Bayesian Inference of High-Dimensional Cluster-Structured Ordinary Differential Equation Models with Applications to Brain Connectivity Studies

Qiannan Yin

Baoji, Shaanxi, China

M.S., University of Virginia, USA, 2015

B.S., Peking University, China, 2013

A Dissertation Presented to the Graduate Faculty of
University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
December, 2017

Abstract

We build a new ordinary differential equation (ODE) model for the directional interaction, also called effective connectivity, among brain regions whose activities are measured by intracranial electrocorticography (ECoG) data. In contrast to existing ODE models that focus on effective connectivity among only a few large anatomic brain regions and that rely on strong prior belief of the existence and strength of the connectivity, the proposed high-dimensional ODE model, motivated by statistical considerations, can be used to explore connectivity among multiple small brain regions. The new model, called the modular and indicator-based dynamic directional model (MIDDM), features a cluster structure, which consists of modules of densely connected brain regions, and uses indicators to differentiate significant and void directional interactions among brain regions. We develop a unified Bayesian framework to quantify uncertainty in the assumed ODE model, identify clusters, select strongly connected brain regions, and make statistical comparison between brain networks across different experimental trials. The prior distributions in the Bayesian model for MIDDM parameters are carefully designed such that the ensuing joint posterior distributions for ODE state functions and the MIDDM parameters have well-defined and easy-to-simulate posterior conditional distributions. To further speed up the posterior simulation, we employ parallel computing schemes in Markov chain Monte Carlo steps. We show that the proposed Bayesian approach outperforms an existing optimization-based ODE estimation method. We apply the proposed method to an

auditory electrocorticography dataset and evaluate brain auditory network changes across trials and different auditory stimuli.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor, Professor Tingting Zhang, for her continuous support in my graduate research and study. Her patience, expertise, and guidance helped me a lot in the process of conducting research and working on my dissertation. She taught me not only the professional knowledge and skills, but also the rigorous and diligent attitude towards research and life. Her assistance and suggestions helped me develop my statistial career plan. I am so lucky and proud to become one of her students.

Secondly, I would like to thank the rest of my committee members, Professor Karen Kafadar, Professor Chao Du, and Professor Yanjun Qi, for being so supportive and taking extra time and efforts on helping me with my dissertation proposal. I would also like to thank the faculty at Department of Statistics for their instruction and assistance both inside and outside of the classroom, and the administrator, Karen Dalton, for making our department like a family and responding promptly to all kinds of questions and requests.

Last but not least, I want to thank my family, especially my parents, for loving me with their whole hearts and accompanying me in both laughters and tears. I cannot achieve where I am without them. Moreover, I want to thank all my friends. They are great and intelligent people who will always be there for me. It is a great pleasure to team up or spend wonderful gatherings with them, which makes my life much more colorful.

Contents

Chapter 1

Introduction

This study focuses on modeling and making inferences about directional interactions among human brain regions. Three types of brain connectivity are investigated in the literature: anatomical connectivity, functional connectivity, and effective connectivity. According to a review by Daunizeau et al. (2011), anatomical connectivity, "i.e., the anatomical layout of axons and synaptic connections, determines which neural units interact directly with each other." Functional connectivity "subsumes non-mechanistic (usually whole-brain) descriptions of statistical dependencies between measured time series." Effective connectivity "refers to causal effects, i.e., the directed influence that system elements exert on each other" (p. 313). Because we aim to understand the causal architecture of brain networks, this thesis focuses on studying effective connectivity among brain regions.

1.1   Brain Activity Measurements

In general, the three types of brain activity time series that are used most commonly in brain studies are: fMRI (functional Magnetic Resonance Imaging) data, EEG (electroencephalogram) data, and ECoG (electrocorticorgraphy) data.

1.1.1   fMRI Data

fMRI is a functional neuroimaging procedure that measures brain activity by using the blood-oxygen-level dependent (BOLD) contrast mechanism. It relies on the fact that activated brain areas demand more oxygen, which results in an increased blood flow to those areas (Figure 1.1). Then, the local field in and around the blood vessels will change because of the change in blood oxygen levels, and the fMRI series can measure those field variation effects. [1] So it is not a direct measurement of the brain activity, and the data is collected non-invasively.

During the data collecting process, the subject must place his or her head inside a scanner, which scans the brain and records the data slice by slice (Figure 1.2). As a result, the data of the whole brain is not collected simultaneously. Each slice has a small time lag with the previous slice. Moreover, this data collection technique also places a light limitation to the spatial resolution of the data, because higher spatial resolution means more slides to scan and thus a longer time for the subject to remain immobile inside the scanner, which may make the subject feel uncomfortable. [52]
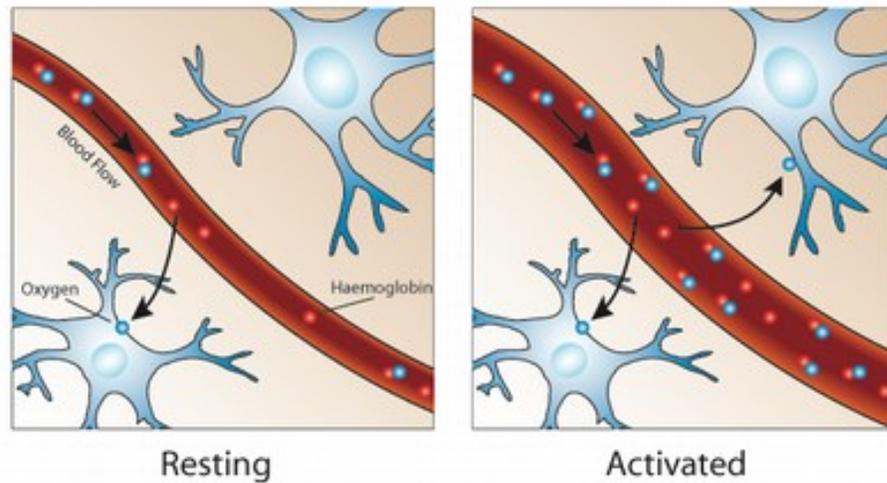
Figure 1.1: *Activated neurons are coupled with increased blood flow for more oxygen. [28]*

Although limited by the number of slices scanned in the experiment, fMRI data still has a relatively high spatial resolution of 4 to 5 mm to 1 mm. However, it has a relatively low temporal resolution of 1 s compared to other brain imaging data in milliseconds. One reason is that it takes a while for the neural activity to be reflected in the vascular system, and one has to wait for the net magnetization to recover before the next measure. Also, all the slices of the whole brain need to be finished before scanning the brain for another time. Finally, fMRI data collection usually involves much noise, which may come from the scanner, the subtle movement of the subject, some random thoughts in the brain, and so on. [52]
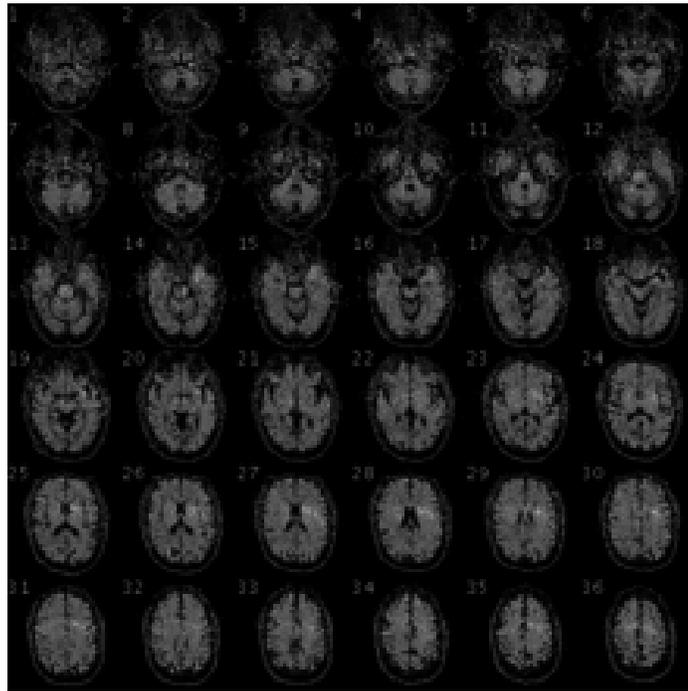
Figure 1.2: *An example of fMRI image of the brain.* [28]

1.1.2 EEG Data

EEG is a physiological method to measure the voltage fluctuations caused by the brain activity from electrodes placed on the surface of the scalp. The brain has millions of neural cells and they pass information from one to another via synapses, whose activity generates a subtle electrical impulse. Although one single impulse is too tiny to be captured, a group of neurons, if firing in sync, can generate an electrical field that is strong enough to be measured outside the skull. Thus, the electrodes placed on the scalp surface are able to measure the electrical activities non-invasively (Figure 1.3(a)) [2]. Further, the data is output in the format of voltage signal series (Figure

1.3(b)). Since all the electrodes can work simultaneously, EEG is able to collect data from all the brain regions at the same time, an advantage over fMRI.

(a) EEG Electrodes   (b) EEG Voltage Flows
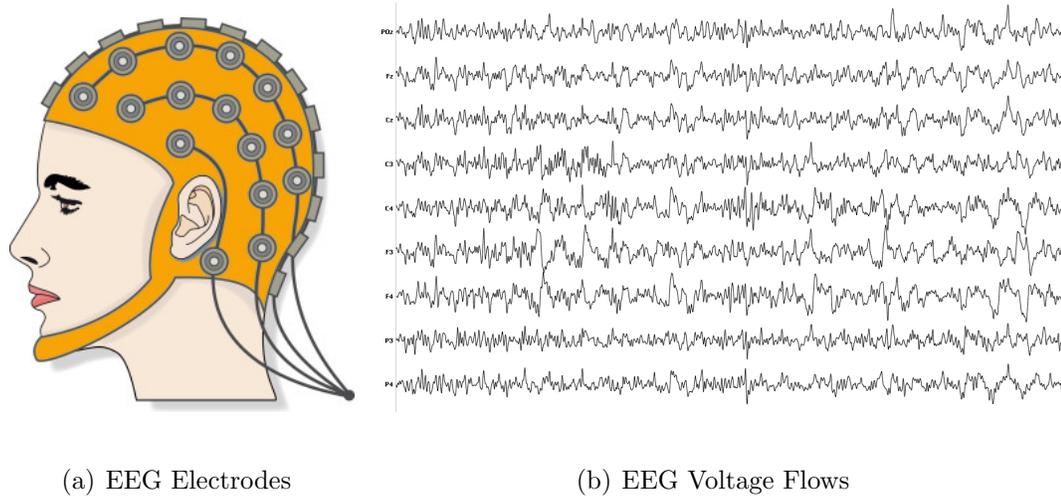
Figure 1.3: 1.3(a) *Electrodes are placed on the scalp surface for EEG.* 1.3(b) *The flow of voltages output by EEG.* [2]

EEG has a high time resolution at millisecond timescales, so it is possible for EEG to follow the process that the brain deals with the onging signal [45]. Since EEG electrodes are placed non-invasively, they are separated from the fluctuation sources by the skull, the scalp, and some other biomass. As a result, EEG has poor spatial resolution of 50 mm to 90 mm [83]. Moreover, the signal-to-noise ratio of EEG is low. The electrical field in the environment and that generated by muscle movements near the brain, e.g., chewing, can add noise to EEG data [78].

1.1.3 ECoG Data

ECoG, also called intracranial EEG (iEEG), uses electrodes placed directly on the cortical surface of the human brain to record its electrical activity for clinical purposes in the treatment of patients with medically intractable seizures or tumors. Figure 1.4 shows the implantation of ECoG electrodes within the skull of the epileptic patient, both graphically and conceptually. In contrast, EEG places the electrodes outside of the skull on the surface of the scalp. So ECoG is an invasive procedure, different from fMRI and EEG.
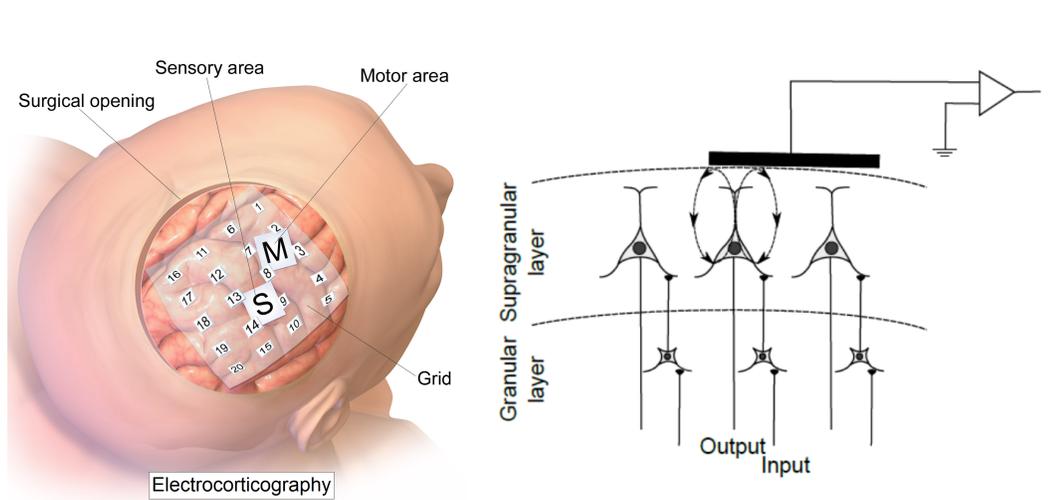
Figure 1.4: *Left: Intracranial electrodes for ECoG. [84] Right: A graphical illustration of ECoG mechanism.*

The special data collection technique leads to nice properties of ECoG data, including simultaneous recordings of the neuronal electrical activity among many small brain regions, combined high temporal and spatial resolutions (Figure 1.5),

and strong signal-to-noise ratio (SNR). First, ECoG has a temporal resolution in milliseconds as EEG. Second, ECoG has a much higher spatial resolution than EEG, of 10 mm [5]. The reason is that, to reach EEG electrodes, the electrical signal has to pass many layers of biological organizations including the solid skull bone, where the signal attenuates rapidly. So EEG electrodes must record over wide areas with a low spatial resolution [46]. Last, ECoG has a high SNR since it is an invasive method.
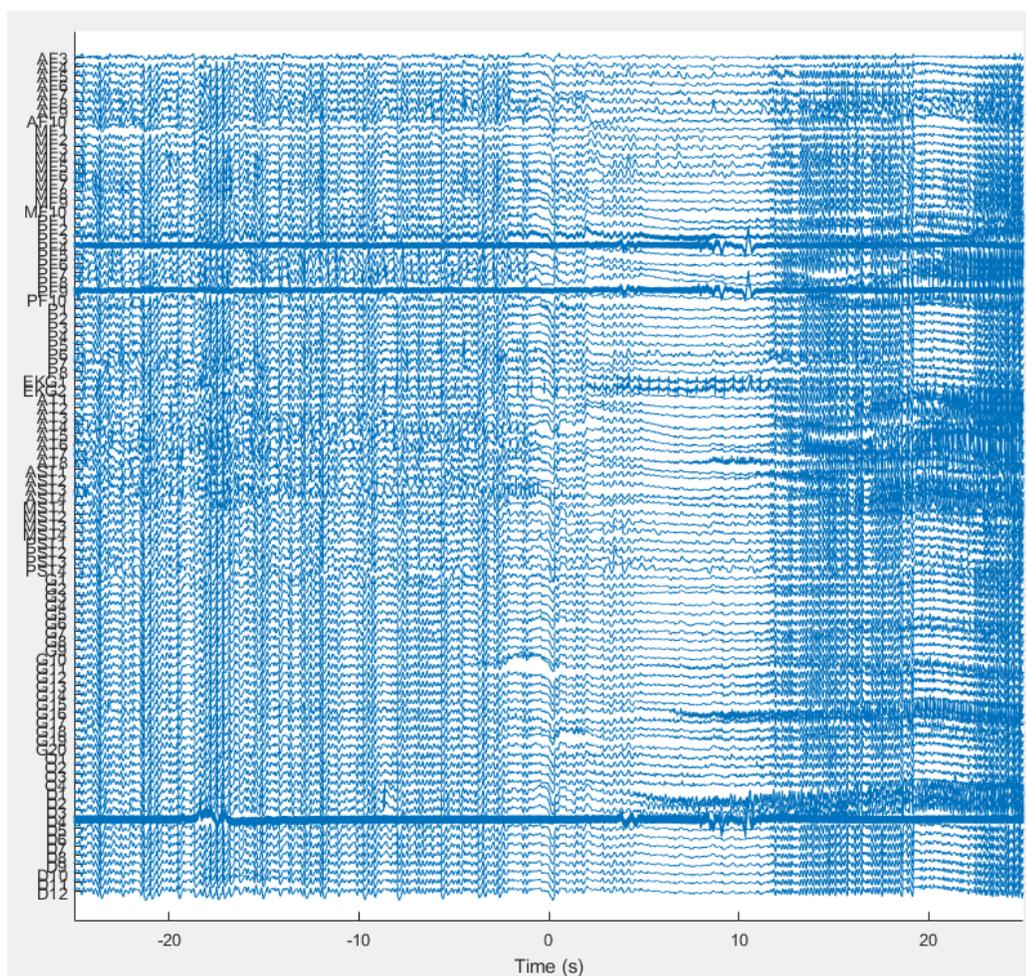
Figure 1.5: *Voltage Series by ECoG.*

Finally, ECoG data is a highly reliable and reproducible measurement of brain activity, which is unavailable in non-invasive measurements of brain activity, e.g., fMRI and EEG data. Studies have shown that, for repeated ECoG data collected over multiple days, comparison analysis based on within and between subject variances have demonstrated that auditory results are robust accross different recordings, regardless of medical conditions or seisure focus. This conclusion suggests the reliability of ECoG data to be used in research [20].

In this project, we leverage the good properties of ECoG data over fMRI and EEG data to study directional interactions among different brain regions and evaluate changes of their interactions across time and stimuli. As intracranial measurements of brain activity, ECoG provides unique information for studying the directional connectivity of the brain.

1.2   Model Development

The human brain is a continuous time dynamic system, so it is biophysically natural to use ordinary differential equations (ODEs) to model the directional effects exerted by each system component (i.e., brain regions) over others. More specifically, since studies have shown that interactions among brain regions occur at the neuronal level [3], we use ODEs to model the neuronal state changes and directional connectivity of the brain.

The study of human brain networks has attracted much attention from many

researchers, and existing popular models include structural equation modeling [60] and dynamic causal modeling [24, 25, 26, 36, 55]. The structural equation model for neuralimaging data can be applied to different measurements of brain activity, either fMRI data or voltage flows, and it models the directional interactions among brain regions in a general linear regression format. As such, the ability to get unique solutions of the model depends on the ratio of unknown parameters to known parameters. The model depends on the researcher to determine the structure of the brain network and where the connections do not exist, which could introduce bias to the model. Also, the model is based on discretization of the brain activity changes over time and could fail to capture the dynamic nature of the brain.

The dynamic causal models (DCMs) for fMRI and EEG data are the most commonly cited ODE models in the literature [24, 25, 26, 36, 55] for directional interactions (effective connectivity) among brain regions. Both fMRI and EEG are noninvasive methods for measuring brain activity with large noise, hence associated DCMs rely on strong prior information of the existence and strength of connections among the brain regions under study to get reliable estimates of brain connectivity. In addition, the DCMs are usually focused on connectivity among only a few large brain regions.

We propose an ODE model for ECoG data, which is motivated by statistical considerations and widely applicable to explore directional connectivity among many different brain regions without relying on strong, specific prior knowledge of the re-

gions. Specifically, we use bilinear ODEs to model directional interactions among brain regions for three reasons. First, the bilinear model, also the simplest ODE model as a low-order Taylor expansion of nonlinear ones, provides general applicability for approximating high-dimensional dynamic systems. This approach is similar to using linear regression models to approximate complex association relationships between various response and predictor variables. As such, linear or bilinear ODEs have been used to model complex dynamic systems in many scientific studies when the underlying dynamic mechanism is elusive and the ensuing detailed model specification is difficult. These studies include gene regulation network [58, 91] and brain effective connectivity studies based on fMRI data [35]. Second, the simple bilinear form provides intuitive scientific interpretation of the model parameters, and enables fast computation for high-dimensional data. Third, taking advantage of ECoG's high temporal resolution, we study the brain activity in response to a simple, short auditory stimulus, and the proposed bilinear model can effectively approximate the brain dynamics within a short period of time.

Within the bilinear formulation, we assume that many model parameters denoting directional interactions among brain regions are zeroes. This is because interactions among brain regions are energy-consuming [4, 33, 68] and sparse connections helps the brain, a biological system, to conserve energy in order to survive and prosper [14, 61]. Moreover, motivated by many reports of brain networks in a cluster structure [63, 64, 66, 82], which consists of clusters, also called modules, of densely

connected brain regions, as shown in Figure 1.6, we build a new bilinear ODE model, called modular and indicator-based dynamic directional model (MIDDM), to characterize sparse connections in the cluster structure in particular. This new ODE model not only has a scientific basis, but also provides intuitive interpretation of different functions of the brain regions in different modules.
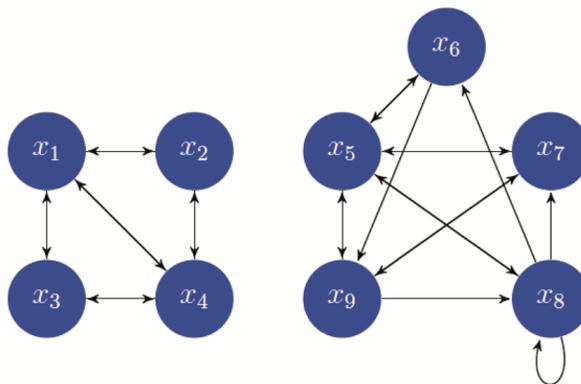
Figure 1.6: *A network in a cluster structure.*

Although an iterative optimization algorithm, called Potts-based iterated principal differential analysis (P-iPDA), has been developed by [100] to identify clusters within a bilinear Potts-based dynamic direction model (PDDM), we develop a new Bayesian approach to estimate the proposed MIDDM, an extension of the PDDM, for three major reasons. First, the Bayesian method provides a unified inference framework for evaluating the statistical significance of identified network edges, each associated with a significantly nonzero model parameter denoting the directional effect between a pair of brain regions. With the proposed Bayesian method, we can

study the brain connectivity changes in response to repetitive events, an important research topic in neuroscience [31, 38]. In contrast, it is difficult to evaluate the statistical significance or compare network results of different trials by the P-iPDA approach (see Section 2.3 for more details). Second, the MIDDM, assuming different properties for connections within and between modules, essentially presents a hierarchical model for the brain network. It is natural and convenient to characterize this multilevel structure and to simultaneously address module identification and directional network edge selection within a unified Bayesian framework, similar to Bayesian methods [29, 56, 86] for simultaneous variable selection and clustering in multiple regression. Third, quantification of the MIDDM model inadequacy for characterizing the complex brain system is natural within a Bayesian framework, an approach unique from most existing ODE estimating methods, as explained in detail below.

The proposed new ODE model, motivated by statistical considerations, is considered as an approximation rather than a principle for the underlying brain mechanism, in contrast to many existing low-dimensional ODE models for simple, well-understood dynamic systems. As such, it is important to account for model uncertainty, defined as the discrepancy between the state functions of the assumed model and the true state functions of the complex brain system, when estimating the MIDDM. Model uncertainty quantification, though straightforward for standard statistical models, is not self-evident for ODEs, since the latter are essentially deterministic models for dy-

namic systems. [22] and [23] developed approaches within the Bayesian framework by [54] to quantify discretization uncertainty of ODE models, which is the discrepancy, caused by limited computation and coarse grids, between the state functions fitted by discretization methods and the exact state functions of the ODE model. However, very few methods in the statistical literature have been developed to quantify ODE model uncertainty. Existing approaches [21, 67] for ODE model uncertainty quantification were mainly developed for specific low-dimensional dynamic systems. Here, we develop a new prior on high-dimensional ODE state functions to quantify the discrepancy between the assumed MIDDM and the underlying brain system.

In summary, we propose a new ODE model for ECoG data to characterize the brain network of effective connectivity in a cluster structure, and develop a new Bayesian framework to quantify the ODE model uncertainty, identify clusters, select significant network edges, and evaluate brain network changes across time and stimulus types. Moreover, we carefully design new priors on the MIDDM parameters to ensure fast posterior simulation of the ensuing hierarchical Bayesian model for high-dimensional ECoG data.

1.3   Outline of Dissertation

The rest of the dissertation is organized as follows.

Chapter 2 reviews the existing models for brain connectivity and methods for ODE model estimation and cluster identification. In this chapter, we will first in-

troduce the structural equation model and the dynamic causal model, two existing models for brain networks, and some of their advantages and limitations. After that, we will discuss some current ODE model estimation methods. Especially, we will talk about the P-iPDA in more details.

Chapter 3 introduces the new ODE model, MIDDM, for cluster-structured directional brain networks, and explains the Bayesian hierarchical method developed to estimate the MIDDM based on the basis representation of ODE state functions. A Markov chain Monte Carlo (MCMC) simulation algorithm for posterior inference is also included in this chapter.

In Chapter 4 we apply the proposed MIDDM and Bayesian method to three simulated examples, compare the results by Bayesian method with those by P-iPDA, and demonstrate the advantages of the former over the latter. The first simulation example is to show that our new method has better performance than P-iPDA when our model assumptions are valid. The second simulation example is designed to demonstrate how to compare networks under our model framework. The last simulation example shows the performance of our model when the bilinear assumptions are moderately violated.

In Chapter 5, we analyze ECoG data collected in an auditory experiment, and evaluate brain network changes across trials and auditory stimuli. The analysis results not only confirm existing results, but also bring new insights into understanding brain connectivity changes in response to repetitive events.

Chapter 6 concludes our work and discusses developments, including some potential directions to further improve the model for ECoG data analysis, some possible ways to accelerate the computation of the algorithm, and some other ECoG data sets or experimental scenarios to which our method may apply.

Chapter 2

Literature Review

The evaluation of effective connectivity of the brain relies on modeling the interactions or effects between pairs of brain components along with understanding the mechanism of the brain measurement process. The most commonly applied and cited models for directional effective connectivity of human brain in literature include structural equation models [60] and dynamic causal models [36]. In the following, a brief review is given for each of them.

2.1 Structural Equation Modeling

Structural equation modeling is a covariance-based method to estimate connections among brain regions based on known neuroanatomical knowledge. The information about neural interactions between different brain areas is extracted by decomposition of the interregional covariances of brain activity measures. Anatomical circuitry

information is combined with the brain imaging data to determine the existence of brain directional effects under a particular task.

Specifically, the relationship between different brain regions is described by a general linear regression equation. For example, a simple structural equation model can be written in the following matrix format [60]:

$$
\begin{bmatrix} X \\ Z \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{zx} & 0 & 0 \\ \beta_{yx} & \beta_{yz} & 0 \end{bmatrix} \begin{bmatrix} X \\ Z \\ Y \end{bmatrix} + \begin{bmatrix} \psi_x \\ \psi_z \\ \psi_y \end{bmatrix},
\tag{2.1}
$$

where X, Y, and Z are measures of the brain activity at three different brain areas; βs are the coefficients representing the direct brain connections, whose absolute values indicate the relative strength of the corresponding interactions between the brain regions; and ψs correspond to the effects that brain components have upon themselves and the mixed effects that come outside the model. The zeros in the β matrix indicate that the corresponding directional effects do not exist in the model, which is decided by the researcher based on their knowledge of the anatomical structure of the brain before the coefficients are estimated. Generally, it is assumed that the direct interaction between a pair of brain components does not exist if they are not anatomically connected. Equation 2.1 also implies how the variance of one brain region is impacted by the variances of the other regions in the model. The coefficients β are then estimated from the variances and covariances of X, Y, and Z. Figure 2.1 graphically shows the brain network described by the above model.
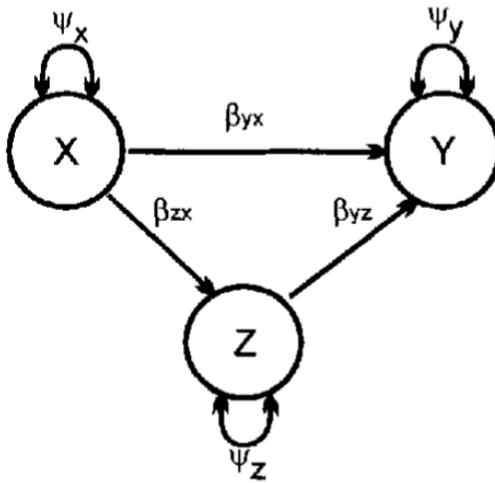
Figure 2.1: *A graphic representation of the structural equation model from equation 2.1. Circles represent three different brain regions, and X, Y, and Z are the brain activity measurements. Uni-directional arrows represent the directional effects from the source to the recipient, and βs are the weights of those connections. Bidirectional arrows with ψs indicate the residual effects with size ψ. [60]*

Maximum likelihood estimates of the parameters in the structural equation model are typically computed through an iterative process. First, some initial values for the unknown parameters are set, with the residual effects ψs set as a fixed amount of the total variance for brain regions depending on the number of connections those brain regions have. Usually, the ratio of ψ to the total variance is lower for brain areas receiving more influence from others, and it is set to 100% for brain regions with zero inputs. Next, for each interation the implied variances and covariances are calculated based on the current parameter values and then compared with the

observed covariance matrix. Then, the parameter values are updated according to the differences between the two covariance matrices, and the whole process stops when those differences are smaller than some reasonable threshold.

The structural equation modeling has the ability to separate the measurement error from the total error by easily adding a measurement model to the original structural model. As such the error of the system is decomposed into two parts: the fitting error in the structural model, and the measurement error in the measurement model. Assume X, Y, and Z in equation 2.1 are latent brain activity variables, and x, y, and z are the corresponding observed data from the experiment device. Then the measurement model can be constructed as the following:

$$x = \lambda_x X + \epsilon_x$$

$$y = \lambda_y Y + \epsilon_y$$

$$z = \lambda_z Z + \epsilon_z,$$

where λs are the extends to which the underlying brain activities are expressed in the measurements at different locations, and ϵs account for the measurement error respectively.

The structural equation model separates the effects among brain regions into direct and indirect effects. The direct effects are those directly estimated by the model coefficients (β), while the indirect effects can be obtained through the paths in the brain network generated by the model. Moreover, the structural equation model has the ability to estimate the asymmetric directional interactions among brain regions

since the coefficient matrix of β is not assumed to be symmetric. Consequently, this model is able to solve loops in the brain network, which reflects a common characteristic of feedback in biological neural networks. In addition, simple t-tests for the coefficients can be used to give statistical inferences on the brain connections since the model is essentially a linear regression. And under the same experimental settings, multiple models can be compared using the goodness-of-fit tests, which eventually improves the fitness of the model.

Despite the advantages described above, the structural equation model also has the following disadvantages. First and foremost, the ability to solve the structural equation model is determined by the complexity of the system. That is, it is impossible to compute a unique solution of the model if the number of the unknown parameters is greater than that of the observations. As a result, the model relies heavily on the researcher to specify the network structure and set zeros in the co-efficient matrix to guarantee that it is solvable, and thus can only be applied to a small number of large brain regions. By this nature, it is a theory-driven method more frequently used to test given hypotheses about the brain network structure. Second, this linear regression format could fail to capture the complex nonlinearity of the biological system, especially the neural system of the brain. Last, the structural equation modeling assumes the interactions between brain regions are instantaneous since it is not a time series model. Hence, it does not consider the dynamism of the brain system, who is changing continuously over time.

2.2   Dynamic Causal Modeling

Dynamic causal modeling (DCM) is an approach to estimate and make inference about the effective connectivity among brain areas using ordinary differential equations (ODEs) based on non-invasive neuroimaging time series, typically fMRI and EEG/MEG data. DCM differs from structural equation modeling as it treats the brain as a deterministic nonlinear dynamic system that produces different brain signals according to different experimental stimuli.

The DCM essentially consists of two parts: the neural state equations, or the ODEs, and the observation equations. The neural state equations, constructed as differential equations, model the effective connectivity of the brain and its changes due to some experimental stimuli. Assume there are d brain regions in the system, and there are J stimuli in the experiment. The stimulus functions are boxcar or stick functions. Then assume that $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))'$ represent the latent neuronal states of the d brain regions, and $\mathbf{u}(t) = (u_1(t), \ldots, u_J(t))'$ correspond to the J stimulus functions designed in the experiment. An arbitrary form of the neural state equations can be written as:

$$\frac{d\mathbf{x}(t)}{dt} = F(\mathbf{x}(t), \mathbf{u}(t), \theta),$$

where F is some unknown nonlinear function describing the interactions among the components of $\mathbf{x}(t)$ and the influence that the stimuli $\mathbf{u}(t)$ give to $\mathbf{x}(t)$. θ includes all the parameters in the model.

A bilinear approximation to the above formula provides a natural and interpretable simplification of the model for the effective connectivity. This bilinear form is:

$$\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t) + \sum_{j=1}^{J} u_j(t) B_j \mathbf{x} + C\mathbf{u}(t), \qquad (2.2)$$

where A is a $d \times d$ matrix representing the brain connections in the absence of the stimuli. The effective connectivity is modeled as the influence that one brain region exerts to the changing rate of another brain area over time $(d\mathbf{x}(t)/dt)$. Each B_j, for j from 1 to J, is then a $d \times d$ matrix standing for the brain connectivity changes because of the stimulus $u_j(t)$. And C is a $d \times J$ matrix corresponding to the effects that the stimuli have on the brain components directly. As already demonstrated in the equation, DCM assumes that the stimuli impact the neural network in two ways: (1) change the interactions between different brain regions; (2) influence the brain activity at certain brain areas directly.

The second part of DCM is the observation equations that relate the neuronal states $\mathbf{x}(t)$ to the observed brain measurements $\mathbf{y}(t)$, with each element of $\mathbf{y}(t) = (y_1(t), \ldots, y_d(t))'$ representing a time series collected from the corresponding brain region. The specific formulation of the model depends on either fMRI or EEG data is used as the response. Nevertheless, it can be generally written as

$$\mathbf{y}(t) = G(\mathbf{x}(t)),$$

where G denotes a known relationship between the neural states $\mathbf{x}(t)$ and the brain measurements $\mathbf{y}(t)$.

Figure 2.2 displays an example of a DCM from [36] with five brain regions $x_1, \ldots, x_5$ represent the neuronal states at the five brain areas, and $y_1, \ldots, y_5$ denote the corresponding brain measurements respectively. x_is are real-valued functions on time t. The black solid arrows between xs indicate the assumed interactions between some pairs of brain components, and the gray solid arrows from xs to ys imply that the latent variable x is reflected as the observed signal y. Besides, there are two stimuli, u_1 and u_2, whose impact on the brain network is depicted as black dotted arrows from us to the network. More specifically, u_1 directly influences the brain component x_1, while u_2 has an effect on the interaction between x_3 and x_2 and that between x_2 and x_4.
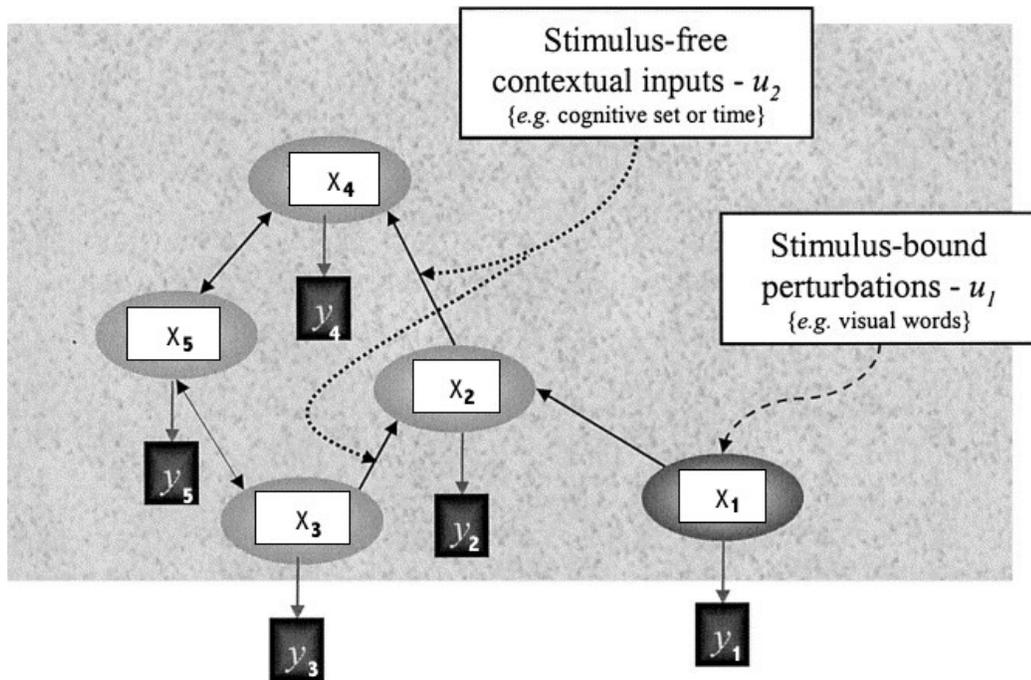

Figure 2.2: *The graphical representation of a DCM example from [36].*

According to the network structure illustrated graphically, we can easily write down the neural state equations of this DCM example as:

$$\frac{dx_1(t)}{dt} = a_{11}x_1(t) + c_{11}u_1(t)$$

$$\frac{dx_2(t)}{dt} = a_{21}x_1(t) + a_{22}x_2(t) + a_{23}x_3(t) + u_2(t)b_{2,23}x_3(t)$$

$$\frac{dx_3(t)}{dt} = a_{33}x_3(t) + a_{35}x_5(t)$$

$$\frac{dx_4(t)}{dt} = a_{42}x_2(t) + a_{44}x_4(t) + a_{45}x_5(t) + u_2(t)b_{2,42}x_2(t)$$

$$\frac{dx_5(t)}{dt} = a_{53}x_3(t) + a_{54}x_4(t) + a_{55}x_5(t).$$

It is quite straight-forward to transform these differential equations into the corresponding matrix format so that they are in accord with equation 2.2. Besides, the observation equations are not shown here, since they depend on the type of the data used and are not of our interest.

The DCM is estimated given the designed stimulus functions $\mathbf{u}(t)$ and the observed brain measurements $\mathbf{y}(t)$. Yet, the specific way to solve DCM is up to which brain measurement is used as well. An expectation maximization (EM) method is developed to estimate the DCM for fMRI data in [36], and a variational Bayes approach is constructed to fit the DCM for EEG/MEG data in [26].

Although DCM has been widely cited and applied to fMRI and EEG/MEG data [26, 55], it has the following limitations. First, similar to structural equation modeling, it relies on strong prior knowledge to determine the existence and strength of connections among brain regions. This would require deep understanding of human brain networks and might introduce significant bias into the model. Second, it only

focuses on the interactions among a few large brain areas, thus the number of brain regions in the model is usually very limited.

2.3   Existing ODE Model Estimation Methods

In the statistical literature, three major categories of approaches have been developed for estimating ODE models: basis-function-expansion techniques in which state functions are represented by functional bases [8, 13, 27, 71, 72, 74, 75, 90], discretization methods using numerical approximation to state functions [6, 9, 17, 18, 39, 43, 47, 50, 51, 57, 59, 96], and Bayesian procedures using a Gaussian process prior for state functions [16, 22, 85].

The strategies mentioned above usually concern low-dimensional dynamic systems with only a few ODEs. For high-dimensional systems, [58, 94, 95] proposed to use penalization-based variable selection methods, which were originally developed for high-dimensional regression problems [32, 88, 93, 99, 101, 102], to estimate sparse ODEs. These methods are suitable for identifying sparse networks in which each component has only a few connections with other components and all components are connected directly or indirectly.

An optimization algorithm P-iPDA, Potts-based iterated Principal Differential Analsis, was developed by [100] to search for the optimal modules or clusters of densely connected brain regions, based on a bilinear Potts-based dynamic directional model (PDDM).

In the algorithm, the Potts model is assumed for the cluster structure of the brain network. Suppose there are d brain regions in the model, and m_i, which only takes integer values ranging from 1 to d, represents the cluster label of the brain region i. Then $m_i = m_j$ indicates that brain components i and j are in the same cluster. The Potts model is constructed for the cluster structure by restricting the cluster labels in the following way:

$$P(m_1, \ldots, m_d) \propto \exp\left\{-\mu \sum_{i,j=1}^{d} \delta(m_i, m_j)\right\},$$

where $P(.)$ represents the probability of an event; μ is a nonnegative tuning parameter in the algorithm; and $\delta(m_i, m_j) = 1$ if and only if $m_i = m_j$; $\delta(m_i, m_j) = 0$ otherwise. Thus, the equation above shows that if the clusters in the network are large, there will be many pairwise edges within the clusters, indicating large value of the sum in the equation, and thus the corresponding probability is small. So the model essentially implies a preference for small clusters.

This Potts model is then transformed into a Potts-based penalty term in the final optimization objective function. Assume H_P is the original objective function of sum of squared errors to be minimized without penalty, then the penalized objective function is:

$$H_P + \mu \sum_{i,j=1}^{d} \delta(m_i, m_j).$$

Minimizing this penalty term is equivalent to maximizing the probability of the cluster labels in the Potts model. Furthermore, this Potts-based penalty term is essentially an L_0 penalty.

The P-iPDA provides a way to estimate the high-dimensional ODE models developed for the dynamic brain system with a cluster structure given the ECoG data. However, the P-iPDA has several limitations. First, the P-iPDA, relying on a crucial assumption that regions within the same cluster are all pairwise connected, does not distinguish connected brain regions from disconnected ones within the same cluster, and thus may lead to many false positives whenever the P-iPDA fails to identify the smallest clusters and groups separate clusters together. Second, the P-iPDA essentially minimizes a log-likelihood based criterion with an L_0 penalty. Statistical inference of the ensuing parameter estimates is challenging, and thus the P-iPDA cannot be used to compare brain networks across different trials and stimulus types. Third, the network results by the P-iPDA are highly sensitive to the tuning parameters. As such, the P-iPDA needs to perform a time-consuming cross-validation procedure on quite a few candidate values to select ideal penalty parameters.

Chapter 3

Approach

In this study, we propose a bilinear ODE model, MIDDM, for the effective connectivity of the human brain based on the intracranial ECoG data. Unlike the structural equation modeling or DCM, we construct our model on a large number of small brain areas, given the good properties of ECoG data. And we do not specify the existence of the connections subjectively. A cluster structure of the network is assumed through the cluster labels and the edge indicators in the model, which introduces sparsity into the model. Thus, our model is data-driven, with the network structure determined by the data.

Coupled with MIDDM, a Bayesian framework is developed to estimate the proposed ODE model. In the Bayesian model, we take advantage of the functional basis representations of the state functions. Due to the hierarchical nature of MIDDM, it is natural to introduce the hierarchical Bayesian model. And once the posterior

samples are obtained, the interpretation and statistical inference of the results would be simple and easy to understand.

With carefully designed prior distributions, we formulate a proper joint posterior distribution in our Bayesian model which guarantees that the conditional posterior distributions for all the parameters are well-defined. So a Gibbs sampler method is applied to get the posterior estimates for the model parameters and it makes the model solvable for the high-dimensional brain imaging data. In the remaining of this chapter, we will talk in more detail about the three parts of the proposed model — the MIDDM, the Bayesian framework, and the posterior simulation.

3.1   MIDDM for ECoG Data

Our proposed model consists of two parts: the neural state equations and the observation model. The neural state equations are differential equations developed to model the effective connections among different brain regions, and the observation model connects the latent neural state variables to the observed brain activity measurements. We start introducing the whole model from the simpler part — the observation model.

Let $\mathbf{y}(t) = (y_1(t), \ldots, y_d(t))'$ be the observed ECoG time series of d brain regions' neuronal activity at time t. The observed data $\mathbf{y}(t)$ are measured at discrete time points $t = 1, 2, \ldots, T$. Let $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))'$ be the neuronal state functions of d brain regions at time t. Since ECoG is an invasive approach, the observation

model for ECoG data is given by

$$\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t), \tag{3.1}$$

where $\boldsymbol{\epsilon}(t) = (\epsilon_1(t), \dots, \epsilon_d(t))'$ is a d-dimensional vector of errors with mean zeroes, which represents the measurement error in the model.

The second part of the model is the ODEs that estimate the brain interactions given the designed stimulus, which could potentially influence the brain system either by changing the directional connections among brain regions or by impacting the neural states of brain components. Suppose the brain system consisting of the d regions under study received one stimulus input in the experiment. Let $u(t)$ be an experimental stimulus function taking values 1 and 0 only, e.g., a boxcar or stick stimulus function. The stimulus function indicates whether the stimulus (e.g., a tone) is present at time t or not. Since brain regions interact with each other at the neuronal level, we use the following bilinear model to characterize directional interactions among the d regions:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\,\mathbf{x}(t) \cdot (1 - u(t)) + \mathbf{B}\,\mathbf{x}(t) \cdot u(t) + \mathbf{C}\,u(t) + \mathbf{D}, \tag{3.2}$$

where $\mathbf{A} = (A_{ij})_{d\times d}$ with entry A_{ij} denoting the effect of region j on region i exerted at the current state without the stimulus; $\mathbf{B} = (B_{ij})_{d\times d}$ with B_{ij} denoting the stimulus-dependent effect exerted by region j on region i; $\mathbf{C} = (C_1, \dots, C_d)$ with C_i denoting the stimulus effect on region i; and $\mathbf{D} = (D_1, \dots, D_d)$ denoting the intercepts for d regions. As a low-order Taylor approximation of the underlying system, the simple

bilinear model (3.2) provides general applicability and facilitates fast computation for a high-dimensional dynamic system.

The stimulus function $u(t)$ is a boxcar or stick stimulus function only taking values 1 and 0, instead of an arbitrary function. First, it is a convention to do so in literature. In DCM, the stimulus function takes values of 1 and 0 as well ([25, 26, 55, 24]). In addition, it is hard to quantify the stimulus in the context of neuronal activities. Thus, it is not straight-forward to generalize the format of the stimulus function to any formula. However, our model can be easily extended to a dynamic system with J types of stimuli:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\ \mathbf{x}(t) + \sum_{j=1}^{J} \mathbf{B}_j\ \mathbf{x}(t) \cdot u_j(t) + \sum_{j=1}^{J} \mathbf{C}_j\ u_j(t) + \mathbf{D},$$

where $\mathbf{A}$ represents the brain connections without any stimulus; each $\mathbf{B}_j$ indicates the changes of the interactions among brain regions due to the stimulus j; $\mathbf{C}_j$ denotes the direct stimulus effect on brain components from the stimulus $u_j(t)$; and $\mathbf{D}$ is still the intercepts of the model. Without loss of generality, we assume that there is only one stimulus into the system for simplicity.

Despite the simple bilinear formulation, estimates of many parameters in the high-dimensional ODE model are often unreliable with large variances. To increase the estimation efficiency, we utilize a scientifically motivated assumption that many connections among system components are void, and corresponding model parameters denoting directional interactions in matrices $\mathbf{A}$ and $\mathbf{B}$ are zeros. This is because the directional interactions among brain regions are energy-consuming [4], and bio-

logical organizations tend to be energy-efficient with sparse connections to survive and grow. With the sparsity assumption, it is of scientific interest to identify interactive brain regions associated with significantly nonzero parameters in the ODE model for the brain system, and to construct the ensuing sparse brain network of effective connectivity among brain regions.

A complication with sparse networks is that they can arise in different forms. For example, Figure 3.1 shows two different forms of sparse networks, implying vastly different functional organizations of the brain. Figure 3.1(a) displays a general sparse network, while Figure 3.1(b) depicts a network with a cluster structure. A cluster in this context indicates a group of nodes in the network that are densely connected by edges, and there are two clusters in Figure 3.1(b). As discussed in Introduction (Chapter 1), it is reasonable to believe that many brain regions are not directly connected and the ensuing sparse brain network with each network edge denoting one directional effect between a pair of regions is in a cluster structure.

(a) Sparse Connections   (b) Cluster Structure
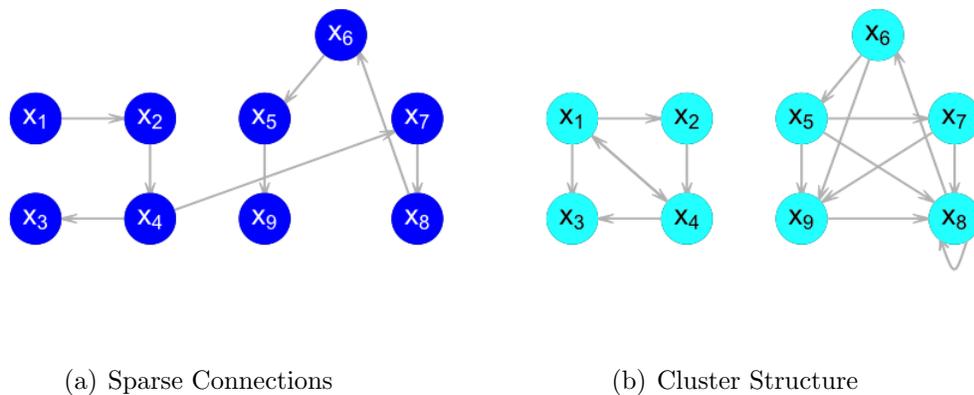
Figure 3.1: *Two typical examples of sparse network structure.*

To characterize the cluster structure, we introduce module labels into the model. Let $\boldsymbol{m} = \{m_1, \ldots, m_d\}$ represent the cluster labels for d brain regions, which take integer values between 1 and d. It implies that the system can have at most d clusters, when each region forms one cluster itself. However, the number of clusters can be less than d when some integers between 1 and d are not used as cluster labels. In addition, following the formulation in the Bayesian stochastic search variable selection (SSVS) framework [12, 41, 42, 97] that uses the spike and slab prior [53, 62, 87], we use indicators γ_{ij}^A and γ_{ij}^B — which take values 0 and 1 only for $i, j = 1, \ldots, d$ — to differentiate strong and void directional effects without and with the stimulus, respectively. We modify equation 3.2 and propose the following ODE model. For $i = 1, \ldots, d$,

$$
\begin{aligned}
\frac{dx_i(t)}{dt} &= \sum_{j=1}^{d} \delta(m_i, m_j) \cdot \gamma_{ij}^A \cdot A_{ij} \cdot x_j(t) \cdot (1 - u(t)) \\
&+ \sum_{j=1}^{d} \delta(m_i, m_j) \cdot \gamma_{ij}^B \cdot B_{ij} \cdot x_j(t) \cdot u(t) + C_i \cdot u(t) + D_i,
\end{aligned}
\tag{3.3}
$$

where $\delta(m_i, m_j)$ is the Kronecker delta, which equals 1 whenever $m_i = m_j$ and 0 otherwise. Under equation 3.3, brain component j has a nonzero directional effect on component i or a directional network edge from brain region j to region i exists if and only if the two brain areas are in the same cluster, i.e., $m_i = m_j$, and either γ_{ij}^A or γ_{ij}^B is nonzero.

To incorporate the cluster structure and the directional interactions, we choose to use cluster labels $\boldsymbol{m}$ and indicators $\{\gamma_{ij}^A\}$ and $\{\gamma_{ij}^B\}$, instead of imposing some

constraints directly on matrices $\mathbf{A}$ and $\mathbf{B}$ in model 3.2 as a penalized optimization method similar to P-iPDA, for the following reasons. First, as discussed in Section 2.3, it is quite challenging to do statistical inference using a penalty model. That is, it is hard to find the significant connections using estimated $\{A_{ij}\}$ and $\{B_{ij}\}$, which on the other hand is directly captured by cluster labels $\boldsymbol{m}$ and directional indicators $\{\gamma_{ij}^A\}$ and $\{\gamma_{ij}^B\}$. Additionally, to control both the cluster structure and the directional effects, we need two penalty parameters, which makes the penalty parameter selection process even more complex and time-consuming.

The ODE model 3.3 together with the observation model 3.1 is referred to as the MIDDM, or modular and indicator-based dynamic directional model. The ODE model 3.2 and the sparse network, in which each node is sparsely connected with the rest of nodes and all nodes are connected directly or indirectly, are special cases of the MIDDM with only one single module and with either mostly nonzero or mostly zero indicators. The MIDDM is also an extension of the existing ODE model for ECoG, PDDM [100], as the latter assumes all regions within clusters to be pairwise connected while the former uses indicators to distinguish nonzero directional interactions within clusters from void ones.

Under the MIDDM, the inference of effective connectivity of the brain is equivalent to identifying modules, selecting statistically significant directional interactions, and estimating the model parameters $\mathbf{A}$ and $\mathbf{B}$, which denote the strength of effective connections among brain regions.

3.2   Hierarchical Bayesian Model for MIDDM

We propose a Bayesian method to estimate and make inferences about the MIDDM. And we elaborate on building the hierarchical Bayesian framework in the remaining of this section.

First, we represent the neural state functions $x_i(t)$ of each brain component i by a vector of B-spline basis functions $\boldsymbol{b}(t) = (b_1(t), \ldots, b_L(t))'$ defined on an equally spaced partition $\{t_1 = 1, t_2, \ldots, t_q = T\}$ of the time interval $[1, T]$ of the observed data. For $i = 1, \ldots, d$,

$$x_i(t) = \boldsymbol{b}(t)' \, \eta_i, \tag{3.4}$$

where η_i is an $L \times 1$ vector consisting of the basis coefficients for $x_i(t)$. The choice of the number of bases L, as a hyperparameter, is discussed in section 3.3.2.

The basis notation is widely applied in many fields with functional data analysis. Simple basis functions, such as monomial basis $(1, \, t, \, t^2, \, \ldots, \, t^L)$, are of limited use when dealing with functions with complex shapes. In addition, two systems of basis functions are most commonly used: spline basis and Fourier series [73]. The Fourier series are usually used for periodic data and the corresponding analysis often focuses on the frequency domain. Therefore, we choose to use the B-spline basis, the most popular spline basis, to represent the neural state functions for the following reasons. First, it is flexible enough to represent the complex shapes of the neural state functions. And it can be specified to make sure that the basis functions have

smooth first derivatives. Last, we concentrate on the temporal domain of the data in this analysis, as the previous ODE model DCM did [36].

3.2.1 Nonparametric Model for Observed Data

Let $Y_i = (y_i(1), \ldots, y_i(T))'$ for $i = 1, \ldots, d$ and $\mathbf{Y} = (Y_1', \ldots, Y_d')'$. Then Y_i represents the observed ECoG time series for brain region i, and $\mathbf{Y}$ is a long vector containing all the ECoG data. Based on basis representations 3.4, we assume that the Y_is are independently distributed with multivariate normal distributions:

$$Y_i \mid \eta_i, \sigma_i^2 \stackrel{\text{ind}}{\sim} \text{MN}(\mathbf{\Phi}\eta_i, \sigma_i^2 \mathbf{I}_T), \tag{3.5}$$

for $i = 1, \ldots, d$. Here, $\text{MN}(\mu, \Omega)$ stands for a multivariate normal distribution with mean μ and covariance matrix Ω, and $\mathbf{\Phi}$ is a $T \times L$ matrix with element $\mathbf{\Phi}[t, l] = b_l(t)$ for $t = 1, \ldots, T$ and $l = 1, \ldots, L$. Though an AR(1) or AR(2) model can be assumed for $\boldsymbol{\epsilon}(t)$, $t = 1, \ldots, T$, in the observation model (3.1), for simplicity we assume them to be independently Gaussian distributed with zero means. For data with a strong SNR, such as ECoG [10, 11, 100], accounting for autocorrelation in the model does not improve the estimation much.

3.2.2 Prior Specification for Basis Coefficients

To simplify notations, let $\boldsymbol{\gamma}^A = \{\gamma_{ij}^A, \ i, j = 1, \ldots, d\}$ and $\boldsymbol{\gamma}^B = \{\gamma_{ij}^B, \ i, j = 1, \ldots, d\}$ represent all the indicators corresponding to matrices $\mathbf{A}$ and $\mathbf{B}$ respectively. Let $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ include all the real-value parameters in the bilinear ODE model.

And let $\boldsymbol{\Theta}_I$ denote all the MIDDM parameters:

$$\boldsymbol{\Theta}_I = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \boldsymbol{m}, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B\}.$$

We propose a prior for basis coefficients $\boldsymbol{\eta} = (\eta_1', \ldots, \eta_d')'$, which depends on the MIDDM parameters $\boldsymbol{\Theta}_I$ through the MIDDM model fitting error:

$$p(\boldsymbol{\eta}|\boldsymbol{\Theta}_I, \tau) \propto \exp\left\{-\frac{1}{2\tau}\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)\right\}, \qquad (3.6)$$

where τ is a pre-specified positive constant and $\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)$ denotes the MIDDM model fitting error:

$$\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I) = \sum_{i=1}^{d} \int_0^T \left[\frac{dx_i(t)}{dt} - \sum_{j=1}^{d} \delta(m_i, m_j) \cdot \gamma_{ij}^A \cdot A_{ij} \cdot x_j(t) \cdot (1 - u(t))\right.$$
$$\left. - \sum_{j=1}^{d} \delta(m_i, m_j) \cdot \gamma_{ij}^B \cdot B_{ij} \cdot x_j(t) \cdot u(t) - C_i \cdot u(t) - D_i\right]^2 dt.$$

Note that $x_i(t)$, $i = 1, \ldots, d$, in $\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)$ is represented by basis functions in equation 3.4, and accordingly we can get

$$\frac{dx_i(t)}{dt} = \boldsymbol{b}^{(1)}(t)' \ \eta_i,$$

where $\boldsymbol{b}^{(1)}(t)$ represents the first derivative of the basis functions $\boldsymbol{b}(t)$. With the linear basis representation of $\mathbf{x}(t)$, $\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)$ can be rewritten as

$$\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I) = \sum_{i=1}^{d} \int_0^T \left[\boldsymbol{b}^{(1)}(t)'\eta_i - \sum_{j=1}^{d} \delta(m_i, m_j) \cdot \gamma_{ij}^A \cdot A_{ij} \cdot (1 - u(t)) \cdot \boldsymbol{b}(t)'\eta_j\right.$$
$$\left. - \sum_{j=1}^{d} \delta(m_i, m_j) \cdot \gamma_{ij}^B \cdot B_{ij} \cdot u(t) \cdot \boldsymbol{b}(t)'\eta_j - C_i \cdot u(t) - D_i\right]^2 dt.$$

Then, it is obvious that the formula in brackets is a linear equation of η_is. Thus, $R(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)$ is fundamentally quadratic of $\boldsymbol{\eta}$ given $\boldsymbol{\Theta}_I$:

$$R(\boldsymbol{\eta}, \boldsymbol{\Theta}_I) = \boldsymbol{\eta}' \boldsymbol{\Omega}_{\boldsymbol{\Theta}_I} \boldsymbol{\eta} - 2\boldsymbol{\Lambda}'_{\boldsymbol{\Theta}_I} \boldsymbol{\eta} + \boldsymbol{\Xi}_{\boldsymbol{\Theta}_I}, \tag{3.7}$$

where $\boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}$, $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}_I}$, and $\boldsymbol{\Xi}_{\boldsymbol{\Theta}_I}$, respectively, are a $dL \times dL$ matrix, a $dL \times 1$ vector, and a scalar, whose values depend on the MIDDM parameters $\boldsymbol{\Theta}_I$. The exact formulas of $\boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}$, $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}_I}$, and $\boldsymbol{\Xi}_{\boldsymbol{\Theta}_I}$, as functions of $\boldsymbol{\Theta}_I$, are provided in the Appendix B.VI. As such, the prior (3.6) is equivalent to a normal distribution:

$$\boldsymbol{\eta} | \boldsymbol{\Theta}_I, \tau \sim \mathrm{MN}(\boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\Theta}_I}, \tau \cdot \boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}^{-1}),$$

and the prior on basis coefficients $\boldsymbol{\eta}$ is equivalent to a Gaussian process prior on the state functions $\mathbf{x}(t)$.

The prior (3.6) quantifies the deviation of the state functions from the assumed ODE model by a probability measure and suggests a preference for state functions with a small model-fitting error given $\boldsymbol{\Theta}_I$. This prior essentially gives a multivariate normal distribution for the basis coefficients $\boldsymbol{\eta}$. Consequently, the state function $\mathbf{x}(t)$ is no longer solely determined by the ODE model. As linearly represented by the basis functions, $\mathbf{x}(t)$ changes as $\boldsymbol{\eta}$ changes. Thus, it has a distribution centered at the function determined by the ODE model. That is, the state function is very likely to be the function determined by the ODE model, but is also possible to be deviated from the function determined by the ODE model. This distribution quantifies the uncertainty of the ODE model.

3.2.3 Prior Specification for MIDDM Parameters

We specify a joint prior for MIDDM parameters $\mathbf{\Theta}_I$:

$$p(\mathbf{\Theta}_I | \tau) \;\propto\; \det(\mathbf{\Omega}_{\mathbf{\Theta}_I})^{-1/2} \cdot \exp\left\{ \frac{1}{2\tau}(\mathbf{\Lambda}'_{\mathbf{\Theta}_I} \mathbf{\Omega}^{-1}_{\mathbf{\Theta}_I} \mathbf{\Lambda}_{\mathbf{\Theta}_I} - \mathbf{\Xi}_{\mathbf{\Theta}_I}) \right\} \tag{3.8}$$

$$\cdot \exp\left\{ -\mu \cdot \sum_{ij}^{d} \delta(m_i, m_j) \right\} \cdot p_0^{\sum_{i,j} \gamma^A_{ij} + \sum_{i,j} \gamma^B_{ij}} \cdot (1 - p_0)^{2d^2 - \sum_{i,j} \gamma^A_{ij} - \sum_{i,j} \gamma^B_{ij}}$$

$$\cdot \prod_{i,j=1}^{d} \phi\left(\frac{A_{ij}}{\xi_0}\right) \cdot \prod_{i,j=1}^{d} \phi\left(\frac{B_{ij}}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{C_i}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{D_i}{\xi_0}\right),$$

where the prior probability p_0 is pre-specified by the user to input the prior belief of the average degree of connectivity within clusters, μ is a given nonnegative constant, ϕ is the standard normal density function, and ξ_0 is a large positive constant to give an almost flat prior for $\boldsymbol{\theta}$ in a wide domain. It is possible to use different p_0 for γ^A and γ^B. For simplicity, we assume identical prior probabilities for them. The discussion of choosing μ and p_0 is deferred to Section 3.3.2.

The above prior essentially combines a multivariate Bernoulli distribution [42] for indicators $\boldsymbol{\gamma}^A$ and $\boldsymbol{\gamma}^B$, the Potts model [44, 70] for module labels $\boldsymbol{m}$, and almost flat multivariate normal distributions for parameters $\boldsymbol{\theta}$. We specify the prior for MIDDM parameters in the above form, a main thrust of the proposed Bayesian framework, for two reasons. First, using the prior (3.8), the full posterior conditional distributions of parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are multivariate normal, which are easy to simulate from; this is a crucial advantage of the proposed Bayesian framework, especially for analyzing high-dimensional ECoG data. Second, in contrast to independent priors on the

MIDDM parameters, the prior (3.8) makes use of the proposed model information by incorporating the model-fitting error. A similar prior has been proposed by [98] in an empirical Bayesian approach for variable selection and estimation in linear models.

3.2.4 Priors for Data Variances

We impose an uninformative prior on $\boldsymbol{\sigma}^2 = \{\sigma_i^2, \; i = 1, \ldots, d\}$:

$$p(\boldsymbol{\sigma}^2) \propto \prod_{i=1}^{d} \frac{1}{\sigma_i^2}. \tag{3.9}$$

3.2.5 Joint Posterior Distribution

In summary, equations 3.5, 3.6, 3.8, and 3.9 jointly define a hierarchical Bayesian model for the MIDDM, referred to as Bayesian MIDDM (BMIDDM) in the following. The joint posterior distribution of the BMIDDM is given by

$$
\begin{aligned}
p(\boldsymbol{\eta}, \boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2 | \mathbf{Y}, \tau, \mu) \propto & \prod_{i=1}^{d} \frac{1}{\sigma_i^T} \exp\left\{-\frac{(Y_i - \boldsymbol{\Phi}\,\eta_i)^2}{2\sigma_i^2}\right\} \cdot \exp\left\{-\frac{1}{2\tau}\mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)\right\} \quad (3.10) \\
& \cdot \exp\left\{-\mu \sum_{i,j=1}^{d} \delta(m_i, m_j)\right\} \cdot p_0^{\sum_{i,j} \gamma_{ij}^A + \sum_{i,j} \gamma_{ij}^B} \\
& \cdot (1 - p_0)^{2d^2 - \sum_{i,j} \gamma_{ij}^A - \sum_{i,j} \gamma_{ij}^B} \cdot \prod_{i,j=1}^{d} \phi\left(\frac{A_{ij}}{\xi_0}\right) \cdot \prod_{i,j=1}^{d} \phi\left(\frac{B_{ij}}{\xi_0}\right) \\
& \cdot \prod_{i=1}^{d} \phi\left(\frac{C_i}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{D_i}{\xi_0}\right) \cdot \prod_{i=1}^{d} \frac{1}{\sigma_i^2}.
\end{aligned}
$$

It can be shown that with a fixed positive constant τ and nonnegative μ, the above joint posterior distribution is proper as long as $T > L$. The proof is provided in the Appendix A.

3.3 Posterior Simulations

We present an MCMC algorithm for posterior simulations in this section. Let $\boldsymbol{m}_{-i} = \boldsymbol{m} \setminus \{m_i\}$, $\boldsymbol{\gamma}^A_{-ij} = \boldsymbol{\gamma}^A \setminus \{\gamma^A_{ij}\}$, and $\boldsymbol{\gamma}^B_{-ij} = \boldsymbol{\gamma}^B \setminus \{\gamma^B_{ij}\}$ for $i,j = 1,\ldots,d$. We use a partially collapsed Gibbs Sampler [PCGS; 89] to sample from the joint posterior 3.10 with given μ and τ (omitted in the posterior conditional distributions below). Specifically, $\boldsymbol{\theta}$ is integrated out when drawing posterior samples of cluster labels $\boldsymbol{m}$ and indicators $\boldsymbol{\gamma}^A$ and $\boldsymbol{\gamma}^B$, and the PCGS is performed in the following order to maintain the target stationary distribution.

1. Sequentially update m_i by a draw from $p(m_i|\boldsymbol{m}_{-i}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y})$, which is a discrete distribution, for $i = 1,\ldots,d$.

2. Sequentially update γ^A_{ij} by a draw from $p(\gamma^A_{ij}|\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A_{-ij}, \boldsymbol{\gamma}^B, \mathbf{Y})$, which is a discrete distribution, for $i,j = 1,\ldots,d$.

3. Sequentially update γ^B_{ij} by a draw from $p(\gamma^B_{ij}|\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B_{-ij}, \mathbf{Y})$, which is a discrete distribution, for $i,j = 1,\ldots,d$.

4. Draw $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y})$, which is a multivariate normal distribution.

5. Draw $\sigma^2_1, \ldots, \sigma^2_d$ from $p(\boldsymbol{\sigma}^2|\boldsymbol{\Theta}_I, \boldsymbol{\eta}, \mathbf{Y})$, which is a product of independent inverse-gamma distributions.

6. Draw $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|\boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2, \mathbf{Y})$, which is a multivariate normal distribution.

We defer technical derivations of posterior $p(\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \boldsymbol{\sigma}^2 | \mathbf{Y}, \tau, \mu)$ with $\boldsymbol{\theta}$ being integrated out and the posterior conditional distributions of each parameter to the Appendix B.

3.3.1 Parallel Computing

To speed up MCMC posterior simulations of the BMIDDM, we employ a parallel computing scheme similar to that developed by [15] in three major MCMC steps: (a) simulation of basis coefficients $\boldsymbol{\eta}$ for regions in different clusters, (b) calculation of the posterior conditional probabilities of cluster labels $\boldsymbol{m}$, and (c) simulation from the posterior conditional distributions of indicators $\boldsymbol{\gamma}^A$ and $\boldsymbol{\gamma}^B$. Specifically, in (a), since given $\boldsymbol{m}$, η_is of brain regions in different clusters are conditionally independent, we employ the same number of process cores as the number of different clusters of $\boldsymbol{m}$, and use each core to simulate η_is in one unique cluster. In (b), we use the same number of process cores as the number of different values that m_i can take given the rest of the parameters, with each core computing one posterior conditional probability for m_i taking one unique value. In (c), since indicator variables $\tilde{\boldsymbol{\gamma}}_i = \{\gamma_{ij}^A, \gamma_{ij}^B, j = 1, \ldots, d\}$, conditional on the rest of the parameters, are independent for $i = 1, \ldots, d$, we employ d process cores, each sequentially simulating every element of one $\tilde{\boldsymbol{\gamma}}_i$ from the element's posterior conditional probability.

The reduction of computational time by using parallel computing in Steps (a) and (b) depends mostly on the size of the largest cluster at each iteration, and parallel

computing is most efficient when all the clusters are small and of similar sizes. The use of parallel computing in Step (c) reduces the computational time for the posterior simulation of indicators from $O(d^2)$ to $O(d)$.

Parallel computing can also be used in other MCMC steps in a similar manner, including simulation of MIDDM parameters $\boldsymbol{\theta}$ of regions in different clusters and simulation of σ_i^2 for $i = 1, \ldots, d$. However, the efficiency of parallel algorithm depends on the computing time of the parallelized part as well. If the original program does not take much time to finish, the time shortened by parallel will not make up the time spent by the manipulation and communication of multiple cores. In that scenario, parallel computing will not accelerate the algorithm. Therefore, we only employ parallel computing in the three major steps described above, instead of in every possible step in MCMC.

3.3.2 Hyperparameter Selection

The choice of hyperparameter τ is the most crucial, because it balances between the data and the model information for inferring directional connections among brain regions. Specifically, a small τ, compared to the data variance, can impose an incorrect strong prior belief that the assumed model fits the underlying dynamic system well. We found that with a small τ , only a few regions' temporal activity can be jointly fitted by the assumed model, and thus only a few brain regions are identified to be connected. On the other hand, if τ is too large, the model information in the posterior

is too weak to be useful for differentiating strong directional connections from weak or void ones. Then, all the brain regions are identified to be connected. In summary, an appropriate value of τ depends on how well the assumed ODE model can fit the data.

Given the above consideration, we decide to choose τ based on the model-fitting error of the observed data. Although cross-validation-based methods [76, 77, 92] are straightforward for choosing hyperparameters, they are time-consuming within a Bayesian framework, since MCMC algorithms can take a long time to converge, especially for high-dimensional ODE models with long time series data. Instead, we propose an easy-to-implement approach to determine the value of hyperparameter τ. Since τ can be regarded as the variance of ODE fitting errors, we set it to the estimated ODE fitting errors. Specifically, we first fit $\mathbf{x}(t)$ nonparametrically with $\boldsymbol{b}(t)$ to the observed data; regress estimated $\frac{d\hat{x}_i(t)}{dt}$ versus $\hat{\mathbf{x}}(t) \cdot (1 - u(t))$, $\hat{\mathbf{x}}(t) \cdot u(t)$, and $u(t)$; and obtain regression mean squared errors, denoted by $\hat{\tau}_i$, for $i = 1, \ldots, d$. The range of $\hat{\tau}_i$, $i = 1, \ldots, d$, gives the range of the variances of model fitting errors for the observed data. Then we choose τ to be $\max\{\hat{\tau}_i\}_{i=1}^d$, which leads to the least informative prior for basis coefficients among all the candidate values. Through some preliminary analysis, we found that this value can help us effectively identify the underlying modules, select true network edges, and provide scientifically interpretable results.

We let $\mu = 0$ to give a non-informative prior on the cluster structure $\boldsymbol{m}$. For

choosing the prior probability p_0 for nonzero network edges within modules, we have tried different p_0 values, from 0.9 to 0.7, which reflect the prior belief that regions within the same cluster are densely connected. We evaluated the network edge selection performance through simulation studies, and found that $p_0 = 0.9$ is the most effective with the highest power for selecting network edges, especially for the cluster structure where all regions within the same cluster are pairwise connected. Considering that the connections within modules are usually short-range, strong, and dense [69], we let $p_0 = 0.9$ to ensure a high power for selecting within-module network edges. Though it is possible to assign a prior to p_0 or tune its value based on the data, we choose to specify its value, because this approach directly uses the existing scientific knowledge of the brain network and thus reduces uncertainty in the model estimation.

The choice of the number of basis functions L directly affects the posterior computational time: The larger L, the more computational time needed for simulating the state functions, which is the most computationally intensive MCMC step. Considering this, we choose a small L without compromising the flexibility of representing the state function $\mathbf{x}(t)$, as suggested in [75]. For the real data under study, we found that the data at every three consecutive points take similar values, and thus we choose $L = \lceil T/3 \rceil$.

3.3.3 Analysis of Posterior Samples

Given hyperparameters (μ, τ) and data $\mathbf{y}(t)$, we conduct posterior simulations of the BMIDDM. Let S be the total number of MCMC iterations excluding the burn-in ones, which are the iterations at the beginning of the MCMC run before the Markov chain converges to the stationary distribution. Let $\theta^{(s)}$ be the value of BMIDDM parameter θ simulated at the sth iteration. Based on the posterior draws of BMIDDM parameters, for each pair of regions (i, j), we estimate the posterior clustering probability of the two regions being in the same cluster, and the posterior probabilities of nonzero directional effects exerted by region j on region i, also called the posterior selection probabilities of directional network edges from component j to component i, without and with the stimulus respectively, by

$$\hat{P}_{ij}^m = \frac{1}{S} \sum_{s=1}^{S} \delta(m_i^{(s)}, m_j^{(s)}) \tag{3.11}$$

$$\hat{P}_{ij}^A = \frac{1}{S} \sum_{s=1}^{S} \delta(m_i^{(s)}, m_j^{(s)}) \cdot (\gamma_{ij}^A)^{(s)}$$

$$\hat{P}_{ij}^B = \frac{1}{S} \sum_{s=1}^{S} \delta(m_i^{(s)}, m_j^{(s)}) \cdot (\gamma_{ij}^B)^{(s)}.$$

We use $\hat{P}_{ij}^m$ to identify clusters of components and $\hat{P}_{ij}^A$ with $\hat{P}_{ij}^B$ to select directional network edges without or with stimulus. Specifically, for module identification, we first rank $\hat{P}_{ij}^m$ for all $i, j = 1, \ldots, d$ and select a set of pairs of components $\mathcal{S} = \{(i, j) : \text{rank}(\hat{P}_{ij}^m) > \mathbf{h}, \; i, j = 1, \ldots, d\}$ for some predetermined threshold $\mathbf{h}$, for example, a top 5% rank. Given $\mathcal{S}$, we identify modules $\mathcal{C}_k$, $k = 1, \ldots, K$, also

a partition, of the d components, such that for any two components i and j in the same module, there exists a set of pairs $\{(k_0 = i, k_1), (k_1, k_2), \ldots, (k_{l-1}, k_l = j)\}$, which is a subset of $\mathcal{S}$. By using a high threshold $\mathbf{h}$, this procedure identifies clusters of components that have high posterior probabilities of being connected directly or indirectly.

For network edge selection, which is performed after the module identification process with a threshold $\mathbf{h}$, we first set $\hat{P}_{ij}^A$, $\hat{P}_{ji}^A$, $\hat{P}_{ij}^B$, and $\hat{P}_{ji}^B$ for i and j in two different clusters to zeroes. Then we order all the $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ for $i, j = 1, \ldots, d$ decreasingly and select network edges whose $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ have ranks higher than the threshold $\mathbf{h}$.

Similar to Bayesian variable selection for the linear regression, we use the ROC curve to summarize the performance of the proposed network edge selection procedure, which is computed as follows: for each given threshold, the percentages of true directional edges and null directional edges whose posterior probabilities are greater than the threshold are calculated as true positive rate (TPR) and false positive rate (FPR) of the selection procedure; the ROC curve summarizes pairs of TRPs and FPRs for different thresholds. The closer to $(0, 1)$ the ROC curve, i.e., the larger TPR for each FPR, the better performance of the selection procedure.

We select network edges based on the ranks rather than the exact values of their posterior selection probabilities, the same strategy used in Bayesian variable selection problems, because given $\mathbf{x}(t)$, identifying connected components is equivalent to

simultaneously solving multiple high-dimensional variable selection problems. With limited data information and many candidate predictors, the difference between posterior selection probabilities of void and true variables is small. As such, the ranks are more informative than the exact values of the posterior selection probabilities regarding the underlying network structure. In addition, it is possible that posterior probabilities of most true directional edges are smaller than 0.5 as well, because of the model uncertainty. Then it is more reasonable to select network edges based on their ranks.

One can select directional network edges based on original $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ without adjusting them for identified modules. However, we found that the network edge selection based on adjusted $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ — obtained through the procedure described above — outperforms that based on unadjusted $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ by having a higher ROC curve. We attribute this finding to two possible reasons. First, within the ODE framework, the components in the same cluster tend to have similar temporal behaviors, because the instantaneous change of each component directly or indirectly depends on the states of others in the same cluster. Then this temporal similarity among components within the same cluster causes difficulty in the directional edge selection due to multicollinearity, but facilitates module identification. Consequently, module identification tends to be easier and more accurate than network edge selection, and utilizing the information from the former can enhance the accuracy of the latter. Second, the network edge selection based on unadjusted $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ treats

each directional edge separately, while the selection based on adjusted probabilities utilizes information across different components.

Chapter 4

Simulation Examples

In this section, we generate some simulation examples to evaluate the performance of the proposed model mainly in three different scenarios. In Section 4.1, we apply the proposed BMIDDM to one simulated dataset in [100], where components within the same cluster are all pairwise connected, and compare the results with those by the P-iPDA. We demonstrate the good performance of our model when model assumptions are satisfied in the first example. We generate another time series data of the same dimension in Section 4.2 and use the BMIDDM to detect differences in the networks of these two simulated examples. The second example, combined with the first one, displays the power of our model in distinguishing between dissimilar networks. Last, we try the BMIDDM on some data generated from a nonlinear ODE model, instead of a bilinear one, in Section 4.3. In this example, we observe the robustness of our model when some model assumptions are moderately violated.

4.1 Example 1: Data from a Bilinear Model

The simulated dynamic system has 4 clusters of size 6, 4, 6, and 4. Each cluster is pairwise connected, and there is no edge between clusters. We let $T = 250$ and $u(t) = 1$ for $100 \leq t \leq 150$ and 0 otherwise, which are identical to those of the auditory ECoG data. For simplicity, parameter $\mathbf{B}$ is set as twice of $\mathbf{A}$ in equation 3.2, meaning that during the stimulus the network structure does not change but the strength of the directional interactions doubled. We generate 20 time series $\mathbf{x}(t)$ by using numerical approximation based on discretized bilinear model (3.3) with given parameters $\mathbf{\Theta}_I$, and generate 20 independent error time series $\boldsymbol{\epsilon}(t)$, each following an AR(1) model with a lag-one correlation of 0.5. The SNR — defined as $\text{var}(x_i(t))/\text{var}(\epsilon_i(t))$ — of each time series $y_i(t)$, the sum of $x_i(t)$ and $\epsilon_i(t)$, was set at 10. Figure 4.1 displays four representative time series from the four responding clusters in this example before error time series are added. The system is carefully designed so that time series do not explode, and curves in different clusters have different frequencies and magnitudes. Before applying the proposed Bayesian approach, we standardize the observed time series to unit variance, such that time series of different components are in the same scale. In the following, $\mathbf{y}(t)$ is referred to as the standardized data.

We applied the proposed BMIDDM with $\mu = 0$ and $\tau = \max_{i=1}^{d} \hat{\tau}_i$ to this simulated dataset for which the P-iPDA failed to identify all the clusters. The network by the P-iPDA is shown in Figure 4.2(a), which has only 36.5% TRP. In contrast, the

(a) Cluster I   (b) Cluster II   (c) Cluster III   (d) Cluster IV
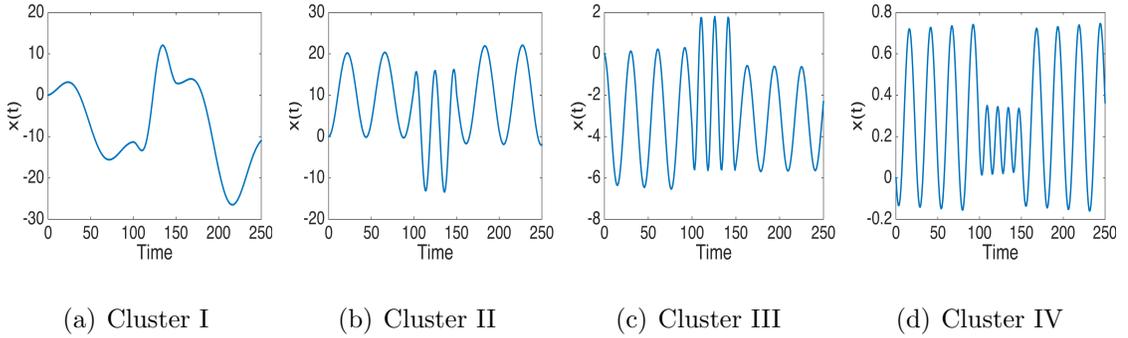
Figure 4.1: *Plots 4.1(a) to 4.1(d) show one representative time series $\mathbf{x}(t)$ from each cluster in Example 1. The curves correspond to $\mathbf{x}(t)$ before error terms $\boldsymbol{\epsilon}(t)$ are added.*

BMIDDM identified all the nonzero directional effects with zero FPR, as shown in Figure 4.2(c), in which directional edges are corresponding to the posterior probabilities $\hat{P}_{ij}^{B}$ with top 26% ranks, the exact percentage of true edges among all possible ones. In this example, the posterior selection probabilities of true network edges and null ones have a sharp difference, as illustrated by Figure 4.2(b) of $\hat{P}_{ij}^{B}$ for $i, j = 1, \ldots, d$. The plot of $\hat{P}_{ij}^{A}$ is similar and not shown here.

The better performance of the Bayesian method than the P-iPDA for this example is possibly due to three reasons. First, though components within the same cluster are indeed pairwise connected, the average degree of nodes, however, in the network is small. Consequently, the BMIDDM, allowing for more sparsity, has a better selection efficiency. Second, the P-iPDA, an optimization algorithm for an L_0 penalized criterion, very likely outputs a network corresponding to a local mode of the criterion, especially since the ODE model estimation is sensitive to noise. In contrast, the Bayesian method, evaluating the posterior probabilities of different cluster struc-

(a) Network by P-iPDA (b) Sel. Prob. with Stimulus (c) Network with Stimulus
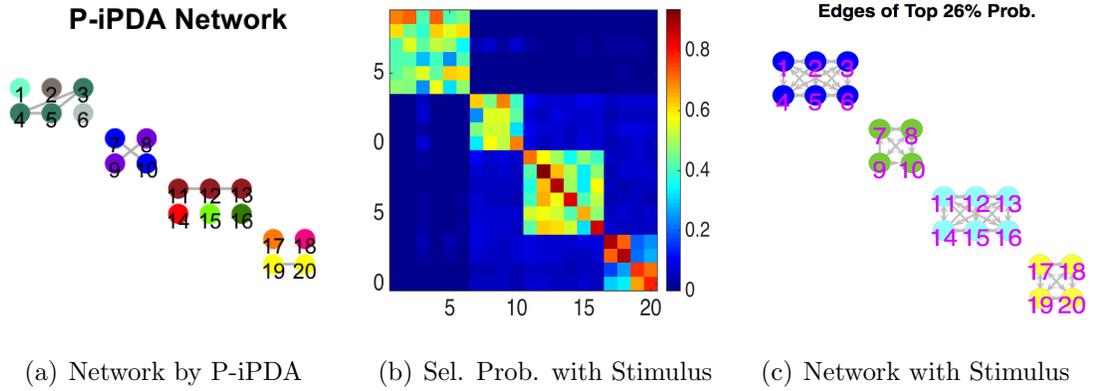
Figure 4.2: *4.2(a) The network result by the P-iPDA in Example 1. 4.2(b) shows $\hat{P}_{ij}^B$ for $i,j = 1,\ldots,20$. 4.2(c) shows network edges with top 26% (i.e., the percentage of true directional edges among all possible ones) posterior selection probabilities under the stimulus. Nodes in the same color correspond to components in the same cluster identified by the used method, either the P-iPDA or the Bayesian method.*

tures, outputs a posterior "average" and thus is more stable. Third, the performance of the P-iPDA is sensitive to the choice of penalty parameters. It is possible that the penalty parameters used by the P-iPDA are not optimal for the presented dataset, while the Bayesian method using $\tau = \max_{i=1}^d \hat{\tau}_i$ is more adaptive to the specific data being analyzed, and thus has a better performance.

We have compared the computational complexities and times of the proposed Bayesian method with the P-iPDA. MCMC simulations of module labels $\boldsymbol{m}$ and basis coefficients $\boldsymbol{\eta}$ from their posterior conditional distributions, the two most time-consuming steps in the PCGS, have similar computational complexity as the two optimization steps updating $\boldsymbol{m}$ and $\boldsymbol{\eta}$ in the P-iPDA, because they all require matrix

inversion of large matrices of similarly high dimensions. In particular, conditional on the same cluster structure m in the previous step, the computational complexity in the following iteration of the P-iPDA and PCGS is the same.

As searching for a mode is computationally much easier than exploring the entire posterior distribution, the P-iPDA takes much fewer iterations to converge than the PCGS. However, the P-iPDA relies on computationally extensive cross-validation, calculating prediction errors for every left-out data point and for a large number of candidate penalty parameter values, to find the best combination of penalty parameters with the smallest prediction error. Specifically, for the simulated dynamic system in Example 4.1, at least 6 cross-validations at 50 left-out points were needed to find the optimal penalty parameters. One usually needs to perform 35 iterations of the P-iPDA for each left-out point given one combination of penalty parameters. As such, the total number of iterations needed for penalty parameter selection is 10,500, close to the number of iterations in the PCGS. For dynamic systems of larger dimensions and with longer time series, more iterations of the P-iPDA and cross-validation on more candidate penalty parameters using more left-out points are needed, because the number of potential cluster structures is larger and the network result by the P-iPDA is more sensitive to the choice of penalty parameters. Overall, if accounting for the penalty parameter selection time, the PCGS and P-iPDA use similar amounts of computational time. For the simulation example under study, it took 1.2 hours for the proposed Bayesian method to finish 10,000 MCMC iterations on a personal

laptop using one i7 core.

We let

$$E_{ij} = \delta(m_i, m_j) \cdot \gamma_{ij}^A \cdot A_{ij},$$

$$G_{ij} = \delta(m_i, m_j) \cdot \gamma_{ij}^B \cdot B_{ij}.$$

With S posterior draws of BMIDDM parameters, for each pair of components i and j, we estimate the directional effects E_{ij} and G_{ij} exerted by region j over region i without and with the stimulus, respectively, by their posterior means, which are given by

$$\hat{E}_{ij} = \frac{1}{S} \sum_{s=1}^{S} \delta(m_i^{(s)}, m_j^{(s)}) \cdot (\gamma_{ij}^A)^{(s)} \cdot A_{ij}^{(s)},$$

$$\hat{G}_{ij} = \frac{1}{S} \sum_{s=1}^{S} \delta(m_i^{(s)}, m_j^{(s)}) \cdot (\gamma_{ij}^B)^{(s)} \cdot B_{ij}^{(s)}.$$

Then we evaluate the mean squared errors (MSE) of $\mathbf{E}$ and $\mathbf{G}$:

$$\text{MSE}(\mathbf{E}) = \sum_{i,j=1}^{d} (\hat{E}_{ij} - E_{ij})^2 / d^2,$$

$$\text{MSE}(\mathbf{G}) = \sum_{i,j=1}^{d} (\hat{G}_{ij} - G_{ij})^2 / d^2,$$

which are summarized in Table 4.1. For comparison, we also present the MSEs of the estimates by the P-iPDA.

Though the underlying cluster structure exactly matches the model assumption of the P-iPDA, i.e., the components in the same cluster are all pairwise connected, when the connectivity degree of each component is small, the Bayesian method gives much smaller errors, suggesting that incorporating indicator variables for significant

	Bayesian	P-iPDA		Bayesian	P-iPDA
MSE($\mathbf{E}$)	0.09	0.11	MSE($\mathbf{G}$)	0.11	0.20

Table 4.1: *The MSEs of estimated model parameters by the P-iPDA and the Bayesian method in Example 1.*

directional effects in the model improves both selection accuracy of network edges and estimation efficiency of model parameters. Furthermore, we can also find that the difference of MSEs of $\mathbf{G}$ between the Bayesian method and P-iPDA is much larger than that of MSEs of $\mathbf{E}$, showing that the Bayesian method has a much better parameter estimation than P-iPDA during the stimulus, when the number of observations is more limited.

4.2   Example 2: Network Comparison

We generate a dynamic system of 20 dimensions with 3 clusters of size 6, 4, and 10 from the bilinear model (3.3). The state functions $\mathbf{x}(t)$ of the first two clusters are generated using exactly the same parameters as those in Example 1. The third cluster consists of 10 densely connected components, which are similar to those in the last two clusters in the previous example. To mimic the network status changes of brain systems, we use values of $\mathbf{x}(t)$ at the last time point in Example 1 as the starting point in this example. Figure 4.3 shows three representative curves from the three clusters.

(a) Cluster I   (b) Cluster II   (c) Cluster III
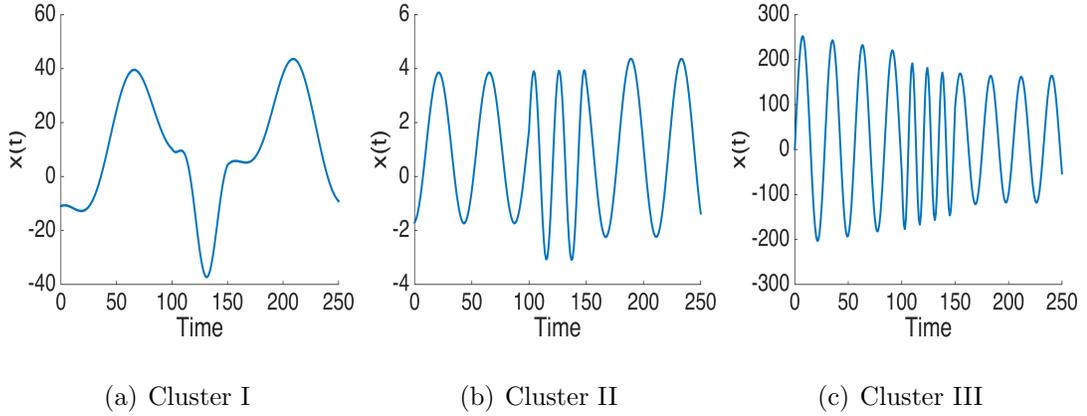
Figure 4.3: *Plots 4.3(a) to 4.3(c) show one representative time series $\mathbf{x}(t)$ from each cluster in Example 2. The curves correspond to $\mathbf{x}(t)$ before error terms $\boldsymbol{\epsilon}(t)$ are added.*

We apply our BMIDDM with $\mu = 0$ and $\tau = \max_{i=1}^{d} \hat{\tau}_i$ to this simulation example independently. Figure 4.4(a) shows estimated network edge selection probabilities $\hat{P}_{ij}^{B}$, and Figure 4.4(b) shows the network edges with top 38% posterior selection probabilities under the stimulus, where 38% is again the true percentage of edges in the underlying network.

We develop a simple approach to compare networks of Examples 1 and 2. Let m_{1i} and m_{2i} denote module labels of the ith component in networks 1 and 2, respectively, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be the observed time series data of networks 1 and 2, respectively, and

$$p_1 = P\left(\delta(m_{1i}, m_{1j}) = 1 | \mathbf{Y}_1\right),$$

$$p_2 = P\left(\delta(m_{2i}, m_{2j}) = 1 | \mathbf{Y}_2\right).$$

So p_1 represents the probability that region i and region j are in the same cluster in network 1 given the first time series, and p_2 then indicates the corresponding proba-

(a) Estimated $\hat{P}_{ij}^B$ (b) Network with Stimulus
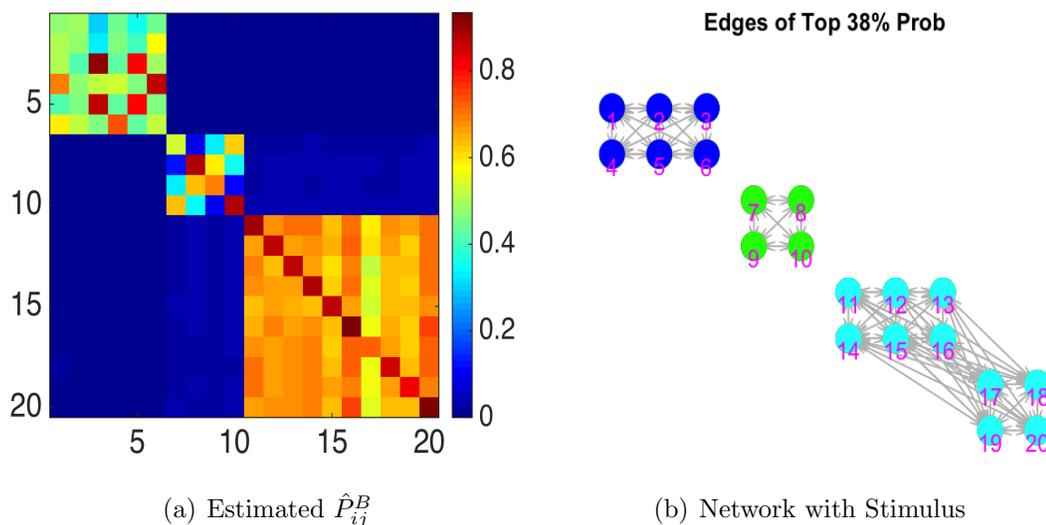
Figure 4.4: *4.4(a) Estimated $\hat{P}_{ij}^B$ by the BMIDDM in Example 2. 4.4(b) shows network edges with top 38% (i.e., the percentage of true directional edges among all possible ones) posterior selection probabilities under the stimulus. Nodes in the same color correspond to components in the same cluster identified by BMIDDM.*

bility in network 2 given the second data set. We compare the clustering probabilities of two components in two separate networks through evaluating the probability

$$P_{ij}^d = P\left(\delta(m_{1i}, m_{1j}) \neq \delta(m_{2i}, m_{2j}) | \mathbf{Y}_1, \mathbf{Y}_2\right),$$

which denotes the probability that whether component i and component j are in the same cluster is different in networks 1 and 2 given the two sets of time series. There are only two possible situations for that to happen: region i and region j are in the same cluster in network 1 but in two different clusters in network 2, or the other way around. Since the time series data of two networks are analyzed independently,

$\delta(m_{1i}, m_{1j})$ and $\delta(m_{2i}, m_{2j})$ are independent and thus

$$
\begin{aligned}
P_{ij}^d &= P\left(\delta(m_{1i}, m_{1j}) = 1 \ \& \ \delta(m_{2i}, m_{2j}) = 0 | \mathbf{Y}_1, \mathbf{Y}_2\right) \\
&+ P\left(\delta(m_{1i}, m_{1j}) = 0 \ \& \ \delta(m_{2i}, m_{2j}) = 1 | \mathbf{Y}_1, \mathbf{Y}_2\right) \\
&= p_1 + p_2 - 2\, p_1 \cdot p_2.
\end{aligned}
$$

In practice, we evaluate p_1 and p_2 by the corresponding clustering probabilities $\hat{P}_{ij}^m$ (3.11) in their respective network. Figure 4.5(a) shows P_{ij}^d for comparing networks in Examples 1 and 2. It is observed that the areas corresponding to the difference between the two networks have clearly higher P_{ij}^d than the others, just as expected. Following typical Bayesian decision procedure [40], we use 0.5 as the threshold for P_{ij}^d, and the two components i and j are deemed to have different clustering probabilities in the two networks if $P_{ij}^d > 0.5$. The green cells in Figure 4.5(b) indicate a pair of components i and j with P_{ij}^d smaller than 0.5; the red cells indicate components with P_{ij}^d larger than 0.5 and $p_1 > p_2$; and the blue cells indicate components with P_{ij}^d larger than 0.5 and $p_1 < p_2$. The proposed approach identified components 11 to 20 to have stronger connectivity in Example 2 than those in Example 1. Overall, 80 edges (400 total edges) are identified significantly different, among which 48 edges are true positives with 100% TPR and 9% FPR. Figure 4.5(c) shows the ROC curve for selecting component pairs with different clustering probabilities in the two simulated examples by using different thresholds on P_{ij}^d.

(a) P_{ij}^d (b) P_{ij}^d with a cutoff 0.5   (c) ROC Curve
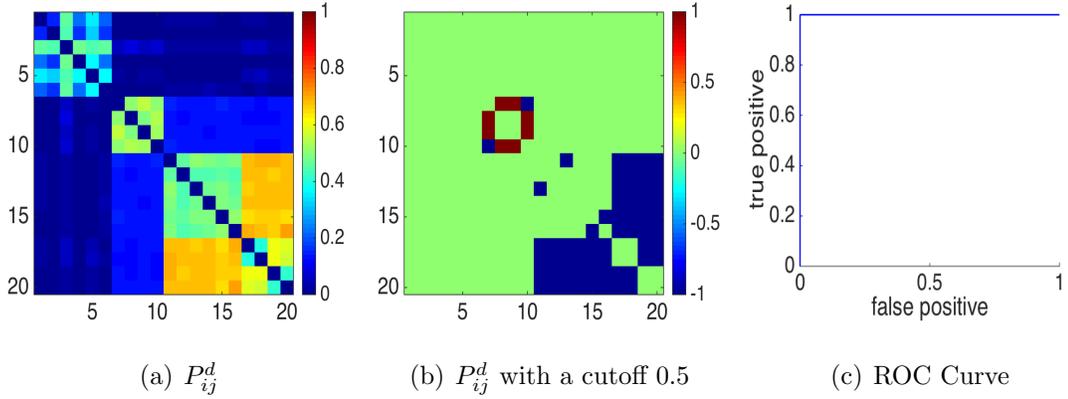
Figure 4.5: *4.5(a) shows P_{ij}^d for $i, j = 1, \ldots, 20$. 4.5(b) shows identified pairs of components with significantly different clustering probabilities using a cutoff 0.5 on P_{ij}^d. Specifically, if $P_{ij}^d < 0.5$, the corresponding cells (a pair) in the Figure are green; if $P_{ij}^d > 0.5$ and $p_1 > p_2$, the corresponding cells are red; and if $P_{ij}^d > 0.5$ and $p_1 < p_2$, the corresponding cells are blue. 4.5(c) shows the ROC curve for selecting pairs of components with different clustering probabilities in two networks using various cutoffs on P_{ij}^d.*

4.3   Example 3: Data from a Nonlinear Model

We use this nonlinear example to evaluate the performance of the proposed Bayesian approach when the bilinear model assumption is moderately violated. A dynamic system of 20 dimension with 3 clusters of size 5, 5, and 10 is generated, from a nonlinear ODE model:

$$\frac{d\mathbf{x}(t)}{dt} = \sin\left(\mathbf{E}\mathbf{x}(t)\right) \cdot (1 - u(t)) + \sin\left(\mathbf{G}\mathbf{x}(t)\right) \cdot u(t) + \mathbf{C}u(t) + \mathbf{D}, \qquad (4.1)$$

where $\mathbf{G} = 2\mathbf{E}$. Among the three clusters, the first two are generated using the same parameter values, so their corresponding time series curves are similar. Figure 4.6

displays what those curves look like in different clusters. The input function $u(t)$, the length of observed data T, and the observed data $\mathbf{y}(t)$ are generated in the same manner as in Example 1 and 2.

(a) Cluster I & II   (b) Cluster III
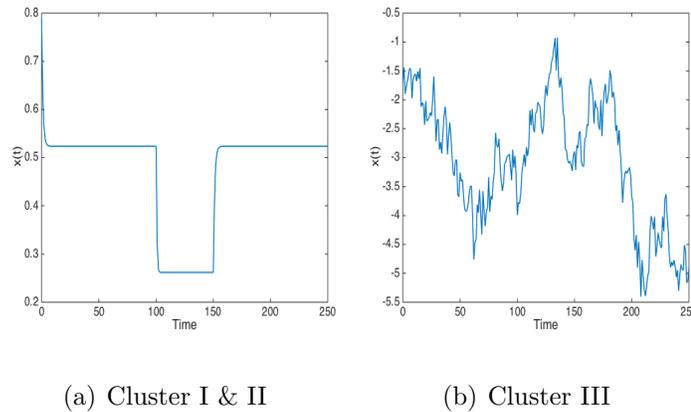
Figure 4.6: *Plots 4.6(a) and 4.6(b) show representative time series $\mathbf{x}(t)$ from the clusters in Example 3. The curves correspond to $\mathbf{x}(t)$ before error terms $\boldsymbol{\epsilon}(t)$ are added.*

Additionally, $\mathbf{E}$ and $\mathbf{G}$ in this nonlinear example (4.1) are highly sparse with only 9.5% nonzero elements. The associated network is shown in Figure 4.7.
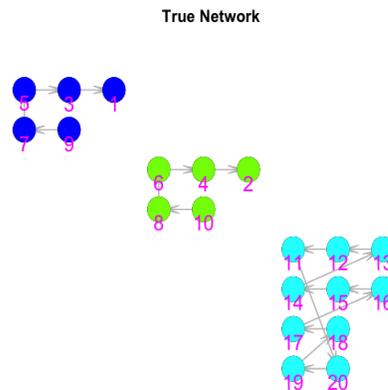

Figure 4.7: *True directional network of Example 3.*

The posterior clustering probabilities $\hat{P}_{ij}^m$ (3.11) for every pair of components (i, j) are shown in Figure 4.8(a). The plots for $\hat{P}_{ij}^A$ and $\hat{P}_{ij}^B$ adjusted for the identified clusters are similar, and thus omitted here. If selecting directional edges with top 30% highest posterior probabilities, four components, two from each of the first two clusters, are incorrectly clustered, as shown in Figure 4.8(b). This is because the model parameters for the firsts two clusters are almost identical, and thus the state functions of these two clusters have similar temporal patterns and tend to be clustered together due to multicolinearity. If using a higher threshold, e.g. 15%, for the posterior probabilities, the Bayesian method identified the largest cluster, but failed to identify the other two clusters, as shown in Figure 4.8(c).


(a) Post. Clustering Prob.   (b) Estimated Network   (c) Estimated Network
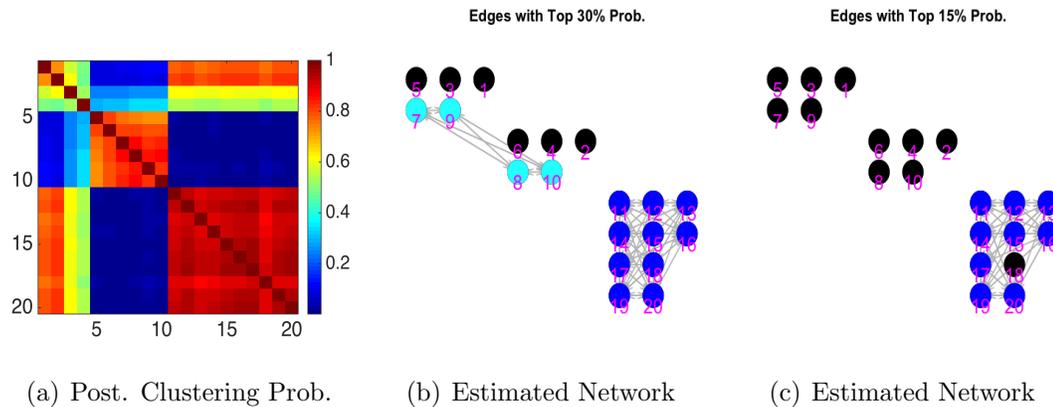
Figure 4.8: *4.8(a) The posterior probabilities for each pair of components being in the same cluster. 4.8(b) and 4.8(c) Two identified directional networks with edges corresponding to top 30% and 15% posterior probabilities $\hat{P}_{ij}^A$. Nodes in the same colors of either light blue or blue, correspond to regions in the same cluster identified by the BMIDDM, and nodes in black correspond to regions in the clusters with only one component.*

Nevertheless, the Bayesian method achieves almost 60% TPF for 0% FPR, as shown by the ROC curves of network edge selection with and without the stimulus in Figure 4.9(a). We also apply the P-iPDA to the same dataset for comparison. The P-iPDA incorrectly clusters four components in the first two clusters together with four components in the third cluster, as shown in Figure 4.9(b), which leads to a FPR larger than 20% and a smaller TPR than that of the proposed Bayesian method given the same FPR, as demonstrated in Figure 4.9(a).


(a) ROC Curve   (b) P-iPDA Network
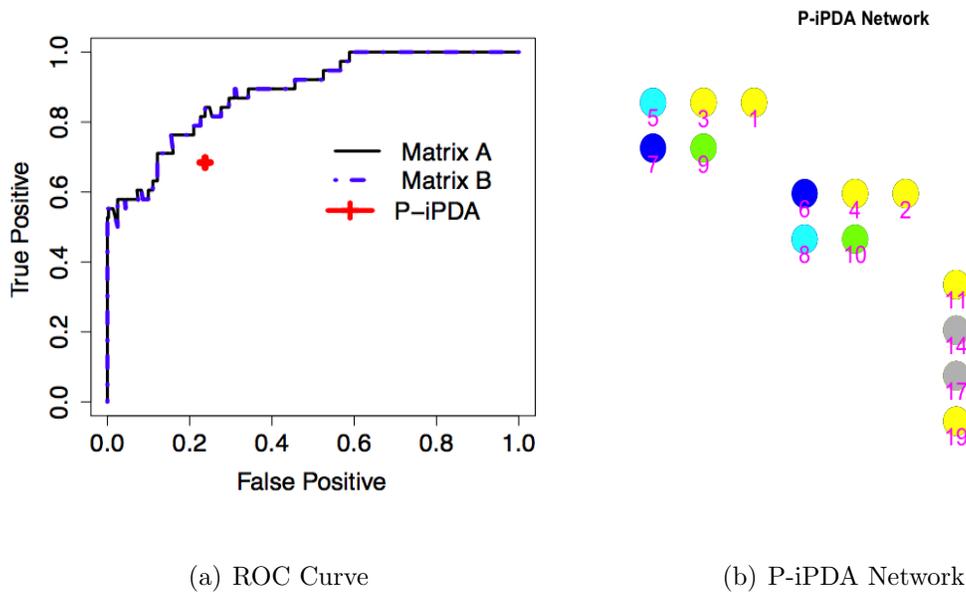
Figure 4.9: *4.9(a) The ROC curves for network edge selection with and without the stimulus. The red cross corresponds to the TPR and FPR of the network result by the P-iPDA, which is shown in 4.9(b). The directional edges are omitted in 4.9(b), because components in the same cluster identified by the P-iPDA are all pairwise connected. Nodes in the same color correspond to components in the same cluster identified by the P-iPDA.*

Overall, in comparison with the P-iPDA, the Bayesian method can have a higher TPR with a threshold that permits similar FPR as P-iPDA. Furthermore, with a relatively high threshold for the posterior selection probabilities of network edges, BMIDDM tends to identify clusters with a small FPR, even if the model assumptions are violated. This implies that when using the proposed BMIDDM, the edges found with a high enough threshold are quite reliable, with a pretty small portion of them being false edges.

Chapter 5

Application to ECoG Data

In this chapter, we apply the proposed model, BMIDDM, to a set of ECoG time series in real life. We briefly explain the experimental setup about how the data is collected, and then discuss the analysis of the data using our model and the scientific meanings of the results.

5.1   Data Acquisition

The auditory ECoG data acquisition and processing methods have been described previously [100]. Briefly, continuous ECoG signals were recorded simultaneously from a 6×8 array of electrodes (2.3 cm diameter, 9 cm spacing) implanted over the lateral left hemisphere of an adult epilepsy patient for clinical purposes of localizing seizures prior to resection surgery. Figure 5.1 shows the spatial placement of ECoG electrodes on the epileptic patient whose ECoG data are analyzed in this study. The

experimental paradigm was a 300-trial passive listening task using pure tones (50 ms duration). Since electrodes 47 and 48 are used as the reference and ground electrodes and electrode 32 contains excessive noise, recordings from a total of $d = 45$ electrode channels were analyzed. Auditory responses, computed using time-domain and time-frequency analyses [10, 30, 34, 81] were identified at three electrode sites (electrodes 14-16) consistent with the location of auditory cortex in the posterior temporal lobe. Seven electrode sites 1-4, 9-10, and 18 located in the inferior anterior temporal lobe were identified as the primary seizure focus based on clinical recordings.
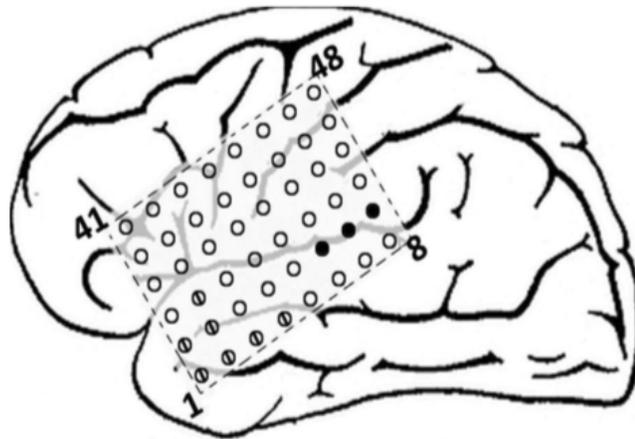

Figure 5.1: *Spatial placement of ECoG electrodes on an epileptic patient.*

The recording data contain 246 trials of ECoG recordings using a frequently presented, standard 1000 Hz tone stimulus of 50 ms duration and 54 trials using a different, infrequently repeated, i.e., deviant, 1200 Hz tone stimulus also of 50 ms duration. All tone stimuli were presented sequentially at 1400 ms inter-stimulus interval. Following common practice in the literature, we focus on brain activity in

an early cortical auditory processing time range, which corresponds to the first 150 ms after the stimulus onset. As such, each trial of data is of 250 ms duration: 100 ms pre-stimulus (0-100 ms), 50 ms for stimulus presentation (100-150 ms), and 100 ms post-stimulus (150-250 ms). Figure 5.2 gives a sense of how the data looks like by displaying curves from 3 electrodes in the first trial using the standard stimulus. We applied the BMIDDM to each 250-ms window independently to allow for variation of brain networks across trials and also to ensure that the assumed bilinear ODE model can approximate the underlying nonlinear system effectively. As such, in the MIDDM for the ECoG data, $d = 45$, $T = 250$, and $u(t) = 1$ for $100 \leq t \leq 150$.
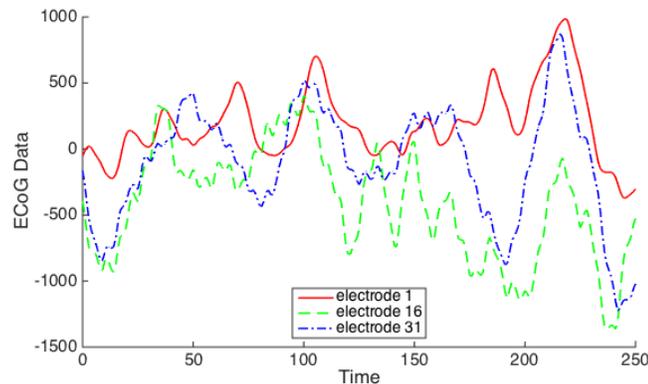

Figure 5.2: *ECoG time series at electrodes 1, 16, and 31 from the first trial using the frequently presented, standard 1000 Hz stimulus.*

5.2 Data Analysis

We applied the proposed Bayesian model to the observed ECoG time series and obtained the posterior selection probabilities for all the trials.

We compared the first five trials with the rest of trials using 1000 Hz stimulus by calculating numbers of region pairs, denoted by $R_{l_1 l_2}$ for the pairwise comparison of trials l_1 and l_2, with different clustering probabilities (threshold 0.5 on P_{ij}^d), and found that the brain network in response to the first stimulus is distinct from those in response to the rest of the stimulus sequence. Figures 5.3(a) and 5.3(b) show histograms of the percentages of region pairs (among all possible pairs) with P_{ij}^d larger than 0.5 in the network comparison between the first two trials ($l_1 = 1, 2$) and the rest of trials using 1000 Hz stimulus, i.e., histograms of $R_{l_1 l_2}/d(d-1)$ for $l_1 = 1$ or 2 and $l_2 \neq l_1$. Most pairs of regions in the first trial have much larger clustering probabilities than those in the rest of the trials, in line with the discovery by [7] that connectivity strength between regions is the strongest in the first of several repetitive auditory events. In addition, by exploring connectivity among many regions, we found that the brain network of the first trial has stronger connections than the networks in the rest of trials (with 1000 Hz tone stimulus).

We evaluated networks in response to infrequent, deviant 1200 Hz stimulus. Figure 5.3(c) shows the histogram of the proportions of region pairs with P_{ij}^d larger than 0.5 in the network comparison between the first and the rest trials using 1200 Hz stimulus. In comparison with distinct networks of the first and of the rest trials using 1000 Hz stimulus, the difference among networks in response to the deviant 1200 Hz stimulus is much less pronounced. We have compared the network of the first trial using 1000 Hz stimulus with those of the 54 trials using 1200 Hz stimulus, as shown

First Frequent Trial vs. The Rest

Second Frequent Trial vs. The Rest

(a) 1000 Hz Trial 1 vs. Rest

(b) 1000 Hz Trial 2 vs. Rest

First Infrequent Trial vs. The Rest

First Frequent Trial vs. Infrequent Trials

(c) 1200 Hz Trial 1 vs. Rest

(d) 1000 Hz Trial 1 vs. 1200 Hz Trials
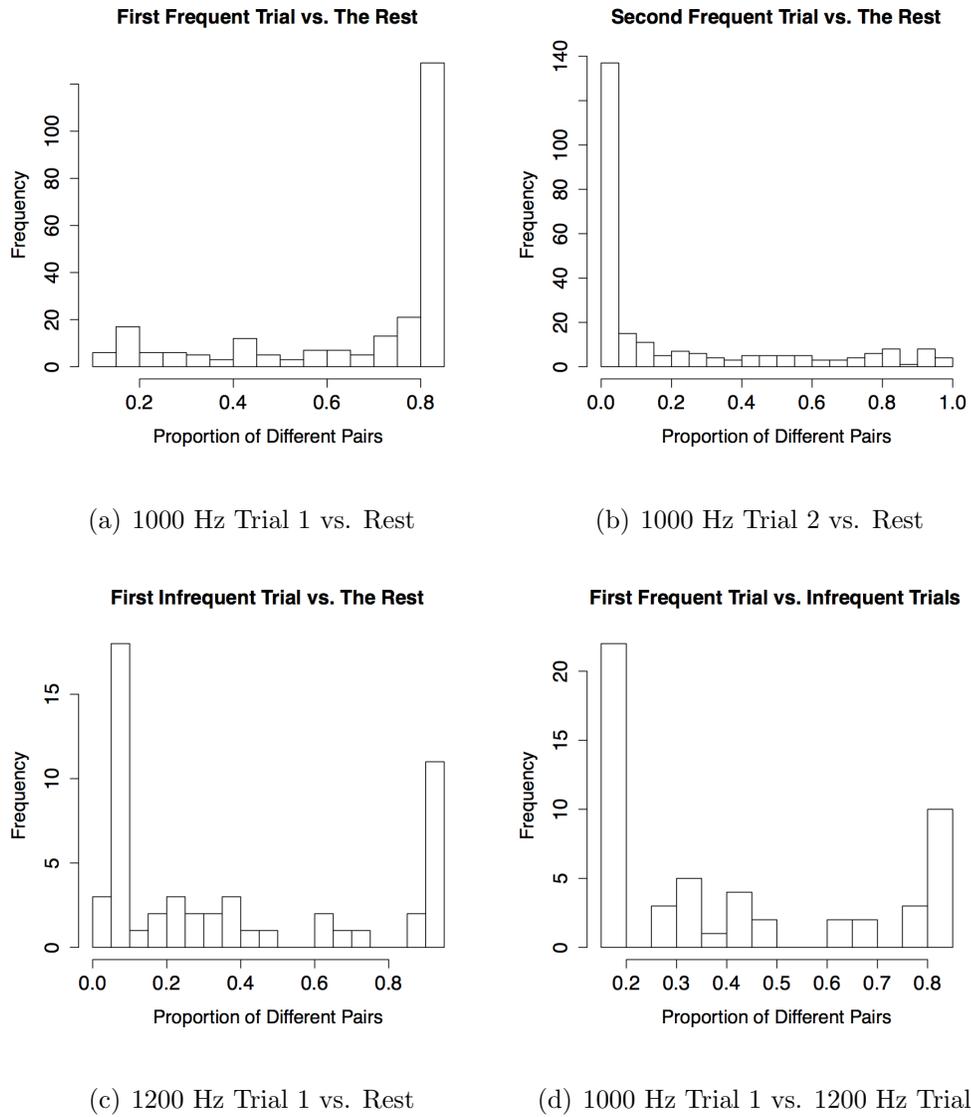
Figure 5.3: *5.3(a) and 5.3(b) show histograms of proportions of region pairs with significantly different clustering probabilities (i.e., $P_{ij}^d > 0.5$) in the network comparison of the first two trials with the rest trials using 1000 Hz stimulus. 5.3(c) shows the histogram that compares the first trial with the rest trials using 1200 Hz stimulus. 5.3(d) shows the histogram that compares the first trial using 1000 Hz stimulus with the trials using 1200 Hz stimulus.*

in Figure 5.3(d): the network of the first trial is similar to networks in most trials using infrequent, deviant stimuli. Similar results have been reported in the literature [31, 48, 49] that the observation of decreasing responses to repetitive stimuli does not apply to different or deviant stimuli.

To summarize analysis results of trials associated with the two different stimulus types, we calculated average ranks of posterior network-edge-selection probabilities $\hat{P}_{ij}^A$ of 246 trials using 1000 Hz stimulus and of 54 trials using 1200 Hz stimulus, respectively, and presented the two average networks with top 5% average ranks in Figures 5.4(a) and 5.4(b). We selected directional edges with top 5% posterior probabilities to identify most closely connected components with a small FPR. Moreover, with this high threshold, the identified clusters are small in size and thus, easier to examine visually.

For trials using regular 1000 Hz stimuli, the identified network consists of two modules of closely connected brain regions that are believed to be specialized in different brain functions. The auditory responsive regions, electrodes 14-16, interact closely with regions in the posterior temporal lobe, involved in auditory perception. Regions in the inferior frontal lobe have dense interactions, which is in line with existing findings of short frontal lobe connections [19] and the implication of this brain area in predictive coding (generating expectations based on stimulus presentation probability) [7]. In comparison, for trials using 1200 Hz stimulus, regions in the frontal lobe show the strongest connections, consistent with the role of the frontal

(a) 1000 Hz Average Network   (b) 1200 Hz Average Network
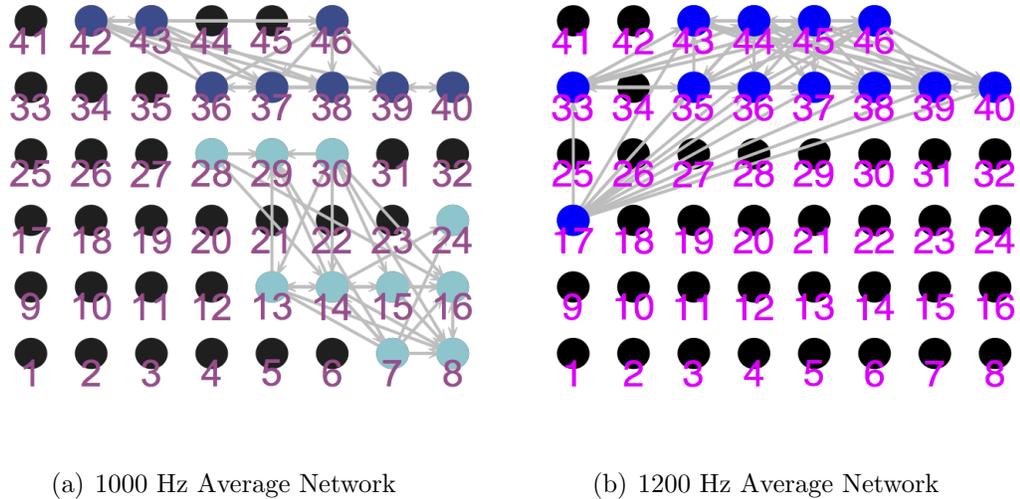
Figure 5.4:   *5.4(a) and 5.4(b) show the network edges with top 5% average ranks of $\hat{P}_{ij}^A$ across trials using 1000 Hz and 1200 Hz stimuli, respectively. Nodes in the same colors of either light blue, blue, or dark blue, correspond to regions in the same cluster identified by the BMIDDM, and nodes in black correspond to regions in the clusters with only one component.*

lobe in detecting novel or different auditory events [65, 80].

We estimated $\mathbf{E}$ and $\mathbf{G}$ of 300 trials, and the average $\hat{\mathbf{E}}$ of trials using two different stimuli are shown in Figures 5.5(a) and 5.5(b). Estimates of parameters denoting brains effective connectivity in response to deviant stimuli are much larger in absolute values than those associated with regular, repetitive stimuli, suggesting stronger effective connectivity among brain regions in the former scenario. This result is in line with the finding of stronger brain responses to deviant stimuli in the literature [31].

(a) 1000 Hz Average **E**
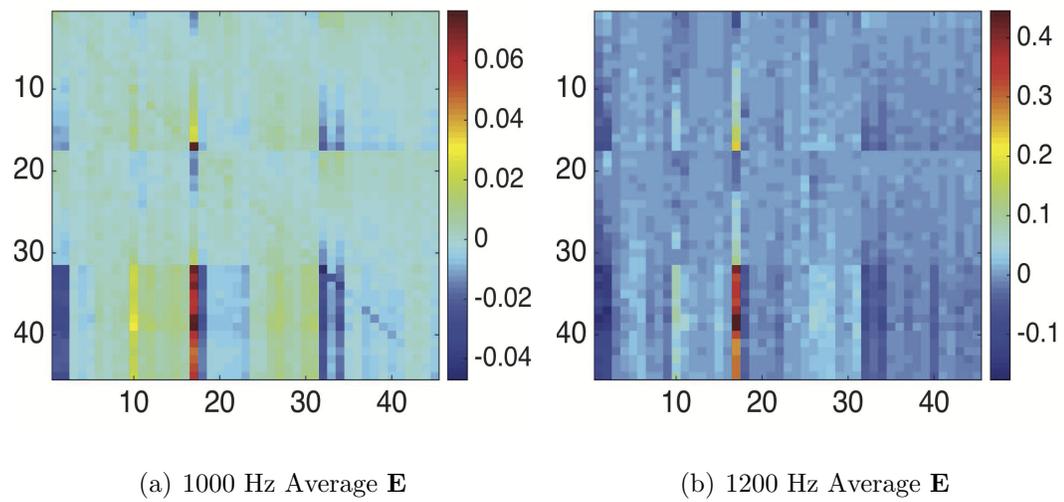
(b) 1200 Hz Average **E**

Figure 5.5: *5.5(a) and 5.5(b) show average posterior means of* **E** *across trials using 1000 Hz and 1200 Hz stimuli, respectively.*

Chapter 6

Conclusions and Future Study

In this chapter, we first summarize the advantages of the proposed model, BMIDDM, over the previous model, P-iPDA, and how our model can help us better understand the functional organizations of the brain. After that, we discuss some potential directions for future study of this topic, including three aspects: model development, computational improvement, and data application.

6.1 Conclusions

The BMIDDM brings three crucial advantages over the existing optimization method P-iPDA. First, one can detect different strengths of directional interactions among brain regions by the BMIDDM, and thus identify different levels of connectivity in the brain network. In the presented real data analysis of trials using 1000 Hz stimulus, given different posterior selection probabilities of network edges, the BMIDDM

identified two small modules, which consist of spatially close brain regions and are specialized for different functions. In contrast, the P-iPDA grouped the two modules into one cluster and treated all connections within this cluster equally. Second, by using a high threshold for the posterior probabilities of network edges, the Bayesian method identifies most closely connected regions with a small FPR. Moreover, the resulting small clusters are much easier to understand and interpret than large clusters outputted by the P-iPDA, especially in connectivity studies of hundreds of brain regions. Third, given different posterior clustering probabilities of brain regions in different trials, the BMIDDM can be used to detect network changes.

The application of the proposed method to brain connectivity studies will potentially enhance our understanding of the brain's functional organization for two reasons. First, since brain components in the same module usually have a similar function, the module identification can be used to determine the module's functional role in the brain network, especially if the function of some brain regions in the module is already known. Taking the network result of the real data analysis, shown in Figure 5.4(a), as an example, the regions (electrode sites 7, 8, 13, 24, 28-30) connected with the auditory responsive regions (electrode sites 14-16) are mostly likely also specialized for primary auditory perception. Second, the module identification can significantly simplify the study of the brain's functional organization. It is much easier, both in terms of computation and interpretation of the results, to evaluate connections between and within modules separately than to evaluate connections be-

tween every pair of brain regions. Moreover, the proposed cluster-structured ODE model and the associated estimation method are scalable to a system with hundreds of components by first identifying modules locally and treating modules as components of the large-scale system.

The proposed Bayesian framework can also be applied for the inference of general high-dimensional ODE models. High-dimensional ODE models are difficult to estimate and make statistical inferences, usually due to computational difficulties or complex relationships between components. Our Bayesian approach, although developed under a medical scenario, can be used for other ODE models with a cluster and indicator structure as well.

6.2 Future Study

In this section, we discuss some potential directions for future study from three perspectives. First, we bring up some ways that we have considered to improve the current model. Second, we put forward some computational techniques that we may want to try since the speed of the program limits the application of the method. Finally, we suggest some other problems in real life that the proposed approach could possibly help with.

6.2.1 Future Development of Models

In the brain network, connections within modules are usually short-range, strong, and dense, while those between modules tend to be sparse and long-range to ensure integration among different specialized areas [69]. It is very likely that the two modules specialized for different functions shown in Figure 5.4(a) are integrated through weak, long-range connections. In the future study, we will extend the MIDDM to accommodate interactions among clusters. Then the ensuing ODE model uses two types of indicators for connectivity within and between modules, and the prior on indicators (3.8) is expanded to incorporate probabilities on between-module indicators. To distinguish between within- and between-module connections, the prior probability for the former should be much larger than that for the latter to reflect the prior belief that connections within modules are dense and connections between modules are sparse. The priors of the rest of model parameters are unchanged, and a similar PCGS algorithm can be developed for ensuing posterior simulations.

Since spatially close brain regions tend to have stronger connections, we can also incorporate regions' spatial structure into the Bayesian framework for inferring effective connectivity between the regions. For example, the prior on module labels $\exp\{-\mu \sum_{ij} \delta(m_i, m_j)\}$ can be modified to $\exp\{-\mu \sum_{ij} W_{ij} \delta(m_i, m_j)\}$, where the weight W_{ij} is proportional to the distance between regions i and j. Similarly, the prior probability for γ_{ij}^A and γ_{ij}^B can be changed to p_{ij}, which is proportional to the distance between regions i and j, leading to higher prior probabilities for short-range

network edges. The choices of W_{ij} and p_{ij} and their effect on the ensuing posterior inference will be evaluated in the future research.

If one is interested in identifying components on which the stimulus has a direct effect, he/she can include indicators γ_i^C for coefficient $C_i, i = 1, \ldots, d$, in the MIDDM. Then the prior on indicator variables in 3.8 is modified to incorporate the prior probability p_c for γ_i^C for $i = 1, \ldots, d$. The same PCGS algorithm can be used to draw samples from the ensuing posterior distribution. Note that if the goal is to identify nonzero C_is, the observed data $\mathbf{y}(t)$ should be standardized to unit norm without changing the center of the time series and the signs (positive, negative, or zero) of C_is in the ODE model for standardized data.

Activities of the brain system are oscillatory, which has been reported to appear in various important brain functions [79]. To better characterize those oscillatory activities, the current first-order ODE model can be extended to a second-order ODE. Specifically, the second-order derivatives of the state functions are modeled as a bilinear equation of both the state functions and their first-order derivatives. The whole framework will be very similar to the current one, except that there are more ODE model parameters to estimate. The new ODE coefficients can be dealt with in the same way as $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}$ in the MIDDM with a flat Gaussian distribution as the prior.

6.2.2   Potential Computational Improvement

The algorithm is computationally challenging for three reasons. First, in some simulation steps, we have to invert large matrices to calculate covariance matrices for Gaussian distributions from which the parameters are simulated from. This matrix inversion step takes cubic time and can be quite slow especially when the matrix is high-dimensional. Second, in MCMC procedure, each iteration can only start after the previous iteration is finished because it depends on the results from the previous one. This makes parallel computing at this level impossible. Last, we need to run sufficiently large number of iterations to arrive at the stationary distribution of the Markov chain.

One potential method to further speed up the algorithm is the asymptotic Bayes [37]. It releases the dependency between iterations so that parallel computing can be applied on the between-iteration level, which could possibly increase the efficiency of the algorithm. It is actually straightforward to implement asymptotic Bayesian method given our current Bayesian framework. However, careful mathematical proof is required beforehand to theoretically show that the Markov chain will converge to the same posterior distribution.

Another potential direction to tackle the computational challenge is the variational Bayesian method. Intuitively, it shortens the computing time greatly by sacrificing some accuracy of the results. However, it would take much more efforts to construct the variational Bayesian framework compared with deriving the Gibbs sam-

pler equations. In addition, further experiments and simulation studies are needed to test the accuracy of the outcome.

6.2.3   Applications to Other ECoG Data

In addition to the auditory ECoG data set that we have analyzed in this work, the MIDDM can be used on the ECoG data for other primary brain experiments (e.g., visual or auditory tasks) as well, in order to estimate and compare the brain networks in response to those particular tasks. Besides, the proposed model can also be applied to the seizure ECoG data, which measures the brain responses over time with several seizure onsets from epilepsy patients, to study brain network changes before and after seizure onsets.

Currently after the patient agrees the surgical resection, the epilepsy specialists will inspect the ECoG data collected from that patient visually and determine the part of the brain identified as the seizure focus, or the initiation point of the seizure, which will be removed in the following surgery. However, different doctors looking at the same ECoG data can have different opinions on where the seizure focus is, and some patients can still have seizure even after the surgery, meaning that either the wrong part of the brain is removed, or the actual seizure focus involves a larger area than the removed one.

Since the MIDDM method analyzes brain networks and does pairwise network comparisons, it could possibly help with the identification of the seizure focus in the

following way. Since the seizure focus is the initiation point of abnormal brain activities, we hope to find it isolated from other normal brain regions when constructing the network for all the brain regions right after the seizure onset. Moreover, the abnormality of the seizure focus will spread to other brain areas after a while. Then we expect to find some patterns of network changes and to observe the seizure focus connected with other brain areas.

One potential challenge of applying the MIDDM to the seizure ECoG data is the long computing time. The dimension of the seizure data is larger than that of the auditory data, so it will take longer for the algorithm to complete for the seizure data. However, the method would be less helpful in practice if it is too slow. Furthermore, different patients can have different seizure focuses and different brain conditions. As such, it is hard to develop a general method to find the seizure focus, with the false positive rate being strictly controlled at the same time. We will further explore the potentials of the MIDDM in helping the treatment of epilepsy patients in the future research.

Appendix

A Proof of Proper Posterior Distribution

Let $\mathfrak{N}$ be the normalizing constant for $p(\boldsymbol{\eta}, \boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2 | \mathbf{Y}, \tau, \mu)$ in equation (3.10). Because $\exp\left\{-\frac{1}{2\tau} \mathrm{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I)\right\} \leq 1$ and $p_0 < 1$, then we have:

$$p(\boldsymbol{\eta}, \boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2 | \mathbf{Y}, \tau, \mu) \leq \mathfrak{N} \cdot \prod_{i=1}^{d} \frac{1}{\sigma_i^T} \exp\left\{-\frac{(Y_i - \boldsymbol{\Phi}\,\eta_i)^2}{2\sigma_i^2}\right\} \cdot \prod_{i,j=1}^{d} \phi\left(\frac{A_{ij}}{\xi_0}\right) \qquad (6.1)$$
$$\cdot \prod_{i,j=1}^{d} \phi\left(\frac{B_{ij}}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{C_i}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{D_i}{\xi_0}\right) \cdot \prod_{i=1}^{d} \frac{1}{\sigma_i^2}.$$

The above inequality gives an upper bound for the posterior joint density. In the following, we integrate out parameters in this upper bound step by step, and show that the upper bound is integrable.

Since $\boldsymbol{m}, \boldsymbol{\gamma}^A$, and $\boldsymbol{\gamma}^B$ are discrete and take a finite number of different values, after integrating these parameters out in (6.1), the ensuing joint posterior of $\boldsymbol{\theta}$ and

$\boldsymbol{\sigma}^2$ follows

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2|\mathbf{Y}, \tau, \mu) \leq \mathfrak{C}_d \cdot \mathfrak{N} \cdot \prod_{i=1}^{d} \frac{1}{\sigma_i^{T+2}} \exp\left\{-\frac{(Y_i - \boldsymbol{\Phi}\,\eta_i)^2}{2\sigma_i^2}\right\} \tag{6.2}$$
$$\cdot \prod_{i,j=1}^{d} \phi\left(\frac{A_{ij}}{\xi_0}\right) \cdot \prod_{i,j=1}^{d} \phi\left(\frac{B_{ij}}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{C_i}{\xi_0}\right) \cdot \prod_{i=1}^{d} \phi\left(\frac{D_i}{\xi_0}\right),$$

where $\mathfrak{C}_d$ is some positive constant depending on d. After integrating out $\boldsymbol{\theta}$ in (6.2), we have

$$p(\boldsymbol{\eta}, \boldsymbol{\sigma}^2|\mathbf{Y}, \tau, \mu) \leq \xi_0^{2d^2+2d} \cdot \mathfrak{C}_d \cdot \mathfrak{N} \cdot \prod_{i=1}^{d} \sigma_i^{-T-2} \exp\left\{-\frac{(Y_i - \boldsymbol{\Phi}\,\eta_i)^2}{2\sigma_i^2}\right\}.$$

Then as long as the number of basis coefficients L is smaller than the number of time points T for each component, the formula on the right of the above inequality is integrable.

B Technical Details of PCGS Algorithm

B.I Joint posterior distribution $p(\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B|\mathbf{Y}, \tau, \mu)$

In the following, we use $p(\theta|-)$ to denote the full posterior conditional distribution of θ. Based on the formulation of the joint distribution (3.10), given the rest of the parameters, $\{A_{ij}, B_{ij}, C_i, D_i\}_{j=1}^{d}$ are independent for $i = 1, \ldots, d$, so we will first derive the posterior conditional distribution of $\{A_{ij}, B_{ij}, C_i, D_i\}_{j=1}^{d}$.

Let $\mathcal{G}_i^A = \{j : \delta(m_i, m_j) \cdot \gamma_{ij}^A \neq 0 \text{ and } j = 1, \ldots, d\}$ and $\mathcal{G}_i^B = \{j : \delta(m_i, m_j) \cdot \gamma_{ij}^B \neq 0 \text{ and } j = 1, \ldots, d\}$. Define a $d \times d$ diagonal matrix $\mathbf{I}_i^A$ where diagonal entries corresponding to $\mathcal{G}_i^A$ equal 1, and the rest diagonal entries equal 0. We define $\mathbf{I}_i^B$

associated with $\mathcal{G}_i^B$ in the same manner. We use $\mathbf{M}[, \mathcal{G}]$ to denote the matrix consisting of columns indexed by $\mathcal{G}$ of $\mathbf{M}$, and $\mathbf{M}[\mathcal{G},]$ to denote the matrix consisting of rows indexed by $\mathcal{G}$ of $\mathbf{M}$. Let $X_i^A(t) = \mathbf{I}_i^A \mathbf{x}(t) \cdot (1 - u(t))$ and $X_i^B(t) = \mathbf{I}_i^B \mathbf{x}(t) \cdot u(t)$, so $X_i^A(t)$ and $X_i^B(t)$ are vectors whose elements are functions of time t. Let $\Lambda_i(t) = ((X_i^A(t))', (X_i^B(t))', u(t), 1)$ and $\boldsymbol{\theta}_i = (\boldsymbol{A}[i, \mathcal{G}_i^A], \boldsymbol{B}[i, \mathcal{G}_i^B], C_i, D_i)'$. We have

$$p(\mathbf{A}[i,], \mathbf{B}[i,], C_i, D_i | -) \propto \exp\left\{ -\frac{1}{2\tau} \int_0^T \left(\Lambda_i(t)\boldsymbol{\theta}_i - \frac{dx_i(t)}{dt} \right)^2 dt \right\}$$
$$\cdot \prod_{j=1}^d \phi\left(\frac{A_{ij}}{\xi_0} \right) \cdot \prod_{j=1}^d \phi\left(\frac{B_{ij}}{\xi_0} \right) \cdot \phi\left(\frac{C_i}{\xi_0} \right) \cdot \phi\left(\frac{D_i}{\xi_0} \right),$$

where $dx_i(t)/dt = \left(\boldsymbol{b}^{(1)}(t) \right)' \eta_i$.

After integrating out A_{ij} and B_{ij} corresponding to zero indicator values in the above equation, we have

$$p(\boldsymbol{\theta}_i | \boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y}, \tau, \mu) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\theta}_i' \left(\frac{1}{\tau} \int_0^T \Lambda_i'(t)\Lambda_i(t)dt + \frac{1}{\xi_0^2}\mathbf{I} \right) \boldsymbol{\theta}_i \right\}$$
$$\cdot \exp\left\{ \frac{1}{\tau} \int \frac{dx_i(t)}{dt} \Lambda_i(t)dt \; \boldsymbol{\theta}_i \right\}$$
$$\cdot \exp\left\{ -\frac{1}{2\tau} \int_0^T \left(\frac{dx_i(t)}{dt} \right)^2 dt \right\}, \tag{6.3}$$

where $\mathbf{I}$ denotes an identity matrix.

Let $\mathbf{M}_i = \frac{1}{\tau} \int_0^T \Lambda_i'(t)\Lambda_i(t)dt + \frac{1}{\xi_0^2}\mathbf{I}$ and $\mathbf{V}_i = \frac{1}{\tau} \int_0^T \frac{dx_i(t)}{dt} \cdot \Lambda_i'(t)dt$. Based on equation

(6.3), we have

$$p(\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B | \mathbf{Y}, \tau, \mu) \propto \prod_{i=1}^{d} \sigma_i^{-T-2} \exp\left\{ -\frac{(Y_i - \boldsymbol{\Phi}\,\eta_i)^2}{2\sigma_i^2} \right\}$$

$$\cdot \prod_{i=1}^{d} \det(\mathbf{M}_i)^{-1/2} \cdot \exp\left\{ \sum_{i=1}^{d} \mathbf{V}_i' \mathbf{M}_i^{-1} \mathbf{V}_i / 2 \right\}$$

$$\cdot \exp\left\{ -\frac{1}{2\tau} \sum_{i=1}^{d} \int_0^T \left(\frac{dx_i(t)}{dt} \right)^2 dt \right\}$$

$$\cdot \exp\left\{ -\mu \sum_{i,j=1}^{d} \delta(m_i, m_j) \right\}$$

$$\cdot p_0^{\sum_{i,j} \gamma_{ij}^A + \sum_{i,j} \gamma_{ij}^B} \cdot (1 - p_0)^{2d^2 - \sum_{i,j} \gamma_{ij}^A - \sum_{i,j} \gamma_{ij}^B}.$$

We have $p(\boldsymbol{m}, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B | \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \mathbf{Y}, \tau, \mu) \propto \mathbb{J}(\boldsymbol{m}, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \boldsymbol{\eta}, \tau, \mu)$, where

$$\mathbb{J}(\boldsymbol{m}, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \boldsymbol{\eta}, \tau, \mu) = \prod_{i=1}^{d} \det(\mathbf{M}_i)^{-1/2} \cdot \exp\left\{ \sum_{i=1}^{d} \mathbf{V}_i' \mathbf{M}_i^{-1} \mathbf{V}_i / 2 \right\}$$

$$\cdot \exp\left\{ -\mu \sum_{i,j=1}^{d} \delta(m_i, m_j) \right\}$$

$$\cdot p_0^{\sum_{i,j} \gamma_{ij}^A + \sum_{i,j} \gamma_{ij}^B} \cdot (1 - p_0)^{2d^2 - \sum_{i,j} \gamma_{ij}^A - \sum_{i,j} \gamma_{ij}^B}.$$

B.II Simulate m_i from $p(m_i | \boldsymbol{m}_{-i}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y}, \tau, \mu)$ for $i = 1, \ldots, d$ sequentially

Let $\mathcal{G}_{-i}$ be the set of distinct values in $\boldsymbol{m}_{-i}$, and g_{-i} be any positive integer smaller than $d + 1$ and not belonging to $\mathcal{G}_{-i}$. Then the posterior conditional distribution of m_i is discrete and has a support of $\{\mathcal{G}_{-i}, g_{-i}\}$. In addition, for each $z \in \{\mathcal{G}_{-i}, g_{-i}\}$,

$$P(m_i = z | \boldsymbol{m}_{-i}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y}) \propto \mathbb{J}(m_i = z, \boldsymbol{m}_{-i}, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \boldsymbol{\eta}, \tau, \mu).$$

B.III Sequentially simulate γ_{ij}^As and γ_{ij}^Bs from their posterior conditional probabilities

Given parameter values $\boldsymbol{m}, \boldsymbol{\gamma}_{-ij}^A, \boldsymbol{\gamma}^B$, and $\boldsymbol{\eta}$, γ_{ij}^A for $i, j = 1, \ldots, d$ follows a Bernoulli distribution with probability

$$\frac{\mathbb{J}(\boldsymbol{m}, \gamma_{ij}^A = 1, \boldsymbol{\gamma}_{-ij}^A, \boldsymbol{\gamma}^B, \boldsymbol{\eta}, \tau, \mu)}{\mathbb{J}(\boldsymbol{m}, \gamma_{ij}^A = 1, \boldsymbol{\gamma}_{-ij}^A, \boldsymbol{\gamma}^B, \boldsymbol{\eta}, \tau, \mu) + \mathbb{J}(\boldsymbol{m}, \gamma_{ij}^A = 0, \boldsymbol{\gamma}_{-ij}^A, \boldsymbol{\gamma}^B, \boldsymbol{\eta}, \tau, \mu)}.$$

Note that if $m_i \neq m_j$, the above probability equals p_0. Similarly, we sequentially simulate γ_{ij}^B conditional on the rest of the parameters.

B.IV Simulate $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y}, \tau, \mu)$

Based on the joint posterior distribution (3.10) and posterior conditional distribution of $\boldsymbol{\theta}_i$ (6.3),

$$A_{ij}|\delta(m_i, m_j) \cdot \gamma_{ij}^A = 0 \quad \overset{iid}{\sim} \quad \mathrm{N}(0, \xi_0^2)$$

$$B_{ij}|\delta(m_i, m_j) \cdot \gamma_{ij}^B = 0 \quad \overset{iid}{\sim} \quad \mathrm{N}(0, \xi_0^2)$$

$$\boldsymbol{\theta}_i|\boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}^A, \boldsymbol{\gamma}^B, \mathbf{Y}, \tau, \mu \quad \overset{ind}{\sim} \quad \mathrm{MN}(\mathbf{M}_i^{-1}\mathbf{V}_i, \mathbf{M}_i^{-1}) \quad \text{for} \quad i = 1, \ldots, d.$$

B.V Simulate $\boldsymbol{\sigma}^2$ from $p(\boldsymbol{\sigma}^2|\boldsymbol{\Theta}_I, \boldsymbol{\eta}, \mathbf{Y}, \tau, \mu)$

From the joint posterior distribution (3.10), we have

$$\sigma_i^2|\boldsymbol{\Theta}_I, \boldsymbol{\eta}, \mathbf{Y}, \tau, \mu \overset{ind}{\sim} \mathrm{Inv\text{-}Gamma}\left(\frac{T}{2}, \frac{(Y_i - \boldsymbol{\Phi}\,\eta_i)^2}{2}\right) \quad \text{for} \quad i = 1, \ldots, d.$$

B.VI Simulate $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|\boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2, \mathbf{Y}, \tau, \mu)$

Define a dT-by-dL matrix $\mathbf{Q} = \begin{pmatrix} \boldsymbol{\Phi} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \boldsymbol{\Phi} \end{pmatrix}$, where $\boldsymbol{\Phi}$ is defined in (3.5). Let $\mathbf{U}$

be a dT-by-dT diagonal matrix with $(i-1)\cdot T$ to $i\cdot T$ diagonal entries equalling $1/\sigma_i^2$,

$i = 1, \ldots, d$. Then

$$
\begin{aligned}
p(\boldsymbol{\eta}|\boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2, \mathbf{Y}, \tau, \mu) &\propto \exp\left\{-\frac{1}{2}(\mathbf{Y} - \mathbf{Q}\boldsymbol{\eta})'\mathbf{U}(\mathbf{Y} - \mathbf{Q}\boldsymbol{\eta})\right\} \\
&\cdot \exp\left\{-\frac{1}{2\tau}(\boldsymbol{\eta}'\boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}\boldsymbol{\eta} - 2\boldsymbol{\Lambda}'_{\boldsymbol{\Theta}_I}\boldsymbol{\eta} + \boldsymbol{\Xi}_{\boldsymbol{\Theta}_I})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\psi})'\mathbf{H}(\boldsymbol{\eta} - \boldsymbol{\psi})\right\}, \qquad (6.4)
\end{aligned}
$$

where $\mathbf{H} = \mathbf{Q}'\mathbf{U}\mathbf{Q} + \boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}/\tau$, and $\boldsymbol{\psi} = \mathbf{H}^{-1}(\mathbf{Q}'\mathbf{U}\mathbf{Y} + \boldsymbol{\Lambda}_{\boldsymbol{\Theta}_I}/\tau)$. From (6.4),

$$
\boldsymbol{\eta}|\boldsymbol{\Theta}_I, \boldsymbol{\sigma}^2, \mathbf{Y}, \tau, \mu \sim \mathrm{MN}(\boldsymbol{\psi}, \mathbf{H}^{-1}).
$$

Notations $\boldsymbol{\Omega}_{\boldsymbol{\Theta}_I}, \boldsymbol{\Lambda}_{\boldsymbol{\Theta}_I}$, and $\boldsymbol{\Xi}_{\boldsymbol{\Theta}_I}$ are introduced in equation (3.7), and we derive

their formulas depending on $\boldsymbol{\Theta}_I$ in the following.

Define vectors with $d \cdot L$ elements:

$$\boldsymbol{\Delta}_i(t) = \Big(A_{i1} \cdot \delta(m_i, m_1) \cdot \gamma_{i1}^A \cdot b_1(t) \cdot (1 - u(t)), \ \ldots,$$

$$A_{i1} \cdot \delta(m_i, m_1) \cdot \gamma_{i1}^A \cdot b_L(t) \cdot (1 - u(t)),$$

$$A_{i2} \cdot \delta(m_i, m_2) \cdot \gamma_{i2}^A \cdot b_1(t) \cdot (1 - u(t)), \ \ldots,$$

$$A_{id} \cdot \delta(m_i, m_d) \cdot \gamma_{id}^A \cdot b_L(t) \cdot (1 - u(t)) \Big),$$

$$\boldsymbol{\Upsilon}_i(t) = \Big(B_{i1} \cdot \delta(m_i, m_1) \cdot \gamma_{i1}^B \cdot b_1(t) \cdot u(t), \ \ldots,$$

$$B_{i1} \cdot \delta(m_i, m_1) \cdot \gamma_{i1}^B \cdot b_L(t) \cdot u(t),$$

$$B_{i2} \cdot \delta(m_i, m_2) \cdot \gamma_{i2}^B \cdot b_1(t) \cdot u(t), \ \ldots,$$

$$B_{id} \cdot \delta(m_i, m_d) \cdot \gamma_{id}^B \cdot b_L(t) \cdot u(t) \Big),$$

$$\boldsymbol{E}_i(t) = \Big(\boldsymbol{0}_L, \ldots, (\frac{d\boldsymbol{b}(t)}{dt})', \ldots, \boldsymbol{0}_L \Big),$$

where $\boldsymbol{0}_L$ is a zero vector with L elements, and the $(i-1) \cdot L + 1$th to $i \cdot L$th elements of $\boldsymbol{E}_i(t)$ are non zero. Then with basis representation, MIDDM (3.3) can be rewritten as $\boldsymbol{E}_i(t)\,\boldsymbol{\eta} - \boldsymbol{\Delta}_i(t)\,\boldsymbol{\eta} - \boldsymbol{\Upsilon}_i(t)\,\boldsymbol{\eta} - C_i \cdot u(t) - D_i = 0$. Let $\mathbf{S}_i(t) = \boldsymbol{E}_i(t) - \boldsymbol{\Delta}_i(t) - \boldsymbol{\Upsilon}_i(t)$. Then we have

$$\mathbf{R}(\boldsymbol{\eta}, \boldsymbol{\Theta}_I) = \sum_{i=1}^{d} \Big\{ \boldsymbol{\eta}' \int \mathbf{S}_i'(t)\mathbf{S}_i(t)dt\,\boldsymbol{\eta} - 2 \int (C_i \cdot u(t) + D_i) \cdot \mathbf{S}_i(t)dt\,\boldsymbol{\eta}$$

$$+ \ \int (C_i \cdot u(t) + D_i) \cdot (C_i \cdot u(t) + D_i)dt \Big\}.$$

Comparing the above to equation (3.7), we have

$$\mathbf{\Omega_{\Theta_I}} = \sum_{i=1}^{d} \int \mathbf{S}'_i(t)\mathbf{S}_i(t)dt,$$

$$\mathbf{\Lambda_{\Theta_I}} = \sum_{i=1}^{d} \int (C_i \cdot u(t) + D_i) \cdot \mathbf{S}'_i(t)dt,$$

$$\mathbf{\Xi_{\Theta_I}} = \sum_{i=1}^{d} \int (C_i \cdot u(t) + D_i) \cdot (C_i \cdot u(t) + D_i)dt.$$

Bibliography

[1] Magnetic resonance, a critical peer-reviewed introduction; functional MRI. European Magnetic Resonance Forum. Retrieved October 17, 2017.

[2] What is EEG (electroencephalography) and how does it work?, 2016. Retrieved on October 17, 2017.

[3] A Aertsen and H Preissl. Dynamics of activity and connectivity in physiological neuronal networks. In HG Schuster, editor, *Nonlinear Dynamics and Neuronal Networks*, pages 281–302. VCH publishers Inc, New York, 1991.

[4] J Anderson. Learning in sparsely connected and sparsely coded system. In *Ersatz Brain Project Working Note*. 2005.

[5] Eishi Asano, Csaba Juhász, Aashit Shah, Otto Muzik, Diane C Chugani, Jagdish Shah, Sandeep Sood, and Harry T Chugani. Origin and propagation of epileptic spasms delineated on electrocorticography. *Epilepsia*, 46(7): 1086–1097, 2005.

[6] Y Bard. *Nonlinear Parameter Estimation*. Academic Press, New York, 1974.

[7] E Basar, A Gönder, and Ungan P. Comparative frequency analysis of single eeg-evoked potential records. *Journal of Biomedical Engineering*, 2:9–14, 1980.

[8] P Bhaumik and S Ghosal. Bayesian estimation in differential equation models. *Preprint*, 1, 2014.

[9] L Biegler, JJ Damiano, and GE Blau. Nonlinear parameter estimation: A case study comparison. *AIChE Journal*, 32:29–45, 1986.

[10] D Boatman-Reich, P Franaszczuk, A Korzeniewska, B Caffo, E Ritzl, S Colwell, and N Crone. Quantifying auditory event-related responses in multichannel human intracranial recordings. *Frontiers in Computational Neuroscience*, 4(4), 2010.

[11] S. Bressler and M. Ding. Event-related potentials. In *The Handbook of Brain Theory and Neural Networks*, pages 412–415. John Wiley & Sons, Inc, 2002.

[12] Philip J Brown, Marina Vannucci, and Tom Fearn. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641, 1998.

[13] NJ Brunel. Parameter estimation of odes via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.

[14] E Bullmore and O Sporns. Complex brain networks: Graph theoretical analysis

of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[15] B Caffo, R Peng, F Dominici, T Louis, and S Zeger. Parallel mcmc for analyzing distributed lag models with systematic missing data for an application in environmental epidemiology. *Handbook of Markov Chain Monte Carlo: Methods and Applications*, 2010.

[16] B Calderhead, M Girolami, and ND Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. *Advances in Neural Information Processing Systems*, 22, 2008.

[17] D Campbell. *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models*. Montreal, 2007.

[18] J Cao, JZ Huang, , and H Wu. Penalized nonlinear least squares estimation of time-varying parameters in ordinary differential equations. *Journal of Computational and Graphical Statistics*, 21:42–56, 2012.

[19] M Catani, F DellAcqua, F Vergani, F Malik, H Hodge, P Roy, R Valabregue, and M Thiebaut de Schotten. Short frontal lobe connections of the human brain. *Cortex*, 48(2):273–291, 2012.

[20] MC Cervenka, PJ Franaszczuk, NE Crone, B Hong, B Caffo, P Bhatt, FA Lenz, and D Boatman-Reich. Reliability of early cortical auditory gamma-band responses. *Clinical Neurophysiology*, 124(1):70–82, 2013.

[21] SH Cheung, TA Oliver, EE Prudencio, S Prudhomme, and RD Moser. Bayesian uncertainty analysis with applications to turbulence modeling. *Reliability Engineering & System Safety*, 96(9):1137– 1149, 2011.

[22] OA Chkrebtii, DA Campbell, B Calderhead, and MA Girolami. Bayesian solution uncertainty quantification for differential equations. Submitted, 2015.

[23] P Conrad, M Girolami, S Särkkä, A Stuart, and K Zygalakis. Probability measures for numerical solutions of differential equations. Submitted, 2015.

[24] J Daunizeau, O David, and KE Stephan. Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *NeuroImage*, 58:312–322, 2011.

[25] O David and KJ Friston. A neural mass model for meg/eeg: Coupling and neuronal dynamics. *NeuroImage*, 20:1743–1755, 2003.

[26] O David, SJ Kiebel, L Harrison, J Mattout, J Kilner, and KJ Friston. Dynamic causal modelling of evoked responses in eeg and meg. *NeuroImage*, 30:1255–1272, 2006.

[27] P Deuflhard and F Bornemann. *Scientific Computing with Ordinary Differential Equations.* Springer, New York, 2000.

[28] H. Devlin. What is functioanl magnetic resonance imaging (fmri)?, 2016. Psych Central. Retrieved on October 16, 2017.

[29] DB Dunson, AH Herring, and SM Engel. Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, 103(482):534–546, 2008.

[30] PJ Durka, D Ircha, C Neuper, and G Pfurtscheller. Time-frequency microstructure of event-related electro-encephalogram eesynchronisation and synchronisation. *Medical & Biological Engineering & Computing*, 39:315–3211, 2001.

[31] SJ Eliades, NE Crone, WS Anderson, D Ramadoss, FA Lenz, and D Boatman-Reich. Adaptation of high-gamma responses in human auditory association cortex. *Journal of Neurophysiology*, 112(9):2147–2163, 2014.

[32] J Fan and R Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[33] P. Földiák and M. P. Young. Sparse coding in the primate cortex. In 895-898, editor, *The Handbook of Brain Theory and Neural Networks*. The MIT Press, 1995.

[34] PJ Franaszczuk and GK Bergey. Application of the directed transfer function method to mesial and lateral onset temporal lobe seizures. *Brain Topogra*, 11:13–21, 1998.

[35] K Friston. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biol*, 7:33, 2009.

[36] KJ Friston, L Harrison, and WD Penny. Dynamic causal modelling. *NeuroImage*, 19:1273–1302, 2003.

[37] Ryohei Fujimaki and Kohei Hayashi. Factorized asymptotic bayesian hidden markov models. *arXiv preprint arXiv:1206.4679*, 2012.

[38] MI Garrido, JM Kilner, SJ Kiebel, KE Stephan, T Baldeweg, and KJ Friston. Comparative frequency analysis of single eeg-evoked potential records. *Neuroimage*, 48(1):269 – 279, 2009.

[39] A Gelman, FY Bois, and J Jiang. Physiological pharamacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91:1400–1412, 1996.

[40] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[41] E George and RE McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

[42] E George and RE McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

[43] M Girolami. Bayesian inference for differential equations. *Theoretical Computer Science*, 408:4–16, 2008.

[44] F Graner and JA Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical Review Letters*, 69:2013–2016, 1992.

[45] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography?theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

[46] Kimiaki Hashiguchi, Takato Morioka, Fumiaki Yoshida, Yasushi Miyagi, Shinji Nagata, Ayumi Sakata, and Tomio Sasaki. Correlation between scalp-recorded electroencephalographic and electrocorticographic activities during ictal period. *Seizure*, 16(3):238–247, 2007.

[47] P Hemker. Numerical methods for differential equations in system simulations and in parameter estimation. *Analysis and Simulation of Biochemical Systems*, pages 59–80, 1972.

[48] Björn Herrmann, Molly J Henry, and Jonas Obleser. Frequency-specific adaptation in human auditory cortex depends on the spectral variance in the acoustic stimulation. *Journal of neurophysiology*, 109(8):2086–2096, 2013.

[49] Björn Herrmann, Nadine Schlichting, and Jonas Obleser. Dynamic range adaptation to spectral stimulus statistics in human auditory cortex. *Journal of Neuroscience*, 34(1):327–331, 2014.

[50] Y Huang and H Wu. A bayesian approach for estimating antiviral efficacy in hiv dynamic models. *Journal of Applied Statistics*, 33:155–174, 2006.

[51] Y Huang, D Liu, and H Wu. Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. *Biometrics*, 62:413–423, 2006.

[52] S.A. Huettel, A.W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer, 2009.

[53] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

[54] Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[55] SJ Kiebel, O David, and KJ Friston. Dynamic causal modelling of evoked responses in eeg/meg with lead-field parameterization. *NeuroImage*, 30:1273–1284, 2006.

[56] S Kim, MG Tadesse, and M Vannucci. Variable selection in clustering via dirichlet process mixture models. *Biometrika*, 93(4):877–893, 2006.

[57] Z Li, MR Osborne, and T Pravan. Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25:264–285, 2005.

[58] T Lu, H Liang, H Li, and H Wu. High dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106:1242–1258, 2011.

[59] R Matteij and J Molenaar. *Ordinary Differential Equations in Theory and Practice*. SIAM, Philadelphia, 2002.

[60] AR McIntosh and F Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Humman Brain Mapping*, 2:2–22, 1994.

[61] S Micheloyannis. Graph-based network analysis in schizophrenia. *World Journal of Psychiatry*, 2(1):1–12, 2012.

[62] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

[63] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[64] R Milo, S Itzkovitz, N Kashtan, R Levitt, S Shen-Orr, I Ayzenshtat, Sheffer M, and U Alon. Superfamilies of evolved and designed networks. *Science*, 303 (5663):1538–1542, 2004.

[65] R Näätänen, P Paavilainen, T Rinne, and K Alho. The mismatch negativity

(mmn) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118:2544–2590, 2007.

[66] MEJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23): 8577–8696, 2006.

[67] TA Oliver and RD Moser. Bayesian uncertainty quantification applied to rans turbulence models. *Journal of Physics: Conference Series*, 318(4):042032, 2011.

[68] BA Olshausen and DJ Field. Sparse coding of sensor inputs. *Current Opinions in Neurobiology*, 14:481–487, 2004.

[69] Hae-Jeong Park and Karl Friston. Structural and functional brain networks: From connections to cognition. *Science*, 342(6158), 2013. doi: 10.1126/science. 1238411.

[70] RB Potts. Some generalized order-disorder transformations. *Mathematical Proceedings*, 48:106–109, 1952.

[71] AA Poyton, MS Varziri, KB McAuley, PJ McLellan, and JO Ramsay. Parameter estimation in continuous dynamic models using principal differential analysis. *Computational Chemical Engineering*, 30:698–708, 2006.

[72] X Qi and H Zhao. Asymptotic efficiency and finite-sample properties of the gen-

eralized profiling estimation of the parameters in ordinary differential equations. *The Annals of Statistics*, 38:435–481, 2010.

[73] James O Ramsay, Giles Hooker, and Spencer Graves. *Functional data analysis with R and MATLAB*. Springer Science & Business Media, 2009.

[74] JO Ramsay and BW Silverman. *Functional Data Analysis*. Springer, New York, 2005.

[75] JO Ramsay, G Hooker, D Campbell, and J Cao. Parameter estimation for differential equations: A generalized smoothing approach (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 69:741–796, 2007.

[76] PT Reiss and RT Ogden. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102:984–996, 2007.

[77] PT Reiss and RT Ogden. Smoothing parameter selection for a class of semi-parametric linear models. *Journal of the Royal Statistical Society, Ser. B*, 71: 505–523, 2009.

[78] Alois Schlögl, Mel Slater, and Gert Pfurtscheller. Presence research and eeg. In *Proceedings of the 5th International Workshop on Presence*, pages 9–11, 2002.

[79] Alfons Schnitzler and Joachim Gross. Normal and pathological oscillatory communication in the brain. *Nature reviews neuroscience*, 6(4):285–296, 2005.

[80] Marc Schönwiesner, Nikolai Novitski, Satu Pakarinen, Synnöve Carlson, Mari Tervaniemi, and Risto Näätänen. Heschl9s gyrus, posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes. *Journal of Neurophysiology*, 97(3):2075–2082, 2007.

[81] A Sinai, NE Crone, HM Wied, PJ Franaszczuk, D Miglioretti, and D Boatman-Reich. Intracranial mapping of auditory perception: Event-related responses and electrocortical stimulation. *Clinical Neurophysiology*, 120:140–149, 2009.

[82] O Sporns. *Networks of the Brain.* The MIT Press, Cambridge, Massachusetts, 2011.

[83] Ramesh Srinivasan. Methods to improve the spatial resolution of eeg. *International Journal of Bioelectromagnetism*, 1(1):102–111, 1999.

[84] Blausen.com staff. Medical gallery of blausen medical 2014. *WikiJournal of Medicine*, 1(2), 2014.

[85] AM Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19: 451–559, 2010.

[86] MG Tadesse, N Sha, and M Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100 (470):602–617, 2005.

[87] HE Theo and EG Mike. Mapping multiple qtl using linkage disequilibrium and

linkage analysis information and multitrait data. *Genet. Sel. Evol*, 36:261–279, 2004.

[88] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B.*, 58:267–288, 1996.

[89] DA van Dyk and T Park. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103:790–796, 2008.

[90] JM Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal of Scientific Computing*, 3:28–46, 1982.

[91] EO Voit. *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, Cambridge, 2000.

[92] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

[93] H Wang and C Leng. A note on the adaptive group lasso. *Computational Statistics and Data Analysis*, 52:5277–5286, 2008.

[94] H Wu, T Lu, H Xue, and H Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109:700–716, 2014.

[95] S Wu, H Xue, Y Wu, and H Wu. Variable selection for sparse high-dimensional

nonlinear regression models by combining nonnegative garrote and sure independence screening. *Statistica Sinica*, Inprint, 2014.

[96] H Xue, H Miao, and H Wu. Sieve estimation of constant and time- varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *The Annals of Statistics*, 38: 2351–2387, 2010.

[97] Nengjun Yi, Varghese George, and David B Allison. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*, 164(3):1129–1138, 2003.

[98] M Yuan and Y Lin. Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472): 1215–1225, 2005.

[99] M Yuan and Y Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[100] T Zhang, J Wu, F Li, B Caffo, and D Boatman-Reich. A dynamic directional model for effective brain connectivity using electrocorticographic (ecog) time series. *Journal of the American Statistical Association*, 110(509):93–106, 2015.

[101] H Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[102] H Zou and T Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2006.

Publication Based on the Dissertation

1. T Zhang, Q Yin, B Caffo, Y Sun, and D Boatman-Reich. Bayesian inference of high-dimensional, cluster-structured ordinary differential equation models with applications to brain connectivity studies. *The Annals of Applied Statistics*, 11(2):868-897, 2017.