigma_{\tau_m}^2$. The hyperparameters from the state sequence are sampled using the iHMM-toolbox, where vague Gamma-priors are placed on them. In each step of the algorithm we save the sufficients statistics to avoid as much unnecessary recomputation as possible, as briefly mentioned in the previous section. In both the IHMM-Wish and the IHMM-MVAR we work in the log-domain when calculating the conditional probabilities. This means that the sufficient statistics are used in a log-determinant calculation, and for that reason we keep the Cholesky factorization of $\mathbf{X}^{(k)}$ in IHMM-Wish and $\mathbf{S}_{\bar{x}\bar{x}}$ in IHMM-MVAR. The Cholesky factorization can be easily rank one up- and downdated as we need, and makes the determinant calculation computationally much cheaper. Furthermore, we add an option in our implementation to switch between a static and a dynamic model. Here we mean static in the sense that the state sequence has been forced to only contain one state, and that we skip the Gibbs- and split-merge sampling steps. This allows us to investigate the differences

```
Input   : X
Output: Clustering of time points, Z
Initialize relevant parameters;
for Number of Iterations do
    Gibbs Sampling: Sample states $z$;
    Split-Merge Sampling: Sample new configuration of states;
    Random-Walk Metropolis-Hastings: Sample hyperparameters for
    observed data-model;
    Gibbs Sampling: Sample hyperparameters for state-model;
end
```

Algorithm 1: Inference procedure for IHMM

between a static and a dynamic model on any data set.

2.6.7 Validation of Implementation

A common problem that arises when implementing large Markov chain Monte Carlo (MCMC) inference procedures, as the one described in section 2.6.6, is how to validate the correctness of the implementation. Since the algorithm itself is non-deterministic, bugs can be hard to find, reproduce and fix. Grosse and Duvenaud [2014] described different approaches for testing MCMC code, one of them being *unit testing*. In this we test that the conditional distribution is consistent with the joint distribution, i.e. if we are sampling any variable, x, from the conditional, $p(x|y)$, then the following equality must hold for all values of x (and y),

$$\frac{p(x', y)}{p(x, y)} = \frac{p(x'|y)}{p(x|y)}. \tag{2.40}$$

In our framework, each time the conditional distribution is calculated, we can compare the conditional with the joint for two random values of the variable in question. The joint is here calculated from scratch as it would be if we initialized the inference procedure. For example, in the Gibbs sampler for the state sequence, we obtain the distribution $p(z_t = k|z_{\setminus t}, \mathbf{X}, \theta)$, where θ is the collection of all relevant parameters. Now we pick two random values for k, k_1 and k_2, yielding two state sequences, $\mathbf{z_1}$ and $\mathbf{z_2}$, only differing on the t-th element. The test is now to evaluate (2.40) in the log-domain given some tolerance ϵ,

$$\log p(z_t = k_1|z_{\setminus t}, \mathbf{X}, \theta) - \log p(z_t = k_2|z_{\setminus t}, \mathbf{X}, \theta) - (p(\mathbf{z_1}, \mathbf{X}, \theta) - p(\mathbf{z_2}, \mathbf{X}, \theta)) < \epsilon. \tag{2.41}$$

This framework has been used extensively for debugging the implementations of IHMM-Wish and IHMM-MVAR, since it can give an indication of where and in which sampler the bug is located. We ran our implementation of the IHMM-Wish and IHMM-MVAR 10 times for 500 iterations to ensure their correctness, with unit testing in the following samplers,

- Gibbs sampler for the state sequence, $\mathbf{z}$
- Metropolis-Hastings sampler for the time-dependent noise, σ_t^2
- Metropolis-Hastings sampler for the scale of the default covariance, η
- *Only for IHMM-MVAR*: Metropolis-Hastings sampler for the lag-specific variance of the VAR coefficients $\sigma_{\tau_m}^2$.

2.7 Predictive Likelihood

To quantify how well the models capture the structure of previously unseen data, we describe a framework that calculates the *predictive likelihood*. The models are fitted to one data set, called the *training* data, and we want to evaluate the likelihood of *test* data given the model trained. Formally we can write this as,

$$p(X^*|\mathcal{M}, X) = \int_{\boldsymbol{\theta} \in \mathcal{M}} p(X^*|\,\boldsymbol{\theta})p(\boldsymbol{\theta}\,|X)d\,\boldsymbol{\theta}, \tag{2.42}$$

in which X is the training data, X^* is the test data and $\mathcal{M}$ is the parameter space for model trained such that $\boldsymbol{\theta}$ is an element of the space. Because this integral in most cases cannot be analytically determined we turn to a sampling scheme where T samples from the posterior $p(\boldsymbol{\theta}\,|X)$, $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(T)}$, are obtained from the parameter posteriors of the model (for the IHMM-MVAR cf. section A.3). The integral can now be approximated by,

$$p(X^*|\mathcal{M}, X) \approx \frac{1}{T} \sum_t p(X^*|\,\boldsymbol{\theta}^{(t)}). \tag{2.43}$$

For the case of the infinite hidden Markov model (IHMM) an element of the parameter space consists of a state sequence $\mathbf{z}$ and a parameter θ_k, $k = 1..K$, that governs the emission probabilities for each of the K states in $\mathbf{z}$. But for the test set we do not have a state sequence available, so the state sequence needs to be integrated out of (2.43), i.e.

$$p(X^*|\mathcal{M}, X) \approx \frac{1}{T} \sum_t \left[\sum_{\mathbf{z}} p(X^*|\,\boldsymbol{\theta}^{(t)})p(\mathbf{z}) \right]. \tag{2.44}$$

The integration can be done using a modified Viterbi algorithm (cf. Bishop et al. [2006]. The original Viterbi algorithm finds the most probable state sequence among all sequences. A sketch of one path through the system is illustrated in figure 2.4.

Starting from an initial distribution, $\pi^{(0)}$, for $t = 1$ we calculate the probability $V_{1,k} = \pi_k^{(0)} p(x_1|z_1 = k)$. From that point on we calculate the probability of each path through the state space, by keeping track of the probabilities

$$V_{t,k} = p(x_t|z_t = k) \sum_{k'} \pi_{k,k'} V_{t-1,k'}, \tag{2.45}$$

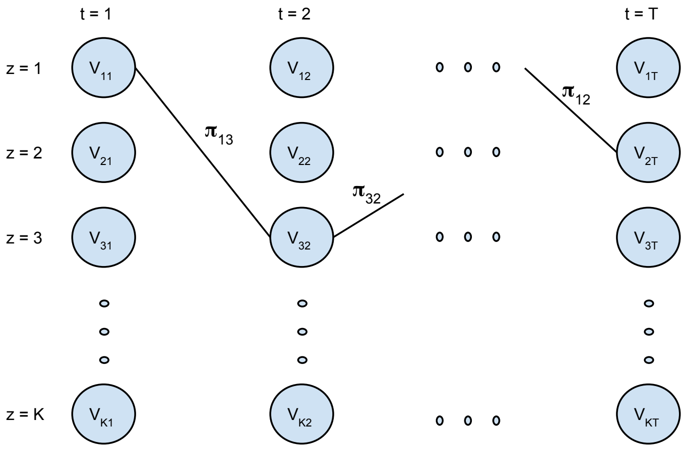


Figure 2.4: A diagram showing one path through the state space. The path is weighted by the transition probability $\pi_{k,t}$. In the Viterbi algorithm we sum over all such paths to completely marginalize the state sequence out.

in which $\pi_{k,k'}$ is the probability of transitioning from state k to k'. $V_{t,k}$ quantifies the probability of transitioning from any state in the previous time step $t-1$ including the probability contribution from all previous time steps, times the probability of emitting x_t from state k. Running through the time steps from $t = 1..T$ sums up all the probability mass from all paths thus yielding the integral in (2.44). So for the integration to be possible we need all state specific parameters - for the IHMM-Wish that is $\Sigma^{(k)}$ and for the IHMM-MVAR $\mathbf{S}_{xx}, \mathbf{S}_{x\bar{x}}$ and $\mathbf{S}_{\bar{x}\bar{x}}$ - and also the transition matrix, $\boldsymbol{\pi}$, all of which we can sample from their respective posteriors.

Chapter 3

Data

In this chapter the data that has been used in the project will briefly be described. In section 3.1 some general aspects of noise and preprocessing in fMRI will be introduced. In sections 3.2 and 3.3 data from Danish Research Center for Magnetic Resonance and from the Human Connectome Project (cf. Van Essen et al. [2012]) will be presented briefly.

3.1 Preprocessing

It is well known that the BOLD signal is noisy due to many factors, and therefore a rather large preprocessing pipeline is needed before any analysis can start (cf. Stippich et al. [2007]). Typical sources of noise include (but is not limited to),

- Scanner drift (cf. Smith et al. [1999]) - local changes in magnetic field caused by scanner instabilities manifests as noise in the measured signal.
- Physiological noise - small movements by the subject can cause voxels to be blurred together over time.
- Spikes - spatial noise-artefacts in the signal not attributed to any physiological meaning
- Psychological noise - the subject might perform the task at hand differently than the experimenter imagined. This challenges how the experi-

ment is set up and also typically narrows the scientific questions that can be answered by an fMRI experiment.

Typical preprocessing steps include,

- Slice time correction - Each slice of the MR scan is not captured simultaneously, so a collection of slices forming the 3D image are taken at slightly different time points. This must be corrected for.
- Head motion correction - the physiological noise described above must be corrected for. This is usually done by assuming that the brain can be mathematically described as a rigid body, and its movements and deformations can be explained using a small number of parameters (varying from 6 to 12). These parameters are then optimized for, and afterwards movement effects can be regressed out.
- Spatial smoothing - the rigid body transformation is typically not enough so the data can be smoothed using a number of the neighbouring voxels (for instance via a Gaussian kernel).

To be able to compare measurements across subjects, measurements from fMRI are transformed into a well-defined coordinate system like the Talairach system or the Montreal Neurological Institute and Hospital (MNI) system. Furthermore, since the spatial resolution often allows for measuring the BOLD signal from over 100.000 voxels most statistical models are not compatible with such a high dimensional space. Thus dimensionality reduction methods like principal component analysis (PCA), and especially independent component analysis (ICA), or atlas-based methods, such as automated anatomical labelling (AAL), are often used. We will in this project use PCA for dimensionality reduction. ICA would be more appropriate if we were to visualize the results with brain maps, since IC's have been shown to yield better physiological meaning (cf. Calhoun et al. [2003]). But since visualization of spatial maps in a dynamic setting is out of scope of this project, we will for convenience stick to PCA (cf. [Bishop et al., 2006, chapter 12.1] for an introduction to PCA).

3.2 Data from Danish Research Center for Magnetic Resonance Imaging

Throughout the project a data set recorded at the Danish Research Center for Magnetic Resonance (DRCMR) was available to us for analysis. The data set is based on two experiments run on 30 subjects, the first being a finger tapping experiment (described in Rasmussen et al. [2012]) from now on denoted, *Motor*, and the second being a resting state experiment (described in Andersen et al. [2014]). The motor task experiment consisted of two conditions, left hand finger tapping and right hand finger tapping, each of length 20 s, and in between each condition there was a period of rest of around 10 s. This cycle was repeated 10 times, and a total of 240 scans was acquired during the whole experiment. In the resting state experiment the subjects were asked to lie in the scanner with their eyes closed trying to refrain from any movement and without falling asleep. The resting state experiment yielded 480 scans per subject.

In the projects mentioned above, a number of preprocessing steps was applied to the data using the statistical parametric mapping toolbox SPM8[1]. For each subject a structural scan was acquired to allow for each image from the time series to be aligned with this, a procedure called co-registration. Head motion correction was applied using a 6-parameter rigid body transformation, and afterwards images were transformed to the MNI coordiante system. To reduce the effects of hardware instability and unwanted physiological effects the data was high-pass filtered. For details on both the preprocessing and the image acquisition parameters cf. Andersen et al. [2014]. Further preprocessing was applied to the resting state data only, most notably despiking using the framework from Campbell et al. [2013]. The fact that the two data sets, motor and resting-state, have not undergone the exact same preprocessing steps is unfortunate and we will have to keep this in mind when discussing results concerning this data.

3.3 Data from the Human Connectome Project

The Human Connectome Project (HCP)[2] is a consortium, led by Washington University, University of Minnesota and Oxford University, that tries to understand and describe human brain function and behaviour using different state-of-the-art neuroimaging techniques. They made a large portion of their data publicly available in a large data release in March 2013 (cf. Human Connectome Project [2014]). The release contained data from 500 subjects from a variety of modalities, tasks and resting-state experiments. The reasons that we are interested in the HCP data for this particular project are two-fold. First we have 6 different task experiments available, in which we expect the functional connectivity to be detectably different. Secondly, the fMRI HCP data is

[1] Available at: http://www.fil.ion.ucl.ac.uk/spm/

[2] Cf. http://www.humanconnectome.org/

sampled at a high sampling rate yielding subsecond resolution, which seems desirable for finding temporal dynamics in the functional connectivity. The task experiments from HCP (cf. [Human Connectome Project, 2014, pp. 40]) that were used in this synthetic study were,

- *Motor* - In this task subjects were asked to either tap their left or right fingers, or squeeze their left or right toes, or move their tongue.
- *Language Processing* - In this task subjects were presented with either a math problem or a story followed by a 2-alternative question about the content.
- *Emotion* - In this task subjects were asked to match images of faces with either an angry or fearful expression.
- *Social* - Subjects are presented with short video clips of objects that either interact in some way or move randomly around, and are afterwards questioned about the interaction.
- *Gambling* - In this task subjects were asked to guess a mystery card value ranging from 1-9 in order to win or lose money. Task blocks are divided into reward, loss or neutral trials, such that it is predetermined by the experimenter whether the participant loses or wins in that trial.
- *Working Memory* - In this task subjects were presented with different pictures of body parts for them to remember.

We used the surface data available, i.e. the data where the volumetric data from around cortex were mapped to surface regions from the MNI surface space. The preprocessing steps applied by the HCP to the volumetric data included motion correction and co-registration using a structural scan, transformation into MNI space and intensity normalization (for details see [Human Connectome Project, 2014, pp. 105]). Notable differences in the data acquisition between the DRCMR data and the HCP, is that the HCP uses finer spatial and temporal resolution, which could lead to more noise artefacts and non-physiological manifestations in the signal. Optimal preprocessing of the data is however out of scope of this thesis.

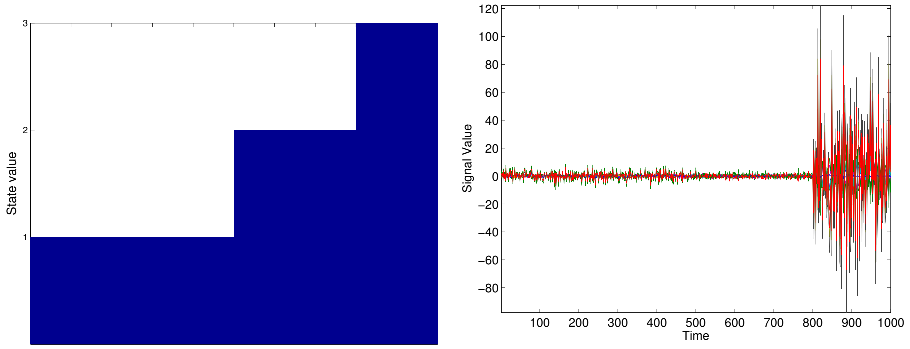
Chapter 4

Results

In this section we present results from a synthetic study in section 4.1. Next we have validated the predictive likelihood framework (cf. section 2.7), the result of which can be seen in section 4.2. In sections 4.3 and 4.4 we present results on real-world data from the Danish Research Center for Magnetic Resonance (DRCMR) and from the Human Connectome Project (HCP), respectively.

4.1 Experiments on Synthetic Data

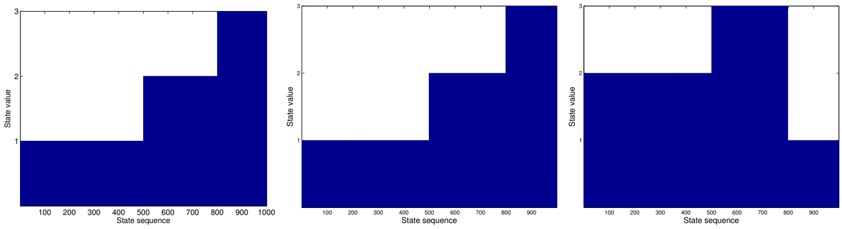
In order to test the models proposed in a synthetic setting that resembles the real world, we generated a synthetic data set in the following way. From one HCP-subject, we took out three different task experiments, appended them together, thereby yielding one long sequence. Then we did a principal component analysis (PCA), and used the first 10 PC's to represent the data. For the first part of the synthetic study we used the 'Motor', 'Language' and 'Emotion' tasks (cf. section 3.3). From this point on we generated data in two ways, from an inverse Wishart mixture model and from a mixture of VAR's. On both data sets we fitted the IHMM-Wish, IHMM-MVAR and the HDPHMM-MVAR, all described in chapter 2.



(a) True state sequence

(b) Data - each color represents one dimension

Figure 4.1: Synthetic data set 4.1b generated from an inverse Wishart mixture with above state sequence 4.1b based on task data from HCP.



(a) IHMM-Wishart

(b) IHMM-MVAR

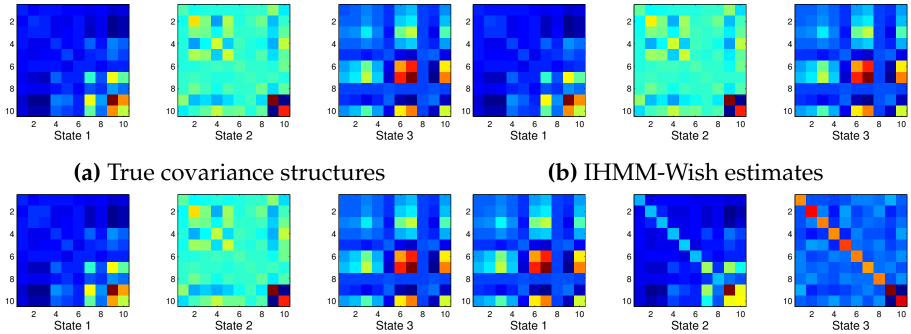
(c) HDPHMM-MVAR

Figure 4.2: Estimated state sequence based on synthetic data shown in figure 4.1

4.1.1 Data from an Inverse Wishart Mixture

We estimated an empirical covariance matrix from the whole data set and used this as the matrix parameter in an inverse Wishart distribution to draw 3 covariance matrices. From these we generated 1000 data points, a plot of which can be seen in figure 4.1 together with the underlying true state sequence.

After fitting the models from section 2 to the data, we saw that they all found the true state sequence (up to a permutation of the state labeling) (cf. figure 4.2). Furthermore all models found something very close to the true covariance structure as seen in figure 4.3. This means that both the IHMM-MVAR and the HDPHMM-MVAR can fit the data even though it has been generated from a simpler generative model, indicating that they both extend the IHMM-Wish.



(a) True covariance structures

(b) IHMM-Wish estimates

(c) IHMM-MVAR estimates

(d) HDPHMM-MVAR estimates

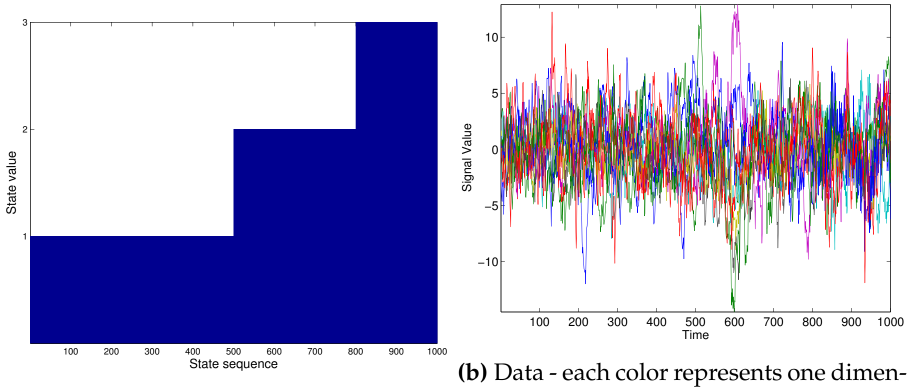
Figure 4.3: The estimated covariance coefficients by the IHMM-Wish, IHMM-MVAR and the HDPHMM-MVAR are compared to the true parameters that generated the data in figure 4.1. Each covariance matrix is plotted as an image where values are represented by a color ranging from blue (low) to red (high) relative to the other values in the matrix.

4.1.2 Data from a Mixture of VAR's

We fitted seperate second order VAR-models to each of the three task blocks, and used these VAR coefficients for 3 seperate states in a mixture of VAR models (as described in section 2.2). From this MVAR we generated a synthetic data set with 1000 timepoints in order to have continuous transition between states. If e.g., we worked directly on the real world data we could have arbitrary large discontinuities between tasks that have been appended together. The true state sequence and data can be seen in figure 4.4.

In figure 4.5 the state sequence estimated by the IHMM-MVAR model and the IHMM-Wish model are shown. As we can see the the IHMM-MVAR and the HDPHMM-MVAR models captures perfectly the true state sequence, whereas the IHMM-Wish model finds 12 states. A closer look at the IHMM-Wish estimate (figure 4.5a) shows that the model has split up the true states into distinct states, and that the estimated states are largely only present within one true state block. Still we must conclude that the IHMM-Wish model does not capture the true underlying dynamics of the MVAR-model data, and that 'dynamicness' in an MVAR model is a much richer term compared to 'dynamics' from a inverse Wishart mixture.

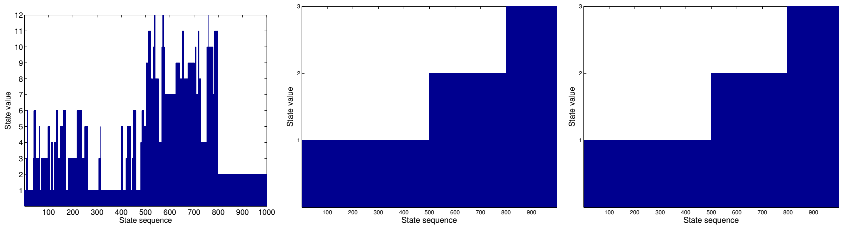
Looking at the parameters estimated by the IHMM-MVAR and the HDPHMM-MVAR compared to the true parameters in figure 4.6, we see that both are able



(a) True state sequence

(b) Data - each color represents one dimension

Figure 4.4: Synthetic data set 4.4b generated from a mixture of VAR's with above state sequence 4.4a based on task data from HCP.

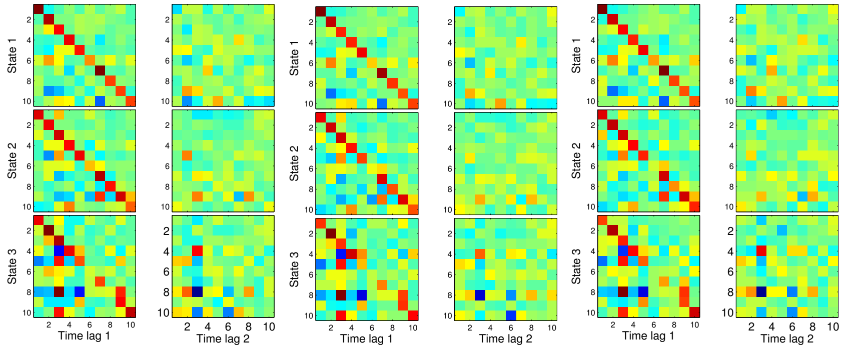


(a) IHMM-Wishart

(b) IHMM-MVAR

(c) HDPHMM-MVAR

Figure 4.5: Estimated state sequence by the IHMM-Wish, IHMM-MVAR and HDPHMM-MVAR on synthetic data shown in figure 4.4



(a) True parameters **(b)** IHMM-MVAR **(c)** HDPHMM-MVAR

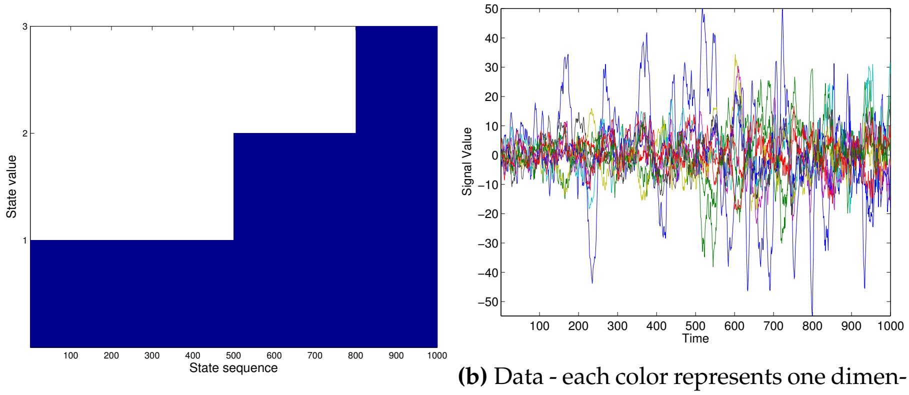
Figure 4.6: The estimated VAR coefficients by the IHMM-MVAR and HDPHMM-MVAR are compared to the true parameters that generated the data in figure 4.4. Each VAR coefficient matrix from each time lag is plotted as an image, where values have been normalized to lie between -1 and 1. Values are represented by a color ranging from blue (low, around -1) to red (high, around 1).

to capture the true coefficients in the data generating process.

Variable Noise

Comparing the IHMM-MVAR model with the switching vector autoregressive model (also denoted HPD-HMM-MVAR) by Willsky et al. [2009] we see that in the generative model the only difference is the 'stickyness' of the HMM and that the IHMM-MVAR has a time-dependent noise parameter controlling the level of the covariance of the state specific noise. To test the use for the the latter, we have generated a data set as in the previous section, but now we vary the innovation that enters the system from having variance 1 to 2 over the course of 1000 time points. The data can be seen in figure 4.7.

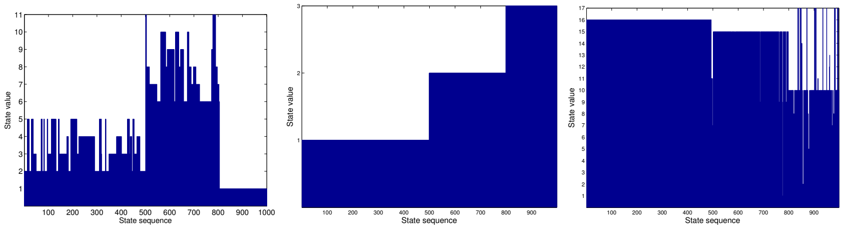
In figure 4.8 we can see that the HPD-HMM-MVAR model overestimates the number of states, especially in the last part of the sequence where the variance on the innovation is at its highest. IHMM-MVAR that finds the true state sequence, whereas the IHMM-Wish still overestimates the number of states, but does it fairly consistent with the previous experiment where the innovation-variance was not varied (cf. figure 4.5a).



(a) True state sequence

(b) Data - each color represents one dimension

Figure 4.7: Synthetic data set 4.7b generated from a mixture of VAR's with above state sequence 4.4a based on task data from HCP. Variance of the innovation that enters the system has been varied to increase linearly over time.



(a) IHMM-Wish (MAP solution)

(b) IHMM-MVAR (MAP solution)

(c) HDPHMM-MVAR (MAP solution)

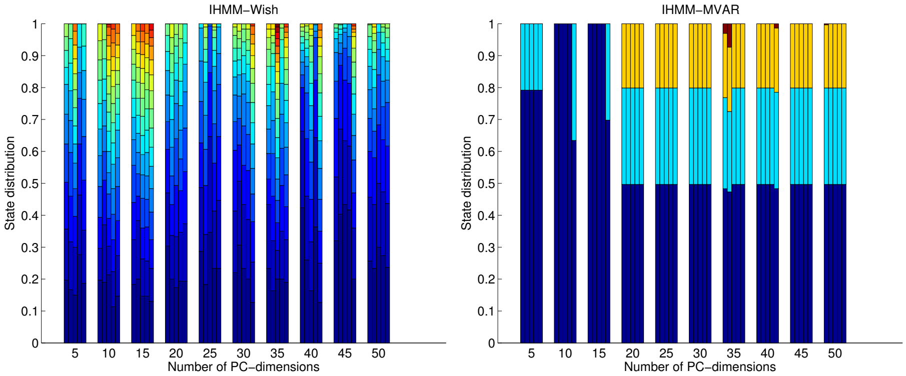
Figure 4.8: The estimated state sequences by the IHMM-Wish, IHMM-MVAR and the HDP-HMM-MVAR respectively on the data in figure 4.7

4.1.3 Dimensionality analysis

The number of states inferred by the model can be influenced by the dimensionality of the data, so we generated synthetic data from a mixture of 3 VAR models as described in the section above with 300 time points. We chose fewer time points in this analysis since 300 time points is closer to the time scale we have available in the real-world data. Now we varied the dimensionality of the data from 5 to 50 incrementally by 5, and ran our two MCMC algorithms, IHMM-MVAR and IHMM-Wish. We restarted each inference procedure 5 times to get an understanding of the influence of initialization. The results can be seen in figure 4.9. We see no general trend in the number of states found by IHMM-Wish, and we see that many of the clusters extracted are very small in size. The IHMM-MVAR model surprisingly does not find the true state distribution (a 3 component model) when using only the first 5-15 PC components. We suspected that this could be caused by the first PC's from the task data not being related to the dynamics that we are synthetically trying to create by a 3 component model. One could imagine that the first couple of PC's are more related to noise like spikes or overall variation not related to task activity, and thereby being static components. We investigated this by running the same experiment with 250 time points, where we skipped the first 10 PC's. The results can be seen in figure 4.10. It turns out that our hypothesis could not be validated by this experiment, since the IHMM-MVAR behaves almost exactly the same way as before. We must conclude that the number of dimensions in the data must be relatively high to find the dynamics we are looking for, especially when we have few data points. Another explanation is that the VAR models extracted from real data are simply not distinguishable from each other in low dimensional spaces on shorter time scales compared to the experiment in section 4.1.2.

4.1.4 Split-Merge Sampling

We investigate the effect of split-merge sampling in the context of hidden Markov models. A dataset from a mixture of VAR models with 3 states and 500 time points in total was generated, and the IHMM-MVAR was run with and without split-merge sampling enabled. In each step of the algorithm we calculate the normalized mutual information (NMI) between the sampled state sequence from the algorithm and the true state sequence. NMI is a measure of mutual dependence among two random variables (cf. Bishop et al. [2006, Chapter 1]), and has the upside for partionings that it is invariant to a permutation of the labels. Furthermore, we report the joint log-likelihood in each iteration. The results of the experiment can be seen in figure 4.11. The general picture is that



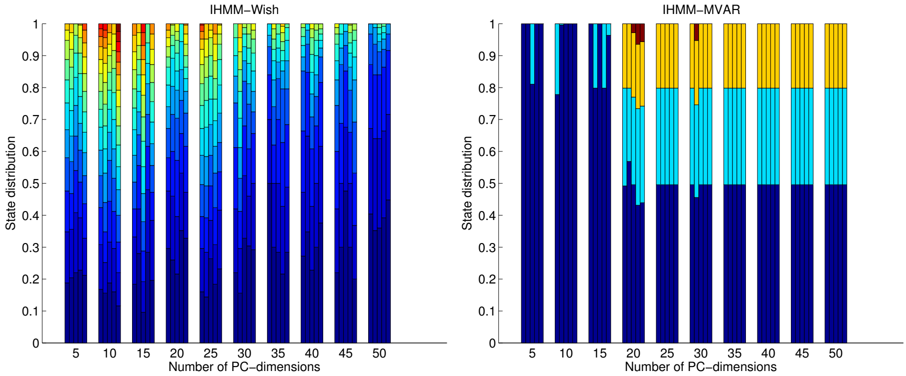
(a) IHMM-Wish **(b)** IHMM-MVAR

Figure 4.9: The number of states found by the IHMM-MVAR and the IHMM-Wish on a synthetic data set created from a mixture of 3 VAR's. The dimensionality of the data has been varied from 5 to 50 PC's (first) with increments of 5 and we ran each model 5 times. Each run is represented by one bar, and the height of the bar determines the number of states found. Each state is represented by a color, and the size of each color in the bar is proportional to the number of data points with that state value.

split-merge sampling has a positive effect on finding the true state sequence, but also has an overall higher joint-likelihood compared to only using Gibbs sampling.

4.1.5 Collating Multiple Data Sets

Ultimately we want to run this framework on real data, preferably resting-state data, in which there is evidence from the literature that dynamics exist. To test the model's behaviour on a scenario where we expect dynamics to exist, we have allowed for multiple data sets to be analysed at the same time in our implementation. This means we can *collate* different tasks and resting state data, and see if states are shared over the different data sets. This is done by keeping track of where a new data block starts and stops, and in those time points we update transition counts accordingly. Specifically, in each time point where a data block starts we subtract one transition from the previous time point. We tested this by taking a synthetic data set containing 6 VAR processes estimated from different tasks in the HCP data set, and collating them together two and two, yielding three tasks blocks. The result of running the IHMM-

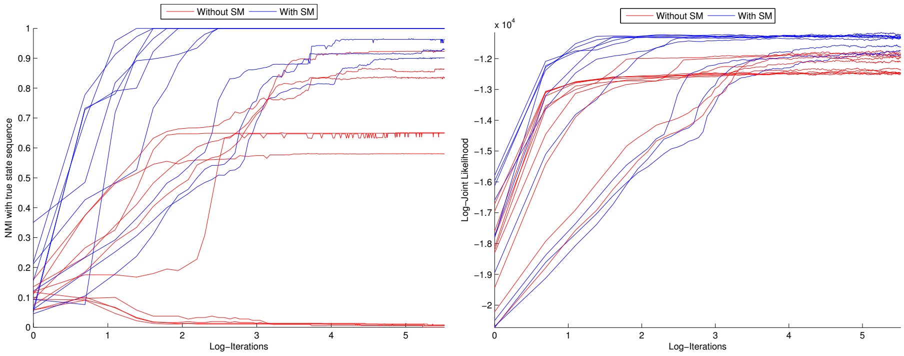


(a) IHMM-Wish

(b) IHMM-MVAR

Figure 4.10: The number of states found by the IHMM-MVAR and the IHMM-Wish on a synthetic data set created from a mixture of 3 VAR's. The dimensionality of the data has been varied from 5 to 50 PC's, skipping the first 10, with increments of 5. We ran each model 5 times. Each run is represented by one bar, and the height of the bar determines the number of states found. Each state is represented by a color, and the size of each color in the bar is proportional to the number of data points with that state value.

MVAR model on the collated data set can be seen in figure 4.12, and as we can see, the model finds the true state sequence.



(a) Normalized Mutual Information

(b) Log-Likelihood

Figure 4.11: Experiment on split-merge sampling. The normalized mutual information was calculated between the true state sequence and the state sequence sampled from an IHMM-MVAR without split-merge and with split-merge moves. Each inference procedure was run 10 times for 500 iterations. The plot 4.11a show the NMI for the 10 runs as a function of the number of iterations (logarithmic). We also report the log-likelihood calculated in each iteration of the algorithm in 4.11b.

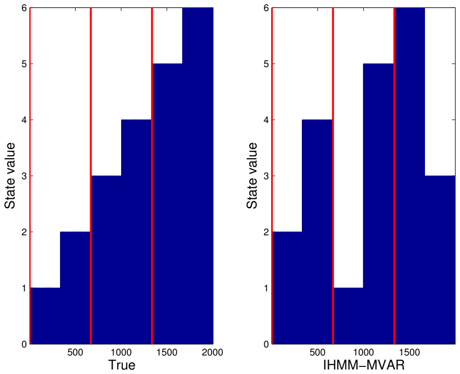


Figure 4.12: A plot of the state sequence found by the IHMM-MVAR if we collate 3 blocks of data together, with the state sequence shown in the left plot. The red vertical lines indicate the beginning of a new data block.

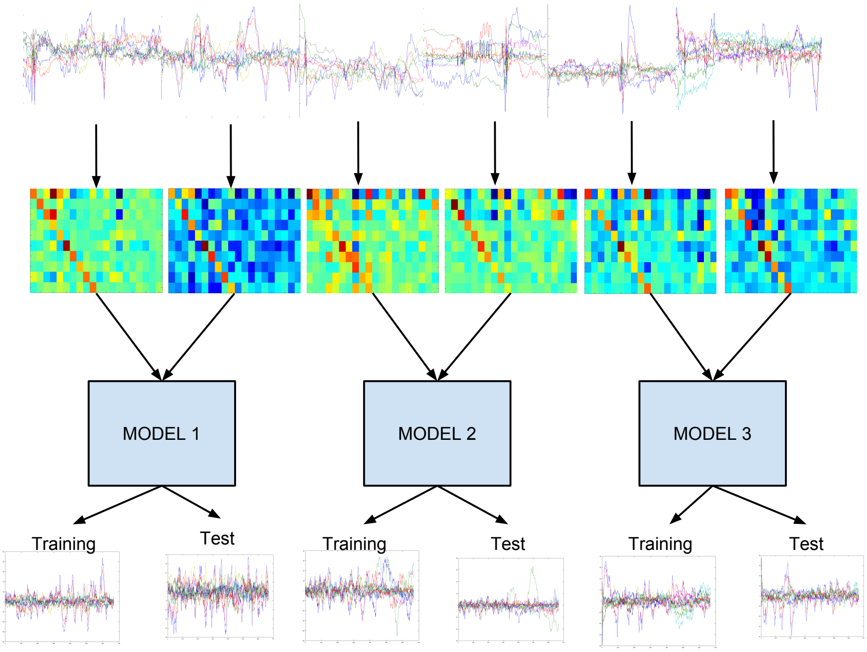
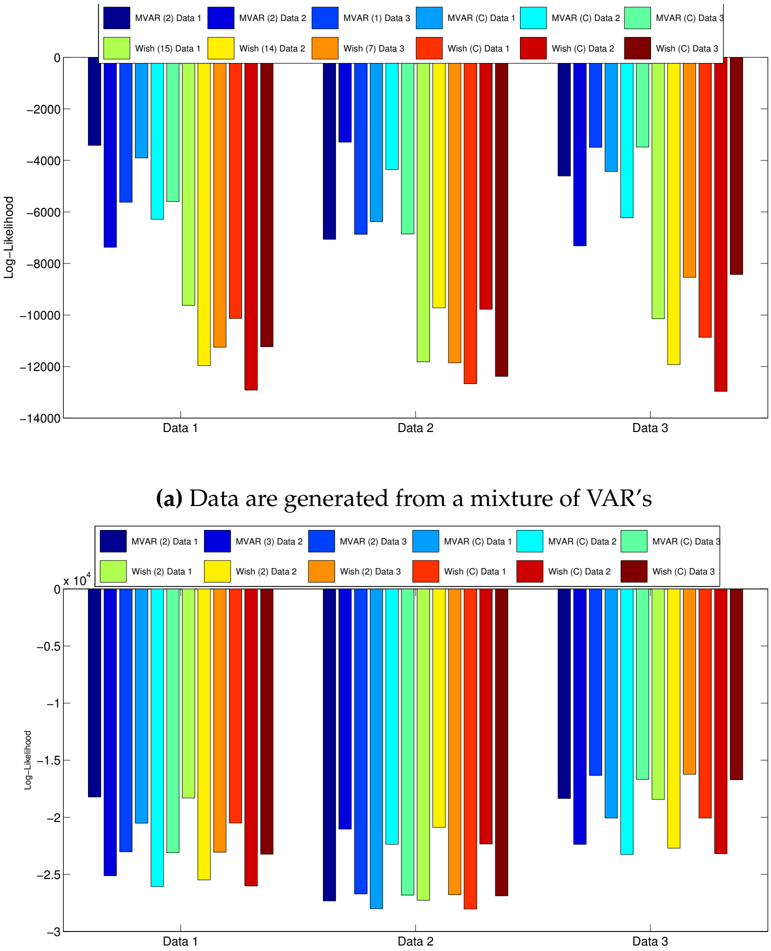


Figure 4.13: An illustration of the setup to validate the predictive framework described in section 2.7.

4.2 Validating the Predictive Likelihood Framework

In order to validate our prediction framework, described in section 2.7, we generated synthetic training and test data from three different models, trained models on the training data and ran our predictive likelihood framework on the test sets. A simplified schematic of the validation procedure can be seen in figure 4.13. We first fitted second order VAR models to all of the 6 task experiments from HCP data. Next, we constructed 3 mixture of VAR models therefrom, each containing two of the previously described VAR models. From each of these, we generated a training and a test dataset, and ran our IHMM models, both dynamic and static, on the three training data sets. Each training and test set comprised around 700 timepoints. Finally, we ran our prediction framework with the posterior samples from each of the models on the three test sets. The results can be seen in figure 4.14a. Similarly, we used the 6 tasks to generate a inverse Wishart data set, replacing the mixture of VAR's with an inverse-Wishart-mixture. The results from this can be seen in figure 4.14b.

Looking first at figure 4.14a, we see that the IHMM-MVAR models outperform the IHMM-Wish models, which seeems natural since the true data in this case is generated from a mixture of VAR's. Looking at each test data set individually, we see that among the MVAR models, the model that allows for dynamics and is trained on the corresponding training set performs the best in terms of predictive log-likelihood. One exception to this is in 'Data 3', where the static MVAR model and the dynamic model perform similarly (with a small advantage to the static model), which could be due to the dynamic model only finding one state. This is most likely again an effect of how the data was generated, namely that some of the tasks' VAR coefficients are not distinguishable with relatively few time points and low dimensionality. Nonetheless, this shows that our predictive likelihood framework is correct; the models that are trained on data generated from a mixture of two VAR's predict well on new data generated from the same mixture of VAR's. Similarly, looking only at the IHMM-Wishart models also in figure 4.14a, we reach the same conclusion, namely that the models trained on one VAR-task block predict relatively well on the corresponding test set from the same task block. Looking at the second figure 4.14b, where the data has been generated from a mixture of inverse-Wishart's, we see a more mottled picture, but patterns emerge. The IHMM-MVAR model performs at the same level in predictive log-likelihood as the IHMM-Wish when the trained model is used on the appropriate test data. As before, when models that are trained on one data set predict on new data generated by the same process as the training data, then they outperform other models trained on other data sets. This validates our predictive implementation, and furthermore gives us a strong indication that the IHMM-MVAR model extends the IHMM-Wish.



(a) Data are generated from a mixture of VAR's

(b) Data are generated from a Wishart-mixture model

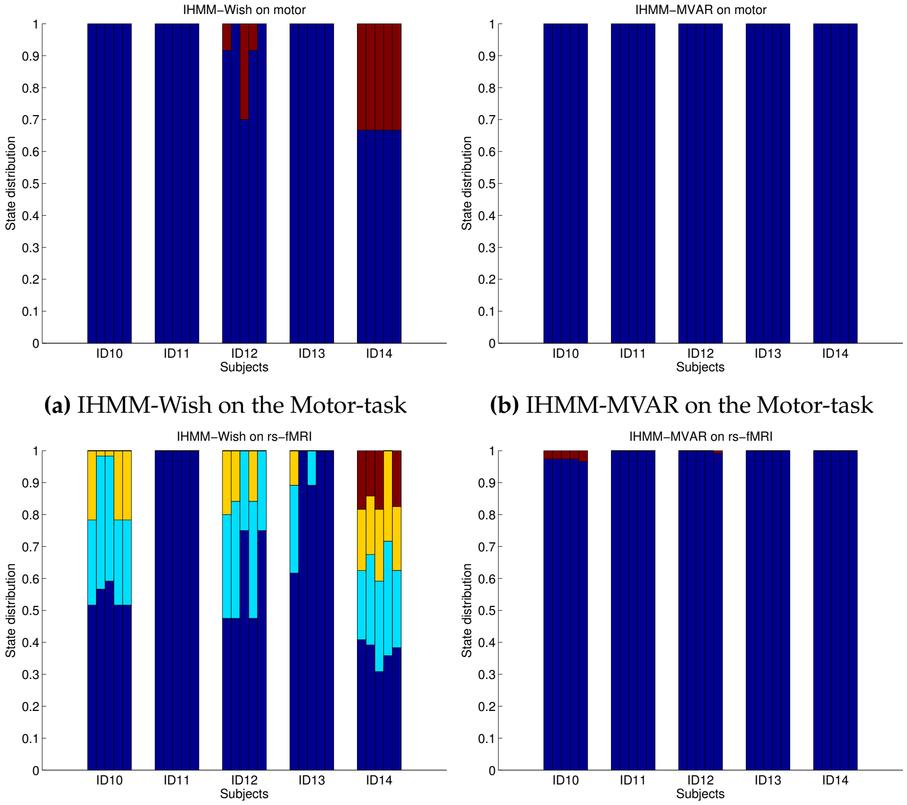
Figure 4.14: The predictive log-likelihood on test data generated by a mixture of VAR's (4.14a) and an inverse-Wishart-mixture (4.14b), respectively. Each bar represents how a model predicts on the test data at hand (the higher the better), and for each model it has been indicated in the legend text what data it has been trained on.

4.3 Experiments on Data from DRCMR

In this section we analyse the data from Danish Research Center for Magnetic Resonance (DRCMR) described in section 3.2. Out of the 30 subjects available to us with motor-task and resting state data, we analysed 5 subjects, one at a time. For each subject we conducted a PCA into 25 dimensions on the concatenated data , both motor and resting state, and afterwards we split each of the two blocks into a training and a test set of equal size (sometimes called split-half). Since the resting state experiment was twice as long, we used half of the time series to estimate a covariance structure, and used that as Σ_0 - the hyperparameter in the prior for the noise covariance in both the IHMM's. Now we trained our models IHMM-Wish and the IHMM-MVAR on the training sets and ran our predictive likelihood framework (cf. 2.7) on both the training and test sets. We ran the each inference procedure 5 times. Figure 4.15 shows the state distribution for each model on both motor-task and resting state data over all subjects and runs. The IHMM-MVAR model seems to consistently find only one state, in both motor and resting state and over subjects. This is what we would expect since the data we are training on are of length 120 time points, which is relatively few. The IHMM-Wish finds more states than the IHMM-MVAR as already hypothesized on resting state data, but only 1-2 states on motor data.

If we look at the predictive likelihood of the models on one subject in figure 4.16, we first notice that the MVAR models perform best in terms of training-scores, which we would expect since it is the most flexible model. Since only one state was found by the IHMM-MVAR in most of the runs there is very little difference between the dynamic and the static version of MVAR. We reach the same (unsurprising) conclusion as in section 4.2, that the more flexible models (here the MVAR) perform better in terms of predictive likelihood on the data that they were trained relative to models trained on another data set. Looking only at the Wish-models the same conclusion can be drawn.

Inspecting figure 4.16b, we see the predictive likelihood by all the models on the split-half test set of the motor and resting-state respectively. It is apparent that the IHMM-MVAR and its static counterpart are the two models that perform best on the two test cases, if they have been trained on the corresponding training set. This indicates that both the motor task and the resting state is characterized better by a MVAR model than it is by an IHMM-Wishart model. Even though the IHMM-Wish finds more states than the IHMM-MVAR, particularly for resting state data, it does not mean that these states characterize the data better (in terms of predictive likelihood) than the one state found by the IHMM-MVAR. This seems fairly reproducible over subjects, and the figures showing the results from the other 4 subjects can be seen in appendix B.2.



(a) IHMM-Wish on the Motor-task

(b) IHMM-MVAR on the Motor-task

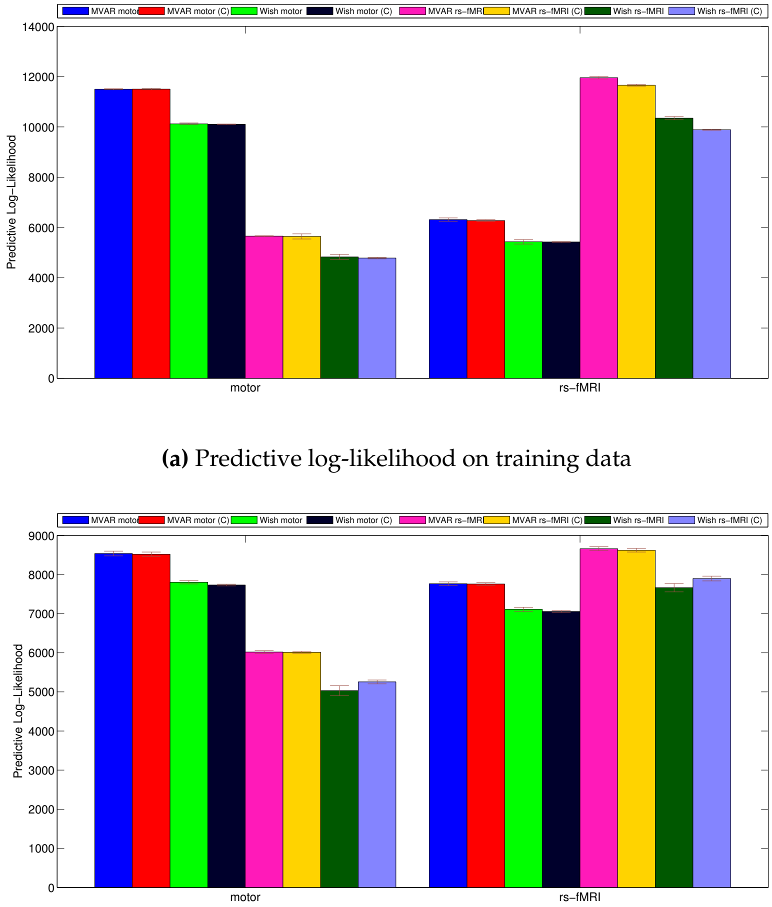
(c) IHMM-Wish on resting state

(d) IHMM-MVAR on resting state

Figure 4.15: In this figure we report the number of states found for each subject on each run on the DRCMR motor and resting state data. Each run is represented by one bar. Each state is represented by a color, and the size of each color in the bar is proportional to the number of data points with that state value.

4.3.1 Collating Task and Resting-State Data

As described in section 4.1.5, we want to investigate what the model infers on multiple real-world data sets from the same subject that have been collated together. We expect that if we collate different tasks-experiments, the models will infer multiple states that each are mainly present in one of the tasks. This means that the states found will be able to characterize the task from which they are inferred. So to investigate this we collated the motor and resting state data from the DRCMR data set for 5 subjects, again excluding half of the resting



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure 4.16: Predictive log-likelihood for 5 runs on both motor and resting-state data from DRCMR for a single subject (ID10). Each bar represents how a model predicts on the test data at hand (the higher the better), and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.

state data for estimation of the prior noise covariance Σ_0. We ran the IHMM-MVAR and the IHMM-Wish for 1000 iterations each on all 5 subjects, and an overview of the results can be seen in figure 4.17. We see that the IHMM-MVAR model almost exclusively (except for the first subject) finds only one state, pointing towards a conclusion that the VAR coefficients are static over motor and resting state. On the other hand, if we look at the results from the IHMM-Wish, we see that multiple states are found in each task block, and that the state sequence is significantly different between motor and resting state. This could either be indicative of the IHMM-Wish being better at discriminating between tasks, or that we simply do not have enough data or the proper preprocessing for the IHMM-MVAR model to find the 'dynamics' we are looking for.

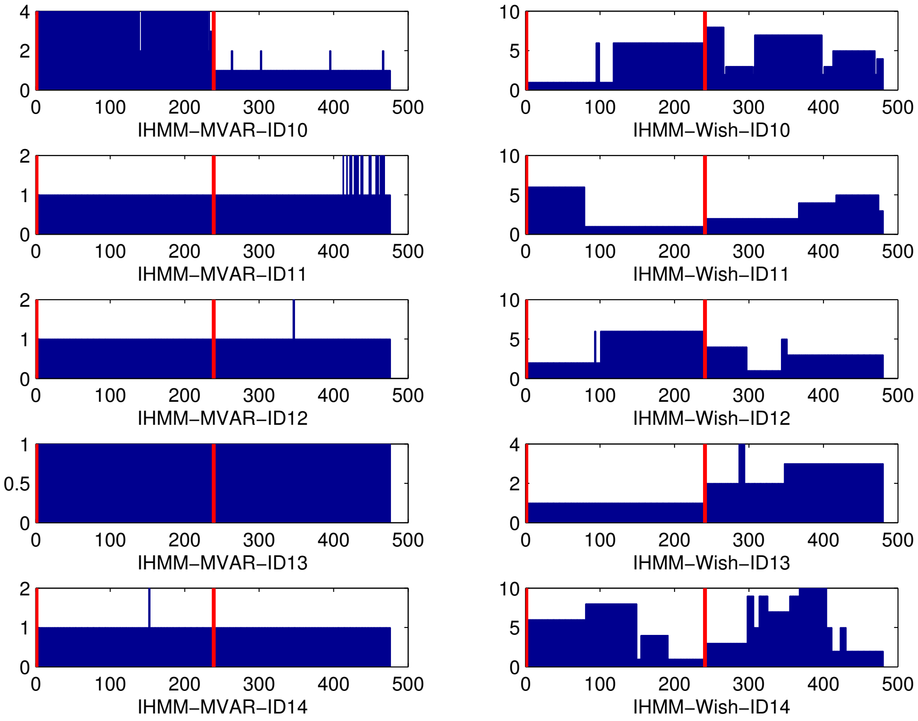
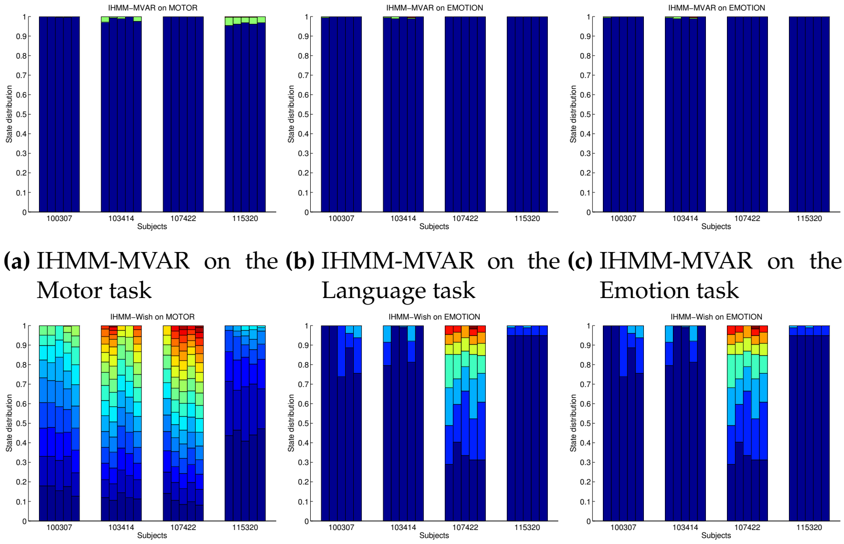


Figure 4.17: The state sequence estimated by the IHMM-MVAR and the IHMM-Wish on the collated motor and resting state data from DR-CMR. We ran the analysis for 5 subjects and each row of the plot corresponds to one subject. The red lines indicates where a transition from motor to resting state occurs.

4.4 Experiments on Human Connectome Project Task Data

In this section we describe the analysis and experiments carried out on data from the Human Connectome Project (HCP) (cf. Van Essen et al. [2012]). Each task and resting state experiment was in the HCP carried out twice, with different phase encoding used by the MRI scanner - 'left-right' denoted LR and 'right-left' denoted by RL. We will use one encoding, LR, for training and another, RL, for testing using our predicitve likelihood framework from section 2.7. We use the same tasks as described in section 4.1, namely 'Motor', 'Language' and 'Emotion', and append the time series from both encodings and resting state data together. As in section 4.1 we do dimensionality reduction by PCA, using principal component 11 to 35 (25 components in total). For 4 subjects we ran the IHMM-Wish and the IHMM-MVAR and restarted each inference procedure 5 times. An overview of the state distribution found by the two models over tasks and subjects can be seen in figure 4.18. We see that the IHMM-MVAR fairly consistent finds only 1 state in all the task experiments both over subjects and runs. Even the IHMM-Wish finds relatively few states (1-3) on the Language and Emotion task, if we ignore subject '107422'. The Motor task seems as the most 'dynamic' as the IHMM-Wish splits the data into many states. In general, the within subject variability over runs seems fairly limited since the same state proportions are roughly found in each restart.

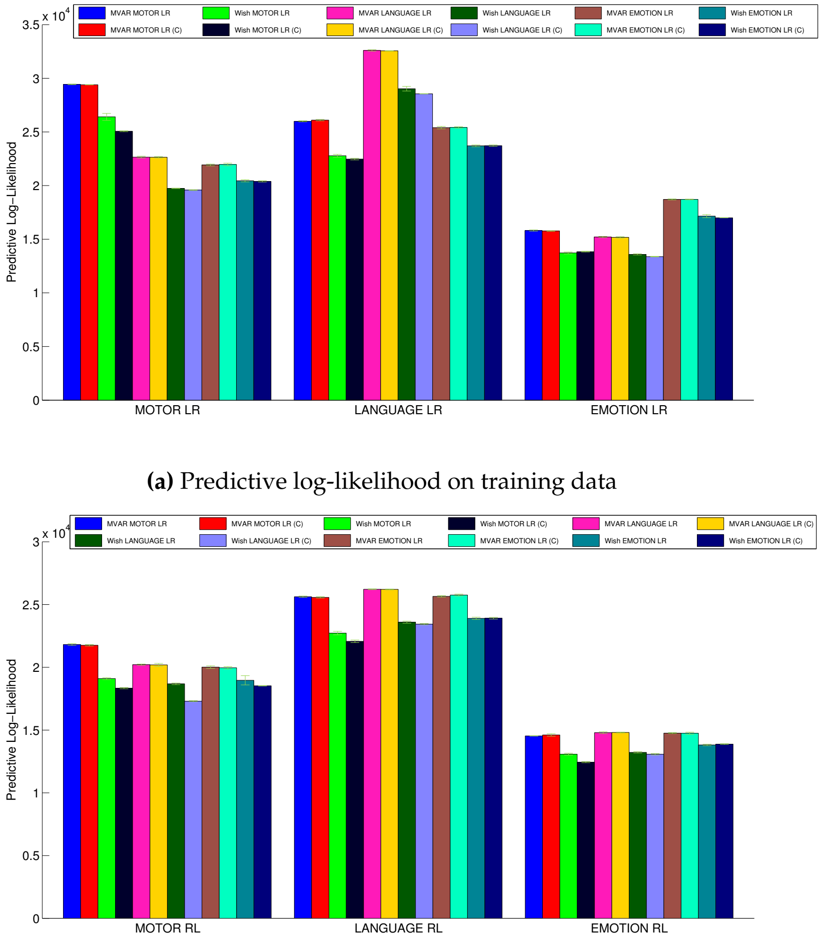
As in the previous experiment on the DRCMR data, we see in figure 4.19a that the IHMM-MVAR model has the best training-score on the task it has been trained on, probably because it is the most flexible model. Looking isolated on how the IHMM-Wish performs over the training data, we see as expected that the model predicts better on the tasks it has been trained on compared to models trained on a different task. The predictive likelihood on the test sets from a different phase encoding seen in figure 4.19b, shows that on the 'Motor' task the IHMM-MVAR model and the static VAR model trained on the motor task are better than the other models. This indicates that the models have found a fairly good characterization of the task. On the 'Language' and the 'Emotion' task it much more mottled, since the IHMM-MVAR model trained on the 'Motor' task and on the 'Emotion' are almost equal in performance. This could be explained by the 'Language' and 'Emotion' task being shorter in time, i.e. making it harder for the models to capture the underlying dynamics. Another explanation could be that our preprocessing has not been good enough, and that the PCA space we are investigating does not reflect the task specific variation we are trying to explain with the models.



(a) IHMM-MVAR on the Motor task **(b)** IHMM-MVAR on the Language task **(c)** IHMM-MVAR on the Emotion task

(d) IHMM-Wish on the Motor task **(e)** IHMM-Wish on the Language task **(f)** IHMM-Wish on the Emotion task

Figure 4.18: In this figure we report the number of states found for each subject on each run on the tasks, Motor, Language and Emotion from the HCP data. Each run is represented by one bar. Each state is represented by a color, and the size of each color in the bar is proportional to the number of data points with that state value.



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure 4.19: Predictive log-likelihood for 5 runs on the Motor, Language and Emotion experiments from one subject from the HCP. Each bar represents how a model predicts on the test data at hand (the higher the better), and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.

Chapter 5

Discussion

In the discussion we will try to address and answer the research questions posed in the introduction. A short recap of the models analysed and their relation to dynamic functional connectivity will be given in section 5.1. Next we will try to address the question of how we can interpret the dynamics we find from such models in section 5.2. In section 5.3 we discuss experiments on real world data, and finally in section 5.4 future work and potential extensions of the current frameworks will be presented.

5.1 Models for Dynamic Functional Connectivity

In this thesis we have analysed and partly implemented two models for dynamic functional brain connectivity. Both are grounded in the Bayesian nonparametric framework proposed by Van Gael [2012], namely the infinte hidden Markov model (IHMM). The first model, denoted IHMM-Wish, was analysed due to its similarities to many of the functional connectivity models in the literature. Functional connectivity is often understood as the correlation of activity between segregated brain regions; thus a correlation matrix can be thought of as a connectivity pattern from which the observed data is generated. Extending this to a dynamic setting, we imagine that the connectivity pattern (or covariance matrix) changes over time, and this is exactly what the IHMM-Wish models. A brain state is thus in this model defined purely by a certain covariance structure in the signal.

Elaborating on this we also analysed a second model, denoted IHMM-MVAR, that on top of the changing covariance matrix also includes a dynamic VAR process, i.e. each latent brain state is connected to a VAR process. We think of the VAR as a process that filters away the 'trivial' connectivity patterns. We could imagine that the activation in a brain-region is distributed to other regions in a certain way modelled by a VAR process. It should be noted here that we do not think of the VAR processes as a direct measure of effective connectivity as in the Granger causality framework (cf. Friston [2011]) or in VAR models described in Friston et al. [2003].

We chose to model dynamic functional connectivity using a non-parametric Bayesian approach. The great upside of using Bayesian non-parametric models is that we avoid choosing the number of states that we expect to find in the data; this parameter is automatically learned from data through the inference procedure. We believe that this is a correct way to model dynamics, since if the number of states is simply chosen by some heuristic we will always find evidence for that number of states. One of the goals when developing and implementing the models in this project has been to use as few heuristics as possible in the inference procedure. We tried to develop models that are not dependent on preprocessing steps such as sliding-windows and removal of noise artefacts.

Comparing the two models considered in this thesis, we look at the IHMM-MVAR as an extension of the IHMM-Wish. This has been validated throughout the results of the report. We saw in the synthetic study in section 4.1, that if we generated data with a varying covariance, mimicking an inverse Wishart mixture model, the IHMM-MVAR was able to infer the true state sequence and covariance structures just as the IHMM-Wish. Furthermore, when we validated our predictive likelihood framework in section 4.2, we saw, when we again generated data with varying covariance structure, that the IHMM-MVAR model obtained a predictive likelihood almost equivalent to the IHMM-Wish.

5.2 On the Interpretation of Dynamic Functional Connectivity

When analysing dynamic functional connectivity it is very important to be aware of what conclusions that can be drawn based on output from a model. We must be able to trust what the model infers, and we must see all results in light of the limitations of the model. We saw in synthetic studies that when generating data from a mixture of VAR's that the IHMM-Wish heavily overestimates the number of mixture components (or states). This points towards the conclusion that a simple model finds arbitrary many 'dynamics' in data generated from a much more complex model. Such a model mismatch can potentially lead to 'synthetic dynamics' if one is not careful. We have not been able to synthetically create a situation where the IHMM-MVAR overestimates the number of states and the IHMM-Wish finds the correct number of states. Such a scenario could happen if the VAR coefficients were (slighty) dynamic and the noise covariance static, but this requires more attention.

On real-world data, both on data from DRCMR in section 4.3 and on HCP data in 4.4, we saw that the the IHMM-MVAR typically found fewer states than the IHMM-Wish. But the IHMM-MVAR was still better to characterize the given task in terms of predictive likelihood on unseen data from the same task. This may imply that the multiple states found by the IHMM-Wish are not useful for characterizing the task at hand, and that some of the dynamics are somewhat synthetic stemming from a model mismatch.

When doing dimensionality reduction, as we have done in this project with PCA, it is also important to understand the influence of the dimensionality in the data. We have particularly seen in a synthetic experiment in section 4.1 that the number of states found, especially by the IHMM-MVAR, is underestimated when using few dimensions and at the same time having relatively few data points. This can be explained by the way we constructed the synthetic data, since it was based on finding VAR coefficients from HCP task data, where we expected that the processes would be significantly different. The result we saw in this dimensionality experiment could suggest otherwise for low-dimensional PCA data (5-15 components). Further investigation is needed into this by finding a structured way to generate stable VAR processes, i.e. by sampling them and not estimating them from data, to ensure that they are different.

5.3 Characterization of Task-Based Brain States

An end-goal for the models describing dynamic functional connectivity, is to run them on resting state data in a fully unsupervised setting. But before we reach that goal, we must first investigate what the models infer on data where we have some kind of 'ground truth' available. In this project, we build our analysis on the assumption that the dynamics extracted from different task-based experiments should be different. We do not expect functional connectivity to manifest in the same way in two different tasks. Working on data from DRCMR and HCP, we tried to quantify how well a model could characterize a given task by the predictive likelihood framework presented in section 2.7.

Looking at the results from the DRCMR, we saw that the VAR models performed the best in terms of predictive likelihood on the test data, we had held out from training. In most cases the IHMM-MVAR only found one state, and we saw that a static VAR model could better characterize any given task than a dynamic IHMM-Wish model in terms of predictive likelihood. With the DRCMR data we tried to mimic a change in dynamics by collating together motor task and resting state data and running the IHMM's on the collated data set. But the IHMM-MVAR still found only one state over both task blocks, indicating that the filtering process for a subject is constant over tasks, and what is left after filtering is static noise. The IHMM-Wish found multiple states in both task blocks in collated data, and the states found were largely only present in one of the blocks. One could interpret this as the IHMM-Wish being better at characterizing different tasks and separating them from each other, which would contradict our result from the predictive likelihood analysis. We must conclude that further analysis is needed into this matter.

In the results from the HCP data analysis, we again saw that the IHMM-MVAR found largely only one state on individual tasks. From the predictive likelihood analysis we saw in some cases that an IHMM-MVAR model trained on one task was not able to capture the same task better than an IHMM-MVAR model trained on a completely different task. This could substantiate our conclusion that the VAR-process parameters are static over tasks. But it could also be an indication that the IHMM-MVAR does simply does not capture the task specific characteristics, and that maybe another model is needed. Yet another explanation for this could be the choice of tasks. We saw for instance in the predictive likelihood that the 'Motor' task was characterizable by the IHMM-MVAR relative to the other models trained, but that the 'Emotion' task was not. It could be that the 'characteristics' we are looking for in a task are not present in the 'Emotion' task.

All of the results and conclusions drawn in this project should be read with caution, due to the lack of preprocessing analysis, which has been out of scope of this thesis. A problem with most fMRI data preprocessing pipelines, is that each experiment (task or resting-state) on a single-subject is processed individually, which could make it harder to distinguish between two tasks. For instance in the DRCMR data the motor task and resting state data was processed slightly different, which is definitely a problem when we are trying to characterize each of them in the same setting. Another preprocessing problem could be the PCA that we carried out. PCA maps the data into a subspace while preserving the maximum amount of variance possible, but can be sensitive to noise outlies and artefacts such as spikes. We therefore cannot be certain that this type of dimensionality reduction is optimal for the problem at hand, because the variance that the PCA preserves is maybe not useful for analysing the functional dynamics. Using ICA could maybe have improved our analysis. We could have chosen IC's by visual inspection that pertained physiological meaning (as it was done by Allen et al. [2012], Yu et al. [2015]) and thereby being certain what IC's we would expect to be 'dynamic' in the context of the task we were analysing.

5.4 Future Work

Concluding the discussion we will describe the outlook and future work regarding dynamic functional connectivity and the models analysed in this thesis. Preprocessing has been a factor of uncertainty in this project, and to get a clearer view of dynamic functional connectivity the influence of preprocessing such as dimensionality reduction methods must be investigated further. Work by Zalesky et al. [2014] suggests that there exists a modular structure in the brain where only relatively few connections are dynamic. Translating this into the context of the thesis we could incorporate a binary variable per dimension in the IHMM (both Wish and MVAR) that controls the 'dynamicness' of each dimension. Learning these variables could helps us understand to what extent dynamics are global at whole-brain level or very localized. Work by Korzen et al. [2014] suggested that the dynamics were not generalizable over subjects, which is why we stuck to analysing one subject at a time. But this could also be because the model does not incorporate subject variability directly. Future work could therefore include extending the IHMM-Wish or IHMM-MVAR to model population differences, such that inferences about dynamics could be made at group level by running the models on multiple subjects at a time.

To validate that the models can extract reasonable brain states, we used data from different task experiments and collated them together in the attempt to create semi-synthetic dynamics. This had the downside that each data set was preprocessed separately and some steps differed from task to task. Optimally we would want a data set where a subject performed a multitude of tasks in the same experiment, to see what the model found. Hopefully, we would find that the states extracted from one task were significantly different from those from another task.

In this project we investigated functional connectivity with fMRI data, but one could consider using other modalities. EEG data seems like the obvious choice for the models we have presented here, due to the low spatial resolution and high temporal resolution.

Chapter 6

Conclusion

In this master thesis we investigated functional brain connectivity, based on functional magnetic resonance imaging (fMRI), in a dynamic setting. We considered a Bayesian statistical approach, where two models were (partly) implemented and analysed. Both models were based on the infinite hidden Markov model (IHMM) first presented by Beal et al. [2001]. In the IHMM each data point is assumed to have a discrete latent representation, a *state* value. All data points with the same state value have associated parameters defining the state. The first model analysed was the IHMM-Wish that models the signal as a normally distributed variable with a changing covariance matrix over time. The second model, the IHMM-MVAR, extends the IHMM-Wish by assuming that the mean of the signal can be explained by a vector autoregressive (VAR) process that can change over time along with the covariance of the signal. Each model represented a way of modeling functional connectivity (FC), the IHMM-Wish being the simpler model that describes FC as the covariance between brain regions, and the IHMM-MVAR the more elaborate model that on top of the covariance between regions also models a signal filtering by a VAR process.

In synthetic studies, where we generated data from the two models, we found that the IHMM-MVAR was able to capture the true parameters in data generated from a mixture of inverse-Wishart's (mimicking an IHMM-Wish). The IHMM-Wish, on the other hand, greatly overestimated the number of states found in data from a mixture of VAR's, displaying that if we use a simple model to estimate the number of states in complex data we can be arbitrarily wrong.

We tried to see how the two models performed on real-world task data from the Danish Research Center for Magnetic Resonance (DRCMR) and the Human

Connectome Project (HCP). The IHMM-MVAR was consistently better at characterizing the task data, compared to the IHMM-Wish, in terms of predictive likelihood on test data from the same task. The IHMM-MVAR mostly found only one state indicating that the VAR coefficients are mainly static in the tasks we analysed. But when running the model on a collated data set with both motor task and resting state data the IHMM-MVAR still only found one state, indicating that the two experiments should be characterizable by the same parameters; a conclusion we find unlikely. Some of the results and conclusions must be read with care and further investigation is needed into the influence of preprocessing, for instance dimensionality reduction by principal component analysis. In general, conclusions about dynamic functional connectivity should be expressed with caution and always be seen in the context of the model used and its limitations.

Appendix A

Derivations

A.1 Mixture of Vector Autoregressive Models: Inference by Expectation-Maximization

The model parameters can be estimated by an expectation maximization algorithm (cf. Bishop et al. [2006]), which works by alternating between two steps. First in the E-step, we calculate the responsibilities, $\gamma_{t,k}$ of the data points to each AR-process by

$$\gamma_{t,k} = p(z_t = k|\mathbf{x_t}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_t|z_t = k, \boldsymbol{\theta})p(z_t = k)}{\sum_{k'} p(\mathbf{x}_t|z_t = k', \boldsymbol{\theta})p(z_t = k')}, \tag{A.1}$$

, in which $\boldsymbol{\theta}$ are all relevant model parameters. Note here that is is only possible to calculate this quantity for $t = M...T$.

In the second step, the M-step, we update all relevant model parameters for each process given their responsibilities by the following maximization problem,

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}), \tag{A.2}$$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \left[\sum_{z} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{z}|\, \boldsymbol{\theta})\right] + \ln p(\boldsymbol{\theta}). \tag{A.3}$$

From Bishop we have that (A.3) is equivalent to,

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \left[\sum_t \sum_k \gamma_{tk} \left(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_t | \mathbf{A}^{(k)} \bar{\mathbf{x}}_t, \sigma_t^2 \mathbf{I}) \right) \right] \\
+ \sum_k \ln \mathcal{N}(vec(\mathbf{A}^{(k)}) | \mathbf{0}, \mathbf{R})
\end{aligned}
\tag{A.4}
$$

, in which $\mathcal{N}(x|\mu, \Sigma)$ is the probability density function of a multivariate Gaussian with mean μ and variance Σ evaluated at point x. All parameters can be estimated by differentiating (A.4) with respect to the parameter in question, equating to zero and solving for the parameter.

The AR-model parameters $\mathbf{A}^{(k)}$ can be estimated as,

$$
\mathbf{A}^{(k)} = \mathbf{X}^{(k)} \mathbf{W}^{(k)} \left(\bar{\mathbf{X}}^{(k)} \right)^T \left(\bar{\mathbf{X}}^{(k)} \mathbf{W}^{(k)} \left(\bar{\mathbf{X}}^{(k)} \right)^T + \mathbf{R} \right)^{-1}, \tag{A.5}
$$

in which, $\mathbf{X}^{(k)}$ is the collection of all data points belonging to cluster k of size , $\bar{\mathbf{X}}^{(k)}$ is the appropriate past matrix of $\mathbf{X}^{(k)}$ and $\mathbf{W}^{(k)}$ is a $N_k \times N_k$ diagonal matrix with elements $\frac{\gamma_{t,k}}{\sigma_t^2}$., where N_k is the number of data points assigned to cluster k.

The mixing coefficients can be estimated by,

$$
\pi_k = \frac{\sum_t \gamma_{t,k}}{\sum_{t,k} \gamma_{t,k}} = \frac{\sum_t \gamma_{t,k}}{T - M}. \tag{A.6}
$$

The time dependent noise σ_t^2 can be estimated as,

$$
\sigma_t^2 = \frac{\sum_k \gamma_{t,k} \left(\mathbf{x}_t - \mathbf{f}_k(\mathbf{x_t}) \right)^T \left(\mathbf{x}_t - \mathbf{f}_k(\mathbf{x_t}) \right) + 2\beta_2}{P + 2(\beta_1 + 1)} \tag{A.7}
$$

The whole procedure is summarized in Algorithm 2.

```
Input  : X, K, M
Output: Clustering of time points into AR-processes
Initialize relevant parameters;
while not converged do
    E-Step: Estimate responsibilities;
    Update $\gamma_{t,k}$ by (A.1) ;
    M-Step: Estimate model parameters for each cluster;
    for $k = 1...K$ do
        Update $\mathbf{A}_k$ by (A.5);
        Update $\pi_k$ by (A.6);
        Update $\sigma_t^2$ by (A.7);
    end
    Evaluate likelihood;
end
```

Algorithm 2: EM-procedure for mixture of VAR's

A.2 Marginalization: IHMM-MVAR

The joint likelihood of the observed data and the coefficients of the VAR-processes can be written as,

$$\begin{aligned}
p(\mathbf{A}, \mathbf{X}, \mathbf{\Sigma}|\mathbf{Z}) = &\prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} |\Sigma^{(z_t)}|^{-\frac{p}{2}}| \exp\left(-\frac{1}{2}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)^T(\sigma_t^2\Sigma^{(z_t)})^{-1}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)\right) \\
&\prod_k (2\pi)^{\frac{-ppM}{2}} |\mathbf{R}|^{\frac{-p}{2}} |\Sigma^{(k)}|^{\frac{-pM}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{R}^{-1}\mathbf{A}^{(k)T}\Sigma^{-(k)}\mathbf{A}^{(k)})\right) \\
&\prod_k \frac{|\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})} |\Sigma^{(k)}|^{\frac{-v_0+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(\eta\Sigma_0\Sigma^{-(k)}\right)} \\
= &\prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k (2\pi)^{\frac{-ppM}{2}} |\mathbf{R}|^{\frac{-p}{2}} \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})} |\Sigma^{(k)}|^{\frac{-(v_0+n_k+pM)+p+1}{2}} \\
&\exp\left(-\frac{1}{2}\operatorname{tr}((\mathbf{X}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{X}}^{(k)})^T\Sigma^{-(k)}(\mathbf{X}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{X}}^{(k)})\right. \\
&\left. +\mathbf{R}^{-1}\mathbf{A}^{(k)T}\Sigma^{-(k)}\mathbf{A}^{(k)} + \eta\Sigma_0\Sigma^{-(k)}\right)
\end{aligned}$$

in which $\mathbf{X}^{(k)}$ is the collection of all data points belonging to process k, $\bar{\mathbf{X}}^{(k)}$ is the appropriate past corresponding to $\mathbf{X}^{(k)}$, the time dependent noise variances σ_t^2 have been multiplied onto the corresponding columns of $\mathbf{X}^{(k)}$ and $\bar{\mathbf{X}}^{(k)}$, n_k is the number of time points belonging to process k, and $\Sigma^{-(k)}$ is the inverse of

$\Sigma^{(k)}$. Looking only at the argument of the exponential we have that,

$$\begin{aligned}
\ln p(\mathbf{A}, \mathbf{X}, \mathbf{\Sigma}|\mathbf{Z}) \propto & \operatorname{tr}\Big(\Sigma^{-(k)}((\mathbf{X}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{X}}^{(k)})(\mathbf{X}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{X}}^{(k)})^T \\
& +\mathbf{A}^{(k)}\mathbf{R}^{-1}\mathbf{A}^{(k)T} + \eta\Sigma_0)\Big) \\
= & \operatorname{tr}\Big(\Sigma^{-(k)}(\mathbf{X}^{(k)}\mathbf{X}^{(k)T} - 2\mathbf{X}^{(k)}\bar{\mathbf{X}}^{(k)T}\mathbf{A}^{(k)T} \\
& + \mathbf{A}^{(k)}(\bar{\mathbf{X}}^{(k)}\bar{\mathbf{X}}^{(k)T} + \mathbf{R}^{-1})\mathbf{A}^{(k)T} + \eta\Sigma_0)\Big) \\
= & \operatorname{tr}\Big(\mathbf{S}_{\bar{x}\bar{x}}(\mathbf{A}^{(k)} - \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1})^T\Sigma^{-(k)}(\mathbf{A}^{(k)} - \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1}) + \hat{\mathbf{S}}\Sigma^{-(k)}\Big),
\end{aligned}$$

in which,

$$\begin{aligned}
\mathbf{S}_{\bar{x}\bar{x}} &= \bar{\mathbf{X}}^{(k)}\bar{\mathbf{X}}^{(k)T} + \mathbf{R}^{-1} \\
\mathbf{S}_{x\bar{x}} &= \mathbf{X}^{(k)}\bar{\mathbf{X}}^{(k)T} \\
\mathbf{S}_{xx} &= \mathbf{X}^{(k)}\mathbf{X}^{(k)T} + \eta\Sigma_0 \\
\hat{\mathbf{S}} &= \mathbf{S}_{xx} - \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1}\mathbf{S}_{x\bar{x}}^T.
\end{aligned}$$

By marginalizing over the VAR-coefficients we can arrive at

$$\begin{aligned}
p(\mathbf{X}, \mathbf{\Sigma}|\mathbf{Z}) = & \int p(\mathbf{X}, \mathbf{A}, \mathbf{\Sigma}|\mathbf{Z}) d\mathbf{A} \\
= & \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k (2\pi)^{\frac{-ppM}{2}} |\mathbf{R}|^{\frac{-p}{2}} \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})} \\
& |\Sigma^{(k)}|^{\frac{-(v_0+n_k+pM)+p+1}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-(k)}\hat{\mathbf{S}})) \\
& \int \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{S}_{\bar{x}\bar{x}}(\mathbf{A}^{(k)} - \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1})^T\Sigma^{-(k)}(\mathbf{A}^{(k)} - \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1})\right)\right) d\mathbf{A} \\
= & \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k |\mathbf{R}|^{\frac{-p}{2}} \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}}\Gamma_p(\frac{v_0}{2})} |\mathbf{S}_{\bar{x}\bar{x}}|^{-\frac{p}{2}} \\
& |\Sigma^{(k)}|^{\frac{-(v_0+n_k)+p+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\hat{\mathbf{S}}\Sigma^{-(k)})\right).
\end{aligned} \tag{A.8}$$

Finally, we can in (A.8) integrate out all Σ's yielding,

$$
\begin{aligned}
p(\mathbf{X}|\mathbf{Z}) &= \int p(\mathbf{X}, \boldsymbol{\Sigma}|\mathbf{Z}) d\boldsymbol{\Sigma} \\
&= \prod_t (2\pi\sigma_t^2)^{\frac{-p}{2}} \prod_k |\mathbf{R}|^{\frac{-p}{2}} \frac{|\eta\Sigma_0|^{\frac{v_0}{2}}}{2^{\frac{v_0 p}{2}} \Gamma_p(\frac{v_0}{2})} \\
& |\mathbf{S}_{\bar{x}\bar{x}}|^{-\frac{p}{2}} \frac{2^{\frac{(v_0+n_k)p}{2}} \Gamma_p(\frac{v_0+n_k}{2})}{|\hat{\mathbf{S}}|^{\frac{v_0+n_k}{2}}}
\end{aligned}
\tag{A.9}
$$

A.3 Parameter Posteriors: IHMM-MVAR

The posterior distributions used to sample the parameters in the model for the predictive likelihood framework will be derived here (cf. Fox [2009] for detailed walkthrough). From (A.8) we see from the exponential expression in **A**, that this has the form of a matrix normal distribution, and if we condition on $\Sigma^{(k)}$ we get,

$$
p(\mathbf{A}^{(k)}|\mathbf{X}, \Sigma^{(k)}) = \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}_{x\bar{x}}\mathbf{S}_{\bar{x}\bar{x}}^{-1}, \Sigma^{(k)}, \mathbf{S}_{\bar{x}\bar{x}})
\tag{A.10}
$$

Integrating out **A** yields (A.8), and from that we can see that the resulting has an inverse Wishart form in $\Sigma^{(k)}$ and so

$$
p(\Sigma^{(k)}|\mathbf{X}) = \mathcal{W}^{-1}(\Sigma^{(k)}; \hat{\mathbf{S}}, v_0 + n_k)
\tag{A.11}
$$

A.4 Predictive Posterior: IHMM-MVAR

Putting an improper 1/X prior on σ_t and conditioning on model parameters $\mathbf{A}, \Sigma^{(k)}$ and the state sequence **z**, we have the following 'joint' likelihood for each time point,

$$
\begin{aligned}
p(x_t, \sigma_t^2|A, \Sigma, z_t) = & \frac{1}{\sigma_t}(2\pi\sigma_t^2)^{\frac{-p}{2}} |\Sigma^{(z_t)}|^{-\frac{p}{2}} \\
& \exp\left(-\frac{1}{2}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)^T (\sigma_t^2\Sigma^{(z_t)})^{-1} (\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)\right) \\
& \propto \left(\sigma_t^2\right)^{\frac{-p-1}{2}} \exp\left(\frac{1}{\sigma_t^2} \cdot \frac{-1}{2}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)^T (\Sigma^{(z_t)})^{-1} (\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)\right)
\end{aligned}
\tag{A.12}
$$

The expression in (A.12) can be identified as an unormalized inverse Gamma distribution, and thus integrating out σ_t yields the inverse normalization constant, i.e.,

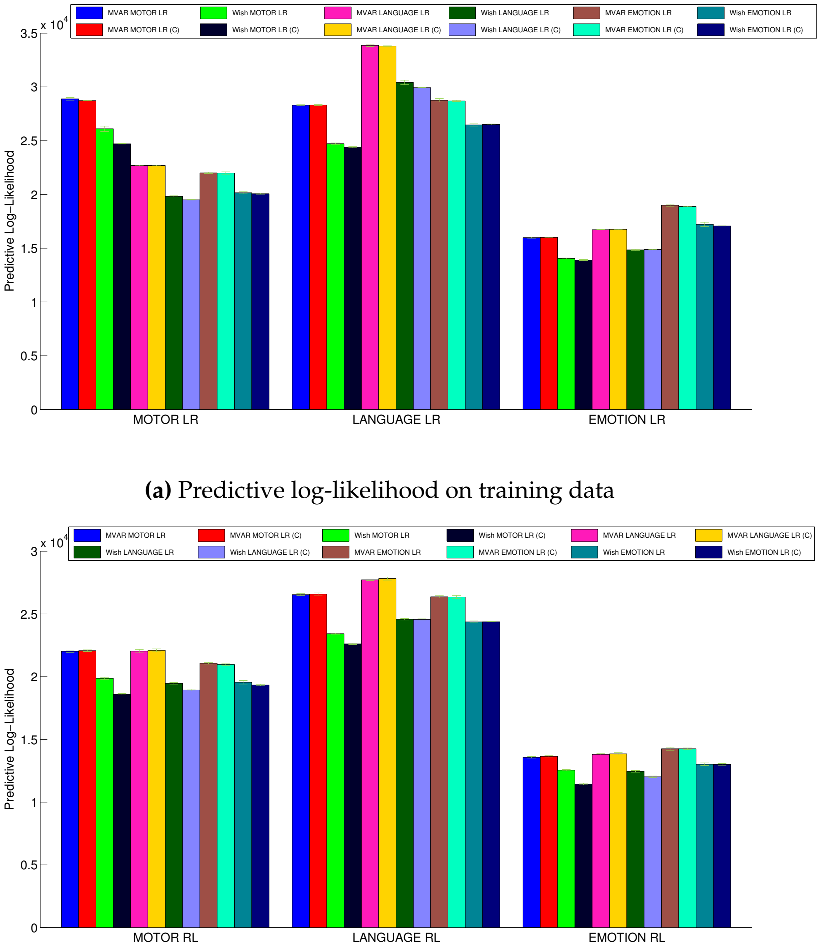
$$\begin{aligned} p(x_t|A,\Sigma,z_t) &= \int p(x_t,\sigma_t^2|A,\Sigma,z_t)d\sigma_t^2 \\ &= (2\pi)^{\frac{-p}{2}}|\Sigma^{(z_t)}|^{-\frac{p}{2}} \frac{\Gamma(\frac{p-1}{2})}{\left(\frac{1}{2}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)^T(\Sigma^{(z_t)})^{-1}(\mathbf{x}_t - \mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t)\right)^{\frac{p-1}{2}}} \end{aligned} \tag{A.13}$$

Appendix B

Results

B.1 Human Connectome Project: Predictive Likelihood Results

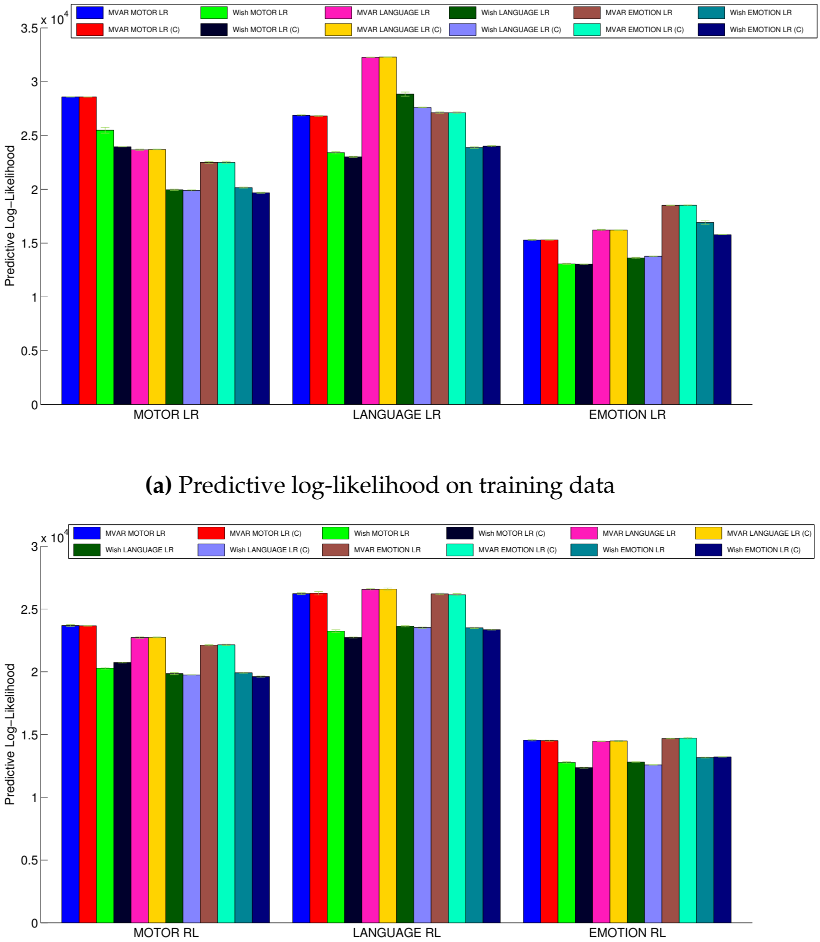
We ran the IHMM-MVAR and the Wish on three task experiments from the Human Connectome Project, a motor task experiment, marked *Motor*, a language processing experiment, marked *Language* and an emotion processing experiment, marked *Emotion*. A total of 500 subjects data was available for analysis, and we ran on 4 of them individually. For all subjects two runs of the same experiment was available each with a different phase encoding. We trained the models on the LR-phase encoding and tested the models using our predictive likelihood framework on the RL-phase encoding data. We have reported the results from one subject in the main report, the rest of the results are shown here.



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure B.1: Predictive log-likelihood for 5 runs on the Motor, Language and Emotion experiment from a subject ('103414') from the HCP. Each bar represents how a model predicts on the test data at hand (the higher the better), and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.



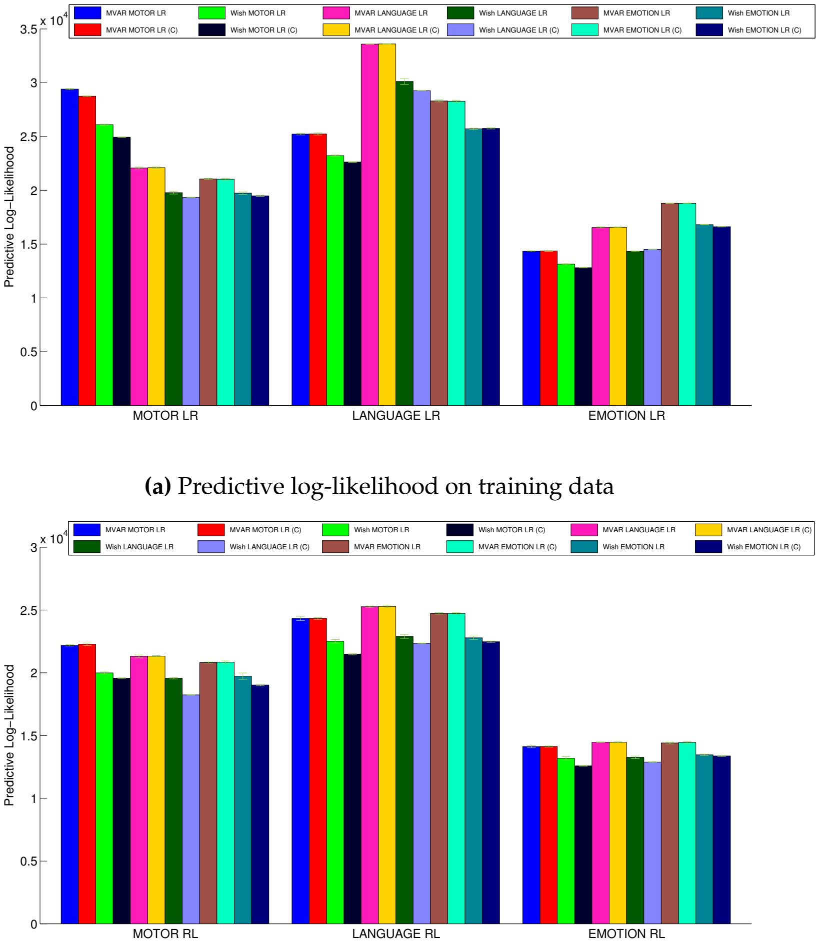
(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure B.2: Predictive log-likelihood for 5 runs on the Motor, Language and Emotion experiment from a subject ('107422') from the HCP. Each bar represents how a model predicts on the test data at hand (the higher the better), and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.

B.2 DRCMR Data: Predictive Likelihood Results

We ran the IHMM-MVAR and the Wish on two data sets, a motor task experiment, marked *motor*, and a resting state experiment, marked *rs-fMRI*, from the

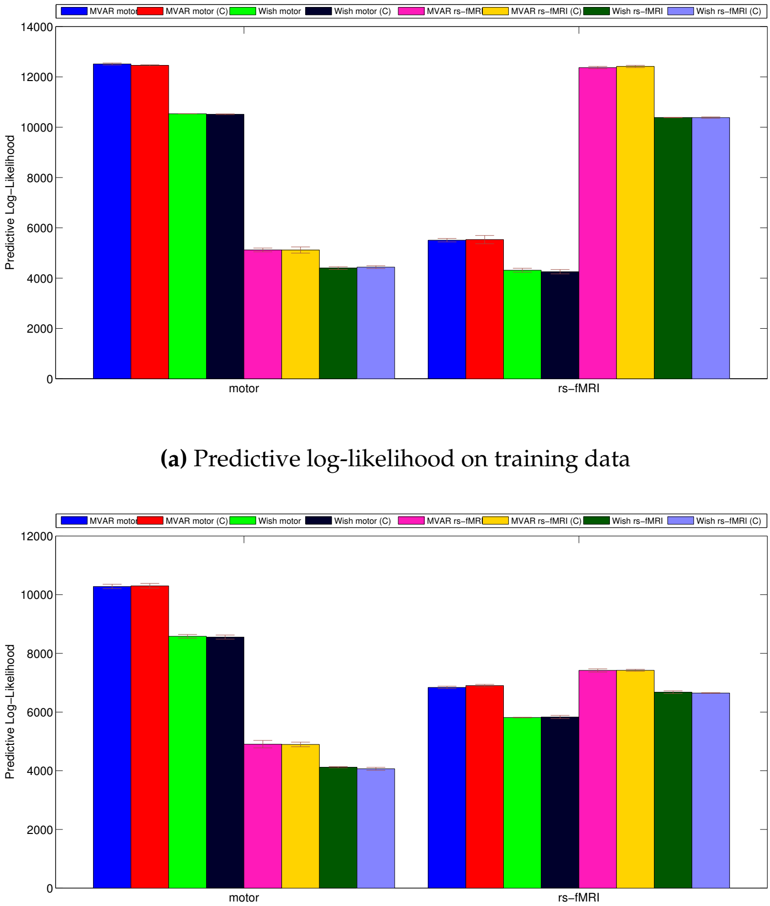


(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure B.3: Predictive log-likelihood for 5 runs on the Motor, Language and Emotion experiment from a subject ('115320') from the HCP. Each bar represents how a model predicts on the test data at hand (the higher the better), and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.

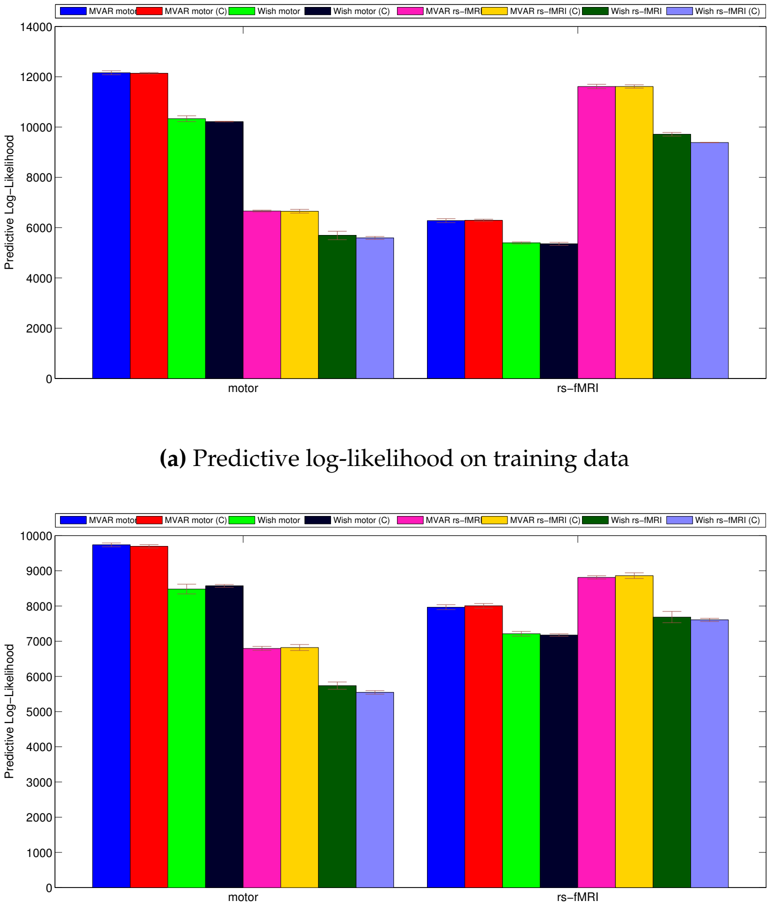
Danish Research Centre for Magnetic Resonance (DRCMR). A total of 30 subjects data was available for analysis, and we ran on 5 of them individually. We split each data set in two equal parts yielding a training and a test set for both experiments. In this section we report the predictive likelihood for each of the 4 subjects not shown in the main report on both the training and the test set.



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

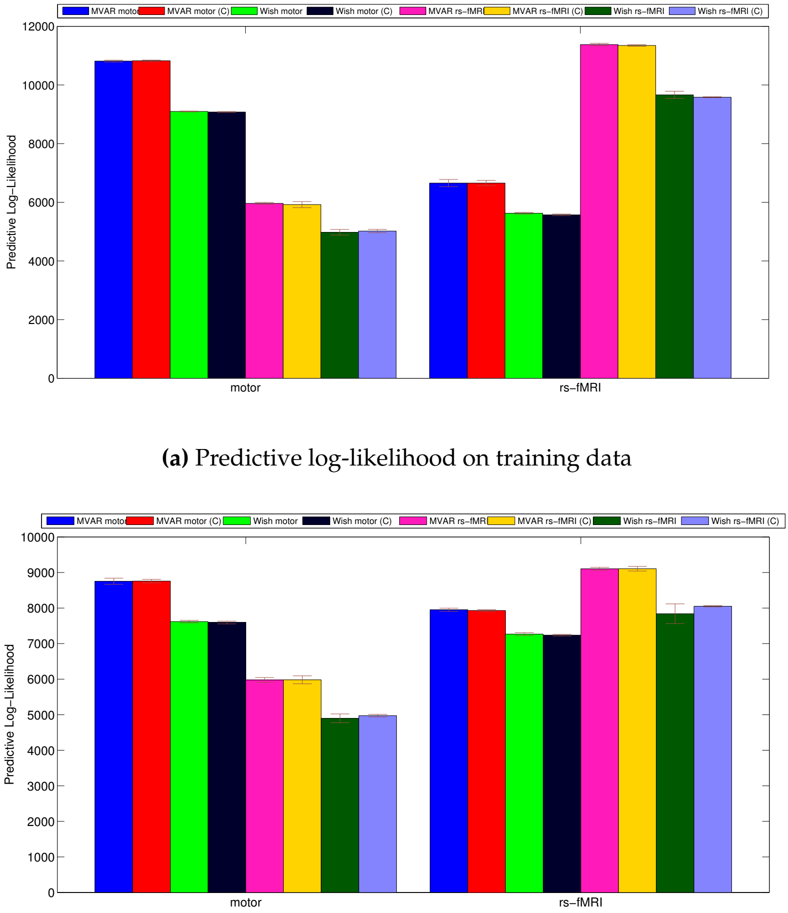
Figure B.4: Predictive log-likelihood for 5 runs on both motor and resting-state data from DRCMR for a single subject (ID11). Each bar represents how a model predicts on the data at hand, and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

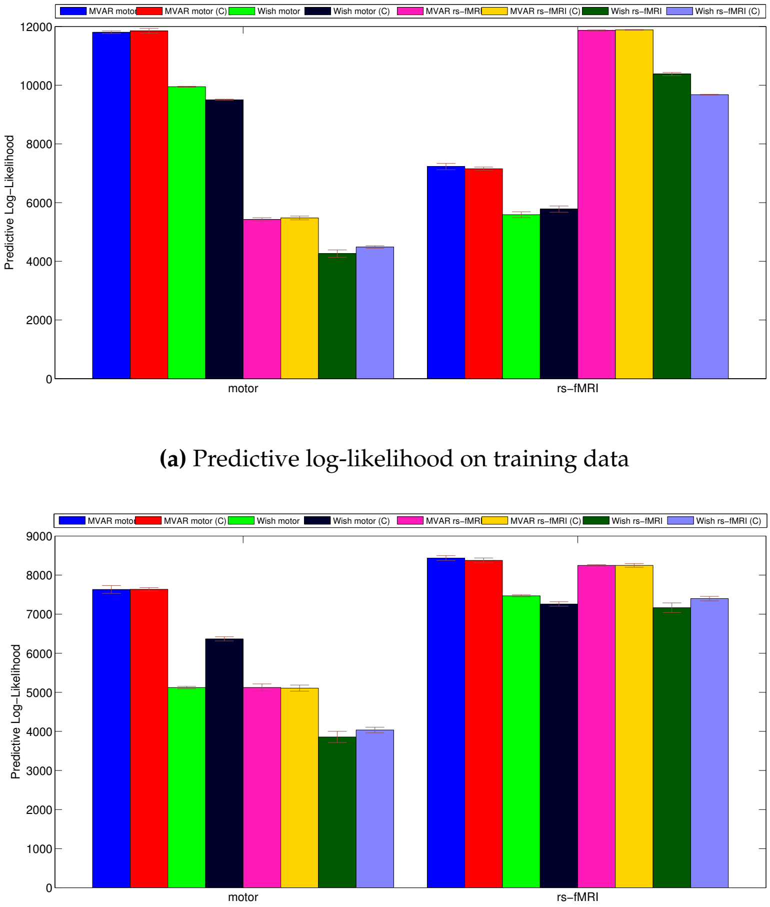
Figure B.5: Predictive log-likelihood for 5 runs on both motor and resting-state data from DRCMR for a single subject (ID12). Each bar represents how a model predicts on the data at hand, and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure B.6: Predictive log-likelihood for 5 runs on both motor and resting-state data from DRCMR for a single subject (ID13). Each bar represents how a model predicts on the data at hand, and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with ′C′ have been forced to be static.



(a) Predictive log-likelihood on training data

(b) Predictive log-likelihood on test data

Figure B.7: Predictive log-likelihood for 5 runs on both motor and resting-state data from DRCMR for a single subject (ID14). Each bar represents how a model predicts on the data at hand, and for each model it has been indicated in the legend text what data it has been trained on. The standard deviation over the 5 runs is represented by the errorbars on top of each bar. The models marked with 'C' have been forced to be static.

Appendix C

Project Plan and Auto-evaluation

C.1 Original Project Plan

Introduction

The human knowledge of how the brain works has grown over the past many decades partly due to advances in neuroimaging methods. Functional magnetic resonance imaging (fMRI) is a neuroimaging technique which relies on the fluctuation of oxygenated blood in the brain. This is used to find functionally correlated regions since neurally active areas will require more oxygen rich blood compared to neurally inactive areas. Blood-oxygen-level dependent (BOLD) signals are measured as a time series throughout different areas of the brain, and this gives a basis for a measurable difference in (indirect) activity both spatially and temporally.

Using fMRI one can study the functional connectivity (FC), which can be defined as the synchronous activity between regions of the brain. Most of the studies over the years have revolved around fMRI data from task-experiments, i.e. visual stimulation, eye movement and so on, and comparing these to each other, but lately a lot of focus has also been given to resting-state experiments.

Up until now most fMRI studies have assumed the measurements to be stationary over time, and in some sense resorting to take a temporal mean of the functional networks. Allen et. al. Hutchison et al. [2013] gives a recent review of the dynamic (as opposed to stationary) approaches for analyzing FC. In particular in Allen et al. [2012] a time-windowing approach is used to yield an connectivity network (correlation matrix) for each time window. This is done for multiple subjects and finally a K-means clustering is performed to find the K most occurring brain networks over time and subjects. K is chosen using a heuristic (in this case the elbow-criterion was used), and a large K indicates very advanced temporal dynamics whereas a low K would point to a more static FC.

The approach used by Allen et al. [2012] raises the question of how we define temporal dynamics. In Zalesky et al. [2014] a vector autoregressive model (VAR) was trained on the pairwise correlation between regions of interest extracted from windowing the original data. Using the VAR model a number of *null*-datasets were generated satisfying the hypothesis of stationarity of the signals, to test against the original data, thus determining what connections that can be deemed dynamic. In Majeed et al. [2011] on the other hand a repeating sequence approach was used to find common FC patterns over time windows, thus defining dynamics as the tendency of a brain network to re-occur. So it does not seem that there is a consensus of how to define temporal dynamics in terms of FC.

Project plan

In this master thesis we will investigate different models for modelling dynamic functional connectivity. The models considered will be extensions of already existing state-of-the-art frameworks for this type of analysis (i.e. Allen et al. [2012], Zalesky et al. [2014], Friston et al. [2003]). The extensions will be based on Bayesian non-parametric methods to overcome choosing certain parameters in the existing models. Furthermore, we will try to analyse what the consequences are of choosing a very simple model for a complex problem by a synthetic study. Finally, the models will be applied to real world data.

The main research questions can be formulated as follows,

- How can functional brain dynamics be modelled?
- What are the model differences? What are the benefits and shortcomings of using one model over the other?

- How does the choice of model influence the interpretation of dynamic functional connectivity?
- Can the models be used to characterize brain states in data from single-subject simple task-based fMRI studies, e.g. from the Human Connectome Project (HCP)Van Essen et al. [2012]?

A time schedule for the project period is given here (22-weeks in total)

Week(s)	Tasks	Milestones
1-6	Litterature study	
	Understanding and validating the IHMM	Week 6: Introduction of report is done together with description of models in theory section
7-10	Implementing other models including baseline	Week 9: All code is working and has been validated
11-16	Analyzing synthetic data	
17-19	Analyzing real world data	Week 17: Preliminary results are in report
20-22	Writing the report	Week 22: The report can be handed in

C.2 Learning Objectives Relevant for the Report

An M.Sc. from DTU:

- **Can identify and reflect on technical scientific issues and understand the interaction between the various components that make up an issue**:

  We model the brains functional connectivity in a dynamic setting, using statistical models.

- **Can, on the basis of a clear academic profile, apply elements of current research at international level to develop ideas and solve problems**:

  We specifically use non-parametric Bayesian models in fusion with current models from the field of neuroimaging.

- **Masters technical scientific methodologies, theories and tools, and has the capacity take a holistic view of and delimit a complex, open issue, see it in a broader academic and societal perspective and, on this basis, propose a variety of possible actions**:

  We try to answer how different models can lead to different interpretations of dynamic functional connectivity, and what consequences certain model choices have.

- **Can, via analysis and modelling, develop relevant models, systems and processes for solving technological problems**:

  We adapt approaches from the literature of functional connectivity to incorporate dynamics using Bayesian non-parametric modeling.

- **Familiar with and can seek out leading international research within his/her specialist area. can work independently and reflect on own learning, academic development and specialisation**:

  All of the approaches and models considered in this project are state-of-the-art in their respective fields.

C.3 Comments and Auto-evaluation

The research carried out in this project was overall kept within the boundaries of the original project plan. However, the HDPHMM framework by Fox et al. [2008] was not discovered before half way through the project period. This resulted in allocating time to understand and compare this framework to the IHMM, yielding less time for analysis on real-world data. It was agreed upon between the supervisors and the student that the statistical analysis and implementation was the important part of this thesis, and therefore the physiological interpretation of the results was deemed out of scope of the thesis.

A general comment to the time schedule for the project is that it took much longer time than anticipated to implement the IHMM-MVAR model and validate the correctness of the code (both IHMM-MVAR and IHMM-Wish). This means that the actual time spent on different parts of the project was more in cycles (implement and validate, implement and validate,...) than the linear time table presented above.

Bibliography

David J Aldous. *Exchangeability and related topics*. Springer, 1985.

Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, page bhs352, 2012.

Kasper Winther Andersen, Kristoffer H Madsen, Hartwig Roman Siebner, Mikkel N Schmidt, Morten Mørup, and Lars Kai Hansen. Non-parametric bayesian graph models reveal community structure in resting state fmri. *NeuroImage*, 100:301–315, 2014.

Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2001.

Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

Randy L Buckner, Jorge Sepulcre, Tanveer Talukdar, Fenna M Krienen, Hesheng Liu, Trey Hedden, Jessica R Andrews-Hanna, Reisa A Sperling, and Keith A Johnson. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to alzheimer's disease. *The Journal of Neuroscience*, 29(6):1860–1873, 2009.

VD Calhoun, T Adali, GD Pearlson, and JJ Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

Vince D Calhoun, Tülay Adali, Lars Kai Hansen, Jan Larsen, and James J Pekar. Ica of functional mri data: an overview. 2003.

Vince D Calhoun, Tom Eichele, and Godfrey Pearlson. Functional brain networks in schizophrenia: a review. *Frontiers in human neuroscience*, 3, 2009.

Karen L Campbell, Omer Grigg, Cristina Saverino, Nathan Churchill, and Cheryl L Grady. Age differences in the intrinsic functional connectivity of default network subsystems. *Frontiers in aging neuroscience*, 5, 2013.

Catie Chang and Gary H Glover. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

Jeff H Duyn. Eeg-fmri methods for the study of brain networks during sleep. *Frontiers in neurology*, 3, 2012.

Emily B Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An hdp-hmm for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319. ACM, 2008.

Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

Rainer Goebel, Alard Roebroeck, Dae-Shik Kim, and Elia Formisano. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging*, 21(10):1251–1261, 2003.

C Gössl, L Fahrmeir, and DP Auer. Bayesian modeling of the hemodynamic response function in bold fmri. *Neuroimage*, 14(1):140–148, 2001.

Michael D Greicius, Gaurav Srivastava, Allan L Reiss, and Vinod Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4637–4642, 2004.

Roger B Grosse and David K Duvenaud. Testing mcmc code. *arXiv preprint arXiv:1412.5218*, 2014.

L Harrison, William D Penny, and Karl Friston. Multivariate autoregressive modeling of fmri time series. *NeuroImage*, 19(4):1477–1491, 2003.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Michael C Hughes, Erik B Sudderth, and Emily B Fox. Effective split-merge monte carlo methods for nonparametric models of sequential data. In *Advances in Neural Information Processing Systems*, pages 1295–1303, 2012.

Human Connectome Project. WU-Minn HCP 500 Subjects + MEG2 Data Release: Reference Manual. http://humanconnectome.org/documentation/S500/HCP_S500+MEG2_Release_Reference_Manual.pdf, 2014. [Online; accessed 02-May-2015].

R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 2013.

Sonia Jain and Radford M Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

Gebhard Kirchgässner, Jürgen Wolters, and Uwe Hassler. *Introduction to modern time series analysis*. Springer Science & Business Media, 2012.

Vesa Kiviniemi, Tapani Vire, Jukka Remes, Ahmed Abou Elseoud, Tuomo Starck, Osmo Tervonen, and Juha Nikkinen. A sliding time-window ica reveals spatial variability of the default mode network in time. *Brain connectivity*, 1(4):339–347, 2011.

Josefine Korzen, Kristoffer H Madsen, and Morten Mørup. Quantifying temporal states in rs-fmri data using bayesian nonparametrics. In *HBM 2014*. Organization for Human Brain Mapping, 2014.

Kenneth K Kwong, John W Belliveau, David A Chesler, Inna E Goldberg, Robert M Weisskoff, Brigitte P Poncelet, David N Kennedy, Bernice E Hoppel, Mark S Cohen, and Robert Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679, 1992.

Xiao Liu and Jeff H Duyn. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 110(11):4392–4397, 2013.

Torben E Lund, Kristoffer H Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E Nichols. Non-white noise in fmri: does modelling have an impact? *Neuroimage*, 29(1):54–66, 2006.

Waqas Majeed, Matthew Magnuson, Wendy Hasenkamp, Hillary Schwarb, Eric H Schumacher, Lawrence Barsalou, and Shella D Keilholz. Spatiotemporal dynamics of low frequency bold fluctuations in rats and humans. *Neuroimage*, 54(2):1140–1150, 2011.

Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

Radford M Neal. Probabilistic inference using markov chain monte carlo methods. 1993.

Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.

David B Phillips and Adrian FM Smith. Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in practice*, pages 215–239. Springer, 1996.

Jim Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability & Computing*, 11(05):501–514, 2002.

Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Marcus E Raichle, Ann Mary MacLeod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

Peter M Rasmussen, Lars K Hansen, Kristoffer H Madsen, Nathan W Churchill, and Stephen C Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, 45(6):2085–2100, 2012.

Marieke L Schölvinck, Alexander Maier, Q Ye Frank, Jeff H Duyn, and David A Leopold. Neural basis of global resting-state fmri activity. *Proceedings of the National Academy of Sciences*, 107(22):10238–10243, 2010.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

Anne M Smith, Bobbi K Lewis, Urs E Ruttimann, Q Ye Frank, Teresa M Sinnwell, Yihong Yang, Jeff H Duyn, and Joseph A Frank. Investigation of low frequency drift in fmri signal. *Neuroimage*, 9(5):526–533, 1999.

Kerri Smith. Brain imaging: fmri 2.0. *Nature News*, 2012. URL http://www.nature.com/news/brain-imaging-fmri-2-0-1.10365.

Klaas Enno Stephan and Alard Roebroeck. A short history of causal modeling of fmri data. *Neuroimage*, 62(2):856–863, 2012.

Christoph Stippich et al. *Clinical functional MRI*. Springer, 2007.

Enzo Tagliazucchi, Pablo Balenzuela, Daniel Fraiman, and Dante Chialvo. Criticality in large-scale brain fmri dynamics unveiled by a novel point process analysis. *Frontiers in Physiology*, 3(15), 2012.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

Garth John Thompson, Matthew Evan Magnuson, Michael Donelyn Merritt, Hillary Schwarb, Wen-Ju Pan, Andrew McKinley, Lloyd D Tripp, Eric H Schumacher, and Shella Dawn Keilholz. Short-time windows of correlation between large-scale functional brain networks predict vigilance intraindividually and interindividually. *Human brain mapping*, 34(12):3280–3298, 2013.

David C Van Essen, Kamil Ugurbil, E Auerbach, D Barch, TEJ Behrens, R Bucholz, A Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

J. Van Gael. The infinite hidden markov model, 2010. http://mloss.org/software/view/205/.

Jurgen Van Gael. *Bayesian Nonparametric Hidden Markov Models*. PhD thesis, University of Cambridge, 2012.

Alan S Willsky, Erik B Sudderth, Michael I Jordan, and Emily B Fox. Nonparametric bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 457–464, 2009.

Qingbao Yu, Erik B Erhardt, Jing Sui, Yuhui Du, Hao He, Devon Hjelm, Mustafa S Cetin, Srinivas Rachakonda, Robyn L Miller, Godfrey Pearlson, et al. Assessing dynamic brain graphs of time-varying connectivity in fmri data: Application to healthy controls and patients with schizophrenia. *NeuroImage*, 107:345–355, 2015.

Andrew Zalesky, Alex Fornito, Luca Cocchi, Leonardo L Gollo, and Michael Breakspear. Time-resolved resting-state brain networks. *Proceedings of the National Academy of Sciences*, 111(28):10341–10346, 2014.